Thesis submitted for the degree of Doctor of Science

University of Aston

Birmingham

High-Quality Audio Systems

Professor Malcolm Omar John Hawksford, BSc, PhD, CEng, FIEE, FAES, FIOA

Submitted July 2007

Contents *page*

0-1 Publications listed by Sections (*all contributing authors indicated*)

Section 2 (publications): Analogue signal processing

2-1 Feedback, error correction and linear circuit techniques

2-1 DISTORTION CORRECTION IN AUDIO POWER AMPLIFIERS, Hawksford, M.J., *JAES*, vol.29, no.1/2, Jan/Feb 1981, pp.27-30 (paper selected from 65th AES Convention for publication)

2-5 DISTORTION CORRECTION CIRCUITS FOR AUDIO AMPLIFIERS, Hawksford, M.J., *JAES*, vol.29, no.7, 8, July/August 1981

2-13 FUZZY DISTORTION IN ANALOG AMPLIFIERS: A LIMIT TO INFORMATION TRANSMISSION?, Hawksford, M.J., *JAES*, vol.31, no.10, pp.745-754, October 1983

2-23 OPTIMIZATION OF THE AMPLIFIED-DIODE BIAS CIRCUIT FOR AUDIO AMPLIFIERS, Hawksford, M.J., *JAES*, vol.32, no.1/2, pp.31-33, Jan/Feb 1984

2-26 REDUCTION OF TRANSISTOR SLOPE DISTORTION IN LARGE SIGNAL AMPLIFIERS, Hawksford, M.J., *JAES*, vol.36, no.4, pp.213-222, April 1988

2-36 TRANSCONDUCTANCE POWER AMPLIFIER SYSTEMS FOR CURRENT-DRIVEN LOUDSPEAKERS, Mills, P.G.L., Hawksford, M.O.J., *JAES*, vol.37, no.10, pp.809-822, October 1989

2-50 DIFFERENTIAL-CURRENT DERIVED FEEDBACK (DCDF) IN ERROR CORRECTING AUDIO AMPLIFIERS, Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 141, no 3, June 1994, pp 227-236

2-60 QUAD-INPUT CURRENT-MODE ASYMMETRIC CELL (CMAC) WITH ERROR CORRECTION APPLICATIONS IN SINGLE-ENDED AND BALANCED AUDIO AMPLIFIERS, Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 143, no 1, Feb. 1996, pp 1-7

2-67 RELATIONSHIP BETWEEN NOISE SHAPING AND NESTED DIFFERENTIATING FEEDBACK LOOPS, Hawksford, M.O.J. and Vanderkooy, J., *JAES*, vol. 47, no. 12, pp 1054-1060, December 1999

2-2 Translinear (variable gain) amplifiers

2-74 LOW-DISTORTION PROGRAMMABLE GAIN CELL USING CURRENT-STEERING CASCODE TOPOLOGY, Hawksford, M.J., *JAES*, vol.30, no.11, November 1982

2-79 TOPOLOGICAL ENHANCEMENTS OF TRANSLINEAR TWO-QUADRANT GAIN CELLS, Hawksford, M.O.J. and Mills, P.G.L., *JAES*, vol.37, no.6, pp 465-475, June 1989

Section 3 (publications): Converter and switching amplifiers systems

3-1 Converters, noise shaping and oversampling

3-1 CHAOS, OVERSAMPLING AND NOISE SHAPING IN DIGITAL-TO-ANALOG CONVERSION, Hawksford, M.O.J., *JAES*, vol. 37, no. 12, pp 980-1001, December 1989

3-23 OVERSAMPLING FILTER DESIGN IN NOISE-SHAPING DIGITAL-TO-ANALOG CONVERSION, Hawksford, M.O.J., Wingerter, W., *JAES*, vol. 38, no. 11, pp 845-856, November 1990

3-35 OVERSAMPLED ANALOG-TO-DIGITAL CONVERSION FOR DIGITAL AUDIO SYSTEMS, Hawksford, M.O.J. and Darling, T.E., *JAES*, vol. 38, no. 12, pp 924-943, December 1990

3-55 DIGITAL-TO-ANALOG CONVERTER WITH LOW INTERSAMPLE TRANSITION DISTORTION AND LOW SENSITIVITY TO SAMPLE JITTER AND TRANSRESISTANCE AMPLIFIER SLEW RATE, Hawksford, M.O.J., *JAES*, vol. 42, no. 11, pp 901-917, November 1994

3-72 TRANSPARENT DIFFERENTIAL CODING FOR HIGH-RESOLUTION DIGITAL AUDIO, Hawksford, M. O. J., *JAES*, vol. 49, no. 6, pp 480-497, June 2001

3-2 Sigma-delta modulation (SDM)

3-90 EXACT MODEL FOR DELTAMODULATION PROCESSES, Flood, J.E., and Hawksford, M.J., *Proc. IEE*, vol.118, pp.1155-1161, 1971

3-97 UNIFIED THEORY OF DIGITAL MODULATION, Hawksford, M.J., *Proc. IEE*, vol.121, no.2, pp.109-115, February 1974

3-104 TIME-QUANTIZED FREQUENCY MODULATION, TIME-DOMAIN DITHER, DISPERSIVE CODES, AND PARAMETRICALLY CONTROLLED NOISE SHAPING IN SDM, Hawksford, M.O.J., *JAES*, vol. 52, no. 6, pp 587-617, June 2004

3-135 PARAMETRICALLY CONTROLLED NOISE SHAPING IN VARIABLE STATE-STEP-BACK PSEUDO-TRELLIS SDM, Hawksford, M.O.J., IEE Proc.-VIS. Image Signal Processing, Vol. 152, No. 1, pp 87-96, February 2005

3-3 Pulse width modulation (PWM) and switching amplifiers

3-145 DYNAMIC MODEL-BASED LINEARIZATION OF QUANTIZED PULSE-WIDTH MODULATION FOR APPLICATIONS IN DIGITAL-TO-ANALOG CONVERSION AND DIGITAL POWER AMPLIFIER SYSTEMS, Hawksford, M.O.J., *JAES*, Vol. 40, no. 4, pp 235-252, April 1992

3-163 LINEARIZATION OF MULTI-LEVEL, MULTI-WIDTH DIGITAL PWM WITH APPLICATIONS IN DIGITAL-TO-ANALOGUE CONVERSION, Hawksford, M.O.J., *JAES*, vol. 43, no. 10, pp 787-798, October 1995

3-175 AN OVERSAMPLED DIGITAL PWM LINEARIZATION TECHNIQUE FOR DIGITAL-TO-ANALOG CONVERSION, Jung, J.W. and Hawksford, M. J., Regular Papers, IEEE Transactions on [see also Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on], vol. 51, no. 9, Issue 9, September 2004-09-15, pp 1781-1789

3-184 MODULATION AND SYSTEM TECHNIQUES IN PWM AND SDM SWITCHING AMPLIFIERS, Hawksford, M.O.J. *JAES*, vol. 54, no. 3, pp. 107-139, March 2006

Section 4 (publications): Loudspeaker systems

4-1 System theory

4-1 APPLICATION OF THE GEOMETRIC THEORY OF DIFFRACTION (GTD) TO DIFFRACTION AT THE EDGES OF LOUDSPEAKER BAFFLES, Bews, R.M., and Hawksford, M.J., *JAES*, vol.34, no.10, pp.771-779, October 1986

4-10 REDUCTION OF LOUDSPEAKER POLAR RESPONSE ABERRATIONS THROUGH THE APPLICATION OF PSYCHOACOUSTIC ERROR CONCEALMENT, Rimell, A. and Hawksford, M.O.J., IEE Proceedings on Vision Image Signal Processing, vol. 145, no 1, Feb. 1998, pp 11-18 [Awarded "The Associates Premium Award" by the IEE to Dr Rimell]

4-18 INTRODUCTION TO DISTRIBUTED MODE LOUDSPEAKERS (DML) WITH FIRST-ORDER BEHAVIOURAL MODELLING, Harris, N. and Hawksford, M.O.J., IEE Proc.-Circuits Devices Systems, Vol. 147, No. 3, pp 153-157, June 2000

4-2 Current drive

4-23 DISTORTION REDUCTION IN MOVING-COIL LOUDSPEAKER SYSTEMS USING CURRENT-DRIVE TECHNOLOGY, Mills, P.G.L., Hawksford, M.O.J., *JAES*, vol.37, no.3, pp.129-148, March 1989

4-3 Crossover and equalization systems

4-43 EFFICIENT FILTER DESIGN FOR LOUDSPEAKER EQUALIZATION, Hawksford, M.O.J. and Greenfield, R., *JAES*, vol. 39, no. 10, pp 739-751, November 1991

4-56 ASYMMETRIC ALL-PASS CROSSOVER ALIGNMENTS, Hawksford, M.O.J., *JAES*, vol. 41, no. 3, pp 123-134, March 1993

4-68 ON THE DITHER PERFORMANCE OF HIGH-ORDER DIGITAL EQUALIZATION FOR LOUDSPEAKER SYSTEMS, Greenfield, R.G. and Hawksford, M.O.J., *JAES*, vol. 43, no. 11, pp 908-915, November 1995

4-76 DIGITAL SIGNAL PROCESSING TOOLS FOR LOUDSPEAKER EVALUATION AND DISCRETE-TIME CROSSOVER DESIGN, Hawksford, M.O.J., *JAES*, vol. 45, no. 1/2, pp 37-62, Jan/Feb 1997. [*Awarded AES Publication prize for the best paper from an author of any age, from JAES volumes 45 and 46.]*

4-102 MATLAB PROGRAM FOR LOUDSPEAKER EQUALIZATION AND CROSSOVER DESIGN, Hawksford, M.O.J., *JAES*, vol. 47, no. 9, pp 707-719, September 1999

4-4 Digital and array loudspeakers

4-116 SMART DIGITAL LOUDSPEAKER ARRAYS, Hawksford, M. O. J., *JAES*, vol. 51, no. 12, pp 1133-1162, December 2003

4-146 SPATIAL DISTRIBUTION OF DISTORTION AND SPECTRALLY-SHAPED QUANTIZATION NOISE IN DIGITAL MICRO-ARRAY LOUDSPEAKERS, Hawksford, M.O.J. *JAES*, vol. 55, no. 1/2, pp. 1-27, January/February 2007

Section 5 (publications): Perceptual and multi-channel audio systems

5-1 Performance assessment

5-1 COMMUNICATIONS IN NOISE - PERFORMANCE RANKING METRIC, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *BT Technology Journal*, vol.10, no.4, pp 109-115, October 1992

5-8 CHARACTERIZATION OF COMMUNICATION SYSTEMS USING A SPEECHLIKE TEST STIMULUS, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *JAES*, vol. 41, no. 12, pp 1008-1021, December 1993

5-22 ERROR ACTIVITY AND ERROR ENTROPY AS A MEASURE OF PSYCHOACOUSTIC SIGNIFICANCE IN THE PERCEPTUAL DOMAIN, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *Proc. IEE of Vision, Image and Signal Processing*, vol. 141, no. 3, pp 203-208, June 1994

5-28 ALGORITHMS FOR ASSESSING THE SUBJECTIVITY OF PERCEPTUALLY WEIGHTED AUDIBLE ERRORS, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R. *JAES*, vol. 43, no. 12, pp 1041-1045, December 1995

5-2 Multi-channel audio

5-33 SCALABLE MULTICHANNEL CODING WITH HRTF ENHANCEMENT FOR DVD AND VIRTUAL SOUND SYSTEMS, Hawksford, M. O. J., *JAES*, vol. 50, no. 11, pp 894-913, November 2002

Section 6 (publications): Measurement systems

6-1 MLS systems

6-1 DISTORTION IMMUNITY OF MLS-DERIVED IMPULSE RESPONSE MEASUREMENTS, Dunn, C. and Hawksford, M.O.J., *JAES*, vol. 41, no.5, pp 314-335, May 1993

6-23 DISTORTION ANALYSIS OF NON-LINEAR SYSTEMS WITH MEMORY USING MAXIMUM LENGTH SEQUENCES, Greest, M. and Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 142, no 5, Oct 1995, pp 345-352

6-2 Volterra modeling

6-29 IDENTIFICATION OF DISCRETE VOLTERRA SERIES USING MAXIMUM LENGTH SEQUENCES, Reed, M.J. and Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 143, no 5, Oct. 1996, pp 241-248

6-37 EFFICIENT IMPLEMENTATION OF THE VOLTERRA FILTER, Reed, M.J. and Hawksford, M.O.J., IEE Proc.-VIS. Image Signal Processing, Vol. 147, No. 2, pp 109-114, April 2000

6-43 SYSTEM MEASUREMENT AND IDENTIFICATION USING PSEUDORANDOM FILTERED NOISE AND MUSIC SEQUENCES, Hawksford, M.O.J. *JAES*, vol. 52, no. 4, pp. 275-296, April 2005

0-2 Declaration of collaboration

Of the 45 papers constituting this thesis, 24 are single-authored (and thus completely my own work) while the remaining 21 have been undertaken in collaboration with either research students under my supervision (associated to my duties as a full-time University academic in the Department of Electronic Systems Engineering, at the University of Essex) or with industrial partners linked to projects again under my supervision. In one case (page 2-67), a paper (where I proposed the core inventive concept) was written in collaboration with Professor John Vanderkooy, a highly distinguished academic visitor from the renowned Audio Research Laboratory at the University of Waterloo.

In all cases I have made substantial contributions to the topics, concepts and content of each project and their related papers.

The following sub-listing states the specific professional relationships with each co-author:

2-36 TRANSCONDUCTANCE POWER AMPLIFIER SYSTEMS FOR CURRENT-DRIVEN LOUDSPEAKERS, Mills, P.G.L., Hawksford, M.O.J., *JAES*, vol.37, no.10, pp.809-822, October 1989

2-79 TOPOLOGICAL ENHANCEMENTS OF TRANSLINEAR TWO-QUADRANT GAIN CELLS, Hawksford, M.O.J. and Mills, P.G.L., *JAES*, vol.37, no.6, pp 465-475, June 1989

4-23 DISTORTION REDUCTION IN MOVING-COIL LOUDSPEAKER SYSTEMS USING CURRENT-DRIVE TECHNOLOGY, Mills, P.G.L., Hawksford, M.O.J., *JAES*, vol.37, no.3, pp.129-148, March 1989

Dr Mills was a PhD student under my supervision.

2-67 RELATIONSHIP BETWEEN NOISE SHAPING AND NESTED DIFFERENTIATING FEEDBACK LOOPS, Hawksford, M.O.J. and Vanderkooy, J., *JAES*, vol. 47, no. 12, pp 1054-1060, December 1999

Professor Vanderkooy was a visiting academic from the University of Waterloo, Canada.

3-23 OVERSAMPLING FILTER DESIGN IN NOISE-SHAPING DIGITAL-TO-ANALOG CONVERSION, Hawksford, M.O.J., Wingerter, W., *JAES*, vol. 38, no. 11, pp 845-856, November 1990

W. Wingerter was a research student under my supervision.

3-35 OVERSAMPLED ANALOG-TO-DIGITAL CONVERSION FOR DIGITAL AUDIO SYSTEMS, Hawksford, M.O.J. and Darling, T.E., *JAES*, vol. 38, no. 12, pp 924-943, December 1990

T. Darling was an MPhil student under my supervision.

3-90 EXACT MODEL FOR DELTAMODULATION PROCESSES, J.E. Flood, J.E., and Hawksford, M.J., *Proc. IEE*, vol.118, pp.1155-1161, 1971

Professor Flood was my academic PhD supervisor at the University of Aston, Birmingham.

3-175 AN OVERSAMPLED DIGITAL PWM LINEARIZATION TECHNIQUE FOR DIGITAL-TO-ANALOG CONVERSION, Jung, J.W. and Hawksford, M. J., Regular Papers, IEEE Transactions on [see also Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on], vol. 51, no. 9, Issue 9, September 2004-09-15, pp 1781-1789

J. Jung was an MSc student under my supervision although this work was undertaken at a later time as the topic was of common interest. The paper was instigated by Mr Jung (resident in Korea) and developed further upon a method which I had previously published (see page 3-145) for linearizing digital PWM.

4-1 APPLICATION OF THE GEOMETRIC THEORY OF DIFFRACTION (GTD) TO DIFFRACTION AT THE EDGES OF LOUDSPEAKER BAFFLES, Bews, R.M., and Hawksford, M.J., *JAES*, vol.34, no.10, pp.771-779, October 1986

Dr. Bews was a PhD student under my supervision.

4-10 REDUCTION OF LOUDSPEAKER POLAR RESPONSE ABERRATIONS THROUGH THE APPLICATION OF PSYCHOACOUSTIC ERROR CONCEALMENT, Rimell, A. and Hawksford, M.O.J., IEE Proceedings on Vision Image Signal Processing, vol. 145, no 1, Feb. 1998, pp 11-18 [Awarded "The Associates Premium Award" by the IEE to Dr Rimell]

Dr Rimell was a PhD student under my supervision.

4-18 INTRODUCTION TO DISTRIBUTED MODE LOUDSPEAKERS (DML) WITH FIRST-ORDER BEHAVIOURAL MODELLING, Harris, N. and Hawksford, M.O.J., IEE Proc.-Circuits Devices Systems, Vol. 147, No. 3, pp 153-157, June 2000

Dr Harris was an external (industrial) PhD student under my supervision, he was also chief scientist for NXT plc.

4-43 EFFICIENT FILTER DESIGN FOR LOUDSPEAKER EQUALIZATION, Hawksford, M.O.J. and Greenfield, R., *JAES*, vol. 39, no. 10, pp 739-751, November 1991

4-68 ON THE DITHER PERFORMANCE OF HIGH-ORDER DIGITAL EQUALIZATION FOR LOUDSPEAKER SYSTEMS, Greenfield, R.G. and Hawksford, M.O.J., *JAES*, vol. 43, no. 11, pp 908-915, November 1995

Dr. Greenfield was a PhD student under my supervision.

5-1 COMMUNICATIONS IN NOISE - PERFORMANCE RANKING METRIC, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *BT Technology Journal*, vol.10, no.4, pp 109-115, October 1992

5-8 CHARACTERIZATION OF COMMUNICATION SYSTEMS USING A SPEECHLIKE TEST STIMULUS, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *JAES*, vol. 41, no. 12, pp 1008-1021, December 1993

5-22 ERROR ACTIVITY AND ERROR ENTROPY AS A MEASURE OF PSYCHOACOUSTIC SIGNIFICANCE IN THE PERCEPTUAL DOMAIN, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *Proc. IEE of Vision, Image and Signal Processing*, vol. 141, no. 3, pp 203-208, June 1994

5-28 ALGORITHMS FOR ASSESSING THE SUBJECTIVITY OF PERCEPTUALLY WEIGHTED AUDIBLE ERRORS, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R. *JAES*, vol. 43, no. 12, pp 1041-1045, December 1995

Dr Hollier was an external (industrial) PhD student from British Telecom Laboratories (BTL) under my supervision. Dr Guard was the industrial supervisor at BTL.

6-1 DISTORTION IMMUNITY OF MLS-DERIVED IMPULSE RESPONSE MEASUREMENTS, Dunn, C. and Hawksford, M.O.J., *JAES*, vol. 41, no.5, pp 314-335, May 1993

Dr. Dunn was a PhD student under my supervision.

6-23 DISTORTION ANALYSIS OF NON-LINEAR SYSTEMS WITH MEMORY USING MAXIMUM LENGTH SEQUENCES, Greest, M. and Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 142, no 5, Oct 1995, pp 345-352

M. Greest was a MSD student under my supervision.

6-29 IDENTIFICATION OF DISCRETE VOLTERRA SERIES USING MAXIMUM LENGTH SEQUENCES, Reed, M.J. and Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 143, no 5, Oct. 1996, pp 241-248

6-37 EFFICIENT IMPLEMENTATION OF THE VOLTERRA FILTER, Reed, M.J. and Hawksford, M.O.J., IEE Proc.-VIS. Image Signal Processing, Vol. 147, No. 2, pp 109-114, April 2000

Dr. Reed was a PhD student under my supervision.

0-3 Statement of authenticity

The material and work presented in this thesis has not been submitted wholly or in part to any university or educational institution with the intention of gaining a degree, neither has any component been submitted and subsequently rejected.

Please note the following two caveats:

(i) The paper presented on *page 3-90* does link with my research while I was a PhD research student at Aston University.

(ii) As some papers (see sub-list in Section 0-2) have formed part of research-student programmes under my supervision, aspects could be deemed to have been submitted for the degrees for which those students were registered. However, this is standard academic practice and a consequence of the role of research supervision, it does not represent any case where I have independently submitted that work to an Institution or other body with the intention of obtaining a degree.

Section 1: Introduction and thesis ethos

1 Introduction and thesis ethos

The research presented in this thesis is derived from a lifetime study spanning the period 1971 until 2007. The content is focussed on high quality audio where the motivation and intention has been to nurture inventive understanding so as to extend the boundaries of electronic systems to match the exacting requirements for high quality audio processing and sound reproduction. Significantly, this subject has also formed a stimulating framework in which to mentor students for research and development within the audio industry and academia. The thesis is composed of 45 refereed Journal papers, where 24 are of single authorship with the remainder co-authored both with research students under my academic supervision and with industrial collaborators. A substantial cohort of additional material is also described in international convention papers; these are listed in Appendix 1. In addition, three book chapters have been written together with a number of articles aimed at a broader readership (Section 2-3, Appendix 2).

My contributions to the Audio Engineering community were recognised in 2006 by the award (Section 1-2, Appendix 2) of the Silver Medal of the Audio Engineering Society (AES); in addition I have received the publication prize of the AES Journal selected from volumes 45 and 46 (page 4-76) for my pioneering work on digital loudspeaker crossover systems. I have published in the AES Journal because it is widely acknowledged within the international audio community and undertakes rigorous peer reviewing. To the best of my knowledge (as of June 2007), I have on record the highest number of journal and convention publications of any author within the AES. In 1996 with Robert Stuart from Meridian and Professor Hiro Negishi from Canon, a technical pressure group was founded called the *Acoustic Renaissance for Audio* (ARA). This group subsequently influenced the specification of the DVD-based high-resolution audio format, designated *DVD-audio*, while in 2001 I was invited to present the keynote address at the 10th AES Regional Convention in Japan (paper C116, Appendix 1). I have also chaired the AES technical committee on High-Resolution Audio which has organized numerous workshops at AES conventions and in 2007 an International conference on *New Directions in High-Resolution Audio* was staged at Queen Mary College, London, where I was papers' chairman.

My research activities divide approximately into six themes which constitute the Sections of this thesis. I have gained substantial expertise in both analogue and digital systems and have made both theoretical and engineering contributions as audio has evolved from the analogue to the digital era. However, my philosophy has been to take a holistic perspective which has enabled me to gain insight into systems including digital-to-analogue conversion and switching power amplifiers that exploit noise shaping and oversampling processes. The foundations for this interest were established in the late 1960s while I was a BBC sponsored research student at Aston University under the learned supervision of Professor John Flood, where I studied delta-modulation (DM) and sigma-delta modulation (SDM). This proved extremely fortuitous and has been a pervasive topic especially as SDM is now widely employed in signal conversion and forms the core technology of the high

resolution audio format designated *Super-Audio CD* (SACD). I have reported (page 3-135) a variation of SDM using a parametric equalizer embedded within a high-order noise shaping loop, this technique when stabilized by a step-back in time procedure has enabled extremely high signal-to-noise ratios (theoretically > 200 dB) within the audio band to be achieved. I have also studied techniques using a digital variant of pulse-width modulation (PWM) that can be used for digital power amplifiers. A paper published in 1992 (page3-145) was one of the first to show how linearization of uniformly-sampled PWM could be achieved using a combination of dynamic inverse digital filtering, oversampling and noise shaping.

In the 1980s it was recognized that digital signal processors (DSP) had evolved to the state where low cost real-time audio processing could be used to implement the crossover filters used in multiple drive unit active loudspeakers. This concept and its subsequent development were facilitated by a research grant from Canon (Tokyo) which enabled a prototype system to be constructed. In 1986 the system was taken to the Canon Research Centre in Japan where it is believed to be the first open demonstration of a digital and active loudspeaker system. In the following years additional theoretical and practical work was undertaken both through project supervision and personal research (Section 4-3), where it was shown that DSP could facilitate sophisticated equalization to correct for both minimum and non-minimum phase linear distortion. In 2003 a digital loudspeaker concept designated a “Smart Loudspeaker” (page 4-116) was described. Conceptually this system merges micro-electroacoustic transduction with digital-to-analogue conversion using SDM techniques, where it has been shown theoretically that multiple, dynamically steer-able beams of sound can be formed. A complementary paper (page 4-146) has extended this work by describing array geometries selected to minimize the polar dependence of quantization distortion in systems where effectively signals are spread dynamically across the array surface as a function of signal amplitude.

A recurrent theme and personal speciality, has been the study of high performance analogue circuit techniques. In the late 1970s a novel power amplifier output stage topology was designed that exploited error feedback and was published subsequently in 1981 (page 2-1). This approach showed that an optimized fast-acting local feedback loop could dramatically reduce crossover distortion while simultaneously achieving low output impedance. Traditional overall negative feedback could then be applied to obtain extremely high degrees of linearity. This technique has enjoyed a wide discussion and was adopted by Meridian Audio, UK for several years in their audio power amplifier products. Other contributions within the analogue domain have included the application of “current drive” in loudspeaker systems (pages 2-36 and 4-23), the design of high performance voltage-controlled amplifiers (pages 2-74 and 2-79) and the refinement of analogue circuits using combinations of local feedback and feedforward (page 2-5).

In the late 1980s the evolution of perceptually-based audio coding triggered new research directions.

In collaboration with British Telecom Laboratories (BTL), I supervised external research student Dr M. P. Hollier in the area of perceptually-motivated objective techniques to perform subjective quality evaluation of coded speech (Section 5-1), where major telecommunications applications were identified. It is gratifying to report, that linked to the successful outcome of this project (pages 5-1, 5-8, 5-22 and 5-28), BTL subsequently spun-out the company Psytechnics headed up by Dr Hollier. Several research contracts from BTL were subsequently secured that enabled studies on spatial audio and echo suppression for teleconferencing. In addition, in 1997 an EPSRC funded research project entitled *A unified environmental human auditory perceptual measurement system* was undertaken. Papers resulting from this work are listed in Appendix 1 (papers C103, C108, C113), where in particular, a co-authored paper (paper C108) with research student Dr D. J. M. Robinson was presented at the EPSRC Conference Prep 2000 and received the best conference paper award.

The final topic (Section 6) embraces measurement techniques that reflect an interest in system evaluation and the desire to enhance the correlation of objective measurement of audio systems with their subjective performance. This work has lead to a number of bespoke computer-based procedures that focus on the use of pseudorandom noise sequences and Volterra system identification procedures. I was invited by the magazine *Hi-Fi News and Record Review* to apply these procedures to evaluate the CD12 compact-disc player from Linn Products, Scotland. These involvements have lead subsequently to further work which has been formally reported (page 6-43). The involvement with the audio media is part of a long-term desire to make my ideas available to the broader audio community and Section 2-3 in my CV (Appendix 2) lists several publications aimed at a more general readership.

The period from the 1970s to 2007 has witnessed radical changes in audio technology where sophisticated high-resolution audio formats have emerged that are complemented by extremely high performance audio electronics the best examples of which harmonize the capabilities of digital processes and carefully crafted analogue electronics. Some of these developments have been touched in some modest ways by my own endeavours either directly through inventive research or indirectly through the mentoring of students who have subsequently sought employment within the audio industry. This latter aspect has certainly been the most rewarding aspect of my academic career. Research continues and there are many new challenges. Current work is investigating efficient means of encoding multi-party teleconference systems that embrace spatial audio together with a radical approach to echo-cancellation that eliminates the need for adaptive filters. There is once again a dramatic paradigm shift ensuing as the audio industry engages with network delivery and distributed media servers. This is facilitating the introduction of network-enabled audio systems that challenge current practice and establish a demand for scalable spatial audio processes linked to enhanced techniques for capturing, processing and sharing of content.

Section 2: Analogue signal processing

2-1 Feedback, error correction and linear circuit techniques

2-1 DISTORTION CORRECTION IN AUDIO POWER AMPLIFIERS, Hawksford, M.J., *JAES*, vol.29, no.1/2, Jan/Feb 1981, pp.27-30 (paper selected from 65th AES Convention for publication)

2-5 DISTORTION CORRECTION CIRCUITS FOR AUDIO AMPLIFIERS, Hawksford, M.J., *JAES*, vol.29, no.7, 8, July/August 1981

2-13 FUZZY DISTORTION IN ANALOG AMPLIFIERS: A LIMIT TO INFORMATION TRANSMISSION?, Hawksford, M.J., vol.31, no.10, pp.745-754, October 1983

2-23 OPTIMIZATION OF THE AMPLIFIED-DIODE BIAS CIRCUIT FOR AUDIO AMPLIFIERS, Hawksford, M.J., *JAES*, vol.32, no.1/2, pp.31-33, Jan/Feb 1984

2-26 REDUCTION OF TRANSISTOR SLOPE DISTORTION IN LARGE SIGNAL AMPLIFIERS, Hawksford, M.J., *JAES*, vol.36, no.4, pp.213-222, April 1988

2-36 TRANSCONDUCTANCE POWER AMPLIFIER SYSTEMS FOR CURRENT-DRIVEN LOUDSPEAKERS, Mills, P.G.L., Hawksford, M.O.J., *JAES*, vol.37, no.10, pp.809-822, October 1989

2-50 DIFFERENTIAL-CURRENT DERIVED FEEDBACK (DCDF) IN ERROR CORRECTING AUDIO AMPLIFIERS, Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 141, no 3, June 1994, pp 227-236

2-60 QUAD-INPUT CURRENT-MODE ASYMMETRIC CELL (CMAC) WITH ERROR CORRECTION APPLICATIONS IN SINGLE-ENDED AND BALANCED AUDIO AMPLIFIERS, Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 143, no 1, Feb. 1996, pp 1-7

2-67 RELATIONSHIP BETWEEN NOISE SHAPING AND NESTED DIFFERENTIATING FEEDBACK LOOPS, Hawksford, M.O.J. and Vanderkooy, J., *JAES*, vol. 47, no. 12, pp 1054-1060, December 1999

2-2 Translinear (variable gain) amplifiers

2-74 LOW-DISTORTION PROGRAMMABLE GAIN CELL USING CURRENT-STEERING CASCODE TOPOLOGY, Hawksford, M.J., *JAES*, vol.30, no.11, November 1982

2-79 TOPOLOGICAL ENHANCEMENTS OF TRANSLINEAR TWO-QUADRANT GAIN CELLS, Hawksford, M.O.J. and Mills, P.G.L., *JAES*, vol.37, no.6, pp 465-475, June 1989

ENGINEERING REPORTS

Distortion Correction in Audio Power Amplifiers*

M. J. HAWKSFORD

Audio Research Group, Department of Electrical Engineering Science, University of Essex, Colchester, UK

An audio power amplifier design technique is presented which has the property of minimizing the nonlinear distortion that is generated in class A and class AB output stages.

A modified feedback technique has been identified that is particularly suited to the design of near-unity gain stages. The technique can linearize the transfer characteristic and minimize the output resistance of the output stage. Consequently it is possible to design a power amplifier that uses fairly modest overall negative feedback, yet attains minimal crossover distortion together with an adequate damping factor.

A generalized feedforward–feedback structure is presented from which a system model is derived that can compensate for both nonlinear voltage and nonlinear current transfer characteristics. From this theoretical model, several circuit examples are presented which illustrate that only circuits of modest complexity are needed to implement the distortion correction technique.

In conclusion a design philosophy is described for an audio power amplifier which is appropriate for both bipolar and FET devices, whereby only modest overall negative feedback is necessary.

0 INTRODUCTION

This paper discusses the problems of minimizing crossover distortion in class A and class AB audio power amplifiers. Traditionally output-voltage-derived negative feedback and appropriate biasing of the output transistors have been applied with varying degrees of success in an attempt to achieve acceptable linearity. However, since all transistors exhibit nonlinearity and as, in particular, the output transistors are generally operated into cutoff, successful suppression of the distortion using these techniques is limited.

There are several fundamental problems that can be encountered when using negative feedback to minimize distortion in power amplifiers:

1) Bipolar power transistors are usually of limited bandwidth (typical f_T = 1–5 MHz); thus if nondynamic behavior is required within the audio band, loop gains of only 30 dB are possible.

2) Since crossover distortion is transient in nature and of wide bandwidth, the inevitably falling high-frequency loop gain, together with the resulting loop delay, severely limits the degree of distortion suppression possible.

3) In output-voltage-derived negative feedback amplifiers the distortion which is generated by the output transistors is fed back to the input circuitry. Consequently the pre-output stages process both the desired input signal and the output stage distortion. Thus intermodulation is impaired, especially as the distortion bandwidth can significantly exceed that of the audio signal.

4) If the output resistance of the output stage is nonzero (independent of any overall feedback), the loudspeaker load is an integral component in the feedback loop. Hence if the load exhibits nonlinearity, then distortion components are again fed back to the amplifier's input stage.

A technique is described in this paper which can dramatically linearize the output device characteristics with respect to both voltage transfer and current transfer. Hence an amplifier philosophy evolves that helps to reduce the problems outlined in 1)–4).

* Presented at the 65th Convention of the Audio Engineering Society, London, 1980 February 25–28.

 0004-7554/81/010027-04$00.75

1 THEORETICAL MODEL

The principle of the distortion cancellation technique can be described by considering the generalized error feedback structure shown in Fig. 1. In this network there is error sensing feedforward as well as feedback applied around the nonlinear element N, where in the most general case the input N is unspecified. The error signal used in the system is defined as the difference between the input and the output of N. Thus if N is ideal (that is, $N = 1$), then the error signal is zero and no correction is applied. However, in all practical amplifiers N will deviate from unity, thus the error signal represents the exact distortion due to N.

1.1 Analysis

Let V_n and $N(V_n)$ be the input and output of the N network. Thus examination of the signals in Fig. 1 reveals:

$$V_{out} = N(V_n) + b\{V_n - N(V_n)\}$$

$$V_n = V_{in} + a\{V_n + N(V_n)\}.$$

Eliminating V_n,

$$V_{out} = N(V_n)\left\{(1 - b) - \frac{ab}{(1 - a)}\right\} + \frac{b}{(1 - a)} V_{in}. \quad (1)$$

If

$$(1 - a) = b \quad (2)$$

then

$$V_{out} = V_{in}. \quad (3)$$

Thus providing that stability is maintained and V_n remains finite, distortion cancellation results when Eq. (2) is enforced.

The result [Eqs. (2) and (3)] indicates that there is a continuum of solutions extending from an error feedback system through to an error feedforward system.

It is interesting to note that the input of N is unspecified. It may therefore be derived directly from V_n or indeed any other point within the structure, providing that stability is maintained. For example, by putting $a = 0$, $b = 1$, the classic feedforward system results, where if the input of N is derived from the output of the error difference amplifier, then the Quad [1], [2] feedback structure results (see dashed connection in Fig. 1).

In this paper we consider the opposite extreme where $a = 1$, $b = 0$, and the input of N is equal to V_n. This system is of the type first discussed by Llewellyn in 1941 [3] in relation to valve amplifiers and later by Cherry [4] in 1978. It will now be shown that this feedback technique is particularly relevant to the design of unity-gain follower-type output stages, where with modest circuitry a dramatic improvement in performance is possible. The theory is extended to show that linearization of devices with nonlinear current gain is also feasible.

2 CIRCUIT TOPOLOGIES FOR OUTPUT-STAGE LINEARIZATION

Power amplifiers generally use bipolar output transistors which exhibit low nonlinear current gain. Consequently when such devices are used in a complementary emitter-follower configuration, the transformed loudspeaker load as seen by the base terminals is rendered nonlinear and therefore contributes to the amplifier distortion.

If distortion correction feedback is configured to include input current sensing, it is possible to compensate for changes in current gain. Thus when combined with voltage error sensing feedback, a unity-gain stage results which can be driven from a stage with a finite-output resistance.

In Fig. 2 the schematic of a system with both voltage- and current-sensing circuitry is shown, where the system is configured to illustrate how a practical circuit (Fig. 3) may be realized.

Analysis shows that when

$$k_1 = 1 + \frac{2R_1}{R_2} \quad (4)$$

$$R_1R_3 = R_2R_4 \quad (5)$$

the voltage gain is unity even when the base currents of T_1 and T_2 are finite and V_{BE}/I_E introduces nonlinearity.

As a point of design interest, the resistor R_1 includes

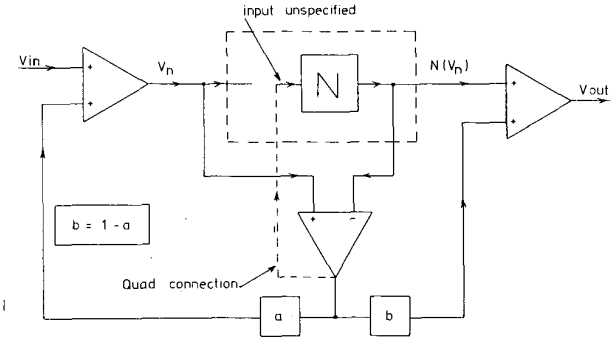


Fig. 1. Generalized feedback-feedforward structure.

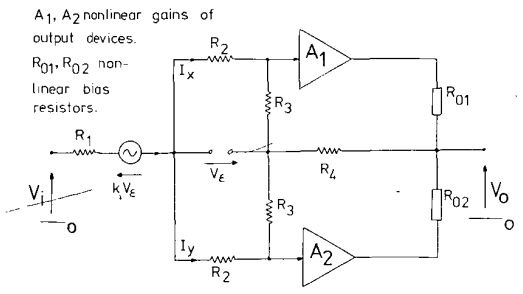


Fig. 2. Current- and voltage-error-sensing feedback.

the output resistance of the driving stage. Consequently the driving amplifier is not required to have zero output resistance.

2.1 Corollary

Since the voltage gain is unity, it follows that the output resistance of the stage is zero, even when the output resistance of the driving stage is finite. As a result, an amplifier that uses this error-correction feedback system does not in principle have to rely upon an overall output-voltage-derived negative feedback loop to achieve adequate loudspeaker damping. Also, the loudspeaker load is then effectively decoupled from the overall feedback loop, and it is this factor that prevents loudspeaker-generated distortion products from reaching the input circuitry of the power amplifier.

Three practical output stage circuits are shown in Figs. 3–5. The circuit of Fig. 3 has both voltage and current sensing and is derived from Fig. 2. However, if the output devices have adequate current gain (such as MOSFET or Darlington transistors), then current sensing is unnecessary. As a result, the much simplified circuits of Figs. 4 and 5 are illustrated to show the modest circuit requirements that are needed to realize only error-voltage sensing. The circuit of Fig. 5 is particularly attractive as the transistors T_3, T_4 form both a complementary error difference amplifier as well as "amplified diodes" for biasing the output transistors.

3 CONCLUSIONS

This paper has described an approach to power amplifier design where the nonlinear distortion generated by the output transistors is compensated by simple fast-acting local circuitry which can result in a high degree of linearity that is appropriate to class A and class AB follower-type output stages.

The technique should find favor among designers who adhere to the low-feedback school of design, as corrective feedback is only applied when distortion in the output stage is generated. If, therefore, the output stage N is designed to be as linear as possible, a fact that can be aided by parallel connection of output transistors, then only minimal error signals result.

Since output stage and loudspeaker generated distortions are in principle isolated from the input stages, these stages are required only to produce modest voltage gains, as large loop gains are not required in an attempt to produce a linear amplifier. Consequently the loop gain is low and the loop bandwidth can be high, enabling a nondynamic loop behavior well in excess of the audio bandwidth.

In practical amplifier design, the sensitivity of adjustment of the balance conditions depends largely on the quiescent bias current of the output transistors, where critical adjustment results only under extremely low biasing. It has been found that for normal bias levels, adjustment is noncritical, also that sensitivity is aided by modest overall feedback.

Several prototype circuits have been investigated where the technique has proved effective. In these amplifiers no stability problems have been encountered other than with the susceptibility to oscillation of power Darlington transistors which appear critical on layout. In fact,

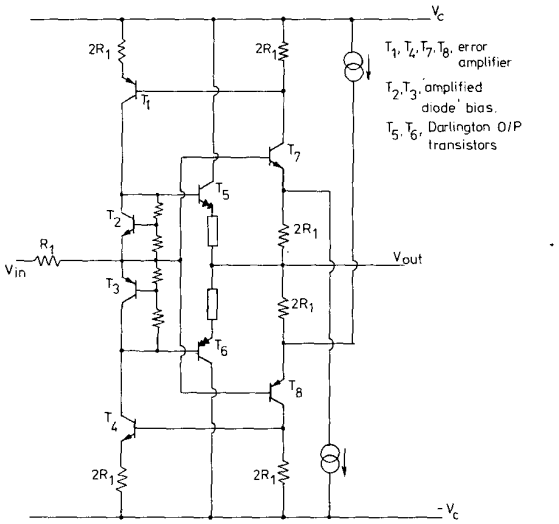


Fig. 4. Example of voltage-error-sensing circuit.

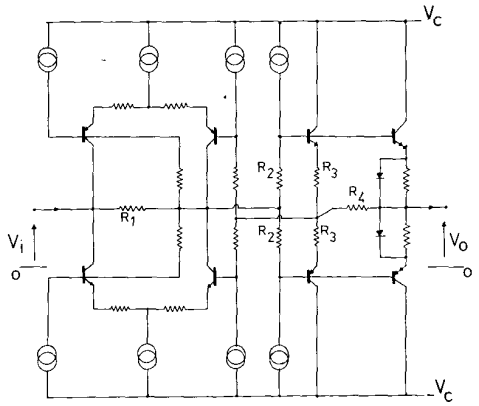


Fig. 3. Circuit schematic of current- and voltage-error-sensing output stage.

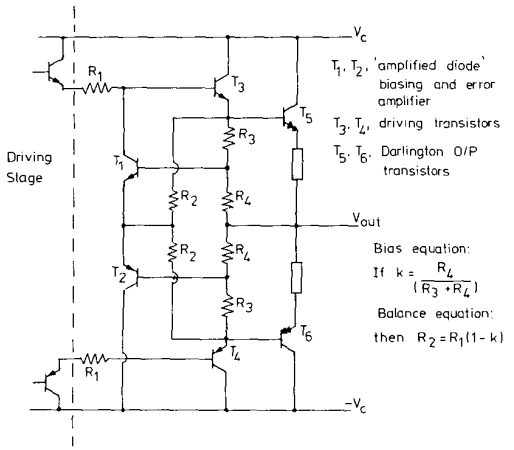


Fig. 5. Voltage error sensing circuit using amplified diodes as error amplifier.

due to the low loop gain, load-dependent instability is minimal, though standard series Zobel circuitry was employed. In practice the bandwidth of the correction circuitry is high which enables fast correction of output-stage nonlinearities. In fact, it is partly the speed of the correction loop that enables a greater suppression of distortion compared with an overall feedback system.

4 REFERENCES

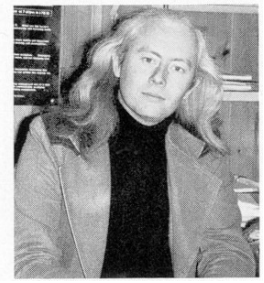
[1] P. J. Walker and M. P. Albinson, "Current Dumping Audio Amplifier," presented at the 50th Convention of the Audio Engineering Society, London, 1975 March 4–7.

[2] P. J. Walker, "Current Dumping Audio Power Amplifier," *Wireless World*, vol. 81, pp. 560–562 (1975 Dec.).

[3] F. B. Llewellyn, "Wave Translation Systems," U.S. Patent 2,245,598, 1941 June 17.

[4] E. M. Cherry, "A New Result in Negative-Feedback Theory and Its Application to Audio Power Amplifiers," *Int. J. Circuit Theory Appl.*, vol. 6, pp. 265–288 (1978 July).

THE AUTHOR



Malcolm J. Hawksford was born in Shrewsbury, England, in 1947. His professional education was at the University of Aston in Birmingham where he studied electrical engineering from 1965–68 and was subsequently awarded a first class B.Sc. degree. In 1968 he obtained a BBC Research Scholarship for three years of postgraduate study at Aston University. His research subject was the application of Delta modulation to color television systems. This work resulted in the award of a Ph.D. degree in 1972.

In 1971 he obtained a lectureship at the University of Essex in the electrical engineering science department where he has taught subjects including electromagnetic theory, audio engineering, digital communications, circuit design and television engineering. At Essex he developed an Audio Research Group where projects on amplifier design, loudspeaker crossover design, analogue-to-digital conversion and music synthesis have been undertaken.

Dr. Hawksford is a member of the Audio Engineering Society, the IEE, the Royal Television Society, and is a chartered engineer.

His hobbies include listening to music, designing audio equipment, home computing and motorcycling.

PAPERS

Distortion Correction Circuits for Audio Amplifiers*

M. J. HAWKSFORD

University of Essex, Department of Electrical Engineering Science, Colchester, CO4 3SQ, United Kingdom

Circuit topologies are introduced which should prove of use to the circuit designer of analog audio amplifiers. The objective is to produce circuits of modest complexity that overcome the nonlinearities inherent in single-transistor and long-tail pair circuits. This allows amplifiers with excellent linearity to be designed without resorting to overall negative feedback with high loop gains. To aid comparison of circuit nonlinear behavior, a parameter called the incremental distortion factor (IDF) is introduced and discussed.

0 INTRODUCTION

Most modern transistor amplifiers use either a single transistor or a pair of transistors in the input circuitry. It is argued that if this stage is cascaded with adequate gain, then by the expedience of overall negative feedback, the input devices will operate within the limits for small-signal operation and thus yield good overall linearity.

Often a consequence of this design philosophy is poor dynamic performance of the input circuitry, where modest input overload can result in gross distortion. There are simple circuit modifications that can be introduced: an increase in device operating current, though possibly at the expense of the noise factor; the introduction of local negative feedback (emitter degeneration) which reduces stage gain but enhances linearity and overload performance, again at the expense of the noise factor.

The aim of this paper is to introduce circuit topologies that enhance the nonlinear performance of amplifier gain cells without recourse to high overall negative feedback. It is considered by this author that the combination of high loop gain together with its inevitable dynamic performance (dominant pole) when compounded with nonlinear elements can result in poor transient distortion characteristics, especially when complex signals are being processed. Since the signals being amplified are rendered more complex due to these nonlinearities falling within a dynamic negative feedback loop, then intermodulation products result which are effectively time smeared. In the limit this must determine the ultimate resolution of an amplifier, which is its ability to transfer fine signal detail in the presence of complex signals.

The only rational methodology to minimize these attributes of nonlinear distortion is to use gain cells that are inherently linear over a wide range of their transfer characteristics and are essentially nondynamic with predictable gain characteristics. Such gain cells can then be used with amplifiers with overall negative feedback without detriment to the intermodulation performance. However, the use of linear circuitry may well render the need for high negative feedback unnecessary.

This paper investigates and catalogs examples of gain cells that generally exhibit good linearity and dynamic range. The circuits should prove of use to designers of both discrete and integrated circuitry, although some design examples which are particularly relevant to integrated-circuit fabrication are included.

In order to facilitate the comparison of various circuit topologies, a parameter called incremental distortion factor (IDF) is introduced. The IDF is related to the change in slope of the transfer characteristic with the input signal and is useful for quantifying nonlinearity under large-signal conditions.

1 PRINCIPLES OF DISTORTION CORRECTION

Three methods are identified in this section to enhance the linearity of gain cells that may already use either local or overall negative feedback within an amplifier structure. (See [1–5] for background.)

1.1 Complementary Nonlinear Stages in Cascade

If a stage has a predictable nonlinearity, then by using a nonlinear stage with a complementary transfer characteristic, overall linearity is possible (Fig. 1). This technique is, for example, used in translinear multiplier stages and in a modified form is the principle of complementary compandors.

* Manuscript received 1981 January 22.

1.2 Device Linearization

This method involves matching device nonlinearities as with the long-tailed pair, where the transconductance is linearized approximately by keeping r_{ep} constant over a wider range of emitter current compared with a single transistor r_e over the same current range. Thus for single transistors,

$$r_{e1} = \frac{\partial V_{be1}}{\partial I_{e1}} \tag{1}$$

$$r_{e2} = \frac{\partial V_{be2}}{\partial I_{e2}} \tag{2}$$

and for a long-tail pair of transistors,

$$r_{ep} = \frac{\partial V_{be1}}{\partial I_{e1}} + \frac{\partial V_{be2}}{\partial I_{e2}} \tag{3}$$

Comparing r_{ep} with r_{e1} or r_{e2} for a given change in emitter current, r_{ep} exhibits greater linearity.

1.3 Error Feedforward and Feedback Distortion Correction

A technique [6] that was recently reported for linearizing near unity gain output stages in analog power amplifiers uses in general a combination of error feedforward and error feedback. Fig. 2 illustrates the method in schematic form.

Analysis shows that when

$$b = (1 - a) \tag{4}$$

then

$$S_{out} = S_{in} \tag{5}$$

where a and b are constrained to values between 0 and 1.

If $a = 1$ and $b = 0$, the system becomes pure error feedback, while if $a = 0$ and $b = 1$, pure feedforward error correction results.

When the balance equation (4) is satisfied, the effects of nonlinearity in the general network N are minimized, and the output parameters S_{out} and S_{in} become linearly related. Though it is inferred that these parameters are voltages, in general they may be any suitable combination of current and voltage, such as voltage in, current out, which is of particular importance for the input stage of an audio amplifier.

Although this principle can be applied to an overall amplifier, it is recommended that the technique be restricted to single stages (which in turn can be compounded to form a complete amplifier), as this permits near nondynamic stage performance and minimizes signal distortion that can be generated in cascaded high gain, low local-feedback amplifier stages.

In practical amplifier design it is possible to compound the techniques outlined in this section to produce amplifier stages of high linearity. It is also possible, within limits, to trade off circuit complexity against performance and to choose a technique that is best suited to a particular amplifier application. In the following sections, circuit examples will be discussed to indicate how predictable amplifiers can be designed and that by the careful choice of design techniques enhanced performance results.

2 INCREMENTAL DISTORTION FACTOR (IDF)

The prime nonlinearity of a transistor which is operated with near constant collector–base voltage is defined by the exponential relationship

$$I_e = I_0 \exp\left(\frac{qV_{be}}{KT}\right) \tag{6}$$

where

I_e = emitter current
I_0 = base-emitter diode saturation current
K = Boltzman's constant
q = charge on electron
T = junction temperature (degrees Kelvin)

Some deviation from this relationship will occur, but is of little consequence here.

Thus when a transistor is used as a transconductance amplifier, nonlinear distortion will result. In order to attempt to quantify the nonlinearity, we introduce the term *incremental distortion factor* (IDF). In essence this term is a measure of the change in incremental gain of a stage to the small-signal gain. In practical circuits the IDF can most simply be expressed as a function of one or more variables. Hence by observing the variation of IDF with these parameters, an accurate measure of nonlinear performance can be made.

To explain the IDF in more detail, we proceed by analyzing first the nonlinear behavior of a simple single-transistor stage with local emitter degeneration and second the performance of a two-transistor long-tail pair. These results are also of use as a reference to allow comparison with the more elaborate gain cell topologies presented in later sections.

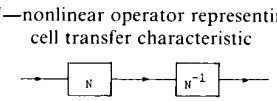


Fig. 1. Complementary linearization.

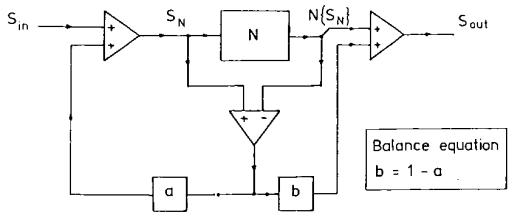


Fig. 2. Error feedforward and feedback distortion correction.

2.1 Distortion Characteristics of a Single-Transistor Cell

A single-transistor cell is shown in Fig. 3. We assume the base current to be negligible. Hence from Eq. (6),

$$V_{be} = \frac{KT}{q} \ln\left(\frac{I_e}{I_0}\right) . \quad (7)$$

Let $\alpha = KT/q$. Therefore

$$V_{be} = \alpha \ln\left[\frac{I_e}{I_0}\right] . \quad (8)$$

Applying Kirchhoff's law to the circuit shown in Fig. 3 and eliminating V_{be} [using Eq. (8)],

$$V_{in} = (i - I_x)R + \alpha \ln\left(\frac{I + i}{I_0}\right) \quad (9)$$

(bias currents I_x, I are shown in Fig. 3). In this simple example V_{in} is a function of a single variable i, that is,

$$V_{in} = f(i).$$

By differentiation we obtain

$$dV_{in} = \left(\frac{dV_{in}}{di}\right) di$$

therefore

$$dV_{in} = R\,di + \frac{\alpha}{I + i}\, di.$$

Extracting linear and nonlinear components,

$$\underbrace{dV_{in}}_{I/P \text{ voltage}} = \underbrace{\left(R + \frac{\alpha}{I}\right)di}_{\text{linear component}} - \underbrace{\left\{\frac{\alpha i}{I(I + i)}\right\}di}_{\text{nonlinear component}} . \quad (10)$$

Eq. (10) relates incremental changes in current and voltage expressed as a function of the bias current I and the present state of signal current i. It is essentially the tangent to the transfer characteristic for transconductance. For linearity, dV_{in} and di must be related by a constant multiplier. However, Eq. (10) reveals that the incremental gain is a function of i, which represents a nonlinear process. We define the IDF $N(\ldots)$ as

$$N(x) = \left[\frac{\text{nonlinear incremental gain component}}{\text{linear incremental gain component}}\right] . \quad (11)$$

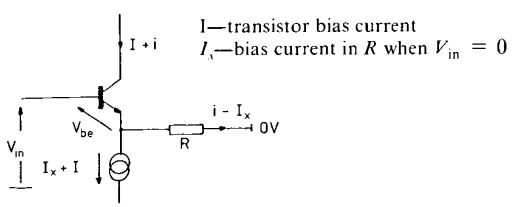


Fig. 3. Single-transistor cell.

$N(\ldots)$ is shown here to be a function of a single variable x. However, in later sections the definition is extended to functions of several variables.

Defining x, the transistor loading factor, as the ratio of signal current i to bias current I for the single-stage transistor amplifier,

$$x = \frac{i}{I} . \quad (12)$$

Hence from Eqs. (10)–(12) we obtain

$$N(x) = \frac{-x}{1 + x}\left(\frac{\alpha}{\alpha + IR}\right) . \quad (13)$$

Eq. (13) reveals that the IDF is an asymmetric function of x, as would be anticipated for a single-transistor nonlinearity. The advantage of this format is that since x is a direct measure of the signal loading of a transistor, then if large values of x result in low values of IDF, this is an expression of near linear performance. In practice x can range from −1 to +1, though usually (except under overload) x will remain well within these limits.

The main advantage of the IDF is that it permits a comparison of circuits with respect to their nonlinear performance, even when complex multiple distorting mechanisms coexist.

2.2 Distortion Characteristic of the Long-Tail Pair Cell

A treatment similar to that presented in Section 2.1 is applied here to the long-tail pair circuit shown in Fig. 4. From Kirchhoff's law,

$$V_{in} = iR + (V_{be1} - V_{be2}) .$$

Applying Eq. (8) to each transistor,

$$V_{in} = iR + \alpha \ln\left[\frac{I + i}{I - i}\right] . \quad (14)$$

Differentiating and extracting linear and nonlinear components,

$$dV_{in} = \left(R + \frac{2\alpha}{I}\right) di + \left\{\frac{2\alpha i^2}{I(I^2 - i^2)}\right\} di. \quad (15)$$

We obtain the IDF using the definition of Eq. (11):

$$N(x) = \frac{x^2}{(1 - x^2)}\left(\frac{2\alpha}{2\alpha + IR}\right) . \quad (16)$$

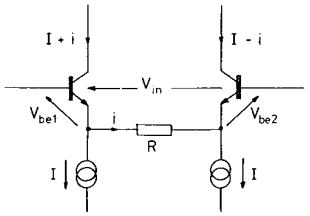


Fig. 4. Long-tail pair circuit.

Comparing Eq. (16) with Eq. (13), the differences in nonlinearity can be compared directly as a function of the transistor loading parameter x. These equations also form a reference for the circuits presented in the following sections.

3 GAIN CELL LINEARIZATION USING FEEDFORWARD ERROR CORRECTION

This section presents a series of circuit topologies that exploit error correction feedforward as outlined in Section 1. Where appropriate, the IDF is evaluated as a means of circuit comparison. All the circuits shown use bipolar transistors, though in most cases adaptation to FET devices should be feasible.

3.1 Single-Stage Feedforward Error Correction

The technique exploited in the circuit of Fig. 5 was derived from Fig. 2, where $a = 0$ and $b = 1$. Essentially when an input signal V_{in} is applied to the base of the input transistor, the resistor R_1 is used as a reference for converting V_{in} to a current. However, due to V_{be1} the voltage across R_1 is less than the input voltage. Hence by using a differential amplifier to measure the error voltage V_{be1}, a corrective current i_2 can be summed with i_1 to compensate almost exactly for the lost current. The transconductance is then almost independent of V_{be1}. Since $V_{be1} < V_{in}$, good linearity results. The main advantage of this circuit is that linearity can be achieved with only modest values of R_1, a fact that increases the transconductance of the cell, yet minimizes Johnson noise due to R_1.

The simplest method of adding the main current i_1 with the error correction current i_2 is to parallel the two collectors. However, if both collector currents of each half of the difference amplifier are used by introducing a current mirror, then either the value of R_2 can be increased, which improves linearity, or the value of R_1 can be reduced, which reduces Johnson noise and increases transconductance.

An example of a more practical amplifier is illustrated in Fig. 6, where biasing requirements and current mirror are shown.

We assume that the output signal current i_0 is derived as

$$i_0 = i_1 + \lambda i_2 \tag{17}$$

where generally λ has a value of 1 or 2 (Fig. 6 assumes $\lambda = 2$). The circuit equations are as follows:

$$V_{in} = (i_1 - I_x)R_1 + V_{be1} \tag{18}$$

$$V_{be1} = (i_2 + I_y)R_2 + (V_{be2} - V_{be3}) \tag{19}$$

$$V_{be1} = \alpha \ln\left[\frac{I_1 + i_1}{I_0}\right] \tag{20}$$

I_x, I_y—bias currents in R_1 and R_2 when $V_{in} = 0$

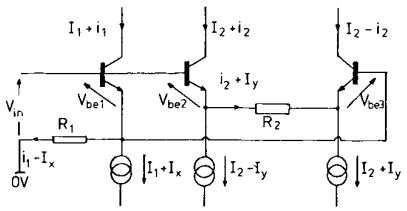


Fig. 5. Single-stage input device with feedforward error correction.

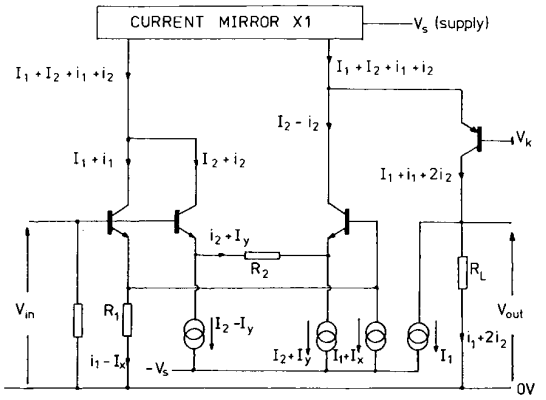


Fig. 6. Practical amplifier stage using a single input transistor with feedforward error correction.

$$V_{be2} - V_{be3} = \alpha \ln\left[\frac{I_2 + i_2}{I_2 - i_2}\right] . \tag{21}$$

Thus

$$V_{in} = (i_1 - I_x)R_1 + (i_2 + I_y)R_2 + \alpha \ln\left[\frac{I_2 + i_2}{I_2 - i_2}\right] .$$

Since

$$V_{in} = f(i_1, i_2)$$

then

$$dV_{in} = \frac{\partial V_{in}}{\partial i_1} di_1 + \frac{\partial V_{in}}{\partial i_2} di_2 .$$

Therefore

$$dV_{in} = R_1 \left[di_1 + \left(\frac{I_2 R_2 + 2\alpha}{R_1 I_2}\right) di_2 \right] + \frac{2\alpha}{I_2}\left[\frac{i_2^2}{I_2^2 - i_2^2}\right] di_2 .$$

By comparison with Eq. (17),

$$\lambda = \frac{I_2 R_2 + 2\alpha}{I_2 R_1} . \tag{22}$$

Expressing di_2 as a function of di_0, we then obtain the IDF

$$N(x, y) = \frac{2\alpha^2 y^2}{\lambda I_1 I_2 R_1^2[(1 - x)(1 - y^2 R_2/\lambda R_1) + (2\alpha/R_1 I_1)(1 - y^2)]} \tag{23}$$

where

$$x = \frac{i_1}{I_1} \quad \text{and} \quad y = \frac{i_2}{I_2}.$$

Since $|y| < |x|$ and $|x| < 1$, then Eq. (23) indicates that a substantial reduction in nonlinearity is possible.

3.2 Symmetrical Long-Tail Pair with Feedforward Error Correction

The primary distortion mechanism of a single transistor is the I_e/V_{be} relationship. If a long-tail pair is chosen, then the primary distortion is reduced, as discussed in Section 2.2, where it was also shown that the nonlinearity is symmetrical about the operating point.

This section investigates the use of feedforward error correction applied to a single long-tail pair. The IDF is stated in Eq. (24), the analysis being similar to that of Section 3.1:

$$N(x, y) = \frac{4\alpha^2 y^2}{\lambda I_1 I_2 R_1^2[(1 - x^2)(1 - y^2 R_2/\lambda R_1) + (2\alpha/I_1 R_1)(1 - y^2)]} \tag{24}$$

where λ, x, y, and i_0 are as defined in Section 3.1. The circuit is given in Fig. 7.

Eq. (24) reveals that the IDF is of a lower order due to the square-law dependence on x and that the nonlinearity is symmetrical about the quiescent operating point.

This particular configuration is applicable to amplifier input stages where offset cancellation of base-emitter junctions is useful in establishing dc biasing of the complete amplifier. Note, however, that there is no requirement for accurate device matching within either the long-tail pair or the error amplifier to achieve useful linearization. (Matching is necessary for accurate dc conditions, but this is a separate problem and may not be of importance in ac-coupled stages.)

3.3 Cascaded Gain Cells to Derive Differential Output Currents

A circuit application may require a differential output current from the gain cell. Since this feature is absent from the circuits presented in Figs. 5, 6 and 7, we consider here modifications that result in differential output currents.

The principle is illustrated in the basic schematic shown in Fig. 8, where identical gain cells are compounded within a cascade topology. This technique results in a fully complementary cell with enhanced IDF due to a reduction in primary distortion by sharing the input signal between cells.

Two circuits are presented in Figs. 9 and 10, which are formed by cascading the respective circuits of Figs. 5 and 7.

3.4 Nested Feedforward Error Correction Amplifier

To conclude this section on feedforward error correction, it should be noted that the error amplifier can be nested to yield even further distortion reduction, where effectively an error amplifier is used to compensate for the main error amplifier. However, in such circuits it is likely that other sources of distortion (other than the V_{be}/I_e nonlinearity) will then be dominant. Also, such circuits become somewhat complex, and the overall improvements are likely to be small. In fact for a given total current consumption in a gain cell, increasing the error amplifier current I_2 will produce a useful reduction in distortion, since the error amplifier loading factor is reduced as a function of y^2. Further enhancement can be obtained by using the modified amplifier cells to be presented in Section 5, in particular the cell shown in Figs. 15 and 16.

4 GAIN CELL LINEARIZATION USING FEEDBACK ERROR CORRECTION

Circuit topologies similar to that of Section 3 can be designed which rely upon the error signal being fed back to the gain cell input. This corresponds to the system

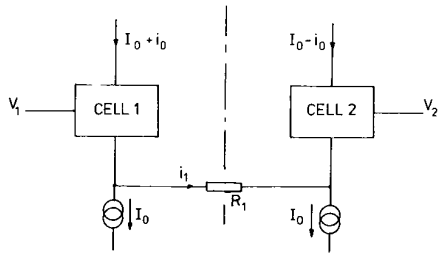


Fig. 8. Basic cascade of two identical gain cells.

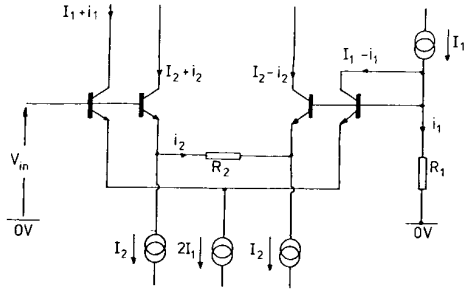


Fig. 7. Single long-tail pair with feedforward error correction.

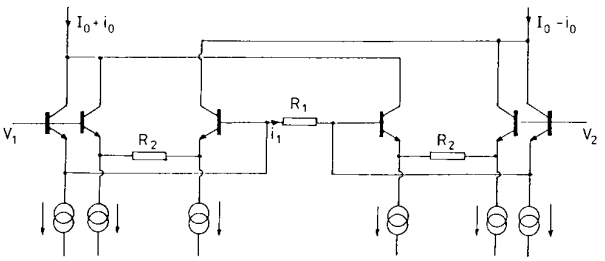


Fig. 9. Single long-tail pair with dual feedforward correction amplifiers (cascade formed from cell shown in Fig. 5).

diagram of Fig. 2, where $a = 1$ and $b = 0$. In these circuits the error signal is generally a current.

Figs. 11 and 12 show two examples which can be compared directly with the feedforward versions illustrated in Figs. 5 and 7.

Compound circuits similar to those described in Section 3.3 also can be derived by cascading gain cells with error correction feedback. These should be proven useful where differential output currents are required in fully symmetrical circuits (see Fig. 8).

The circuit equations are as follows. For the input transistor (Fig. 11),

$$V_{in} = V_{be} + (i_1 - i_2 + I_y)R_1$$

$$V_{be} = \alpha \ln\left[\frac{I_1 + i_1}{I_0}\right]$$

and for the error amplifier,

$$V_{be} = (I_x + i_2)R_2 + \alpha \ln\left[\frac{I_2 + i_2}{I_2 - i_2}\right] .$$

Let

$$R_2 = \frac{R_1 I_2 - 2\alpha}{I_2} . \qquad (25)$$

Differentiating V_{in},

$$dV_{in} = R_1\, di_1 - \left[\frac{2\alpha^2 i_2^2 di_1}{I_1 I_2^3 R_1 (1 + i_1/I_1)(1 - i_2^2 R_2/I_2^2 R_1)}\right]. \qquad (26)$$

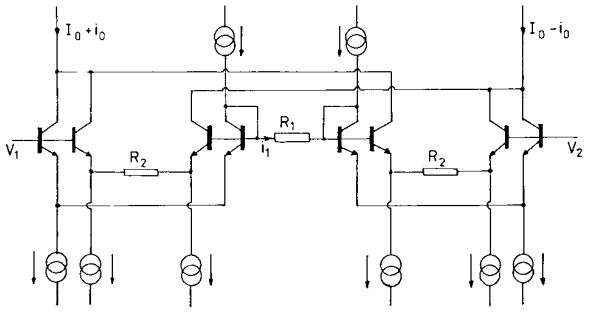


Fig. 10. Dual long-tail pair circuits with dual feedforward correction amplifiers (cascade formed from cell shown in Fig. 7).

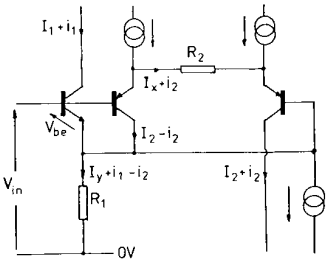


Fig. 11. Single input transistor with error correction feedback.

Therefore

$$N(x, y) = \frac{-2\alpha^2 y^2}{I_1 I_2 R_1^2 (1 + x)(1 - y^2 R_2/R_1)} . \qquad (27)$$

Again the loading parameters x and y are defined as in Section 3.

It is interesting to note that Eq. (25) represents a balance equation which minimizes the output current i_1 dependence on i_2 and allows R_1^{-1} to determine the transconductance exactly.

A similar analysis for the circuit in Fig. 12 gives the IDF as

$$N(x, y) = \frac{-4\alpha^2 y^2}{I_1 I_2 R_1^2 (1 - x^2)(1 - y^2 R_2/R_1)} , \qquad (28)$$

where the balance is again determined by Eq. (25).

These results show that the feedback circuits give virtually the same performance as their feedforward counterparts, and for practical circuits the performance should be essentially identical.

5 INDIRECT DISTORTION CANCELLATION TOPOLOGIES

A significant improvement over the standard long-tail pair can be realized by using matched transistors. In these circuits it is assumed that the I_e/V_{be} characteristics are essentially identical. As examples Figs. 13 and 14 show indirect error correction.

In Figs. 13 and 14 transistors T_1, T_3, and T_2, T_4 are matched, and since they carry the same emitter current (excluding the small base current), the base–emitter voltages are identical. Thus an error-sensing difference am-

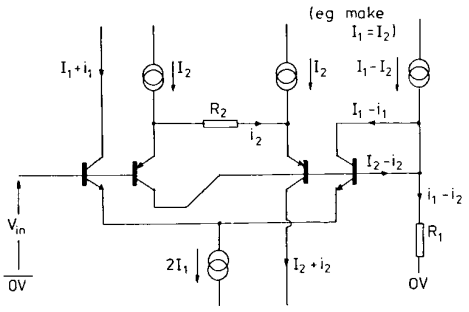


Fig. 12. Long-tail pair with error correction feedback.

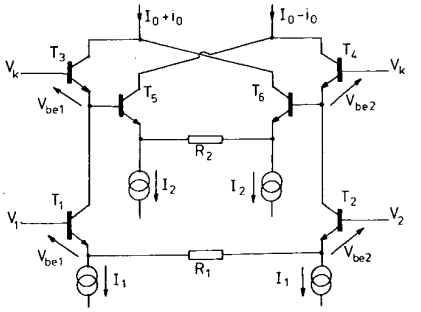


Fig. 13. Indirect error feedforward.

plifier can measure the error voltage ($V_{be1} - V_{be2}$) indirectly and compensate either by feedforward or by feedback.

The advantages of these circuits are that only a single error amplifier need be used, and transistors T_3, T_4 form a cascode configuration, which enhances bandwidth and linearity.

The IDFs for the error feedforward and error feedback circuits should compare with the circuits of Fig. 7 [Eq. (24)] and Fig. 12 [Eq. (28)], respectively, provided that there is accurate transistor matching, and base currents are neglected (i.e., high β transistors).

Finally a circuit is presented in Fig. 15 which combines the advantages of error feedforward with indirect error sensing to minimize nonlinearities.

We assume that all transistors are matched in terms of I_e/V_{be} nonlinearity and collector–base current gain β. The values of currents and voltages are shown in Fig. 15.

$$V_{in} = (V_1 - V_2) = (V_{be1} - V_{be2}) + (V_{be3} - V_{be4}) + i_1 R$$

where

$$(V_{be1} - V_{be2}) = \alpha \ln\left[\frac{I_1 - ki_1}{I_1 + ki_1}\right]$$

and

$$(V_{be3} - V_{be4}) = \alpha \ln\left[\frac{I_1 + i_1}{I_1 - i_1}\right] .$$

Differentiating V_{in} and substituting for base–emitter voltages,

$$dV_{in} = R\,di_1 + \frac{2\alpha(1 - k^2)x^2\,di_1}{I_1(1 - x^2)(1 - k^2x^2)} \qquad (29)$$

where

$$x = \frac{i_1}{I_1} \qquad (30)$$

and

$$k = \frac{\beta - 1}{\beta + 1} . \qquad (31)$$

Therefore

$$N(x) = \frac{2\alpha(1 - k^2)x^2}{I_1R\{(1 - x^2)(1 - k^2x^2)\}} . \qquad (32)$$

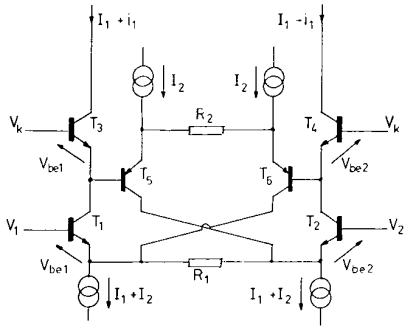


Fig. 14. Indirect error feedback.

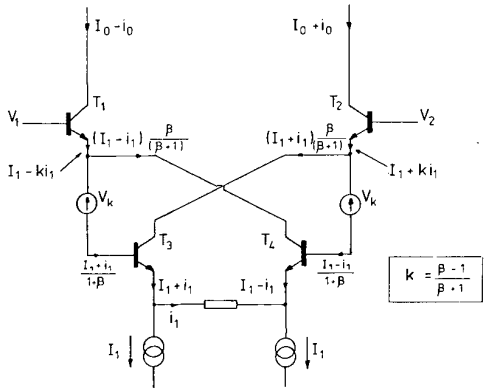


Fig. 15. Modified error feedforward with indirect V_{be} compensation.

This circuit topology reveals that if transistor matching is achieved, the nonlinearities are mainly dependent upon transistor β.

In order to obtain an adequate dynamic range without transistor saturation, constant offset voltages V_k are required (as shown in Fig. 15). However, in situations where the input signal is controlled and small, V_k can be set to zero. Such an application is to use this circuit as the error amplifier for a single long-tail pair, as shown in Fig. 7. This compound circuit is illustrated in Fig. 16.

Defining λ, x, and y as in Section 3.2, then if

$$\lambda = \frac{R_2}{R_1} \qquad (33)$$

we have

$$dV_{in} = R_1\,di_0 + \frac{4\alpha^2(1 - k^2)y^2\,di_0}{[I_1(1 - x^2)\{I_2R_2(1 - y^2)(1 - k^2y^2) + 2\alpha(1 - k^2)y^2\} + 2\alpha\lambda I_2(1 - y^2)(1 - k^2y^2)]} . \qquad (34)$$

Therefore

$$N(x, y) = \frac{4\alpha^2(1 - k^2)y^2}{\lambda I_1I_2R_1^2[(1 - y^2)(1 - k^2y^2)(1 + 2\alpha/I_1R_1 - x^2) + (2\alpha/\lambda I_2R_1)(1 - k^2)y^2(1 - x^2)]} . \qquad (35)$$

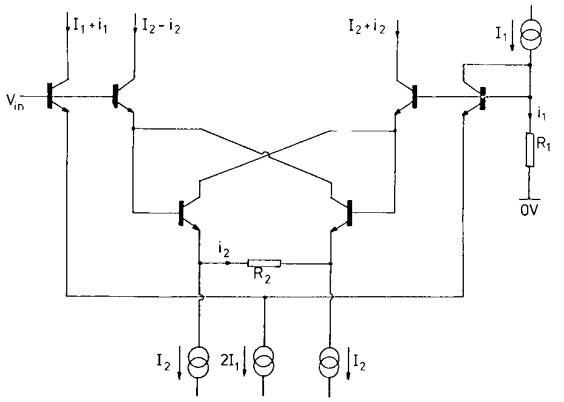


Fig. 16. Single long-tail pair using an error feedforward correction with indirect V_{be} compensation.

Hence comparing Eqs. (35) and (24), there is a further reduction of distortion of approximately $(1 - k^2)$, where k is just less than 1.

6 CONCLUSIONS

A series of circuit topologies were presented which are suitable for the input and subsequent stages of audio amplifiers. The aim has been to show that partial linearization can be achieved without recourse to excessive negative feedback. The circuits require the implementation of a balance condition which is essentially noncritical, provided that only moderate cell transconductance is required.

As an aid to circuit comparison the parameter IDF was introduced which readily expressed nonlinearities as a function of amplifier current loading factors. These expressions can be approximated still further by letting terms of the form $(1 - x^2) \rightarrow 1$ with the assumption of modest loading factors.

It is hoped that some of the circuits will prove useful to the designers of audio amplifiers and allow enhanced performance by minimizing both nonlinearity and loop-gain requirements, which have a strong correlation with transient distortion phenomena.

7 REFERENCES

[1] A. M. Sandman, "Reducing Distortion by Error Add-on," *Wireless World*, vol. 79, p. 32 (1973 Jan.).

[2] D. Bollen, "Distortion Reducer," *Wireless World*, vol. 79, pp. 54–57 (1973 Feb.).

[3] A. M. Sandman, "Reducing Amplifier Distortion," *Wireless World*, vol. 80, pp. 367–371 (1974 Oct.).

[4] P. J. Walker and M. P. Albinson, "Current Dumping Audio Amplifier," presented at the 50th Convention of the Audio Engineering Society, London, England, 1975 March 4–7.

[5] J. Vanderkooy and S. P. Lipshitz, "Feedforward Error Correction in Power Amplifiers," *J. Audio Eng. Soc.*, vol. 28, pp. 2–16 (1980 Jan./Feb.).

[6] M. J. Hawksford, "Distortion Correction in Audio Power Amplifiers," *J. Audio Eng. Soc.*, vol. 29, pp. 27–30 (1981 Jan./Feb.).

Dr. Hawksford's biography was published in the January/February issue.

PAPERS

Fuzzy Distortion in Analog Amplifiers: A Limit to Information Transmission?*

M. J. HAWKSFORD

University of Essex, Department of Electrical Engineering Science, Wivenhoe Park, Colchester, Essex, United Kingdom

A theoretical model is introduced that attempts to emulate a low-level distortion mechanism inherent in bipolar junction transistor amplifiers and, as a consequence, suggests a low-level bound to the transmission of fine signal detail. The model gives positive support to the low-feedback school of design and proposes circuit techniques for maximizing signal transparency. The design principles have particular relevance to low-level signal stages, but should also find an association with all classes of amplifiers.

0 INTRODUCTION

The last decade has seen substantial debate concerning the relationship between objective and subjective assessment of amplifiers. Measurements have frequently been performed with often impressive results [1], yet on extended audition significant audible differences can still be perceptible.

Various investigations have cited, for example, the levels of harmonic distortion as a measure of excellence, where emphasis has been directed to the distribution and relative weights of the harmonic structure. Conclusions have been drawn suggesting that low-order harmonics exhibiting a smooth rolloff in amplitude with frequency [2], [3] are a useful indicator of an amplifier's performance. However, when on this basis the levels of distortion are critically compared, it is generally difficult to assert a high correlation between objective and subjective results. In fact auditioning of amplifier performance suggests that the absolute level of harmonic distortion is, within limits, only a second-order interest, as highlighted during valve/transistor comparison.

A second indicator of potential excellence depends on the assessment of transient intermodulation distortion (TID) [4], [5], a distortion that is prevalent in slow high-loop-gain feedback amplifiers. However, design criteria have been established [6], [7] which minimize the onset of TID. Clearly, TID is only part of the distortion repertoire and is probably of minimal consequence once the probability of its occurrence is low.

Primary and secondary crossover distortion, though predominant in power amplifier circuits, also occur in certain low-level operational amplifiers that use class AB output stages. However, although this nonlinear mechanism can lead to significant signal impairment, there are now a variety of design techniques [8]–[10] that successfully minimize the error signal.

A direct consequence of amplifier nonlinearity and signal interaction is partial rectification, which produces a dynamic shift in the quiescent bias state. If an amplifier incorporates energy storage elements (such as ac coupling and by-pass capacitors), then the error signal is filtered and exhibits "overhang," which is dominant in the lower midrange and bass frequency bands. Amplifiers should therefore minimize energy storage components and be designed to be near aperiodic within the audio band. Research has shown that an asymmetric pulse test is a sensitive method of assessment [11], [12].

Where amplifiers are operated at high signal levels, other mechanisms of dynamic distortion become significant. Nonlinear delay modulation (NLDM) of the

* This paper was the basis of a lecture to the British Section in 1982 October (see *JAES*, vol. 31, no. 3, pp. 164 and 166 (1983 March)). Manuscript received 1982 October 11; revised 1982 November 22.

signal will occur due to the dynamic variation of transistor parameters with signal: Modulation of collector–base capacitance with collector–base voltage, the shift of small-signal bandwidth with collector current, and general parametric changes when devices are thermally exercised are all contributory factors. However, after reviewing the many conventional forms of nonlinearity it is apparent that certain areas of subjective assessment still elude a satisfactory explanation, and it is unclear as to an optimum design strategy. Specifically the area of greatest concern is that of subjective clarity or what may be usefully described as signal transparency: the ability to resolve fine signal detail, especially in the presence of complex high-level signal components. There appears to be a distinction between distortion mechanisms that "color" the signal, thus adding their own character, and distortions that corrupt fine signal detail.

This paper addresses what is believed to be both a significant and a neglected factor of amplifier performance where two basic clues have emerged: first, that amplifiers using low or distributed feedback often audition with higher rank, even though they may exhibit higher levels of error signal, and second, that low-level amplifier stages appear particularly susceptible to signal impairment. A primitive theory is proposed and a design strategy presented as a means of performance optimization.

In preparing the work presented in this paper, a literature survey revealed an embryonic idea first published by West [13] in 1978. However, the idea was not developed to any extent, and its significance with respect to amplifier design was not established in depth. A later discussion by Curtis [14] dismissed the theory as a cause of "transistor sound." The author considers this dismissal somewhat premature and attempts in this paper to extend the theory in more detail, with respect both to the charge-control model of a transistor and to the application of the derived theory to amplifier design.

1 FUZZY NONLINEARITY: THE THOUGHT EXPERIMENT

Classical circuit theory represents current as a continuous function that flows smoothly and can be considered to have infinite precision within an uncorrelated random bound. This viewpoint is taken from a macroscopic stance of electromagnetism where the individual electrical fields of electrons merge to a nongranular continuum that allows near infinite precision in the transmission of information. Account is of course taken of the behavior of partial randomness of electrons, and this is introduced through linear noise analysis where the noise is seen as the limiting factor on low-level signal resolution. In fact basic calculations on the numerousness of electrons would suggest this to be perfectly reasonable and of little consequence to the audio circuit designer. We speculate here that this may well be an invalid assumption which disguises the true limit to the ultimate resolution of a low-noise amplifier stage.

Transistor operation depends in part on the transfer of charge from signal source to device, a theory first proposed by Beaufoy and Sparkes [15]. Essentially the theory shows that the level of collector current in a bipolar junction transistor (BJT) is a linear function of the local stored charge in the base region. The theory also proposes that the continual base current of a BJT provides a "top up" charge to compensate for recombination resulting from a finite carrier lifetime within the base. In equilibrium the rate of recombination is just balanced by the base current to maintain a constant average charge, which in turn determines the collector current.

However, in this paper we shall not be concerned directly with the mechanics of device operation, only a consequence of those mechanisms, namely, the level of charge transfer required in the amplification process. The probable importance of charge levels can be established by the following thought experiment.

In this discussion we shall evaluate the approximate levels of charge that are transferred to the base of a transistor under low-level signal excitation. Fig. 1 shows a basic zero feedback amplifier stage interfaced to a moving-coil transducer with source resistance r_c, where the input impedance of the amplifier is derived directly from the hybrid-π equivalent circuit of a transistor. In Fig. 1 r_{bb}' is the base bulk resistance, $r_{b'e}$ the dc input resistance (modeling small-signal recombination), and $C_{b'e}$ the base region capacitance storing the charge q_b which controls the collector current.

A value of the base storage capacitor can be estimated directly from a knowledge of f_β, the 3-dB bandwidth of h_{fe}, which is the collector–base current gain, assuming a first-order response,

$$C_{b'e} = \frac{1}{2\pi r_{b'e} f_\beta} \Big| V_{CB} \rightarrow \text{constant} , \quad (1)$$

this expression is derived from the observation that the reactance of $C_{b'e}$ is equal to $r_{b'e}$ at the frequency f_β, it also follows that $C_{b'e} \propto I_e$ (emitter current).

Let us further our argument by considering the output voltage v_i of a moving-coil cartridge,

$$v_i = \frac{f}{f_n} V_n \sin(2\pi f t) \quad (2)$$

where V_n is the nominal cartridge output amplitude at a normalized frequency f_n, typically 1 kHz. If the dynamic range of the system is DR, then the minimum

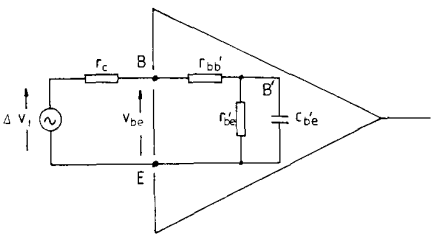


Fig. 1. Basic transistor amplifier stage.

resolvable signal level Δv_i is

$$\Delta v_i = \frac{f}{f_n}\frac{V_n}{DR}\sin(2\pi ft) . \quad (3)$$

In defining DR we refer to the smallest resolvable change in signal level that can exist without nonlinear corruption from complex high-level signal components. Ideally this change in signal level should be below the noise floor.

We next assign a minimum resolvable time period τ_m estimated by direct reference to the sampling theorem, which conveniently relates τ_m to the audible bandwidth f_a,

$$\tau_m \cong \frac{1}{2f_a} . \quad (4)$$

Assuming a sinusoidal input signal, an expression for the control base charge $q_b(t)$ for an input signal Δv_i is derived as

$$q_b(t) = \frac{1}{2\pi r_{b'e} f_\beta}\frac{f}{f_n}\frac{V_n}{DR}\sin(2\pi ft) \quad (5)$$

where

$$r_c + r_{bb'} << \left| r_{b'e} \,//\, \frac{1}{2\pi f C_{b'e}} \right| .$$

Hence the change in control charge that occurs over a time τ_m is

$$\Delta q_{b,\min} = q_b\left(t + \frac{\tau_m}{2}\right) - q_b\left(t - \frac{\tau_m}{2}\right) \quad (6)$$

where, aligning the difference equation to maximize Δq_b (in this sense our estimation is optimistically high) and assuming $\sin(\pi f\tau_m) \cong \pi f\tau_m$, we have

$$\Delta q_{b,\min} = \frac{V_n}{2r_{b'e}DR}\frac{f^2}{f_n f_a f_\beta} . \quad (7)$$

We note from standard transistor theory that

$$r_{b'e} = (1 + h_{fe})r_e \quad (8)$$

$$r_e = \frac{0.025}{I_e} , \ (I_e \text{ in amperes}) \quad (9)$$

$$f_T \simeq (1 + h_{fe})f_\beta . \quad (10)$$

Hence

$$\Delta q_{b,\min} = \frac{20V_n I_e}{DR}\frac{f^2}{f_n f_a f_T} .$$

To estimate typical values of changes in the base charge consider the following data base:

V_n	= 200 μV	medium output moving-coil cartridge,
I_e	= 10^{-3} A	transistor emitter bias current,
f_T	= 50 MHz	bandwidth to unity h_{fe},
DR	= 10^4	80-dB dynamic range,
f_n	= 1 kHz	normalizing frequency for cartridge,
f_a	= 20 kHz	audible bandwidth,
e	= 1.96×10^{-19} C	charge on electron,
h_{fe}	= 500	small-signal collector–base current gain (V_{CE} constant)

whereby the minimum change in base charge is evaluated as

$$\Delta q_{b,\min} \simeq (2 \times 10^{-6} f^2)e \quad \text{[coulombs]} \quad (11)$$

Eq. (11) shows a remarkably low level of average charge transfer that occurs for small signals observed over the minimum resolvable time period (here assumed to be 25 μs).

It is also instructive to estimate the change in the number of electrons transferred into the base region through recombination over the minimum time period τ_m, due only to the minimum signal component Δv_i.

If we assume $r_{b'e}$ to be the dominant input resistance of the transistor, the base input current Δi_i associated with Δv_i is

$$\Delta i_i = \frac{f}{f_n}\frac{V_n}{r_{b'e}DR}\sin(2\pi ft) . \quad (12)$$

The charge Δq_r transferred from source to input due to recombination in time τ_m is calculated by integration,

$$\Delta q_r = \int_{t-\tau_m/2}^{t+\tau_m/2} \Delta i_i \, dt . \quad (13)$$

Aligning the integration window to maximize Δq_r and again assuming that $\pi f\tau_m$ is small,

$$\Delta q_r = \frac{20V_n I_e}{(1 + h_{fe})DR}\frac{f}{f_n f_a} . \quad (14)$$

Using the same data base,

$$\Delta q_r = 0.2fe \quad \text{[coulombs]}. \quad (15)$$

Eqs. (11) and (15) show that low-level signals in transistor stages are associated with an extremely small transfer of charge into the base of the input transistor. The basic analysis indicates that within τ_m the signal amplitude generally has greater effect on the charge transferred for recombination than that charge having direct control of the collector current (according to charge control theory). Nevertheless both calculations yield results of only a few electrons.

We therefore propose a theory that partial signal quantization is the fundamental process that sets an inherent bound to signal transparency through a transistor stage. Both Eqs. (11) and (15) support the probable existence of significant granularity where Eq. (11) suggests a form of amplitude quantization and Eq. (15) an association with $1/f$ noise.

It is also proposed that signal interaction with inherent

nonlinearities in transistors, together with even small levels of interference from power supplies, neighboring circuitry, or undesired signal coupling (such as poor ground line design), can easily corrupt such minute signals and that such corruption should be interpreted as modifications to these low charge levels.

We conclude this preliminary discussion by giving in Table 1 typical levels of charge transferred to the base of a transistor within the minimum time period $\tau_m = 25$ μs against various signal levels to illustrate the potential dynamic range available. The example already cited in this section is used as a data base.

2 FUZZY MODELS

In this section we build upon the observations made of quantization and the relative magnitudes of low-level signals by introducing a basic model of the distortion process. It is emphasized that although the model is primitive, it is a natural extension of our thought experiment.

The proposed model is to be classed as "fuzzy" and the resulting distortion as fuzzy distortion due to its strong stochastic association. We commence by establishing two distinct groups of nonlinearity.

1) *Deterministic nonlinearity.* Classic system nonlinearity can be envisaged using a continuous model incorporating static or dynamic transfer characteristics. The main attribute of this broad distortion classification is repeatability where, assuming no time-dependent system parameters, the same error waveform will result under repeated tests. We note in particular that when measuring such distortion a degree of signal averaging is often used to suppress random events.

2) *Fuzzy nonlinearity.* A distortion process that results in an error signal with a strong stochastic element that does not include any uniform sampling function is defined here as fuzzy distortion. Such distortion will not exhibit exact error waveform replication under repeated tests. We note in particular that when measuring such distortion, any signal averaging will tend to mask the error waveform.

We proceed by further reference to the charge control model of a BJT [15] and attempt to produce a primitive model that matches the input impedance characteristic of a BJT transistor (see Fig. 1), exhibits the correct frequency response when observing h_{fe}, introduces a degree of charge quantization, and maintains the proper static relationship between base and collector currents.

The proposed model is illustrated in Fig. 2(b) and is configured so that it replaces directly the standard hybrid-π circuit shown in Fig. 2(a). A simplified notation is illustrated in Fig. 2(c).

Table 1. Typical levels of charge transferred.

	Charge Transfer in 25 μs	
	Δq_r	$\Delta q_{b,\,min}$
Bias current of 1/500 mA	$2.56 \times 10^8 e$	—
Input signal of 200 μV	$2 \times 10^6 e$	$2 \times 10^4 e$
Input signal 80 dB below 200 μV at 1 kHz	$200e$	$2e$

The model consists of an integrator to convert input signal current to charge, cascaded with a uniform quantizer with an associated dither source $n(t)$ to scatter the quanta. The integrator and quantizer are enclosed within a negative-feedback loop, which together emulate the process of recombination and quantization of the stored base charge. The quantized base–emitter voltage $V_{b'e}$, which is proportional to the stored base charge, is converted to collector current by a transconductance stage with mutual conductance g_m. From standard transistor theory,

$$g_m = \frac{h_{fe}}{(1 + h_{fe})r_e} \tag{16}$$

$$r_e = \frac{\partial V_{BE}}{\partial I_E} = \frac{kT}{eI_e} \tag{17}$$

where k is Boltzmann's constant, T the junction temperature (kelvins), e the charge on an electron, and I_e the emitter bias current.

The model shown in Fig. 2(b) has a strong resemblance to certain classes of analog-to-digital encoder, in particular feedback (pulse-code modulation) and multilevel delta sigma modulation (DSM) [16], [17]. Since these encoding schemes combine integration and quantization within a feedback loop, they form useful vehicles for comparison. A major distinction between

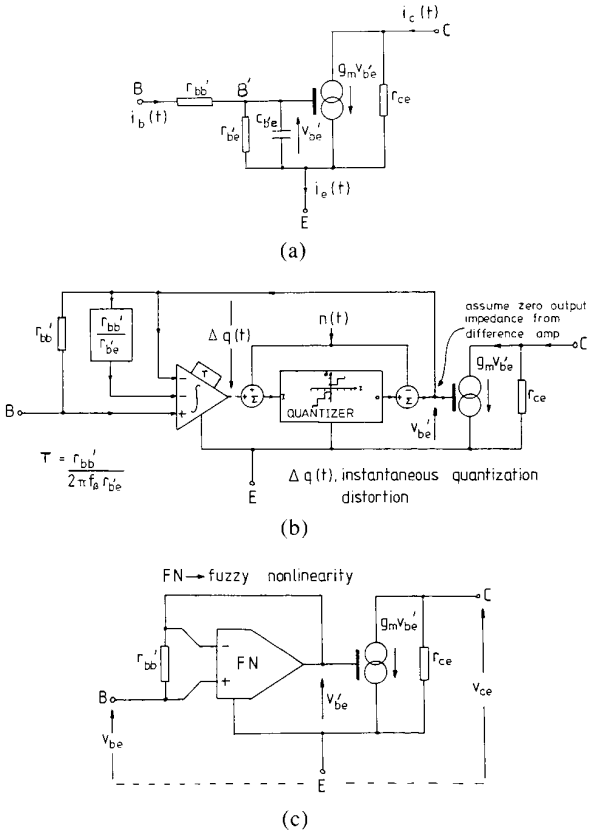


Fig. 2. Basic hybrid-π equivalent circuit of a BJT. (a) Standard circuit. (b) Circuit with modification to incorporate charge quantization. (c) Simplified functional presentation, including charge quantization.

the fuzzy model and digital encoders is that the former excludes a uniform sampling process. However, a random sampling function is permissible where the mean sampling frequency corresponds to the mean rate of recombination within the base of the transistor, which is determined by the base bias current (that is, a base bias current of 2 μA corresponds to a mean sampling rate of $\cong 10^{13}$ Hz). We note also from Eq. (2) that the mean sampling rate will undergo frequency modulation due to the instantaneous change in recombination current with change in base–emitter voltage.

We estimate the approximate frequency characteristic of the distortion spectra for the model of Fig. 2(b) by assuming the loop to be essentially linear and by representing the quantization distortion as a sinusoidal error signal added within the loop where, for purposes of analysis,

$$q(t) = Qe^{j2\pi ft} . \tag{18}$$

Thus the collector error current $I_{q,c}e^{j2\pi ft}$ follows as

$$I_{q,c} = \frac{jQg_m f/f_\beta}{\left[1 + \dfrac{r_{b'e}}{r_c + r_{bb'}}\right]\left[1 + \dfrac{jf}{[1 + r_{b'e}/(r_c + r_{bb'})]f_\beta}\right]} \tag{19}$$

where r_c is the source resistance between base and emitter, as shown in Fig. 1.

From Eq. (19) we infer the basic form of error spectrum, which is illustrated in Fig. 3. Note the effect a low source impedance has on the break frequency in the approximate error spectrum.

The error spectrum shown in Fig. 3 compares with the general trend of pulse-code-modulation-type systems [16], [17] where quantization is dominant at high frequencies. The curve ignores other forms of random noise, such as the noise associated with $r_{bb'}$. Thus in general this effect will be at or below the device noise level.

The results show that the source resistance plays a dominant role in shaping the error spectrum where optimum performance is obtained when r_c is minimized. This compares favorably with the more common noise model of a transistor where the noise sources are represented as equivalent input noise voltage and current generators. An interesting by-product of the model structure is that it includes a mechanism that modifies the output noise as a function of source resistance, this being achieved by modifying the feedback factor around the internal feedback loop. Essentially when $r_c = 0$, maximum feedback is applied, and when $r_c = \infty$, minimum feedback results. Examination of the model schematic illustrated in Fig. 2(b) should clarify this operation.

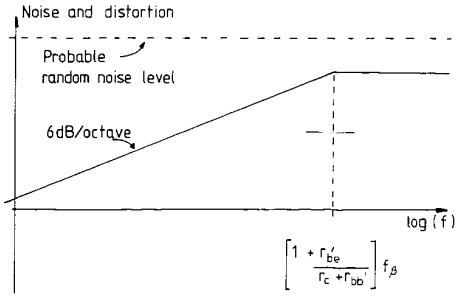


Fig. 3. Approximate error spectrum of collector current due only to quantization effects.

In this section we have established a modification to the basic hybrid-π model of a transistor which includes the effect of charge quantization. We now proceed to examine some implications of these observations.

3 IMPLICATIONS OF FUZZY DISTORTION IN AMPLIFIER DESIGN

If we accept that a low-level nonlinear mechanism exists in transistors which has a different nature from deterministic nonlinearity, then we can make some basic observations as to the correct global strategy toward amplifier design.

Where a transistor operates with very-low-level signals that approach the noise floor, the artifacts of charge quantization will generate significant fuzzy distortion. The method to minimize this effect can be summarized as follows:

1) It is essential that low-noise devices be used that exhibit low $r_{bb'}$ and low $1/f$ (recombination noise). The device should be chosen so as to maximize $C_{b'e}$. Thus large integrated arrays of transistors where many matched devices are paralleled should prove the best choice (such as the LM394).

2) Operate transistors so that $C_{b'e}$ is maximized. From Eq. (1) this infers a substantial level of emitter bias current which in turn will lower the device input impedance.

3) Eq. (1) infers that $C_{b'e}$ is an inverse function of $r_{b'e}$ (for given f_β). Thus a device should be chosen with a low value of h_{fe} [see Eq. (8)].

4) Selection of f_β is more complex. A low level of f_β will increase $C_{b'e}$, but at the expense of lowering the break frequency in the distortion spectra (see Fig. 3). It is suggested that f_β should be sensibly in excess of 20 kHz.

5) Design the transistor stage so as to maximize the device loading factor [18]. This will maximize the changes in charge for a given signal.

6) Minimize resistance in the input mesh of a transistor. This will reduce low-frequency fuzzy distortion. For the input stage of Fig. 1 this implies a low value of r_c. The effect of r_c can be observed by reference to Eq. (19) and the error spectrum in Fig. 3.

Consider by way of example a low-level disk preamplifier stage for use with a low-output moving coil cartridge. In Fig. 4 a moving-coil cartridge with source resistance r_c and generator signal $e_c(t)$ is interfaced to

a disk amplifier which has an input resistance r_{in} and a voltage gain A_v.

The classical viewpoint would not expect r_{in} to play an important role other than providing an optimum load for the cartridge. (This may affect the frequency response, for example, when coupled with the generator source inductance.) However, for low-output-impedance moving-coil cartridges this generally has minimal effect. Indeed in selecting an "optimum" load resistance, it is normal to use an input shunt resistor.

However, fuzzy nonlinearity suggests that the level of i_{in} is of fundamental importance, and that this current must be maximized and flow into the base of the input transistor. A shunt input resistance is not an acceptable solution, as current will by-pass the transistor.

It therefore follows that the input signal must be considered in terms of both input voltage and input current. It is the input signal power that is fundamental.

3.1 Corollary 1

If we accept the notion of maximizing the signal power that flows into the base of the input transistor, then a transducer for an analog disk system must be selected such that

1) It converts a relatively high proportion of platter rotational energy into mechanical signal energy, as seen at the cantilever of the cartridge.

2) It exhibits a high mechanical-to-electrical power conversion.

It is possibly in these areas of performance where many moving-coil cartridges offer a significant performance advantage.

3.2 Corollary 2

In selecting a matching transformer/input circuit topology, the aim must be to maximize the flow of signal power into the base of the input transistor.

The proposal to maximize the input power is open to some debate. However, if it is realized that we wish both to maximize input signal current to the base of each transistor and to minimize source resistance r_c, then the notion of power maximization is a reasonable target.

Corollary 2 has profound ramifications in the choice circuit topology. Consider the classical amplifier configuration shown in Fig. 5. The circuit shows an input signal generator e_c with source impedance r_c. Again the amplifier has an input impedance r_{in} and voltage gain A_v, but a negative-feedback loop is included where the feedback factor is B with a Thévenin source impedance (seen by the inverting input) of r_f.

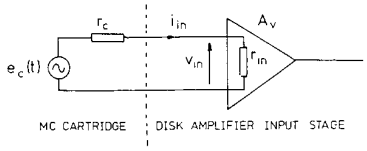


Fig. 4. Moving-coil cartridge–amplifier interface.

We proceed by calculating the instantaneous input signal power $p_{in}(t)$ to the differential input of the amplifier,

$$p_{in}(t) = \left[\frac{e_c(t)}{(r_c + r_f) + r_{in}(1 + AB)}\right]^2 r_{in} . \quad (20)$$

Differentiating $p_{in}(t)$ ωrt r_{in},

$$\frac{\partial p_{in}(t)}{\partial r_{in}} = \left[\frac{e_c(t)}{(r_c + r_f) + r_{in}(1 + AB)}\right]^2 \times \left[1 - \frac{2r_{in}(1 + AB)}{(r_c + r_f) + r_{in}(1 + AB)}\right] \quad (21)$$

and setting $\partial p_{in}(t)/\partial r_{in} = 0$ to maximize the input power, the optimum r_{in} (for maximum input power) follows as

$$r_{in}\Big|_{opt} = \frac{r_c + r_f}{1 + AB} . \quad (22)$$

This gives the maximum input power as

$$p_{in}(t)\Big|_{max} = \frac{e_c^2(t)}{4(r_c + r_f)(1 + AB)} . \quad (23)$$

Eq. (23) shows the need to minimize all extraneous resistances within the input signal mesh, which is also a requirement for good noise design (that is, minimize r_f). [See also Eq. (19) and the discussion in Section 2 concerning input mesh resistance and fuzzy distortion.]

However, a more fundamental observation shows the maximum power flow to be an inverse function of the feedback parameter. Thus although classical feedback theory would suggest an improvement by operating the device well into its linear region of operation, it in fact forces the signal to within a relatively few quanta, thus exaggerating any effects of quantization.

To illustrate the process further, consider the combined systems of Figs. 2(c) and 5, as shown in Fig. 6.

Any amplification which follows the quantization process must by necessity amplify the quantized signal, together with additional random noise sources. The effect of negative feedback on a purely linear system will reduce the levels of additional noise resources that are injected within the feedback loop by a factor of $(1 + AB)$. However, this process is not true of a loop that includes quantization. In fact in this system the feedback will again reduce the additive noise, but it will only

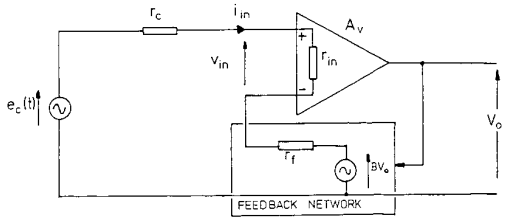


Fig. 5. Classical feedback amplifier structure.

partially reduce the effects of quantization. Thus fuzzy distortion components will be partly exposed by negative feedback, together with a complex process of intermodulation between signal additive noise and quantization distortion, which in general must include time smearing due to limitations in loop bandwidth.

4 MINIMIZATION OF FUZZY NONLINEARITY

We conclude our discussion on fuzzy distortion by suggesting a design method and basic circuit topologies that in principle meet the requirements of both high-level and low-level nonlinearities. In particular we emphasize low-level signal stages as these are potentially more susceptible to fuzzy nonlinearity.

Following the design aims discussed in Section 3, we must choose a low-noise transistor with a low value of collector–base current gain. This device should be operated at a collector current commensurate with noise considerations such that (ideally) the input impedance between base and emitter matches the source impedance of the transducer or presents an optimum load to the transducer. Provided the source signal is of suitable magnitude, the signal should be coupled directly to the base–emitter junction (assuming that high-level distortion will not be problematic), and preferably no ac coupling component should be used.

Coupled with this requirement, the input transistor should ideally use no feedback (local or overall), since Eq. (23) indicates a reduction in signal power. If the transducer output is too great, resulting in a high level of deterministic distortion, then a step-down transformer should be selected to permit using a zero feedback input stage. Ideally the input impedance should be designed to match the transformed source impedance of the transducer. This process will not change (in principle) the level of power extracted from the source (assuming a power match), but it will minimize high-level distortion and eliminate a loss of input power through the use of negative feedback. In many instances it will not be practical to design for a power match as high operating currents or many parallel devices may be necessary, though investigation into the LM394-type device should be encouraged.

In general an amplifier system will include several cascaded transistor stages within the signal path. Potentially each stage is a cause of low-level distortion, but as with noise design, the first transistor should be the dominant offender. To guarantee this ideal we must arrange for a progressive increase in input signal power as we proceed along the cascade, as well as attempting to minimize the number of series-connected transistors in the signal path.

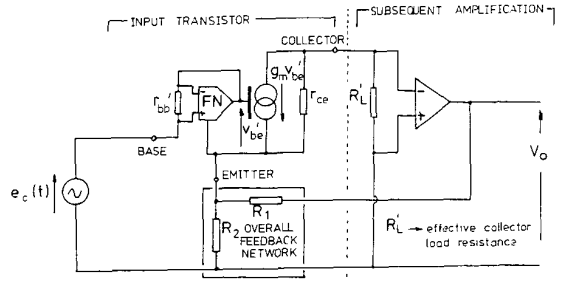


Fig. 6. Simplified feedback amplifier with quantizer model of Fig. 2(c).

In order to control deterministic distortion as signal levels are amplified, a degree of negative feedback will become mandatory, which will consist generally of a combination of distributed and multiple-feedback loops. However, in selecting the topology for the feedback structure, an increasing input-signal-power progression should be observed.

To examine this design strategy, consider $N + 1$ cascaded transistor stages, as shown in Fig. 7. Stages $1 \rightarrow N$ use distributed feedback, while the input stage 0 is optimized for fuzzy nonlinearity by using zero feedback. A single feedback loop encloses stages $1 \rightarrow N$, thus modeling a typical amplifier.

Let

$A_0, \ldots, A_N$ = amplifier gains
$B_0, \ldots, B_N$ = feedback factors[1]
$e_0, \ldots, e_N$ = amplifier input signals
} see Fig. 7

$r_0, \ldots, r_N$ = amplifier input resistances
$p_0, \ldots, p_N$ = input signal powers to amplifiers
r_c = transducer source impedance.

From Eq. (20) we calculate the rth-stage input signal power p_r. For stages $r = 1, \ldots, N$. (Assume that the source resistance is small compared with r_r.)

$$p_r = \left[\frac{e_r}{1 + A_r B_r}\right]^2 \frac{1}{r_r} \tag{24}$$

and for stage $r = 0$,

$$p_0 = \left(\frac{e_0}{r_c + r_0}\right)^2 r_0 \ . \tag{25}$$

4.1 Design Criterion

To minimize signal degradation caused by fuzzy nonlinearity in a cascade of transistor stages,

$$p_r = G_{fr} p_{r-1} \tag{26}$$

where ideally the interstage power gain $G_{fr} > 1$. We calculate the voltage gain relating e_r and e_{r-1}, for $r = 1$,

$$\frac{e_1}{e_0} = \frac{A_0}{\left\{1 + B_0 \prod_{p=1}^{N} [A_p/(1 + A_p B_p)]\right\}} \tag{27}$$

and for $r = 2, \ldots, N$,

$$\frac{e_r}{e_{r-1}} = \frac{A_{r-1}}{1 + B_r A_r} \ . \tag{28}$$

Hence we establish the constraints on the choice of feedback parameters by reference to Eqs. (24)–(28).

[1] The feedback networks are assumed to exhibit zero Thévenin source impedances.

If

$$\delta_{f1} = \left\{1 + B_0 \prod_{p=1}^{N} \frac{A_p}{1 + B_p A_p}\right\}\left\{\frac{\sqrt{G_{f1} r_0 r_1}}{r_c + r_0}\right\} \quad (29)$$

$$\delta_{fr} = \sqrt{\frac{G_{fr} r_r}{r_{r-1}}}\Bigg|_{r=2,\ldots,N} \quad (30)$$

we have

$$A_{r-1} = \delta_{fr}(1 + B_r A_r)\Big|_{r=1,\ldots,N} \quad (31)$$

where we define $\delta_{fr}|_{r=1,\ldots,N}$ as the set of fuzzy gain parameters of the amplifier system.

Examination of Eqs. (29)–(31) reveals the design criterion that will ensure a progressive power increase along the cascade of transistor stages (noting that calculated power levels refer to the input power to each transistor, not the associated circuitry).

In practice there will be a limit to the input power to a transistor that will be dependent on the acceptable levels of deterministic distortion. We note that for a bipolar transistor which adheres to the form of Eq. (17) the fractional error component of emitter current is independent of I_{EO} for a given V_{BE} (where the subscript EO infers quiescent values),

$$V_{BE} = V_{BEO} + \Delta V_{BE}$$

which corresponds to $I_E = I_{EO} + \Delta I_E$. Then

$$\frac{\Delta I_E}{I_{EO}} = e^{(q\Delta V_{BE}/KT)} - 1 . \quad (32)$$

Since the base–emitter voltage of a transistor is directly dependent upon the input power and input resistance, the input resistance should be minimized to reduce high-level distortion, for a given power level.

It is constructive to reflect upon a common circuit arrangement where a discrete transistor stage is cascaded with a BJT operational amplifier with local feedback. We will assume for simplicity that there is no overall feedback and proceed by suggesting typical circuit parameters:

Discrete stage		
input impedance	1 kΩ	(transistor)
voltage gain	20	
Operational amplifier stage		
input impedance	1 MΩ	(operational amplifier)
closed-loop gain	20	
open-loop gain	1000	(conservative estimate)

Hence from Eq. (31) $\delta_{fr} \cong 0.4$, whereby the power gain follows from Eq. (30) as 1.6×10^{-4}.

This result shows a substantial reduction of input signal power presented to the second stage, the consequence being that any quantization effects will be significantly increased.

This circuit arrangement has often been used for disk preamplifiers and is a good illustration of a potential hazard of using high-gain high-input-impedance operational amplifiers.

We conclude this section by suggesting how it is possible to use a combination of low distributed feedback with feedforward error correction as a compromise to the distortion dichotomy existing between deterministic and fuzzy nonlinearities.

Germane to the design strategy is the selection of a distributed feedback system where the appropriate gains and feedback factors are calculated according to our now established fuzzy nonlinearity criterion. In so doing we accept that deterministic nonlinearity inherent in devices will potentially increase. However, by using nested feedforward error correction we can partially compensate the deterministic error signals and achieve acceptable linearity with high loading factors, even when local negative feedback is low.

In Fig. 8 we illustrate a two-stage feedforward amplifier where the error due to base–emitter nonlinearity in T_1 is partially corrected by the differential amplifier formed by T_2 and T_3. Further error-correction stages can be used to compensate for T_2 and T_3 nonlinearity using a nested configuration. The performance of such stages as a function of loading factor was considered in a previous paper [18].

The dominant advantages of this approach is that only very modest local negative feedback need be applied via R_1 and that the high-level distortion is partially compensated by the error amplifier. Such a technique allows good signal power coupling to T_1, yet permits an acceptable high-level distortion characteristic. It therefore follows that T_1 is exercised over a wide range of its operating characteristic while retaining good overall linearity.

The example just discussed illustrates how ideas of fuzzy nonlinearity could influence amplifier design. A second area of application concerns the construction and layout of circuits. Once the very small signal levels are appreciated and the point of view of "counting electrons" is taken, such factors as metal–metal con-

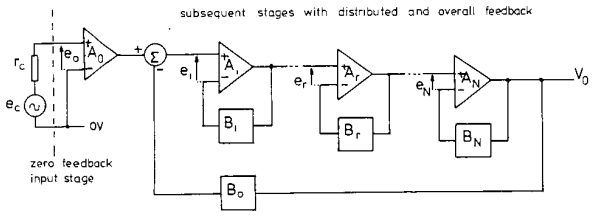


Fig. 7. Basic multiple-loop feedback amplifier topology.

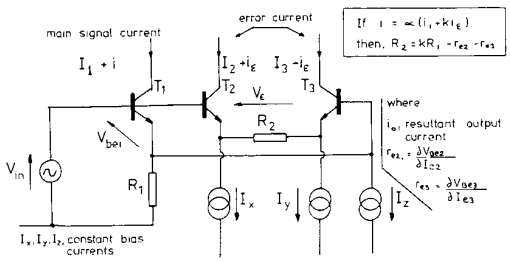


Fig. 8. Basic feedforward error-correction stage.

tacts, interference from adjacent circuits, and the displacement of charge through the dielectric of a capacitor require careful attention. These secondary factors will not be discussed in this paper, but they are influential in setting potential limits to signal transparency.

5 CONCLUSIONS

This paper has speculated on the existence of low-level nonlinearity inherent within BJT devices due to the quantization of charge carriers and has drawn attention to the relative magnitude of low-level signals. A model was introduced which forms a vehicle for comparison between an analog device and a class of digital modulation. This comparison is useful in that it is possible to speculate upon the nature and characteristics of the distortion.

Consideration of the mechanism of fuzzy distortion drew attention to the role of the input signal current at the base of a transistor and the need to maximize its value. This led directly to the usefulness of input signal power as a parameter in establishing levels of fuzzy distortion. A target design objective suggested the need to maximize this power flow, where the flow must be directly into the base–emitter junction and not into an external shunt resistor.

The role of negative feedback was then debated, where it was shown that the input signal power was an inverse function of amplifier loop gain. It was therefore concluded that levels of feedback should be minimized and that extraneous resistance within the input mesh should also be minimized. However, it was noted that the role of negative feedback offers a contribution to high-level signal distortion, but that its application must be considered with great care from the viewpoint of fuzzy nonlinearity.

A brief discussion was presented where the bounds on the selection of feedback factor and forward amplification were established. Finally a circuit technique using low levels of distributed feedback with feedforward error correction was introduced as a means of circumventing the distortion dichotomy, thus allowing both good low-level and high-level distortion characteristics, that is, the dynamic range.

It is satisfying to see some of the design objectives compatible with established design techniques which are used to minimize the artifacts of TID and also as support to the low-feedback school of design, in particular since there are now several good-quality amplifiers which adhere in part to these design objectives and also have excellent subjective ratings.

Finally it must be emphasized that the ideas presented here are the extension of a thought experiment into the approximate nature and behavior of low-level signals in amplifiers. Clearly such parameters as the physical size of transistors and the relative amounts of total charge stored in the base region are of importance. However, such considerations put in doubt the application of BJT operational amplifiers with their high open-loop gains, very high differential input impedance due to the low collector bias currents in the input transistors, and low real estate. If such devices exhibit low-level quantum effects, they are not suitable for use in high-quality audio amplifiers where precision of control of fine signal detail is mandatory. In fact, applying the thought experiment discussed in Section 1, the implication of Eq. (23), and the example of the BJT operational amplifier in Section 4, the potential consequences should at least be of concern to the circuit designer.

6 REFERENCES

[1] N. Keywood, "Amplifier Review," *Hi-Fi News*, vol. 27, pp. 37–45 (1982 Aug.).

[2] J. Hiraga, "Amplifier Harmonic Distortion Spectrum Analysis," *Hi-Fi News*, vol. 22, pp. 41–45 (1977 Mar.).

[3] C. Ray, "Negative Feedback and Non-linearity," *Wireless World*, vol. 84, pp. 47–50 (1978 Oct.).

[4] M. Otala, "Transient Distortion in Transistorized Audio Power Amplifiers," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 234–239 (1970 Sept.).

[5] M. Otala, "Circuit Design Modifications for Minimizing Transient Intermodulation Distortion in Audio Amplifiers," *J. Audio Eng. Soc.*, vol. 20, pp. 396–399 (1972 June).

[6] P. Garde, "Transient Distortion in Feedback Amplifiers," *J. Audio Eng. Soc.*, pp. 314–322 (1978 May).

[7] W. M. Leach, "An Amplifier Input Stage Design Criterion for the Suppression of Dynamic Distortion," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 29, pp. 249–251 (1981 Apr.).

[8] J. Vanderkooy and S. P. Lipshitz, "Feedforward Error Correction in Power Amplifiers," *J. Audio Eng. Soc.*, vol. 28, pp. 2–16 (1980 Jan./Feb.).

[9] M. J. Hawksford, "Distortion Correction in Audio Power Amplifiers," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 29, pp. 27–30 (1981 Jan./Feb.).

[10] S. Takahashi and S. Tanaka, "Design and Construction of a Feedforward Error-Correction Amplifier," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 29, pp. 31–37 (1981 Jan./Feb.).

[11] Y. Hirata, M. Ueki, T. Kasuga, and T. Kitamura, "Nonlinear Distortion Measurement Using Composite Pulse Waveform," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 29, pp. 243–248 (1981 Apr.).

[12] Y. Hirata, "Quantifying Amplifier Sound," *Wireless World*, vol. 87, pp. 49–52 (1981 Oct.).

[13] R. West, "The Great Amplifier Debate, 2-Transistor Sound," *Hi-Fi News*, vol. 23, p. 79 (1978 Jan.).

[14] S. Curtis, "Amplifier Noise and Clipping," *Hi-Fi News*, vol. 23, pp. 81–85 (1978 June).

[15] R. Beaufoy and J. J. Sparkes, "The Junction Transistor as a Charge Control Device," *ATE J.*, vol. 13, pp. 310–327 (1957 Oct.).

[16] R. Steele, *Delta Modulation Systems* (Pentech Press, London, 1975).

[17] M. J. Hawksford, "Unified Theory of Digital Modulation," *Proc. IEE*, vol. 121, pp. 109–115 (1974 Feb.).

[18] M. J. Hawksford, "Distortion Correction Circuits for Audio Amplifiers," *J. Audio Eng. Soc.*, vol. 29, pp. 503–510 (1981 July/Aug.).

THE AUTHOR

Malcolm Hawksford was educated at the University of Aston in Birmingham, England, from 1965 to 1971. In 1968 he obtained a first class honors degree in electrical engineering, and that same year was awarded a BBC research scholarship to investigate the application of deltamodulation to color television. In 1972 he obtained a Ph.D. degree. In 1971 Dr. Hawksford became a lecturer at the University of Essex, England, in the Department of Electrical Engineering Science. During his time at Essex he actively pursued research projects within the field of audio engineering, where projects on power amplifier design, loudspeaker crossover design, analog-to-digital conversion, and music synthesis have been undertaken. He has presented papers at conventions of the Audio Engineering Society. He is currently a member of the AES, IEE, and RTS.

ENGINEERING REPORTS

Optimization of the Amplified-Diode Bias Circuit for Audio Amplifiers*

M. J. HAWKSFORD

University of Essex, Department of Electrical Engineering Science, Colchester, Essex, UK

An economic enhancement to the conventional "amplified diode" bias circuit is presented for use in power amplifier circuit topologies which do not allow precise, temperature invariant control of the operating current of the bias circuit. In essence, the modification minimizes the sensitivity of the derived bias voltage to changes in operating current without compromising the desirable temperature tracking properties when thermally bonded to the complementary follower output cell.

0 BACKGROUND

The output stage of power amplifiers and some operational amplifiers requires circuitry to allow precision control of the output bias current. The classical approach is to use a bias network that consists of either a series of diodes [1] or a transistor with local feedback in a circuit called an amplified diode [2], [3]. In Fig. 1 we illustrate the basic output-cell topology for the complementary follower configuration.

The circuit objective is to produce a dc offset V_B between the base connections of the output devices to compensate for the ON bias voltage that is required to establish the output-device bias quiescent current I_Q.

Although in practice emitter resistors R_e are used as local series feedback elements to help stabilize changes in I_Q with changes in device temperature and circuit parameters, their use only degenerates performance in other areas by increasing the output resistance of the stage and by making the output resistance a nonlinear function of output current (especially in class AB stages).

Consequently, as is well known, the use of a bias network that in principle can track changes in output-device temperature is necessary to control I_Q within reasonable bounds and thus allows low or zero values of R_e to be employed. Such networks are generally of the type shown in Fig. 1, where the bias devices (diodes or transistor) should be in close thermal contact with the output cell for good temperature tracking.

However, one area of bias network design that has been given little attention is the variability of V_B with changes in operating current I (see Fig. 1). We define a function $S_I^{V_B}$, which is a measure of this dependency,

$$S_I^{V_B} = \frac{\partial V_B}{\partial I} \quad (1)$$

In many amplifier circuits the quiescent value of I can change as a function of temperature, where in general the trend is for a positive temperature coefficient, that is, I increases with temperature. To some extent this can be compensated by a suitable choice of feedback structure to the input stage, but this may well compromise other areas of performance and prove impractical to implement in a low-feedback amplifier.

This communication addresses the optimization of $S_I^{V_B}$ and suggests a simple modification to the design of the amplified diode which allows $S_I^{V_B}$ to be zero or indeed negative, thus reducing the tendency for increased output device bias current I_Q with temperature due to changes within the input stage of the amplifier.

1 THE MODIFIED AMPLIFIED DIODE

The modification to the basic amplified diode that enables optimization of $S_I^{V_B}$ is shown in Fig. 2 where

* Manuscript received 1983 Aug.

I_E is the emitter current and V_{BE} the base emitter voltage.

The addition to the circuit is the resistor R_3. To investigate the operation of the modified circuit, we calculate the parameter $S_I^{V_B}$. We proceed by choosing resistors R_1 and R_2 such that the current in R_1 is

$$I_{R1} = \sqrt{I_C I_B} = \frac{I_C}{\sqrt{\beta}} \tag{2}$$

where β is the current gain, I_C the collector current, and I_B the base current of T_1. Thus as an example, if $\beta = 100$, then $I_{R1} \cong 10I_B$ and $I_C \cong 10I_{R1}$. This will therefore realize good bias stability within the amplified diode. Consequently we may assume (for high B) that

$$I_C \cong I \ . \tag{3}$$

Hence,

$$(V_B + IR_3)\frac{R_1}{R_1 + R_2} = V_{BE}$$

and thus,

$$V_B = \left(1 + \frac{R_2}{R_1}\right)V_{BE} - IR_3 \ . \tag{4}$$

Differentiating V_B with respect to I to determine $S_I^{V_B}$, we have

$$S_I^{V_B} = \left(1 + \frac{R_2}{R_1}\right)\frac{\partial V_{BE}}{\partial I} - R_3 \ . \tag{5}$$

Since $I_E \cong I$, then from the diode equation,

$$I_E = I_S e^{qV_{BE}/KT} \cong I$$

where I_S is the transistor saturation current, q is the charge on an electron, K is Boltzmann's constant, and T is the junction temperature.

Differentiating, $\partial V_{BE}/\partial I \cong KT/qI$, and thus,

$$S_I = \left(1 + \frac{R_2}{R_1}\right)\frac{KT}{qI} - R_3 \ . \tag{6}$$

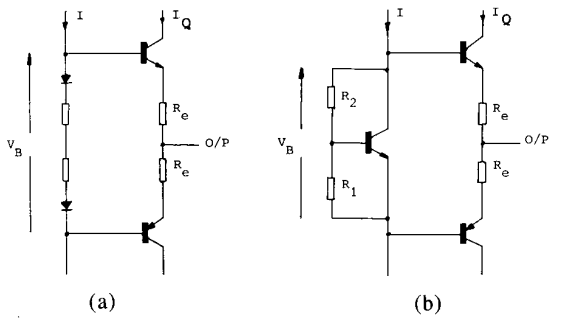


Fig. 1. Basic output-cell biasing circuit. (a) Series diode. (b) Amplified diode.

At room temperature $KT/q \cong 0.025$ V, and

$$S_I^{V_B} = \frac{0.025}{I}\left(1 + \frac{R_2}{R_1}\right) - R_3 \ . \tag{7}$$

Eq. (7) allows R_3 to be selected to minimize $S_I^{V_B}$. Hence for zero S_I, where optimum $R_3 = R_{3\,\mathrm{opt}}$,

$$R_{3\,\mathrm{opt}} = \frac{0.025}{I}\left(1 + \frac{R_2}{R_1}\right) \ . \tag{8}$$

Under this condition the bias voltage V_B is to a good first-order approximation independent of I. However, in practice it is suggested that $R_3 > R_{3\,\mathrm{opt}}$, thus implying a negative S_I. This will counteract tendencies for thermal runaway with increasing ambient temperature. Also, as $R_{3\,\mathrm{opt}}$ is temperature dependent, the value of R_3 should be calculated at the maximum device temperature. Thus for lower temperatures S_I will be slightly negative.

In circuit applications requiring a bias voltage V_B of several V_{BE}, such as where Darlington output devices are used, two transistors may be used as shown in Fig. 3.

2 CONCLUSIONS

This communication has discussed a modification to the basic amplified–diode circuit which will minimize the dependency of the bias voltage V_B on the magnitude of the amplified-diode operating current. The modification is simple, yet has proved to be extremely effective in operation. It is important on two counts. First, it will minimize changes in output-cell bias current due to changes in the driving circuit, which will generally be temperature dependent. Also in circuits that use a differential drive current to the amplified diode, it will

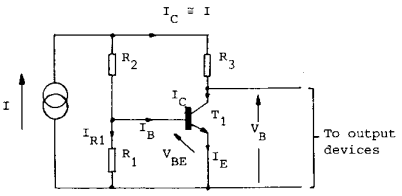


Fig. 2. Modified amplified-diode circuit.

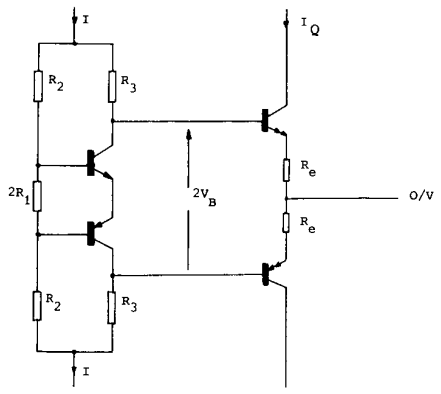


Fig. 3. Two-transistor amplified-diode circuit for use with Darlington output transistor.

minimize common-mode current variations with this drive configuration. Finally it should be noted that the modification in no way compromises the performance of the amplified diode with respect to thermal tracking of the output devices since for constant I, the voltage IR_3 is almost independent of temperature where, from Eq. (4),

$$\frac{\partial V_B}{\partial T} = \left(1 + \frac{R_2}{R_1}\right) \frac{\partial V_{BE}}{\partial T} . \quad (9)$$

3 REFERENCES

[1] P. Horowitz and W. Hill, *The Art of Electronics* (Cambridge University Press, London, 1980), pp. 75–77.

[2] M. J. Hawksford, "Distortion Correction in Audio Power Amplifiers," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 29, pp. 27–30 (1981 Jan./Feb.).

[3] E. M. Cherry, "Feedback, Sensitivity, and Stability of Audio Power Amplifiers," *J. Audio Eng. Soc.*, vol. 30, pp. 282–294 (1982 May).

THE AUTHOR

Malcolm Hawksford was educated at the University of Aston in Birmingham, England, from 1965 to 1971. In 1968 he obtained a first class honors degree in electrical engineering, and that same year was awarded a BBC research scholarship to investigate the application of deltamodulation to color television. In 1972 he obtained a Ph.D. degree. In 1971 Dr. Hawksford became a lecturer at the University of Essex, England, in the Department of Electrical Engineering Science. During his time at Essex he has actively pursued research projects within the field of audio engineering, where projects on power amplifier design, loudspeaker crossover design, analog-to-digital conversion, and music synthesis have been undertaken. He has presented papers at conventions of the Audio Engineering Society. He is currently a member of the AES, IEE, and RTS.

PAPERS

Reduction of Transistor Slope Impedance Dependent Distortion in Large-Signal Amplifiers*

MALCOLM HAWKSFORD

University of Essex, Department of Electronic Systems Engineering, Colchester CO4 35Q, UK

0 INTRODUCTION

The static characteristics of a bipolar transistor reveal that, under large-signal excitation, there are sources of significant nonlinearity. In an earlier paper [1] consideration was given to the I_E/V_{BE} nonlinearity, where a family of techniques was presented to attempt local correction of this error mechanism. However, the collector–emitter and collector–base slope impedance of transistors also result in significant distortion, where under large-signal conditions they can become a dominant source of error [2].

The static characteristics show only part of the problem; a more detailed investigation reveals capacitive components which are dependent upon voltage and current levels. Consequently under finite-signal excitation, modulation of the complex slope impedances results in dynamic distortion. It will be shown that the level of error that results from slope distortion is not strongly influenced by negative feedback once certain loop parameters are established. Also, because of the frequency and level dependency of slope distortion, the overall error will contain components of both linear and nonlinear distortion that are inevitably linked to individual device characteristics. It is therefore anticipated that a change of transistor could, in principle, lead to a perceptible change in subjective performance, even when the basic dc parameters are similar.

In this paper consideration is given to a class of voltage amplifiers employing a transconductance gain cell g_m, a gain-defining resistor R_g, and a unity-gain isolation amplifier, together with an overall negative-feedback loop. This structure is typical of most voltage and power amplifiers. However, although it is more usual to focus attention on input stage and output stage distortion, we shall consider in isolation the distortion due only to slope impedance modulation and assume other distortions are controlled to an adequate performance level. It will be demonstrated that significant distortion results from slope modulation, and a design methodology is presented to virtually eliminate its effect, even when the slope parameters are both indeterminate and nonlinear and when signals are of substantial level.

We commence our study by investigating the role of negative feedback as a tool for the reduction of slope distortion and to show that although effective, in isolation, it is not an efficient procedure.

1 NEGATIVE FEEDBACK AND THE SUPPRESSION OF SLOPE IMPEDANCE DEPENDENT DISTORTION

Consider the elementary amplifier shown in Fig. 1, where the principal loop elements are transconductance g_m, gain-defining resistor R_g, and feedback factor k. The nonideality of the transconductance cell is represented by an output impedance Z_n, where ideally $Z_n = \infty$, but in practice is finite and signal dependent. (Any linear resistive component of Z_n is assumed isolated and lumped with R_g.) In general, Z_n is a composite of the slope parameters of the output transistors in the transconductance cell. It can also include a reflection of any load presented to the amplifier. However, we assume here a perfect unity-gain buffer amplifier to isolate the slope distortion of the transconductance cell.

Although Z_n is signal dependent, our analysis will assume small-signal linearity so that performance sensitivity to Z_n can be established. However, the circuit topologies presented in Sec. 3 are not so restricted and can suppress the nonlinearity due to Z_n modulation.

For a target closed-loop gain γ there is a continuum of k and R_g for a given g_m, where the target closed-loop gain γ for $Z_n = \infty$ is defined,

$$\gamma = \frac{g_m R_g}{1 + k g_m R_g} . \tag{1}$$

Hence for a given k, g_m, and γ, R_g is expressed as

$$R_g = \frac{\gamma}{g_m(1 - \gamma k)} \tag{2}$$

* Manuscript received 1987 June 22.

where, for $0 \leq k \leq 1/\gamma$, then $\gamma/g_m \leq R_g \leq \infty$.

The actual closed-loop gain A, for finite Z_n, is

$$A = \frac{g_m Z_n R_g}{Z_n + R_g + k g_m Z_n R_g} \tag{3}$$

and eliminating R_g defined by Eq. (2) for selected target gain γ and transconductance g_m,

$$A = \frac{g_m Z_n}{1 + Z_n g_m/\gamma} . \tag{4}$$

This result demonstrates that the dependence of the transfer function A on Z_n is independent of the selection of feedback factor k, provided the condition of Eq. (2) is satisfied to set the target gain γ.

The error contribution due to Z_n can be estimated by evaluation of the transfer error function [3], [4] E defined by

$$E = \frac{A}{\gamma} - 1 \tag{5}$$

where E represents the ratio of error signal to primary signal and can be visualized according to Fig. 2.

Substituting A from Eq. (4) into Eq. (5),

$$E = \frac{-\gamma}{\gamma + g_m Z_n} . \tag{6}$$

In practice $g_m Z_n >> \gamma$ for a well-behaved amplifier, whereby

$$E \simeq \frac{-\gamma}{g_m Z_n} . \tag{7}$$

The results of Eqs. (6) and (7) reveal that to reduce the dependence on slope distortion, the product $\{g_m Z_n\}$ must increase. However, it is important to observe that Z_n reduces with increasing frequency due to device capacitance and that g_m also reduces with frequency due to closed-loop stability requirements, so that there are fundamental constraints on the effectiveness of slope distortion reduction using overall negative feedback, particularly at high frequency.

As an aside we are assuming g_m to be linear. In practice a reduction of R_g places a heavier current demand on g_m; thus a greater distortion contribution from g_m is to be anticipated for a given output [1]. Also, in power amplifier circuits, the output stage will exhibit distortion under load, a factor not considered in the present discussion. However, the independence of E on k and R_g for a given γ and g_m is true for distortion resulting only from Z_n, and when considered in isolation, it is an interesting example of a distortion that is not reduced by moving from a zero-feedback to a negative-feedback topology, especially as the choice of R_g is often the principal distinction between low-feedback and high-feedback designs [5].

In the next section the common-emitter amplifier is examined as a transconductance cell and current mirror, and an estimate is made of the output impedance Z_n for a range of circuit conditions.

2 OUTPUT IMPEDANCE OF COMMON-EMITTER AMPLIFIER

The common-emitter amplifier is shown in Fig. 3 in both single-ended and complementary formats. In this section the output impedance of the common-emitter amplifier is analyzed in terms of the small-signal parameters for a range of source resistances R_s and emitter resistances R_E. For analytical convenience, the base and emitter bulk resistances are assumed lumped with R_s and R_E, respectively.

Fig. 4 illustrates a small-signal transistor model of the common-emitter cell, where z_{ce} and z_{cb} represent collector–emitter and collector–base slope impedances, respectively, and h_{fe} is the collector–base current gain.

The output impedance Z_c observed at the collector of the common-emitter cell is given by

$$Z_c = \frac{v_o}{\alpha i_o} = \frac{1}{\alpha}\left\{ z_{ce} + R_E + \frac{z_{ce}}{z_{be}} [R_E + R_s(1 - \alpha)] (1 + h_{fe}) \right\} \tag{8}$$

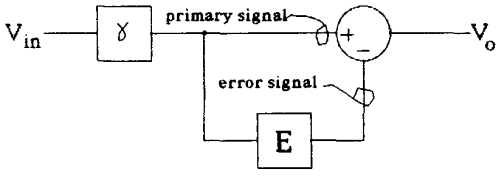


Fig. 2. Transfer error function model of voltage amplifier in Fig. 1.

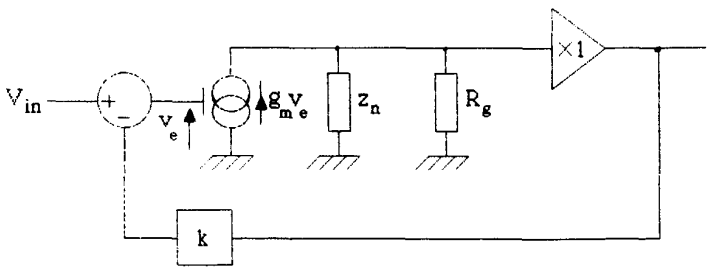


Fig. 1. Elementary amplifier topology using transconductance cell and gain-defining resistor.

where the collector/emitter current division factor is

$$\alpha = 1 + \frac{z_{be}z_{ce} + R_E\lambda}{z_{be}z_{cb} + R_s\lambda} \quad (9)$$

and

$$\lambda = (1 + h_{fe})z_{ce} + z_{cb} + z_{be} \quad (10)$$

or, alteratively, eliminating α,

$$Z_c = \frac{(z_{ce} + R_E)(z_{be}z_{cb} + R_s\lambda) + (1 + h_{fe})z_{ce}(R_Ez_{cb} - R_sz_{ce})}{z_{be}(z_{cb} + z_{ce}) + \lambda(R_s + R_E)} . \quad (11)$$

The expressions for Z_c reveal significant complexity, which is compounded by the signal dependence of the small-signal parameter set $\{z_{ce}, z_{cb}, z_{be}, h_{fe}\}$.

To simplify the results, consider a family of approximations for Z_c for specific cases of R_s and R_E, so that the dominant contributors to the output impedance can be determined.

1) *Case 1*: $R_s = 0$, $R_E = 0$.

Eq. (11) reduces to

$$Z_c \simeq \frac{z_{ce}z_{cb}}{z_{ce} + z_{cb}} \quad (12)$$

that is, Z_c is parallel combination of z_{ce} and z_{cb}.

2) *Case 2*: $R_s = 0$, $R_E >> z_{be}/(1 + h_{fe})$.

Eq. (10) approximates to $\lambda = (1 + h_{fe})z_{ce}$ and the denominator of Eq. (11) reveals $\lambda R_E >> z_{be}(z_{cb} + z_{ce})$. Hence,

$$Z_c \approx z_{cb} . \quad (13)$$

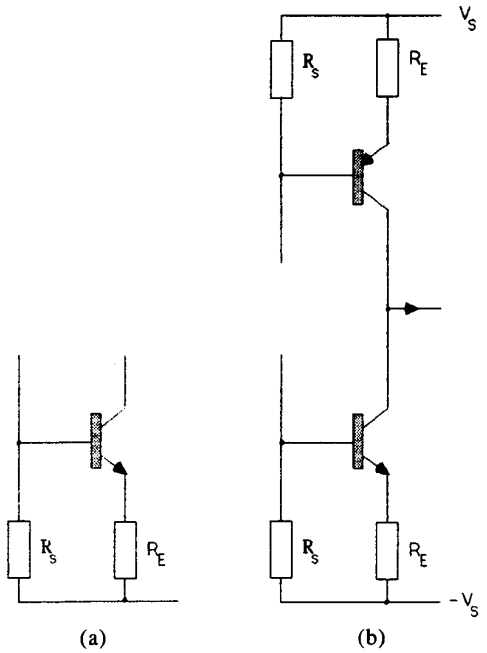


Fig. 3. Common-emitter gain cells. (a) Single-ended current mirror. (b) Complementary current mirror.

This case is typical of the current source and grounded-base amplifier as used in the cascode configuration.

3) *Case 3*: $R_s >> z_{be}$, $R_E = 0$.

From Eq. (11),

$$Z_c \simeq \frac{z_{cb}}{\frac{z_{cb}}{z_{ce}} + \frac{z_{be} + (1 + h_{fe})R_s}{z_{be} + R_s}} \quad (14)$$

where, for $R_s >> z_{be}$, Z_c is z_{ce} in parallel with $z_{cb}/(1 + h_{fe})$ and represents the worst-case output impedance condition.

4) *Case 4*: $R_s >> z_{be}$, $R_E >> z_{be}/(1 + h_{fe})$.

Applying inequalities to Eq. (11), and noting $z_{be} << z_{ce}, z_{cb}$,

$$Z_c \approx \left[\frac{z_{ce}\,z_{cb}}{(1 + h_{fe})z_{ce} + z_{cb}}\right]\left[\frac{R_s + (1 + h_{fe})R_E}{R_s + R_E}\right] + \frac{R_sR_E}{R_s + R_E} . \quad (15)$$

In selecting a circuit topology it should be noted that $z_{cb} > z_{ce}$; thus the grounded-base stage as used in the cascode will offer superior results in terms of output impedance. Nevertheless, z_{cb} is still signal dependent and represents a significant distortion mechanism where large signals are encountered, especially as z_{cb} falls with frequency. Such distortion is demonstrated in Sec. 5.

In Sec. 4 a new form of distortion correction is proposed that reduces output impedance dependence on both z_{ce} and z_{cb} even when nonlinear, and results in lower overall distortion that is virtually frequency independent.

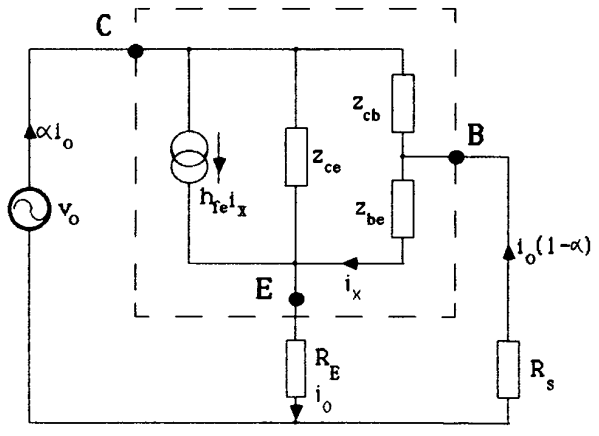


Fig. 4. Small-signal model of common-emitter amplifier showing slope impedances z_{ce} and z_{cb}.

3 REDUCTION OF NONLINEAR SLOPE IMPEDANCE DEPENDENT DISTORTION

The output impedances of the grounded-base and common-emitter amplifier cells are bounded by the device slope impedances z_{cb} and z_{ce}, respectively, as demonstrated by cases 2 and 3 in Sec. 2. However, an examination of Eq. (8) reveals that the factor α in the denominator restricts the output impedance. If a modified circuit topology could be realized such that the base current is summed with the collector current but without incurring an extra load on the collector, then the expression for collector output impedance would become

$$Z_{cu} = \frac{v_o}{\alpha i_o + (1 - \alpha) i_o} = \frac{v_o}{i_o}.$$

Hence from Eqs. (8)–(10) an upper bound on Z_{cu} is established where

$$Z_{cu} = R_E + z_{ce} + \frac{(1 + h_{fe}) z_{ce}(z_{cb} R_E - z_{ce} R_s)}{z_{be}\, z_{cb} + R_s \lambda}. \quad (16)$$

An examination of Eq. (16) reveals that, with typical component values and transistor parameters, a substantial increase in collector impedance is possible and that this is achieved even when z_{ce} and z_{cb} are dynamic. However, this result is an upper bound that assumes that all the base current is returned to the collector. In practical topologies this is compromised by a small margin, so that lower values should be anticipated.

Two circuit approaches have been identified to meet the requirement of base and collector current summation without direct connection to the collector. These are based on a local feedforward and feedback strategy, respectively, and can be used independently or compounded to give further enhancement.

3.1 Feedforward Topology

The feedforward topology is a derivative of the Darlington transistor that is occasionally employed in power amplifier current mirrors [6], [7]. In Fig. 5 two circuit examples are presented which yield similar performance. In each circuit the base current of the output device is returned to the emitter via the emitter–collector of the driver stage. Consequently the advantages of the Darlington are retained, yet with an enhanced output impedance realized by removing the respective currents in z_{ce} and z_{cb} from the output branch of the complementary stage. It should be noted that the collector–emitter voltage variation of the drivers is small, with only the output collectors swinging the full range of output voltage. The conventional Darlington connection of parallel collectors compromises this ideal, with the driver stage adding a degree of slope distortion under large-signal excitation. It is, however, important to note that a small fraction of output transistor base current is not returned to the emitter and is dependent on the ratio of R_E to transistor output impedance as seen at the emitter of the output device. This fractional loss of current will lower the bound suggested by Eq. (16), although there is still substantial advantage.

3.2 Feedback Topology

The conventional cascode as illustrated in Fig. 6(a) offers an output impedance approaching z_{cb}, which is a significant improvement over the common-emitter stage as $z_{cb} > z_{ce}$. A simple modification to the basic circuit can return the base current of the grounded-base stage to the emitter of the common-emitter stage. Consequently signal current flowing in both z_{ce} and z_{cb} now form local loops which do not include the output branch. The new topology is shown in Fig. 6(b), while in Fig. 6(c) the basic current paths are illustrated which apply even when z_{ce} and z_{cb} are nonlinear. Again, it is only the output device whose collector is required to swing over the full output voltage; thus the common-emitter stage offers a minimal slope distortion contribution.

In circuit applications where the common-emitter stages operate at a high bias current to improve I_E/V_{BE} linearity, a bypass current I_x [see Fig. 6(b)] can lower the operating current of the common-base stage. This technique both reduces output device power dissipation and aids a further increase in the slope impedances, while circuit symmetry ensures that noise in I_x does not flow in the output branch. As a practical detail, experimentation has revealed the desirability of ac bypassing of the base bias resistance of the grounded-base stages [see capacitors C in Fig. 6(b)]. This both enhances circuit operation and eliminates any tendency toward high-frequency oscillation due to the positive-feedback loop formed by the base–emitter connections.

3.3 Compound Feedback/Feedforward Topologies for z_{ce}, z_{cb} Reduction

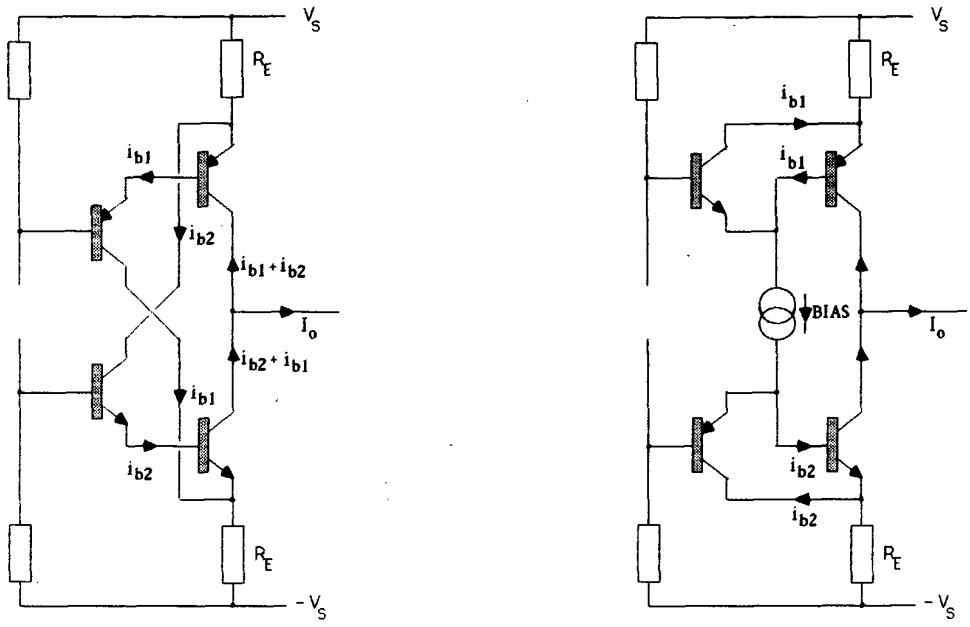
The methods based on feedforward and feedback addition of the output device base current can be compounded to offer further performance advantage. There are many possible topologies offering minor variations, though each uses the same basic concept. It is not intended to analyze each variant, though a family of topologies is presented in Fig. 7 to stimulate development.

4 NOISE CONTRIBUTION OF GROUNDED-BASE STAGE WITH BASE CURRENT SUMMATION

In this section brief consideration is given to the contribution of noise from the common-base stage in the cascode for the two basic topologies shown in Fig. 8.

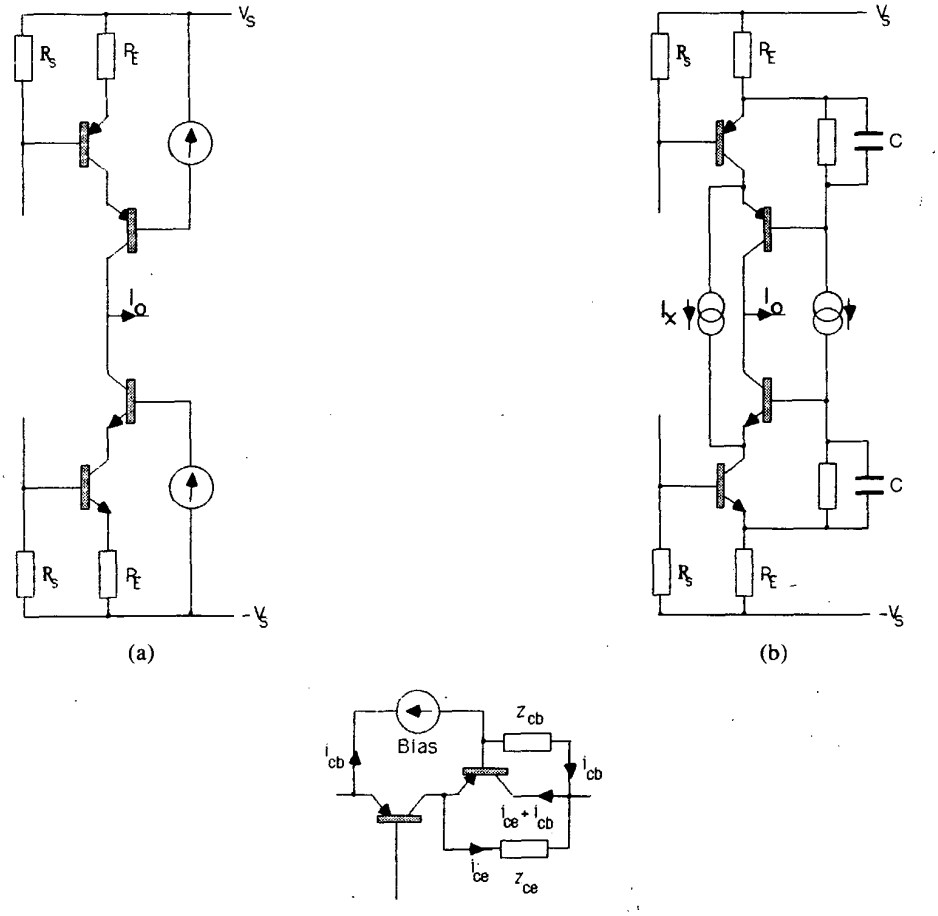
In both cases let $\overline{i_{cn}^2}$ be the mean square noise current in the collector of the common-emitter stage and let the common-base stage have respective noise voltage and noise current sources $\overline{e_n^2}$ and $\overline{i_n^2}$.

It is clear that because the common-emitter stage offers a relatively high output impedance at the collector, the equivalent voltage noise generator of the common-base stage yields a negligible contribution to the output



(a) (b)

Fig. 5. Two examples of feedforward addition of output stage base currents using a two-stage topology. (Observe base current paths i_{b1} and i_{b2}.)



(a) (b)

(c)

Fig. 6. Slope distortion reduction using feedback topology. (a) Conventional cascode. (b) Enhanced cascode. (c) Illustration of signal current paths i_{ce}, i_{cb} in z_{ce}, z_{cb}.

noise current.

However, an inspection of the noise current paths reveals that in Fig. 8(a) almost all $\overline{i_n^2}$ must flow in the collector, hence effective load, while in Fig. 8(b) virtually all the noise current circulates locally through the common-emitter stage, resulting in only a fraction, $\approx \overline{i_n^2}/[1 + 1/h_{fe} + h_{fe}R_E/(R_s + R_E + z_{be})]^2$, appearing in the collector (assuming similar transistor h_{fe}'s). Consequently with the enhanced topology there is virtually no extra noise generated by the addition of the common-base stage. Hence the output noise current is also $\overline{i_{cn}^2}$.

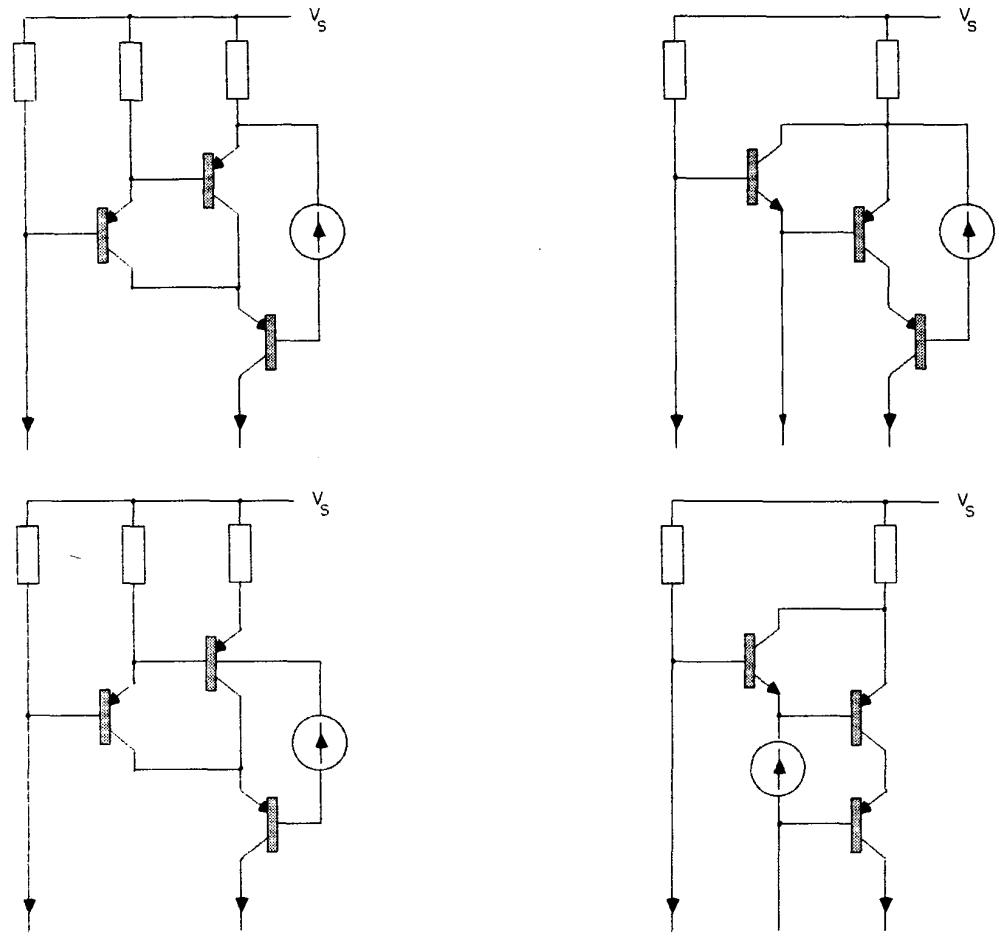


Fig. 7. Circuit examples using two-stage common-emitter amplifier with a common-base output stage.

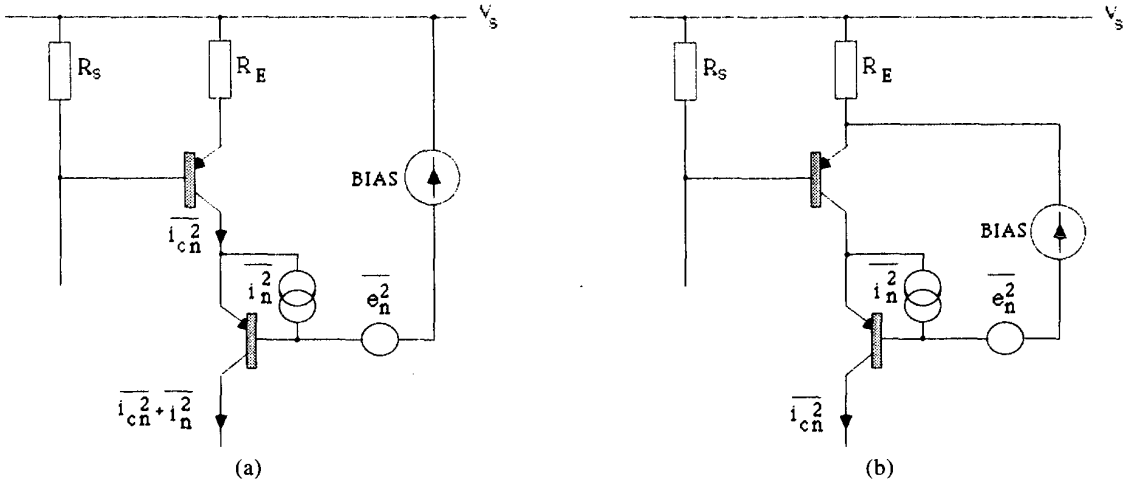


Fig. 8. Noise sources of common-base stage. (a) Conventional cascode. (b) Enhanced cascode.

5 MEASURED PERFORMANCE ADVANTAGE OF ENHANCED TOPOLOGY

To highlight the performance advantage of the modified common-base stage and to demonstrate the significance of slope distortion at large signal levels, a test circuit was constructed to validate the technique and to permit an objective assessment.

Three variants of the circuit were constructed and tested with ascending levels of modification. The enhanced topology is shown in Fig. 9(c), with the comparative output stage variants highlighted in Fig. 9(a) and (b). The circuit is dc coupled and no overall feedback is used. The output voltage is derived using a 10-kΩ gain-defining resistor R_g, and an offset-null potentiometer is provided since no servo amplifier is used. The total harmonic distortion results are given in Table 1. All measurements were performed with a sinusoidal input and an output voltage of 80 V peak to peak.

The results show that the basic circuit exhibits a distortion rising with frequency, reaching an unacceptable 1.9% at 50 kHz. This result is a function of the voltage-dependent nature of the device capacitance and represents a severe dynamic distortion. The conventional cascode exhibits a marked improvement, which reflects the popularity of this topology, where distortions are consistently reduced by 20 dB compared with the no-cascode circuit. However, although distortion products are of a lower order, they are still frequency dependent. This difference in performance arises from the basic common-emitter stage having an output impedance $\approx z_{ce}$, while the common base stage is z_{cb}, where $z_{cb} > z_{ce}$, though they follow the same basic frequency dependence, hence the tracking of the distortion figures.

However, the enhanced cascode, where performance is almost independent of both z_{ce} and z_{cb}, shows a distortion reduction greater than 40 dB at 50 kHz with a very desirable 31.8-dB improvement at 1 kHz over the basic circuit. Of particular significance is the almost frequency-independent nature of the distortion, together with the indication that the two stages of amplification are of inherent low distortion, though clearly they are a limit to linearity for the enhanced circuit. This performance level was masked by slope distortions in the conventional circuit.

These tests are sufficient to validate the technique, especially as the cost overhead is minimal compared with the conventional cascode, and represent a substantial performance enhancement irrespective of whether overall feedback is contemplated in a final design.

6 CONCLUSION

This paper has presented a method of reducing the performance dependence on transistor collector–emitter and collector–base slope impedance parameters, whereby useful distortion reduction can be achieved for large-signal voltage amplifiers.

A theory was presented to demonstrate that for a given input cell transconductance and closed-loop gain, the error signal due to the modulation of output impedance Z_n was not dependent on the level of feedback, provided g_m and target gain γ remained constant. Consequently for the test circuits of Section 5, if overall feedback was applied together with an appropriate increase in the gain-defining resistor R_g, the same level of distortion due to modulation of Z_n should be anticipated. (Note that a unity-gain buffer amplifier would be required.) However, if R_g is raised, the signal current level operating in the transconductance gain stage will fall, resulting in a reduced distortion from modulation in g_m. This later distortion would be particularly evident with the enhanced cascode, where modulation of g_m is now the limiting distortion mechanism.

The enhanced topology has specific application in large-signal voltage amplifiers and, with appropriate circuit additions, to power amplifiers. In particular, MOSFET power amplifiers can benefit by using a more optimum current source to drive the output stage since this reduces dependence on both gate-to-source voltage errors as well as slope impedance modulation errors [8].

A third area of application is RIAA disk preamplifiers that use a transconductance cell and a passive equalization-defining impedance [9], [10]. The more optimum current source will lower distortion and increase EQ accuracy as the current source exhibits a lower output capacitance, together with a higher output resistance, the latter particularly affecting low-frequency performance.

It is interesting to observe that if negative feedback alone were used to reduce error dependence on Z_n by the same factor as the enhanced cascode, at 1 kHz an increase in loop gain of more than 30 dB is required, or at 50 kHz this requirement rises to more than 40 dB. Such factors are often impractical to achieve, thus vindicating the adoption of the enhanced topology. However, more fundamentally, the distortion dependence on transistor slope impedance inevitably rises with both frequency and output voltage level, and moves against the loop gain requirement for stability, thus making negative feedback less effectual in suppressing slope-dependent nonlinearity.

The techniques described in this paper should also find application in circuits that require enhanced supply rail rejection. An appendix outlines how slope impedance distortion reduction can improve the performance of voltage/power amplifiers by enhancing the interface between amplifier stages which alternate their signal reference between ground and supply rail.

Although the reduction of large-signal-related errors arising from slope distortion has been the central thesis, the reduction of linear distortion at lower signal levels is also welcome. Slope distortion has been shown to involve several factors that depend on both transistors and the associated circuit elements in a particular application. Such device-specific distortion can, in principle, contribute to the subjective performance and

reflects the mutual interrelationship of transistors and circuit construction, which results in small deviations from the target transfer function.

The paper has presented a family of primitive circuit topologies based on the same principle as the enhanced cascode, which are candidates for adoption in transconductance-based amplifiers. There are numerous circuit possibilities for enhancement. However, the two

Table 1. Total harmonic distortion.

Test frequency, kHz	No cascode, %	Conventional cascode, %	Enhanced cascode, %
1	0.39	0.039	0.010
10	0.47	0.11	0.011
20	0.51	0.14	0.012
50	1.9	0.16	0.016

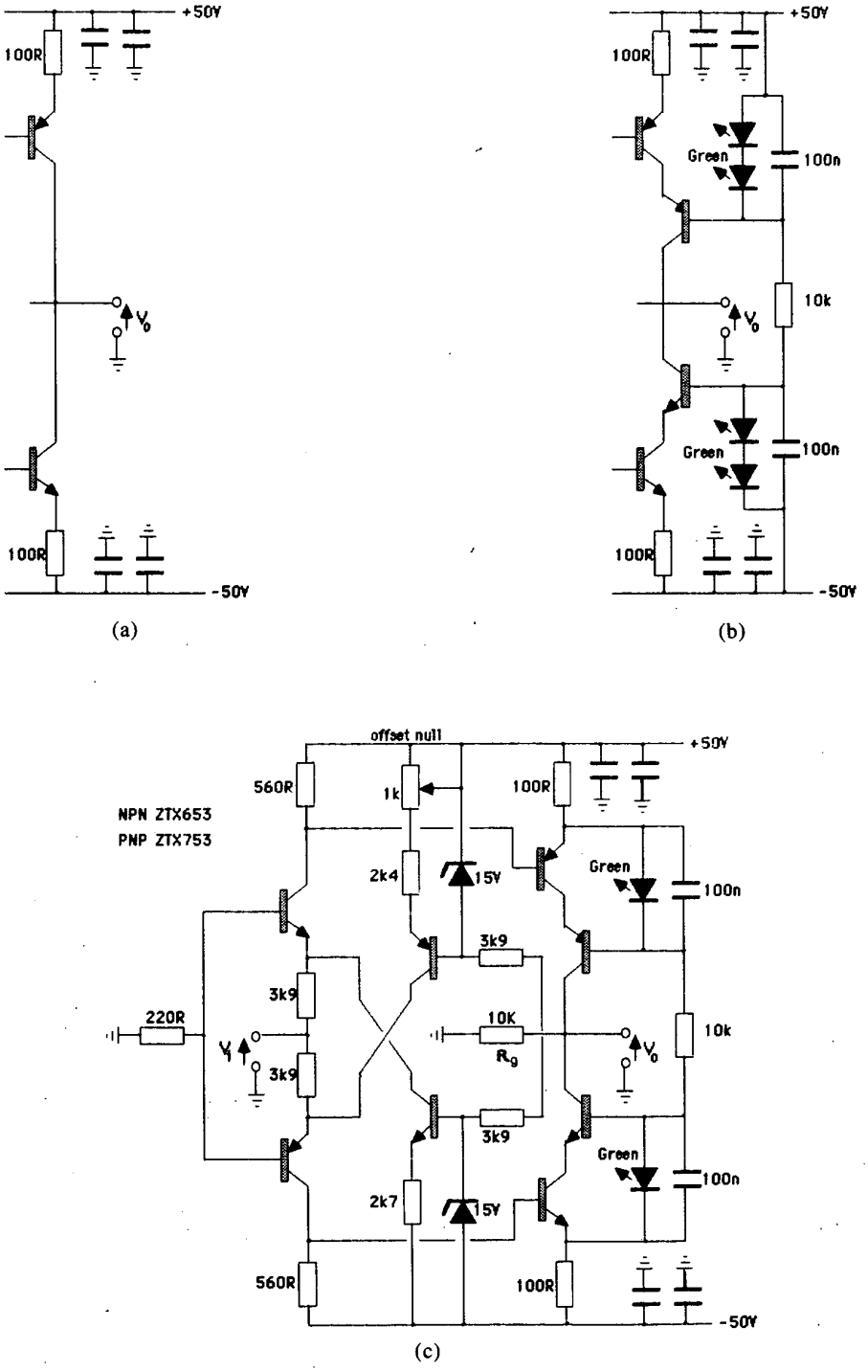


Fig. 9. Test circuit with three output stage variants. (a) Complementary common-emitter output stage. (b) Complementary cascode output stage. (c) Complete test circuit with enhanced cascode.

basic principles to be observed are

1) Adequately high effective emitter resistance R_E to disassociate z_{ce} from the output impedance at the collector;

2) Addition of base current to collector current, without adding extra circuitry to collector, to disassociate z_{cb} from the output impedance at the collector.

Observation of these two principles then enables a transformation of the signal level from low voltage to large voltage without incurring a significant distortion penalty due to dynamic modulation of the transistor slope parameters, together with a distortion characteristic that is considerably less frequency dependent.

7 ACKNOWLEDGMENT

The author wishes to gratefully acknowledge the assistance of Paul Mills from the Department of Electronic Systems Engineering for his support in constructing and compiling the measured data on the three circuit derivatives.

8 REFERENCES

[1] M. J. Hawksford, "Distortion Correction Circuits for Audio Amplifiers," *J. Audio Eng. Soc.*, vol. 29, pp. 503–510 (1981 July/Aug.).

[2] E. M. Cherry and G. K. Cambrell, "Output Resistance and Intermodulation Distortion of Feedback Amplifiers," *J. Audio Eng. Soc.*, vol. 30, pp. 178–198 (1982 Apr.).

[3] M. J. Hawksford, "Power Amplifier Output-Stage Design Incorporating Error-Feedback Correction with Current-Dumping Enhancement," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 960 (1983 Dec.), preprint 1995.

[4] M. J. Hawksford, "The Essex Echo: Reflexions," *HFN/RR*, vol. 30, pp. 35–40 (1985 Dec.).

[5] K. Lang, "The Lang 20W Class-A MOSFET Amplifier," *Audio Amateur*, vol. 2, pp. 7–12 (1986).

[6] E. M. Cherry, "Feedback, Sensitivity, and Stability of Audio Power Amplifiers," *J. Audio Eng. Soc.*, vol. 30, pp. 282–294 (1982 May).

[7] P. J. Walker and M. P. Albinson, "Current Dumping Audio Amplifier," presented at the 50th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 23, p. 409 (1975 June).

[8] R. R. Cordell, "A MOSFET Power Amplifier with Error Correction," *J. Audio Eng. Soc.*, vol. 32, pp. 2–17 (1984 Jan./Feb.).

[9] Y. Miloslavskij, "Audio Preamplifier with no TID," *Wireless World*, vol. 85, no. 1524, pp. 58–60 (1979 Aug.).

[10] O. Jones, " . . . about the genesis of the Pip," *HFN/RR* (Letter to the Editor), vol. 30, no. 12, p. 25 (1985 Dec.).

APPENDIX
SUPPLY RAIL REJECTION AS A FUNCTION OF INPUT STAGE AND CURRENT MIRROR SLOPE IMPEDANCES

In this appendix the sensitivity of a two-stage negative-feedback amplifier is determined as a function of the slope impedances Z_{n1} and Z_{n2} of the two stages. The basic circuit is shown in Fig. 10 where g_m is the transconductance of the input stage, m the current gain of the current mirror, R_g a gain-defining resistor, r_2 the input impedance of the current mirror ($r_2 << Z_{n1}$), and k the feedback factor.

Using linear analysis to express V_o as a function of both V_{in} and V_s,

$$V_o = \frac{mg_mR_gV_{in} + R_g[m/Z_{n1} + 1/Z_{n2} + r_2/Z_{n1}Z_{n2}]V_s}{(1 + r_2/Z_{n1})(1 + R_g/Z_{n2}) + kmg_mR_g} . \quad (17)$$

Let δ be the ratio of output to input transfer functions for inputs V_s and V_{in},

$$\delta = \frac{V_o/V_s}{V_o/V_{in}} \quad (18)$$

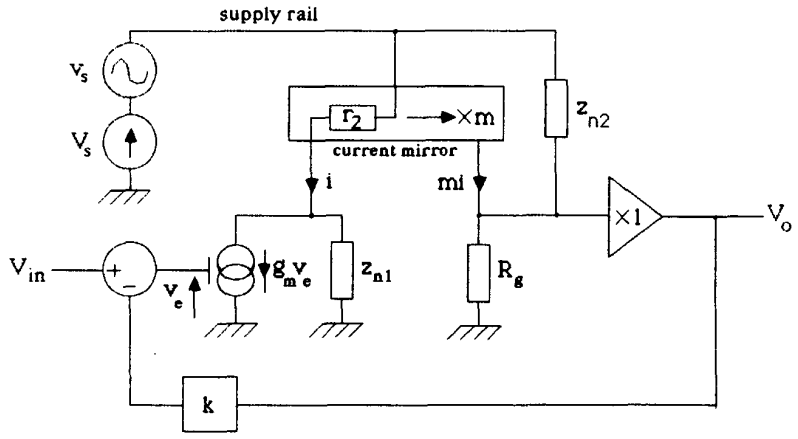


Fig. 10. Two-stage voltage amplifier with v_s representing power supply voltage variation.

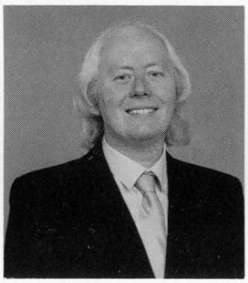
and

$$\delta = \left[\frac{1}{Z_{n1}} + \frac{1}{mZ_{n2}}\left(1 + \frac{r_2}{Z_{n1}}\right)\right]\frac{1}{g_m}. \quad (19)$$

The results show that the slope impedances define the suppression of supply rail rejection together with g_m. This is particularly important in power amplifier applications, where in class AB operation V_s is wide band (>>20 kHz) and a nonlinear function of the input signal due to output stage commutation. The advantages of maximizing both Z_{n1} and Z_{n2} and using separate power supplies for voltage amplifier and output stage in power amplifiers are evident.

Eq. (19) is also shown to be independent of R_g. However, in high loop gain applications where g_m is large, the high-frequency distortion characteristics together with the falling high-frequency gain of g_m may become a limiting factor, particularly if required to suppress wide-band power supply injection. In low-feedback applications, the slope impedance dependent distortion is suppressed more by the presence of R_g than by the presence of g_m. For example, observe how R_g and Z_{n2} form a potential divider to supply injected distortion, but as $R_g \to \infty$, the distortion is processed completely by the feedback loop. Also in low-feedback designs greater local feedback enhances the wide-band distortion characteristics of g_m and helps aid an overall distortion profile which is less frequency dependent.

THE AUTHOR



Malcolm Hawksford is a senior lecturer in the Department of Electronic Systems Engineering at the University of Essex, U.K., where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honors B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship where work on the application of deltamodulation to color television was undertaken.

Since his appointment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing and loudspeaker systems has been undertaken. Dr. Hawksford has had several AES publications that include topics on error correction in amplifiers and oversampling techniques for ADC and DAC systems. His supplementary activities include designing commercial audio equipment and writing articles for Hi-Fi News—activities that integrate well with visits to Morocco and France. His leisure activities include listening to music, motorcycling and motor mechanics. Dr. Hawksford is a member of the IEE, a Chartered Engineer, Fellow of the AES, and a member of the Review Board of the AES *Journal*.

PAPERS

Transconductance Power Amplifier Systems for Current-Driven Loudspeakers*

P. G. L. MILLS

Tannoy Limited, Coatbridge, Strathclyde ML5 4TF, UK

AND

M. O. J. HAWKSFORD

University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK

Moving-coil loudspeakers generally provide a substantial improvement in linearity when current driven, together with the elimination of voice-coil heating effects. Consequently there is a need to investigate low-distortion power amplifier topologies suitable for this purpose. After considering established current feedback approaches, a novel method using a common-base isolation stage is outlined and extended to show a prototype amplifier circuit in detail. In addition, the elements of a two-way active current-driven system are described, with low-frequency velocity feedback control derived from a sensing coil. The coupling error between this coil and the main driving coil is nulled by electronic compensation.

0 INTRODUCTION

The moving-coil drive unit can readily be shown to benefit in terms of linearity when controlled by a current source rather than the more conventional voltage source. Throughout this paper we will term this mode of operation *current drive*, whereby the amplifier source impedance can, to all intents and purposes, be considered infinite compared to the drive unit impedance.

Of the drive unit error mechanisms that can be countered by current drive, the voice-coil resistance is of particular interest. As a result of self-heating in excess of 200°C, the increase in coil resistance leads to sensitivity loss (often referred to as power compression [1], [2]), loss in electrical damping of the fundamental resonance, and crossover filter misalignment. In their paper Hsu et al. [3] concluded that a satisfactory method of compensating for the effect had yet to be found.

At higher frequencies, nonlinearity occurs as the coil inductance is modulated by movement in the magnetic circuit and by other effects such as magnetic hysteresis [4]. Measurements under current drive have shown, in comparison with voltage drive, a high-frequency distortion reduction of typically 20–30 dB for a bass–midrange drive unit.

These performance advantages arise from the coil resistance and inductance being totally eliminated from the system transfer function. The force on the cone is proportional to the voice-coil current, not the applied voltage. Analysis also shows a reduced dependence on nonlinearity within the force factor and mechanical impedance of the drive unit.

Thus as a result of the performance gains that can be demonstrated using current drive, there arises the need to investigate suitable power amplifier topologies to make the best of the technique. This paper therefore aims to review some of the earlier published work on transconductance amplifier design, while presenting new topologies and detailed circuitry of a two-way active prototype system. In addition, due to the loss of voice-coil damping under current drive, control circuitry for restoring damping by means of motional

* Manuscript received 1988 July 6. This paper expands on some areas covered by the authors in "Distortion Reduction in Moving-Coil Loudspeaker Systems Using Current-Drive Technology," volume 37, number 3 (1989 March).

feedback applied to the bass–midrange drive unit is described, along with the low-level crossover circuitry of the prototype system.

1 POWER AMPLIFIER TOPOLOGIES FOR CURRENT DRIVE

1.1 Review of Transconductance Amplifier Techniques

A transconductance power amplifier requires a high output impedance that is linear and frequency independent. It must also possess the attributes of a conventional voltage power amplifier such as high linearity, wide bandwidth, freedom from slewing-induced errors, and insensitivity to load variations (be they linear or nonlinear).

The most commonly used technique to obtain a high output impedance is to apply current feedback around a conventional power amplifier by means of a sensing resistor in the loudspeaker earth return [5], [6], as illustrated in Fig. 1. The transconductance g_m is defined

$$g_m = \frac{1}{R_f} . \quad (1)$$

The method has also been used in high-current industrial applications. There are two main disadvantages with such a system. First, the open-loop gain of the amplifier is frequency dependent as a result of the amplifier's dominant pole, and this is reflected in the output impedance. Second, the loudspeaker impedance, which is both frequency dependent and nonlinear, tends to modulate the transconductance of the amplifier. The fact that the load is not ground referenced may be considered inconvenient in some applications.

A refinement of the basic technique was described in Lewis [7]. The circuit was symmetrical in nature, using two current-sensing resistors, with a ground referenced load fed from MOSFET output devices. Good linearity was indicated at 10-W average power into a 5-Ω load. However due to class A operation, the design would be inefficient at the power levels necessary for a moving-coil drive unit (between 50 and 100 W typically). Care must be taken to minimize output offset current with this scheme.

A further ground-referenced current feedback scheme was described in Nedungadi [8], but this required the complexity of a differential voltage-to-current converter in conjunction with a floating sensing resistor in order to maintain current feedback.

Another technique used in implementing transconductance amplifiers involves the combination of supply current sensing around a follower, together with current mirrors feeding the load [9]. The arrangement is shown in Fig. 2, where R_f is a dummy load and the transconductance is again defined by Eq. (1). Operation of the circuit is typically in class AB.

While suitable for low output currents (<50 mA peak), the approach is difficult to extend to the levels required for driving a loudspeaker (typically 5 A peak or more) due to the linearity of the mirrors and also power loss in R_f. Although the mirrors could be arranged to provide current gain and could be partially linearized by error-correction techniques [10], the technique is not felt to offer a particularly practical solution.

1.2 Methods Using a Common-Base Isolation Stage

The approach devised to overcome the limitations cited as being inherent to existing topologies is illustrated in basic form by Fig. 3. The notable aspect of this strategy is the open-loop grounded base stage, which isolates the load Z_L from the main amplifier A_t while providing a naturally high output impedance without the use of overall current feedback. In addition a cascode configuration is formed in conjunction with the output devices in the main amplifier A_t. Resistor R_f defines the transconductance, driven from amplifier A_t, a voltage source, which may operate with low values of supply voltage $\pm V_{S1}$ to reduce power dissipation. The loudspeaker is referenced to ground and isolated from any feedback loop used to linearize the amplifier A_t. Although a successful prototype based on this scheme has been constructed, with amplifier A_t running in class A and with a class AB output stage, it is to some extent an uneconomical solution due to the need for two pairs of floating power supplies.

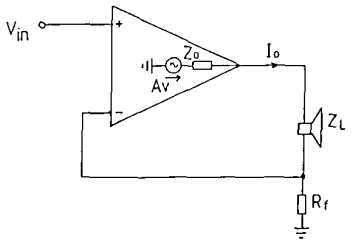


Fig. 1. Basic current-feedback-derived transconductance amplifier.

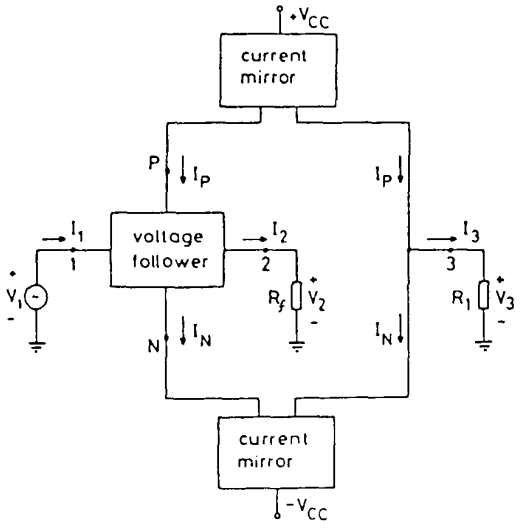


Fig. 2. Voltage-current converter. After Rao and Haslett [9].

A more viable alternative is the revised topology illustrated by Fig. 4. This circuit takes the form of a current amplifier of current gain

$$\alpha = -\frac{R_x}{R_f} \tag{2}$$

The first-stage power supply $\pm V_{S1}$ is ground referenced, unlike the previous case, meaning that several power amplifiers within an active system may share a common supply, thus reducing complexity and cost. Like the previous scheme, the current flowing in the transconductance defining resistor R_f is that which flows in the load Z_L, except for any base current lost to ground in the common-base stage. The fact that the amplifier A_i is referenced to the input of the common-base stage and not to ground tends to decouple it from any distortion appearing at the emitters of the common-base stage.

This topology forms the basis of the prototype system, the detailed circuitry of which is described in Sec. 2. On a practical note, it is important to provide adequate current gain in the common-base stage in order to prevent nonlinear current loss to ground, which introduces distortion.

1.3 Alternative Approaches

All of the circuits described so far rely on a current-sensing resistor to define the overall system transconductance. Even when this resistor is of a low value (about 1 Ω), it still tends to dissipate an appreciable amount of power. This element would at first seem to be fundamental to the design of a transconductance amplifier, but it is interesting to note the possibilities of transformer-derived feedback in perhaps reducing such losses.

Nordholt, in his classification of feedback configurations [11], described how transformer-derived feedback could be used to generate transconductance and current gain functions, as shown in Fig. 5. In Fig. 5(a) resistor R_f is still necessasry in order to define the stage transconductance.

Although no research has been directed in this area and the approach is only conceptual in nature, it may be worth further investigation, given a wide-bandwidth transformer design.

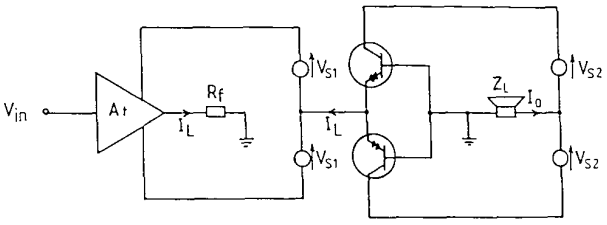


Fig. 3. Basic transconductance power amplifier using grounded-base output stage.

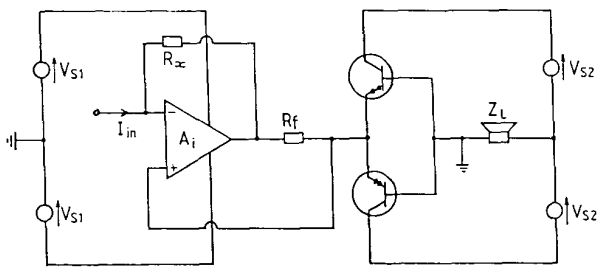


Fig. 4. Alternative configuration for current gain.

2 PROTOTYPE AMPLIFIER SYSTEM

2.1 General Overview

The two-way active loudspeaker system constructed to validate the basic approach proposed for high output impedance power amplifier design was based on the Celestion SL600 loudspeaker. In this section the current gain power amplifier is considered in detail along with the necessary transconductance preamplifier, while Sec. 3 considers the associated motional feedback control circuitry, which is required for the bass–midrange drive unit.

Throughout the design, the underlying philosophy has been to use symmetrical direct-coupled circuitry to give good transfer function linearity without recourse to high levels of overall negative feedback [12], [13]. DC stability is taken care of by servo amplifiers (feedback integrators).

2.2 Transconductance Preamplifier

Fig. 6 shows a two-stage design, the basic topology of which has often been used with overall feedback as

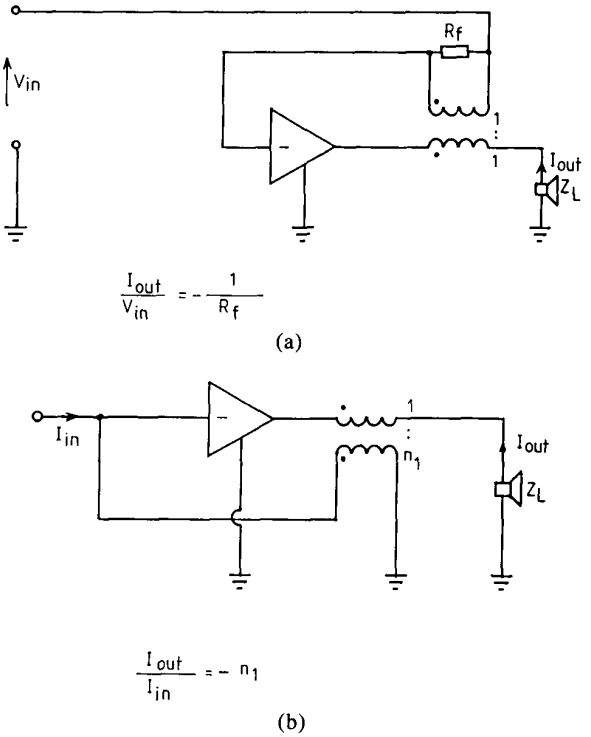


Fig. 5. Transformer-derived feedback systems. After Nordholt [11]. (a) Transconductance stage. (b) Current gain stage.

a voltage gain stage [14]. It is operated here open loop to provide a high output impedance and consequently must be capable of good linearity.

Transistor pairs Q_3/Q_4 and Q_5/Q_6 form cascodes to increase high-frequency linearity and give a high output impedance. Bias arrangements for the cascode are somewhat unusual in that resistors R_{14} and R_{15} are not returned to the supply rails, but are connected to the emitters of the common-emitter part of the cascode, thus avoiding nonlinearity from base current loss in the common-base devices [15]. This reduces high-frequency distortion by typically a factor of 10 at 20 kHz over the conventional bias method.

Operational amplifier IC_1 with associated passive components forms a current-sensing differential servo amplifier to null any output offset current due to imbalances in the main circuit and has no effect on performance within the audio band. This configuration of servo amplifier, to the authors' knowledge, has not been seen before in the literature.

At frequencies within the passband of the amplifier, the transconductance g_m may be approximated by the expression

$$g_m = \frac{I_{out}}{V_{in}} \approx \frac{R_6}{R_{11}(R_8/2 + R_{10})} . \quad (3)$$

With the component values shown, $g_m \approx 4$ mS.

In addition to the main input and output, an auxiliary velocity feedback input is provided along with an error-nulling output. These are only required for low-frequency use, and their function is described in Sec. 3 when considering the velocity feedback control circuitry.

2.3 Current Gain Power Amplifier

The current gain power amplifier, which accepts the output of the transconductance preamplifier, is based on the structure shown in Fig. 4. For the purpose of description, it is split into three sections: input amplifier, power follower, and common-base output stage. Both input amplifier and follower are represented by the gain block A_i in this simplified representation.

We consider first the input amplifier, Fig. 7. This is essentially the same topology as the transconductance preamplifier, but with a few refinements. Input stage biasing is performed with current sources based around transistors Q_1 and Q_2, instead of resistive biasing. This is a result of the need to provide immunity to the greater level of supply rail contamination caused by class AB operation of the power follower stage. The output from the transconductance preamplifier is fed to the emitters of the input devices Q_3 and Q_4, which thus operate in common-base mode. The first and second stages of the amplifier are coupled together by current mirror pairs Q_5/Q_8 and Q_9/Q_{12} to reduce loading effects and interaction between the two stages. These mirrors are themselves linearized by local error feedback correction consisting of transistor pairs Q_6/Q_7 and Q_{10}/Q_{11}. This approach has been previously documented [10], al-

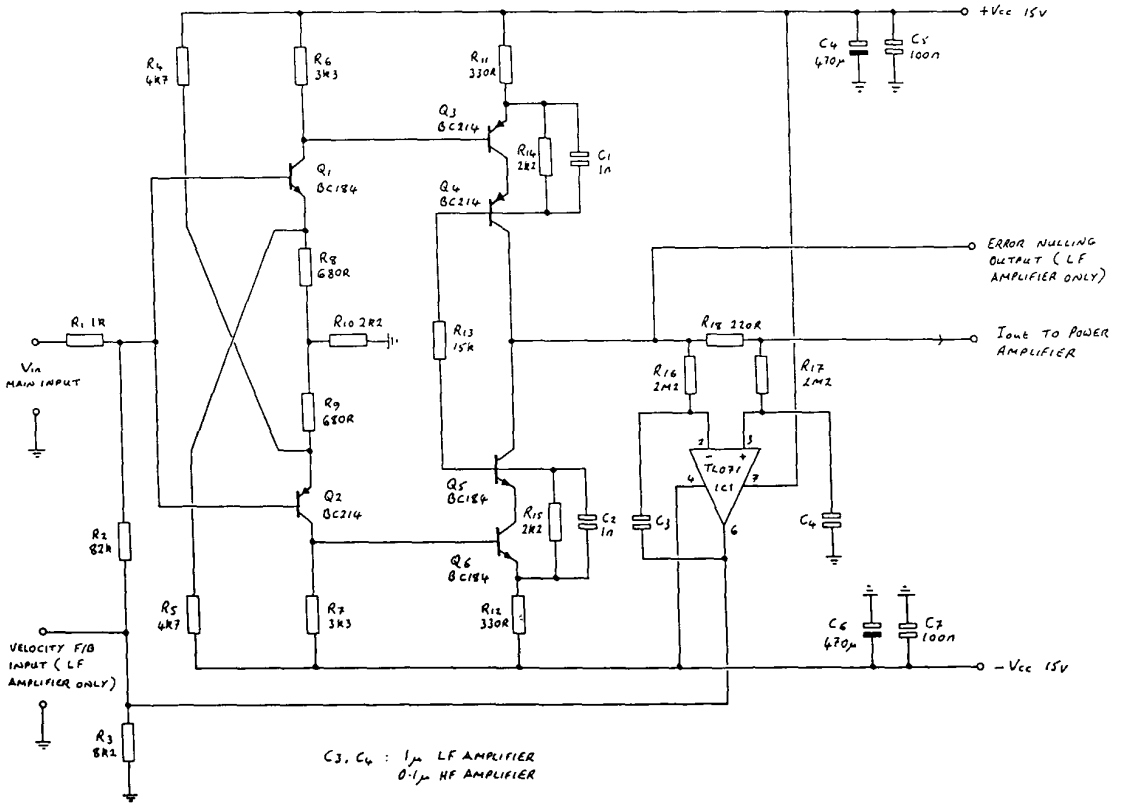


Fig. 6. Transconductance preamplifier.

though in this case some current gain has been introduced into the mirrors to enable correct quiescent operating conditions to be established in the first and second gain stages.

The outputs from the cascode pairs Q_8/Q_{13} and Q_{12}/Q_{14} are displaced ± 4 V about ground by green LEDs D_1–D_4, in order to bias the next stage. Resistor R_{28} and capacitor C_3 are included to define the open-loop gain characteristics of the amplifier, to ensure that stability is maintained under closed-loop conditions.

Fig. 8 shows the next section, which is a follower with extensive error-correction circuitry and is essentially similar to a previously published topology [16], but with improvements to biasing arrangements. It is worth briefly reviewing the principle of operation.

Transistor pairs Q_{16}/Q_{18} and Q_{17}/Q_{19} form a Darlington follower, preventing loading of the previous stage and driving the Darlington output devices Q_{30} and Q_{31}. Transistors Q_{28} and Q_{29} form V_{be} multipliers to bias the output Darlingtons, but are also configured as error amplifiers, which together with Q_{22} and Q_{23} form the main error feedback loop, delivering a correction current through resistors R_{38} and R_{78} in response to any nonlinearity in the output devices Q_{30} and Q_{31}. R_{48} is included as an adjustment to achieve the best distortion null.

In order to linearize Q_{18} and Q_{19}, which have to drive the output Darlingtons, additional error correction in the form of feedforward is applied with the aid of Q_{20} and Q_{21}, in combination with the input transistors Q_{16} and Q_{17}. Further linearization is achieved by current mirror transistors Q_{25} and Q_{26}, which form a negative feedback loop, thus reducing the source impedance seen by output Darlingtons Q_{30} and Q_{31}.

Moving now to Fig. 9, which shows the output common-base stage, the preceding follower drives current through resistor Q_{67}, which in conjunction with R_2 (Fig. 7) sets the midband current gain of the complete amplifier to around 800. Inductor L_2 serves to reduce the high-frequency current gain of the amplifier to ensure stability. The current in R_{67} flows into the common-base output stage, consisting of Darlingtons Q_{32} and Q_{33}, along with driver devices Q_{36} and Q_{37}, the bases of which are referenced to ground. Except for any current loss to ground, such as through the bases of these devices and through the biasing current sources (Q_{34}, Q_{35}), the current in R_{67} flows through the load via floating power supplies $\pm V_{CC2}$.

In order to establish a low-output offset current for the amplifier (typically less than ± 2 mA), a servo based around IC_1 and referenced to the input of the common-base stage, is used to feed a dc compensation current back to the input of the amplifier.

To prevent switch-on and switch-off transients from reaching the load, relay RL_1 is included, controlled by a time-delay circuit on startup and almost instantaneously dropping out on power down. The control circuitry to perform this function is not shown.

The power amplifier together with the transconductance preamplifier was evaluated in terms of standard

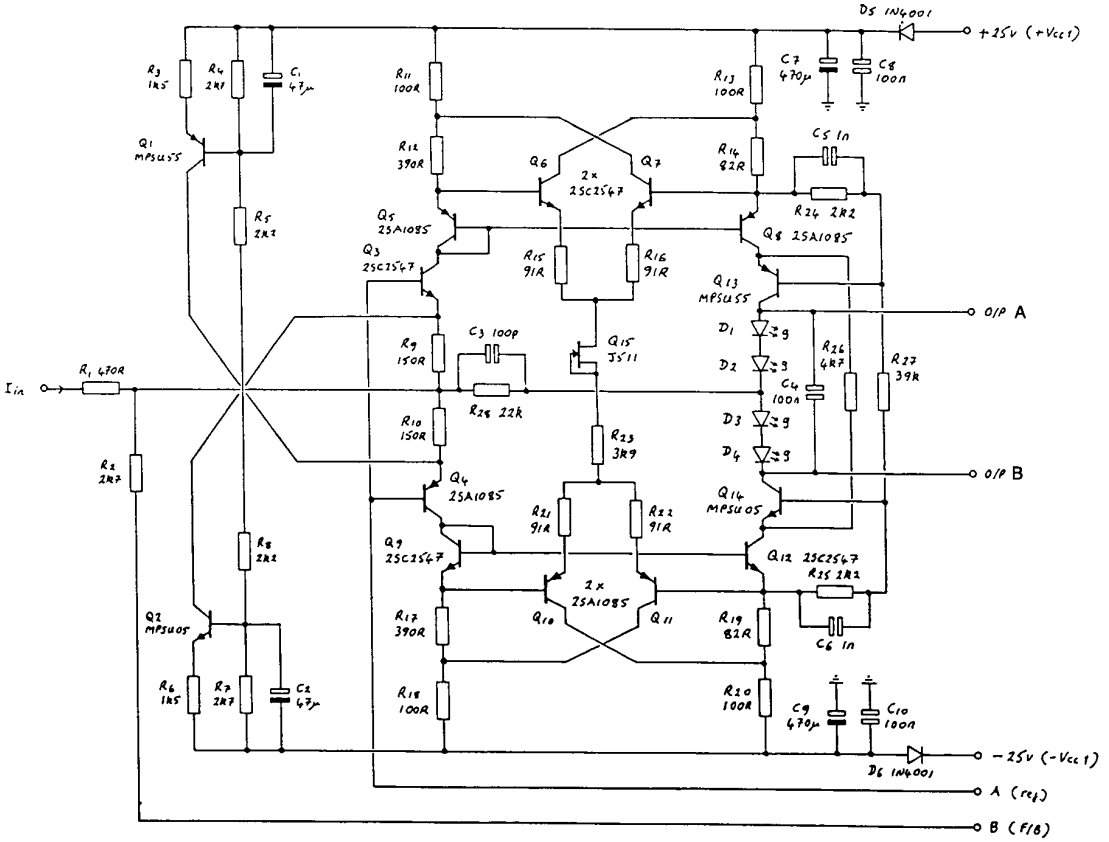


Fig. 7. Prototype amplifier, input stage.

measurements and found to be comparable with a typical high-performance conventional amplifier. The results are as follows:

Rated power output (8-Ω resistive load)	75 W average
Total harmonic distortion at rated power	
20 Hz	−79 dB
1 kHz	−86 dB
20 kHz	−68 dB
Intermodulation distortion (19 and 20 kHz at equal levels at rated power)	−86 dB
Hum and noise (re maximum output)	−90 dB
Small-signal bandwidth, −3 dB	0.1 Hz to 50 kHz
Output impedance*	
20 Hz	4.1 MΩ
1 kHz	106 kΩ
20 kHz	11.4 kΩ

* From computer simulation, due to the difficulty in performing these measurements.

It is interesting to note that the distortion measurements may only easily be made indirectly by converting the output current to a voltage, by means of a resistive load bank. The measurements as shown will thus reflect any nonlinearity in the load.

The protective features, consisting of output fuses and relay contact, should not introduce any degradation in performance, as they are in series with a high source impedance, which is not the case with a conventional power amplifier.

3 VELOCITY FEEDBACK CONTROL SYSTEM

3.1 Outline Approach

In order to compensate for the loss in electric damping of the bass–midrange unit caused by the high amplifier output impedance, velocity feedback was used to restore damping [5], [6]. While many forms of sensing arrangement have been described ([17]–[23], for example), the method adopted here is attractive for reasons of mechanical simplicity and cost effectiveness. The technique used is to wind a sensing coil over the main voice coil of the drive unit. The output voltage of the sensing coil will ideally be defined by

$$V_s = (Bl)_s u \tag{4}$$

where $(Bl)_s$ is the sensing coil Bl product, N/A, and u is the cone velcity, m/s.

Unfortunately an error is induced in the sensing coil by transformer action from the main driving coil. In the previously documented work induced errors were overcome by neutralizing coils or by an altogether more elaborate mechanical arrangement to physically isolate the driving and sensing coils. With the approach considered here, a procedure of electronic compensation has been chosen in order to avoid expensive tooling costs for a specialized drive unit.

The physical arrangement of the assembly is shown in Fig. 10. It should be noted that in this case, the sensing coil follows roughly the same Bl profile as the main driving coil, so the action of velocity feedback

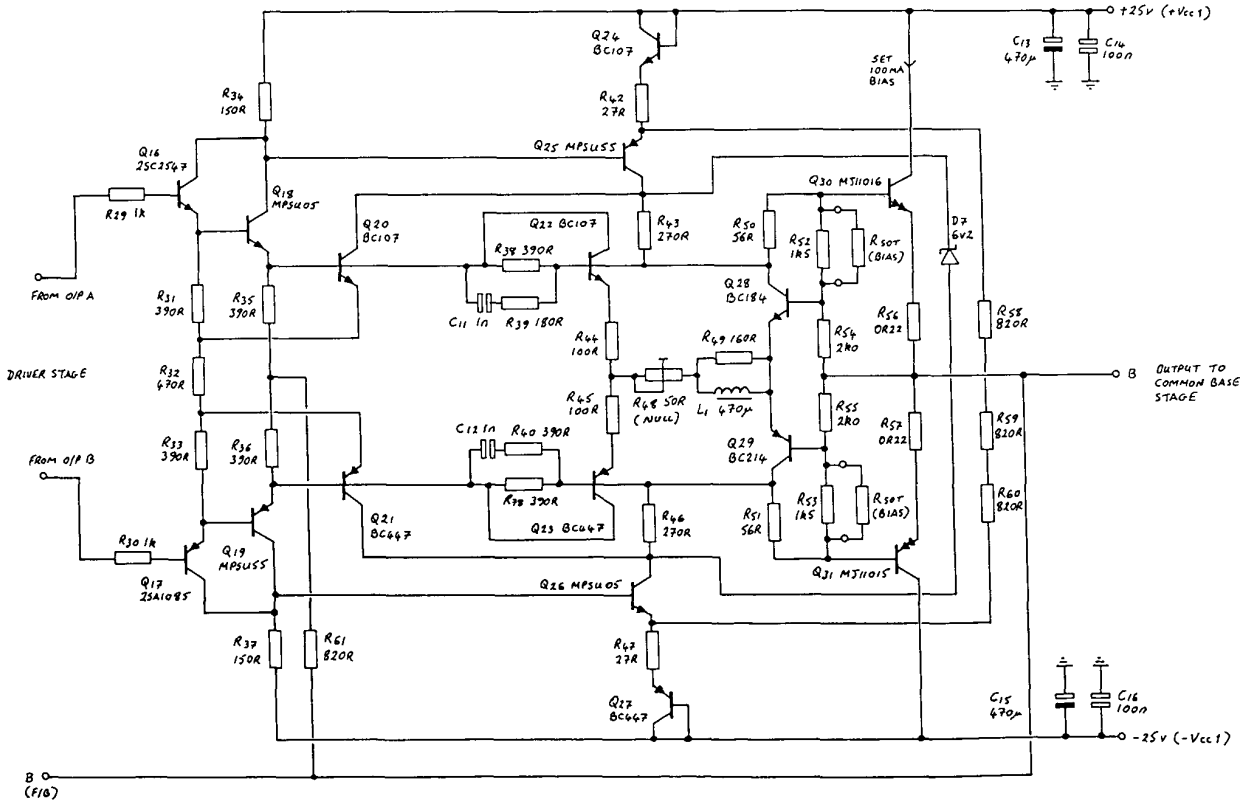


Fig. 8. Prototype amplifier, follower stage.

does not improve linearity above that already afforded by the current drive.

If a longer sensing coil could be accommodated (or indeed a very short coil that remained well within the magnet gap), a further reduction in distortion would be possible.

3.2 Coupling Error Compensation

In order to investigate the nature of the transformer coupling error, Fig. 11 shows the error magnitude with respect to frequency for the coil assembly at equilibrium and also at both extremes of travel. For this measurement the driving coil was powered from the prototype transconductance amplifier system. The level of error is seen to be frequency dependent, rising initially at a rate of approximately 4.6 dB/octave. This unusual characteristic is considered to be a function of pole-piece coupling with the magnetic circuit, but a full analysis of the mechanisms at work has not been undertaken. In addition, some positional dependence of the error magnitude is also apparent. At 100 Hz the coupled error is around 15 dB below the voltage appearing on the driving coil, thus illustrating the need for an effective compensation system.

To implement the compensator, it is necessary to derive a signal proportional to the current in the driving coil and to subject this signal to the same frequency dependence as the error mechanism itself in order to null the error from the sensing coil output. The variation in error level with displacement (typically ± 3 dB) has not been accounted for.

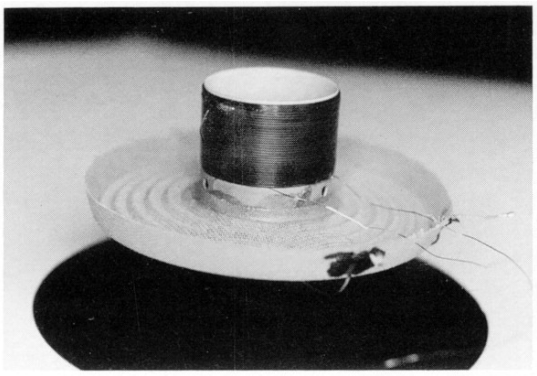


Fig. 10. Sensing-coil assembly.

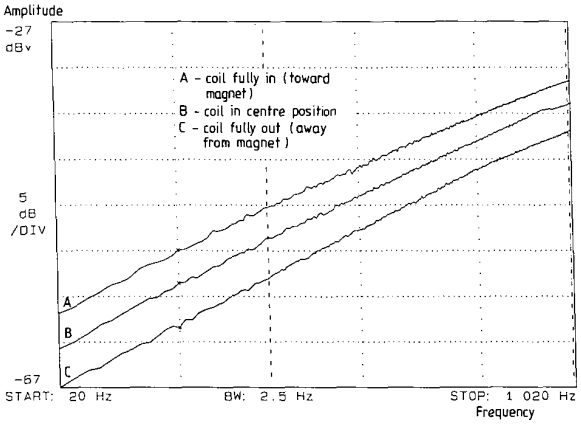


Fig. 11. Measured transformer coupling error.

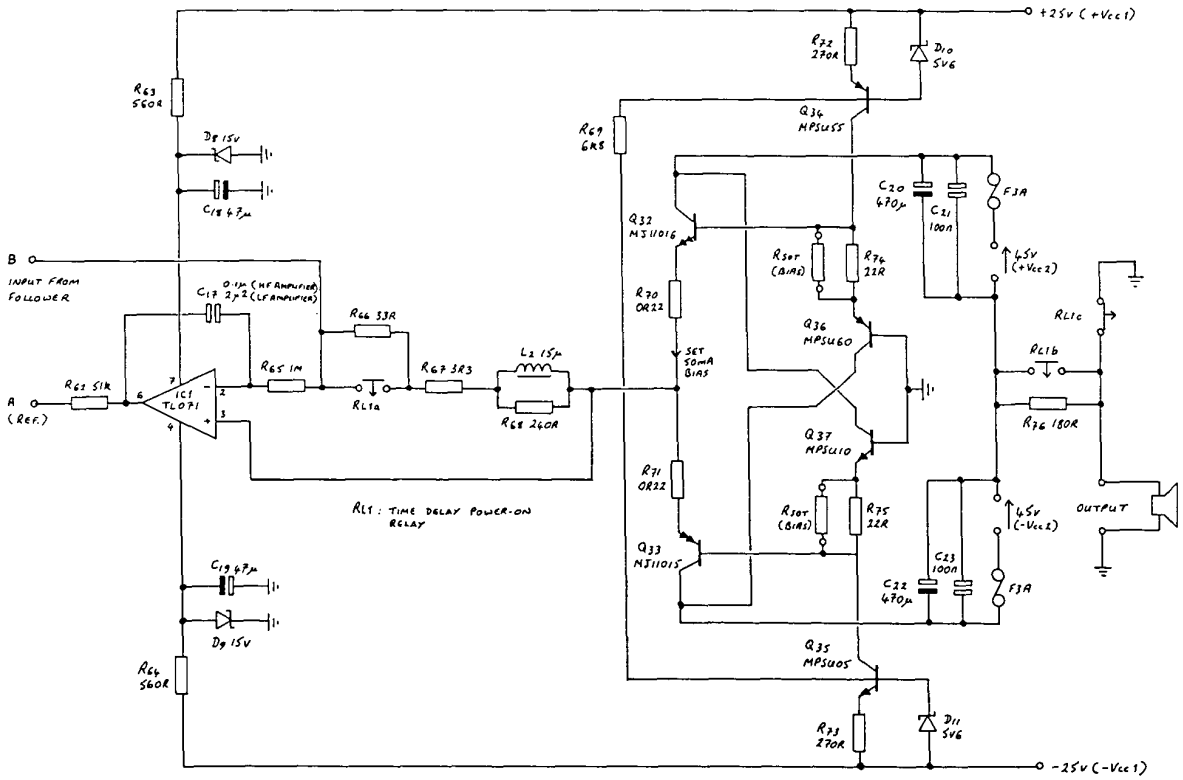


Fig. 9. Prototype amplifier, common-base stage. Protection relay contacts RL_1 are shown with amplifier shut down.

Fig. 12 illustrates the general approach to synthesis of the frequency-dependent element of the compensator. A number of first-order sections are combined, with pole–zero locations set to produce a slope approximating that desired. The general circuit configuration to give this response is shown in Fig. 13 for an nth-order compensator. The transfer function of this circuit is written

$$\frac{V_{out}}{V_{in}} = -\sum_{r=0}^{n}\left(\frac{j\omega RC_r}{1 + j\omega R_r C_r}\right) . \quad (5)$$

A software optimization routine was used to select component values in order to match the 4.6-dB/octave slope required. For a 6th-order compensator, the computer-predicted frequency response is shown in Fig. 14, which also lists the nearest preferred value component values chosen. The result is deemed more than adequate for our purposes, bearing in mind that some positional dependence of the coupling error is present, together with a gradual deviation from the idealized 4.6-dB/octave response with increasing frequency.

3.3 Complete Control System

We continue by considering the complete velocity feedback control system shown in Fig. 15. The sensing coil (source impedance 28 Ω) is connected to a high input impedance buffer stage IC_{2a} via an attenuator network to avoid overload. IC_{2b} forms a summing amplifier in order to subtract the signal derived from the coupling error compensator.

The compensator input is differential, accepting the voltage across the servo current-sensing resistor R_{18} of the transconductance preamplifier (Fig. 6). Thus the input to the compensator is proportional to the drive unit current. This differential signal is converted to single-ended format before the 4.6-dB/octave weighting is applied by the circuitry based around IC_{1d}. R_{22} provides an adjustment to enable the best error null to be obtained with a static motor coil assembly connected to the velocity feedback input.

In order to maintain stability of the closed-loop system, a second-order low-pass filter at 500 Hz is included in the feedback control loop. This also has the benefit of reducing any residual transformer coupling error at high frequencies, where the compensator is no longer as effective due to the changing slope of the error. Finally the output of the controller is summed with the main signal at the velocity feedback input of the transconductance preamplifier, with R_{27} (Fig. 15) providing an adjustment of the low-frequency Q alignment.

To illustrate the performance of the velocity feedback control system, a number of frequency and time domain measurements were obtained. First, Fig. 16(a) shows

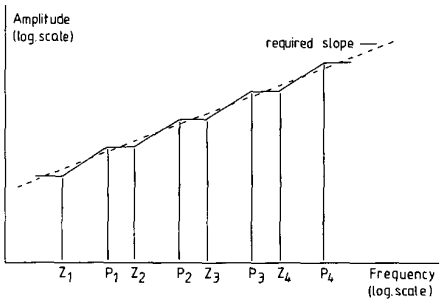


Fig. 12. Basis for synthesis of coupling error compensator.

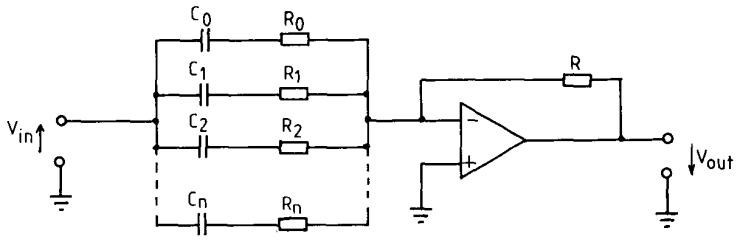


Fig. 13. General configuration for nth-order compensator.

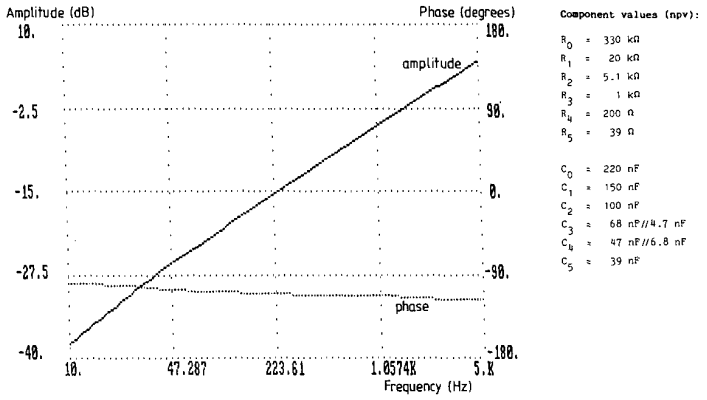


Fig. 14. Computer-predicted response of compensator.

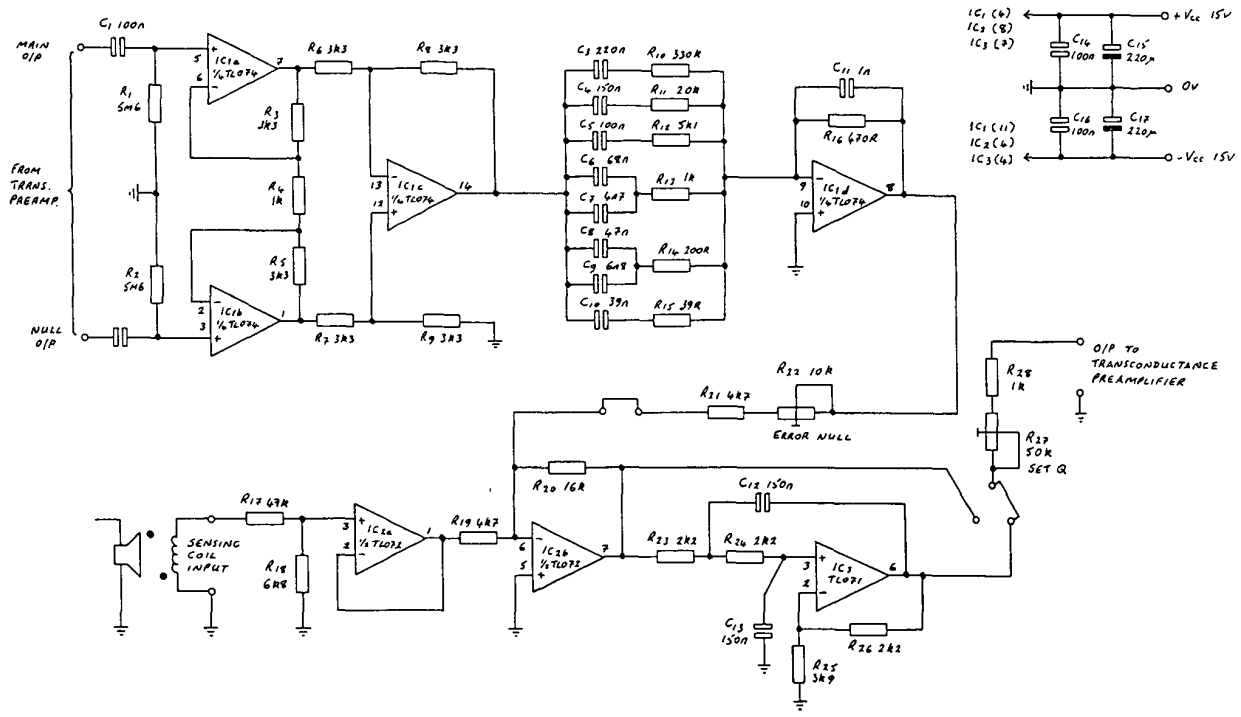


Fig. 15. Velocity feedback control system.

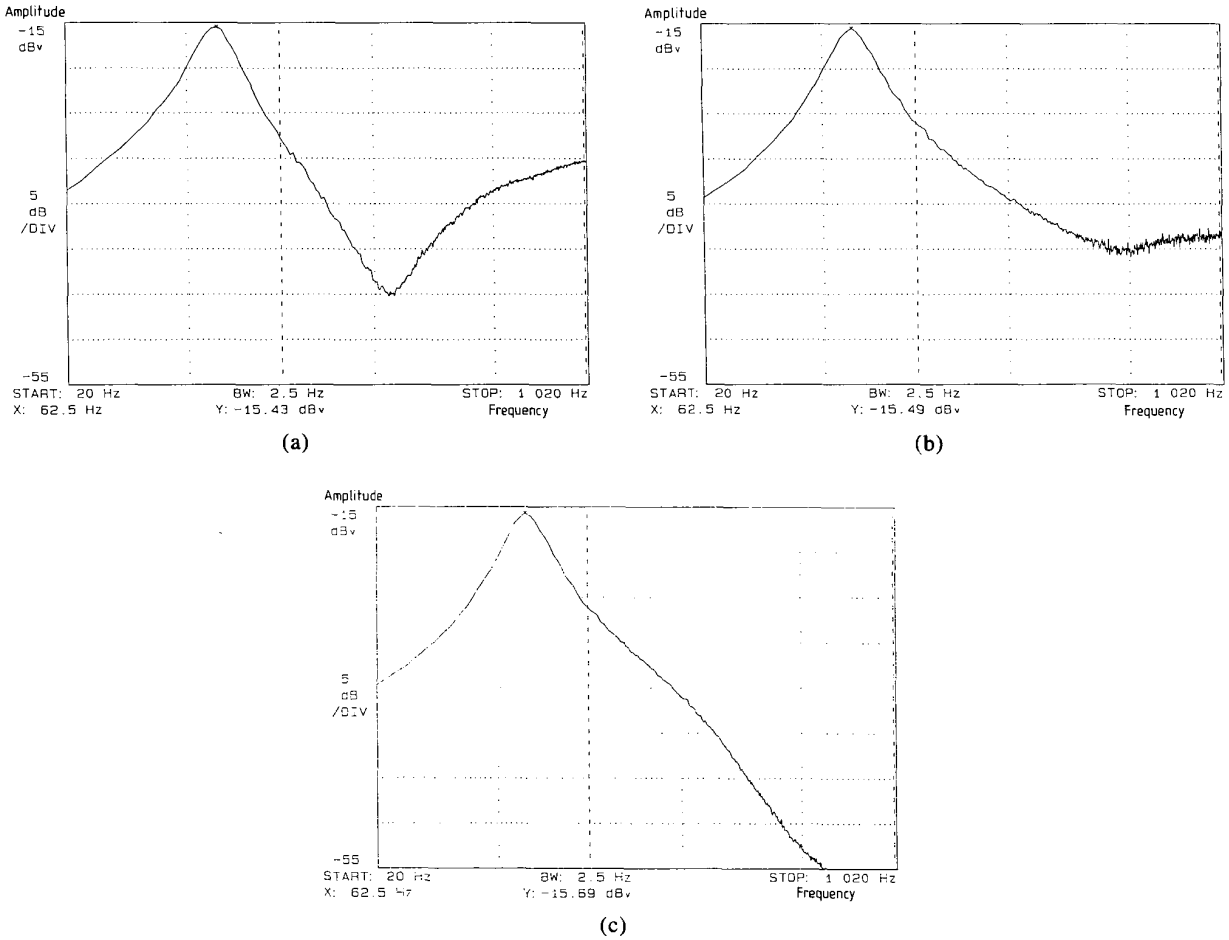


Fig. 16. Velocity feedback control signal. (a) Sensing-coil output voltage. (b) Addition of compensator. (c) Addition of compensator and low-pass filter.

the output of the sensing coil, corresponding to velocity, with frequency. The peak at 62.5 Hz corresponds to the drive unit–enclosure fundamental resonance, while the rising high-frequency output is due to the coupling error between drive and sensing coils. Fig. 16(b) shows the addition of the coupling error compensator, giving a much reduced spurious high-frequency output. The further addition of the second-order low-pass filter at 500 Hz gives the response of Fig. 16(c), which is close to an idealized velocity function.

Steady-state sine-wave measurements of the acoustic output suggest a worthwhile improvement in linearity of the bass–midrange drive unit compared to voltage drive. The following acoustic distortion measurements at a drive current of 1 A peak are illustrative:

	Voltage Drive (dB)	Current Drive* (dB)
Total harmonic distortion at 100 Hz re fundamental	−34.1	−43.3
Total harmonic distortion at 3 kHz re fundamental	−28.4	−55.0

* Under closed-loop conditions.

The effectiveness of the coupling error compensator and filter is confirmed by the fact that no increase in harmonic distortion is measurable up to 3 kHz and beyond (that is, over the full operating range of the drive unit) when the feedback loop is closed. The actual increase in distortion level present on the unfiltered and uncompensated velocity signal, compared to the drive unit acoustic output, ranges from 11 to 23 dB as frequency is increased from 500 Hz to 3 kHz.

The ability to vary the system Q with the velocity feedback control circuit is shown by means of near-field acoustic step response measurements. Fig. 17(a) is without the velocity feedback operational, showing a Q of around 2.5, which is the natural mechanical Q of the drive unit. Fig. 17(b)–(e) shows compensated Q alignments of 1.5, 1.0, 0.7, and 0.5, respectively. A value of $Q = 0.7$ preserves the low-frequency characteristics of the unmodified loudspeaker under voltage drive.

4 LOW-LEVEL CROSSOVER

To complete the two-way prototype system, a second-order low-level high- and low-pass crossover was included to integrate the drive units together, with a nominal crossover point of 3 kHz.

The crossover is implemented by passive RC elements, the time constants of which are individually adjustable to give the flattest frequency response, with buffer amplifiers between stages, to avoid loading effects. The complete system is shown in modular form by Fig. 18. After the input level control, amplifier A_1 provides a low-impedance drive to the first low- and high-pass filters, which are buffered by amplifiers A_2 and A_3 before the second set of filter sections. A_4 and A_5 are the transconductance preamplifiers previously

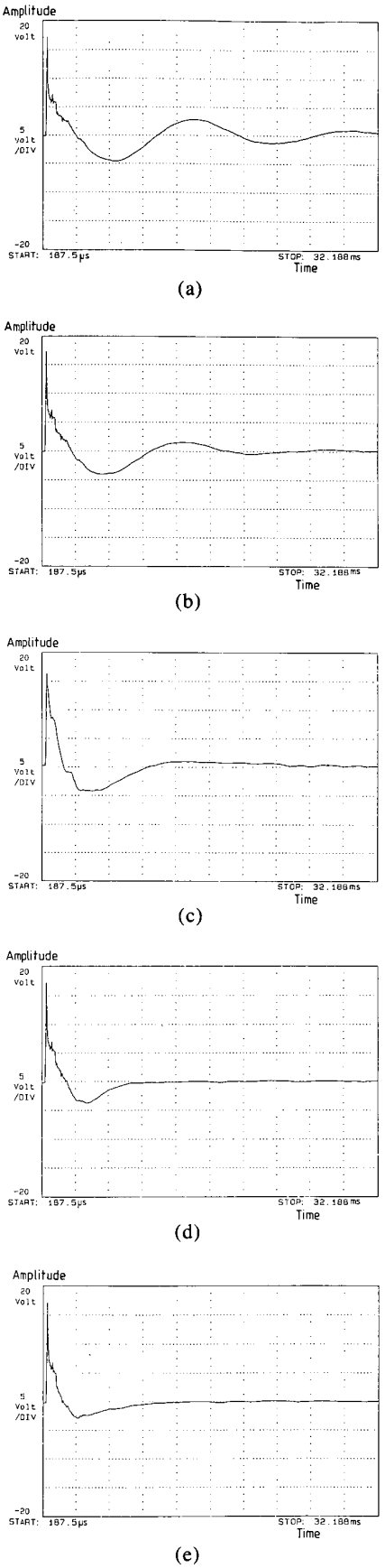


Fig. 17. Measured step responses of bass–midrange drive unit. (a) $Q = 2.5$ (no feedback). (b) $Q = 1.5$. (c) $Q = 1.0$. (d) $Q = 0.7$. (e) $Q = 0.5$.

described, which drive the high- and low-frequency power amplifiers, respectively.

The circuit topology of buffer amplifiers A_1, A_2, and A_3 (Fig. 19) is similar to the transconductance preamplifier, but with the addition of an output follower and overall negative feedback to provide a low output impedance. Certain gain and frequency response defining components are specific to individual amplifier stages as indicated.

The performance of the system is shown by the in-room measured frequency response curve of Fig. 20, with the measurement microphone 1 m on axis. The uneven high-frequency response is a function of the tweeter characteristics, with the resonant peak near 19 kHz being due to the first bending mode resonance of the copper dome. There is no discernible frequency response deviation in moving from voltage drive to current drive with this device, due to its high level of intrinsic damping.

While the low-frequency drive unit benefits sub-

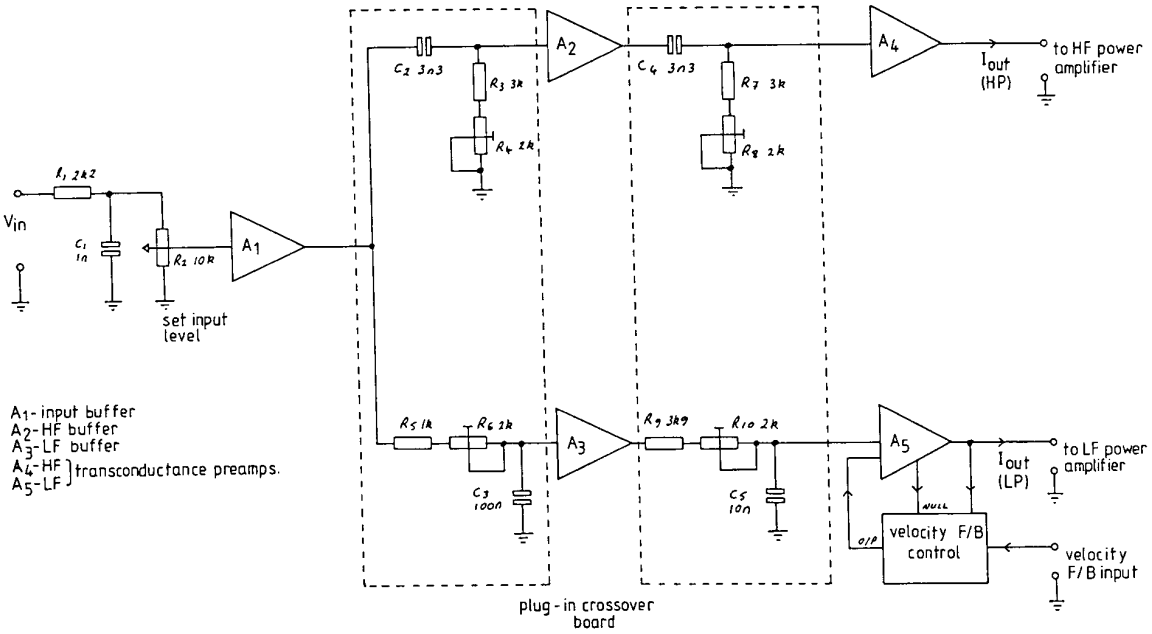


Fig. 18. Block diagram of low-level crossover.

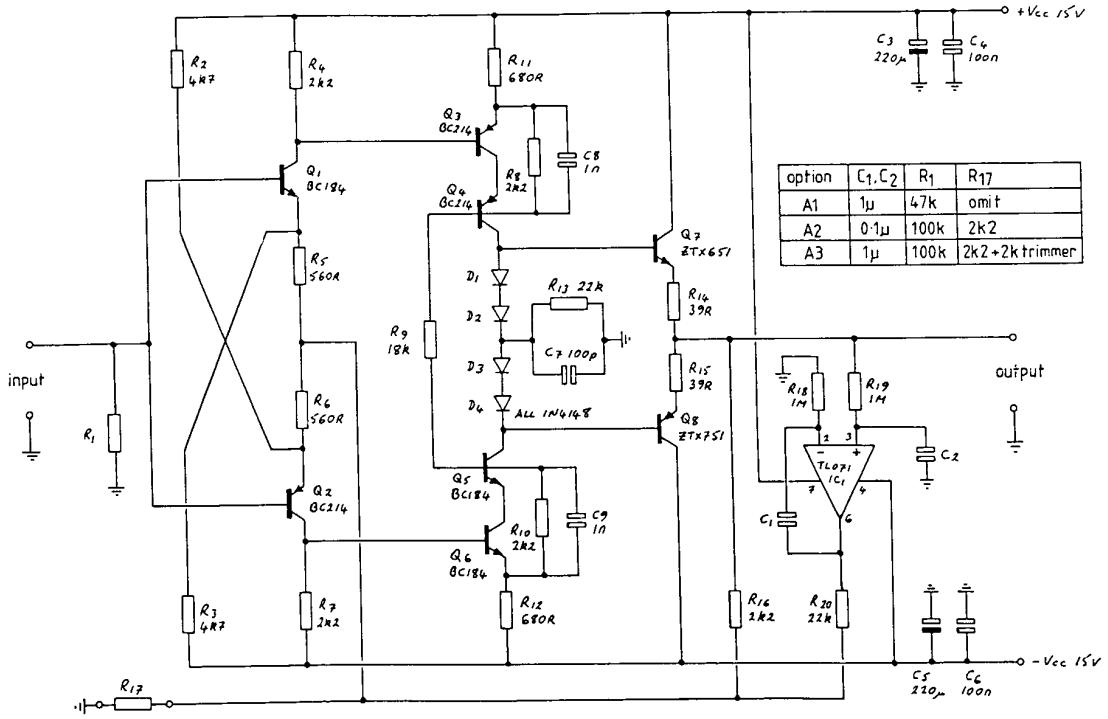


Fig. 19. Buffer amplifiers for crossover network.

stantially from current drive, the improvements in linearity to the tweeter are more modest, largely as a result of a more linear magnetic circuit and lower cone displacements. A distortion improvement of typically 3–7 dB is afforded. However, benefits in terms of the elimination of thermally induced errors are still apparent.

5 DYNAMIC CURRENT AND VOLTAGE DEMAND

Under certain signal conditions, drive units and loudspeaker systems have, under voltage drive, been shown to exhibit an instantaneous impedance modulus lower than might initially be suggested from the steady-state impedance characteristics, thus stressing the power amplifier in terms of current delivery [24]–[28]. In this section we consider the implications of this work in relation to current drive.

As an example, the bass–midrange unit and enclosure combination is considered. The equivalent electrical model is shown in Fig. 21. Under voltage drive, a pulse is applied to the drive unit, the duration of which is set to excite the large negative-going current excursion shown in Fig. 22. The voltage signal has been second-order low-pass filtered at 3 kHz to represent realistic operating conditions. At the point of maximum negative-going current, the instantaneous impedance modulus is 4.15 Ω, lower than the steady-state minimum of 7 Ω.

Under current drive it is only realistic to consider the drive unit when equalized by velocity feedback to give the same Q at fundamental resonance as in the voltage-driven case ($Q \approx 0.7$). Under these conditions, also with a 3-kHz crossover point, the current waveform is seen to be similar to the voltage-driven case, while the voltage waveform shows peaks due to the voice-coil inductance (Fig. 23). The instantaneous impedance modulus is similar to that under voltage drive, but slightly lower at 4.05 Ω.

The main significance of these results is that while the power amplifier is similarly stressed under both voltage and current drive, allowance must be made for sufficient headroom in the power amplifier for the voltage peak resulting from the coil inductance. The problem is worsened by voice-coil heating. The effect of a temperature rise of 200°C (meaning that the coil resistance increases to 13 Ω using copper) is shown by the waveforms of Fig. 24. While the current waveform is identical to that at normal temperature, as expected, the voltage waveform is increased in magnitude in order to keep the current constant. The negative-going excursion is seen to be 1.7 times greater. Although the performance of the drive unit is unaffected by the in-

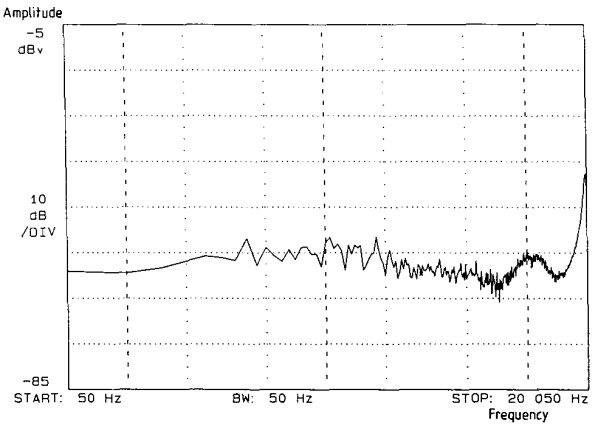


Fig. 20. Measured frequency response of complete system.

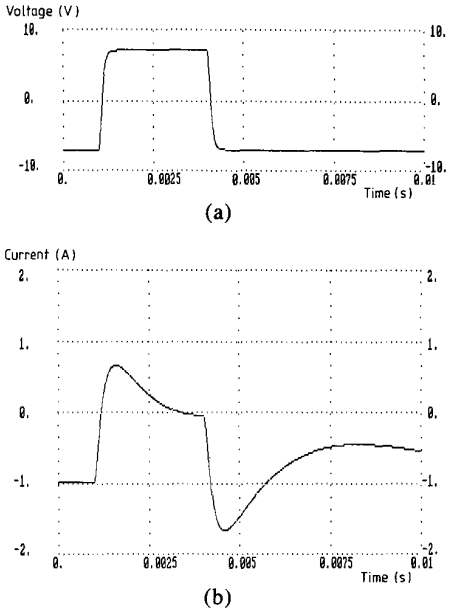


Fig. 22. Bass–midrange unit: dynamic current demand under voltage drive.

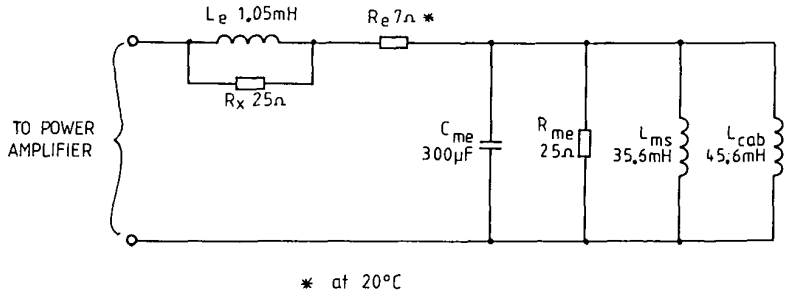


Fig. 21. Electrical model of bass–midrange drive unit and enclosure combination. L_e—voice-coil inductance; R_x—losses due to pole piece coupling; R_e—voice-coil resistance; C_{me}—capacitance due to moving mass; R_{me}—resistance due to mechanical losses; L_{ms}—inductance due to suspension compliance; L_{cab}—inductance due to volume of air in cabinet.

creased temperature, care must be taken in selecting the supply voltages for the power amplifier to avoid voltage clipping, with implications regarding the safe operating area of power devices.

6 CONCLUSIONS

In this paper we have considered a prototype two-way active current-driven loudspeaker system. The main benefits of current drive are seen to be a freedom from thermally induced distortion effects and also high-frequency nonlinearity caused by the voice-coil inductance.

By reviewing earlier work on transconductance power amplifiers, the limitations of these approaches were noted and a novel topology described, employing a common-base isolating stage to drive the loudspeaker. This decouples the nonlinear load from the feedback loop of the amplifier and provides a naturally high output impedance. For low-frequency operation, motional feedback is acknowledged as providing a suitable method of damping the drive unit fundamental resonance. A velocity-sensing coil was employed for this purpose, wound over the main driving coil, with electronic compensation used to null the transformer coupling error between the two. At high frequencies, where motional feedback is no longer feasible, a drive unit with good self-damping properties is recommended, although open-loop compensation could also be considered.

While the power amplifier circuits presented may be regarded as being somewhat complicated, they demonstrate some of the earlier work carried out within the group on distortion correction techniques. The availability of new hybrid gain stage devices (such as the Deltec DH-OA32) with high open-loop bandwidth and good output voltage and current-driving capabilities considerably simplifies the task of power amplifier design.

The provision of velocity feedback control circuitry introduces additional complexity over voltage drive, but this must be considered the price to be paid for what is regarded as a most worthwhile improvement in loudspeaker performance.

7 REFERENCES

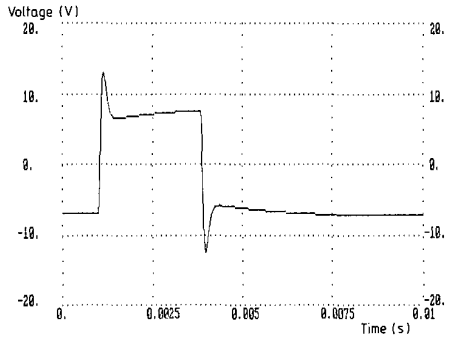
[1] M. R. Gander, "Dynamic Linearity and Power Compression in Moving-Coil Loudspeakers," presented at the 76th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 32, pp. 1008–1009 (1984 Dec.), preprint 2128.

[2] J. R. Gillion, P. L. Boliver, and L. C. Boliver, "Design Problems of High-Level Cone Loudspeakers," *J. Audio Eng. Soc. (Project Notes/Engineering Briefs)*, vol. 25, pp. 294–299 (1977 May).

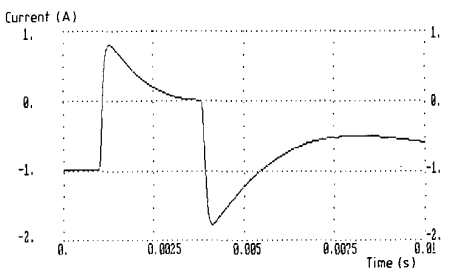
[3] T. S. Hsu, S. H. Tang, and P. S. Hsu, "Electromagnetic Damping of High-Power Loudspeakers," presented at the 79th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, p. 1011 (1985 Dec.), preprint 2297.

[4] W. J. Cunningham, "Nonlinear Distortion in Dynamic Loudspeakers Due to Magnetic Effects," *J. Acoust. Soc. Am.*, vol. 21, pp. 202–207 (1949 May).

[5] J. A. M. Catrysse, "On the Design of Some

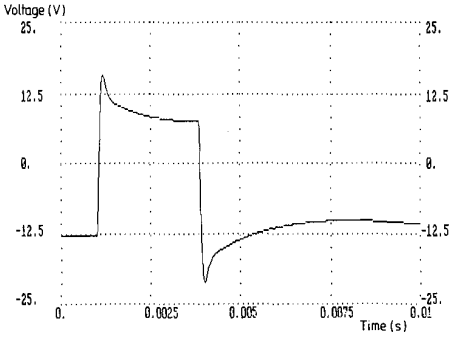


(a)

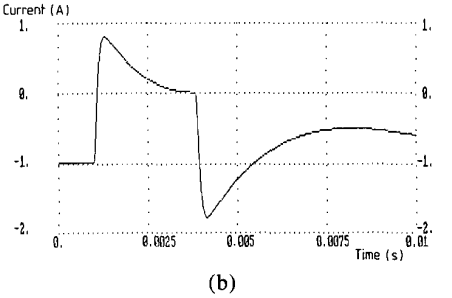


(b)

Fig. 23. Dynamic excitation under current drive.



(a)



(b)

Fig. 24. Dynamic excitation as for Fig. 23, but voice-coil temperature raised to 200°C (R_e = 13 Ω).

Feedback Circuits for Loudspeakers," presented at the 73rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 364 (1983 May), preprint 1964.

[6] R. A. Greiner and T. M. Sims, Jr., "Loudspeaker Distortion Reduction," *J. Audio Eng. Soc.*, vol. 32, pp. 956–963 (1984 Dec.).

[7] K. Lewis, "Transconductance Amplifiers," *Electron. Wireless World*, pp. 580–582 (1987 June).

[8] A. Nedungadi, "High Current Class AB Converter Technique," *Electron. Lett.*, vol. 16, pp. 418–419 (1980 May).

[9] M. K. N. Rao and J. W. Haslett, "Class AB Voltage-Current Converter," *Electron. Lett.*, vol. 14, pp. 762–764 (1978 Nov.).

[10] M. J. Hawksford, "Low-Distortion Programmable Gain Cell Using Current-Steering Cascode Topology," *J. Audio Eng. Soc.*, vol. 30, pp. 795–799 (1982 Nov.).

[11] E. H. Nordholt, *Design of High-Performance Negative-Feedback Amplifiers* (Elsevier, Amsterdam, 1983).

[12] J. J. Davidson, "A Low-Noise Transistorized Tape Playback Amplifier," *J. Audio Eng. Soc.*, vol. 13, pp. 2–16 (1965 Jan.).

[13] J. L. Linsley Hood, "Symmetry in Audio Amplifier Circuitry," *Electron. Wireless World*, pp. 31–34 (1985 Jan.).

[14] R. N. Marsh, "A Passively Equalised Phono Pre-amplifier," *Audio Amateur*, no. 3, pp. 18 (1980).

[15] M. Hawksford, "Reduction of Transistor Slope Impedance Dependent Distortion in Large-Signal Amplifiers," *J. Audio Eng. Soc.*, vol. 36, pp. 213–222 (1988 Apr.).

[16] M. J. Hawksford, "Power Amplifier Output-Stage Design Incorporating Error-Feedback Correction with Current-Dumping Enhancement," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 960 (1983 Dec.), preprint 1993.

[17] P. G. A. H. Voight, "Improvements in or Relating to Thermionic Amplifying Circuits for Telephony," *UK patent 231972* (1924 Jan.).

[18] J. A. Klaassen and S. H. de Koning, "Motional Feedback with Loudspeakers," *Philips Tech. Rev.*, vol. 29, pp. 148–157 (1968).

[19] D. de Greff and J. Vandewege, "Acceleration Feedback Loudspeaker," *Wireless World*, pp. 32–36 (1981 Sept.).

[20] G. J. Adams, "Adaptive Control of Loudspeaker Frequency Response at Low Frequencies," presented at the 73rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 364 (1983 May), preprint 1983.

[21] E. de Boer, "Theory of Motional Feedback," *IRE Trans. Audio*, pp. 15–21 (1961 Jan./Feb.).

[22] A. F. Sykes, "Damping Electrically Operated Vibration Devices," *UK patent 272622* (1926 Mar.).

[23] R. L. Tanner, "Improving Loudspeaker Response with Motional Feedback," *Electronics*, pp. 142 ff. (1951 Mar.).

[24] P. G. L. Mills and M. J. Hawksford, "Transient Analysis: A Design Tool in Loudspeaker Systems Engineering," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 386 (1986 May), preprint 2338.

[25] I. Martikainen, A. Varla, and M. Otala, "Input Current Requirements of High-Quality Loudspeaker Systems," presented at the 73rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 364 (1983 May), preprint 1987.

[26] M. Otala and P. Huttunen, "Peak Current Requirement of Commercial Loudspeaker Systems," *J. Audio Eng. Soc.*, vol. 35, pp. 455–462 (1987 June).

[27] D. Preis, "Peak Transient Current and Power into a Complex Impedance," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 386 (1986 May), preprint 2337.

[28] J. Vanderkooy and S. P. Lipshitz, "Computing Peak Currents into Loudspeakers," presented at the 81st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, pp. 1036–1037 (1986 Dec.), preprint 2411.

Biographies for Drs. Mills and Hawksford were published in the March issue.

Differential-current derived feedback in error-correcting audio amplifier applications

M.O.J. Hawksford

Indexing terms: Feedback, Audio

Abstract: A dominant distortion mechanism in audio power amplifiers results from nonlinearity in the voltage and current transfer functions within the power output stage. A technique is presented that uses a magnetic differential sensor to sense this distortion and to apply a supplementary error-correcting feedback path within the amplifier, in addition to overall negative feedback. Differential-current derived feedback is explored and the technique is shown to be applicable to both voltage and current transfer errors. Analysis reveals the system alignment for obtaining a distortion null and these results are confirmed using nonlinear transient analysis. Basic system topologies are presented using voltage transfer and transconductance gain cells as an aid to complete amplifier synthesis. Comparisons are made with feedforward-feedback error correction where the current-dumping amplifier is reconfigured using differential-current feedback. Generalisations are made to a multi-loop nest of amplifiers using individually aligned error correcting paths to achieve corresponding improvements in distortion reduction.

1 Introduction

A cardinal objective of analogue amplifier design is to identify circuit techniques that can minimise the distortion induced by transistor nonlinearity, especially in the output stage of power amplifier circuits. It has already been demonstrated that error-correction topologies exist that can substantially linearise the earlier stages of an amplifier [1], and the use of an enhanced cascode [2] can further reduce distortion under large voltage swing conditions. It is well understood that feedback alone [3] cannot completely eliminate distortion because of the need to reduce the loop gain at high frequency so as to maintain closed-loop stability. In theory, to eliminate broadband distortion, a feedforward path is required that can either extend beyond the main nonlinear feedback loop or correct locally for a nonlinearity.

This paper proposes differential-current derived feedback (DCDF) as a means of error correction that can accommodate both voltage and current transfer nonlinearities in the amplifier output stage. System validation is by transient analysis, and comparisons are made with the Albinson–Walker current-dumping amplifier [4–14], where it is shown that a dominant differential-error feedback loop is supplemented by feedforward, rather than comprising just feedforward.

Finally, DCDF is extended to an mth-order nest of DCDF loops, which yields greater distortion reduction both by increasing loop gain and by introducing m zeros in the output-stage-related error function.

Paper 1013G (E10), first received 31st August and in revised form 6th December 1993

The author is with the Department of Electronic Systems Engineering, University of Essex, United Kingdom

2 Differential-current sensing

DCDF uses a current transformer to sense the rate of change of a current that is related to an error signal, and it will be shown that this arrangement can compensate for the frequency-dependent and distortion-reduction characteristics of the closed-loop gain. The current transformer shown in Fig. 1 consists of a magnetic circuit and two loosely coupled windings. For the purpose of definition, the single-turn winding carrying the current to be sensed is the primary one, and the multi-turn winding used to derive a voltage proportional to the rate of change of the primary current is the secondary winding.

The open-circuit secondary voltage e_s expressed as a function of primary current I_p follows from standard electromagnetic theory as

$$e_s = L\frac{\partial I_p}{\partial t} = \frac{\mu_0 \mu_r NA}{2\pi R}\frac{\partial I_p}{\partial t} \tag{1a}$$

where the mutual inductance L for the transformer is defined as

$$L = \frac{\mu_0 \mu_r NA}{2\pi R} \tag{1b}$$

where R = mean radius of toroid, μ_0 = permeability of free space (H/m), A = mean cross-sectional area of toroid (m^2), and μ_r = relative permeability of ferrite.

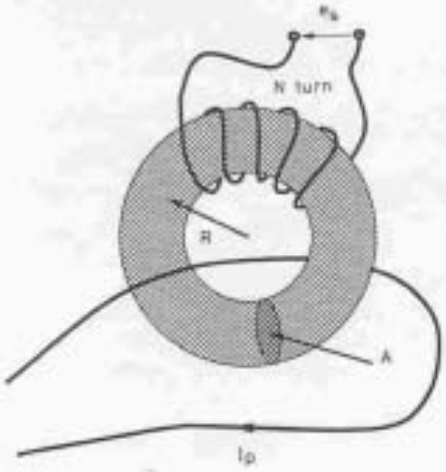
To illustrate the performance of an experimental differential-current sensor using a ferrite toroid, a current transformer was designed using the following (approximate) parameters: internal diameter = 10 mm, secondary turns = 5, external diameter = 40 mm, and primary turns = 1.

When a primary current of triangular waveform was injected, close adherence to a differential operator was demonstrated, as shown in Fig. 2, which was limited only by the coil's resonant frequency, ~10 MHz.

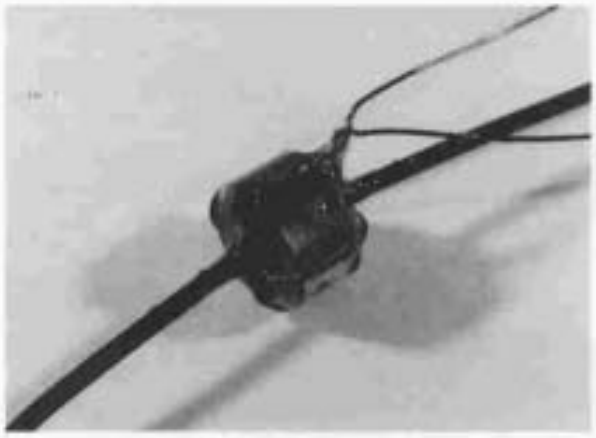
3 DCDF amplifier without error correction

An amplifier is shown in Fig. 3 with a forward gain A, an open-loop output impedance Z_0 and both unity-gain voltage feedback and DCDF. DCDF is modelled with a floating-voltage generator related to the rate of change of

output current I_0 where, using the Laplace operator s, this is described by $sL_0 I_0$ (the notation used in this paper gives the same subscript to the mutual inductance L_x and



a



b

Fig. 1 *Toroidal current sensing element*
a Actual
b Experimental

corresponding current to be sensed I_x). The closed-loop gain is

$$\frac{V_0}{V_i} = \frac{A}{1 + A + \left(\frac{Z_1 - sL_0 A}{Z_L}\right)} \tag{2a}$$

However, by aligning inductance L_0 such that

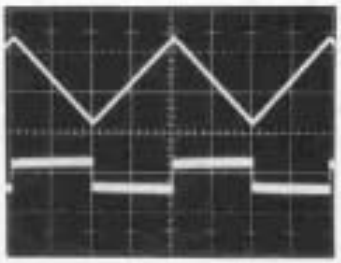
$$sL_0 A = Z_1 \tag{2b}$$

the closed-loop gain becomes independent of the load impedance Z_L and reduces to

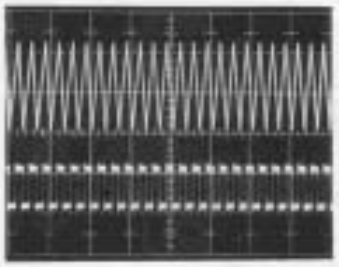
$$\frac{V_0}{V_i} = \frac{A}{1 + A} \tag{2c}$$

Eqn. 2b describes a balance condition that reduces the amplifier's closed-loop output impedance to zero, providing Z_1 is resistive and A takes the form

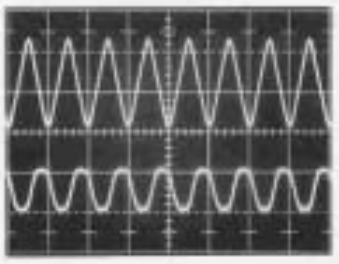
$$A = \frac{1}{sT_0} \tag{2d}$$



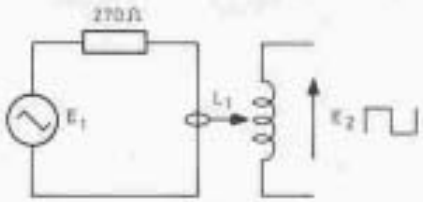
a



b



c



d

Fig. 2 *Example measurements on toroidal transformer for triangular-wave excitation*
a 50 kHz triangular waveform
Top trace: 10 V/div
Bottom trace: 100 mV/div
Timebase: 5 μs/div
b 500 kHz triangular waveform
Top trace: 10 V/div
Bottom trace: 1 mV/div
Timebase: 5 μs/div
c 5 MHz triangular waveform
Top trace: 10 V/div
Bottom trace: 10 V/div
Timebase: 5 μs/div
d Measurement system

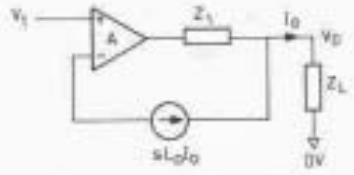


Fig. 3 *Basic amplifier using both voltage and DCDF*

Hence, from eqns. 2*b* and *d*, the optimum mutual inductance L_0 is calculated as

$$L_0 = Z_1 T_0 \tag{2e}$$

A computer simulation of the topology in Fig. 4 was performed where a family of closed-loop gains was as shown in Fig. 5. A gain parameter γ in the DCDF loop is varied from 0 to 1 (1 is optimum) in five steps to demonstrate the effect of varying degrees of current feedback. Without DCDF, a second-order response is visible at a frequency of ~154 kHz, whereas for optimum DCDF the resonance is almost suppressed.

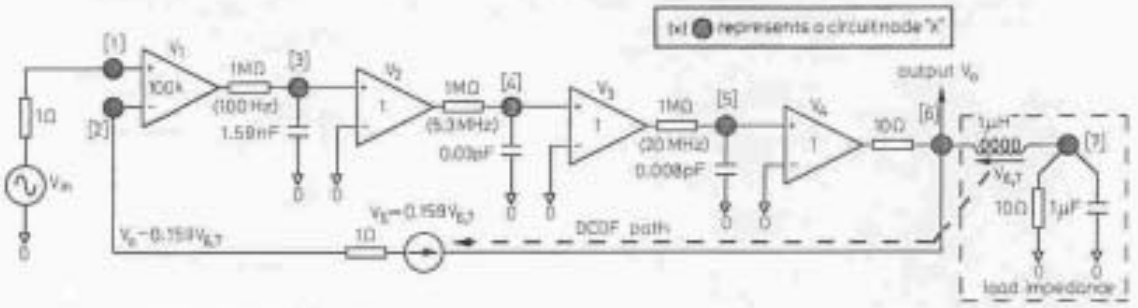


Fig. 4 *Simulation model of 3-pole DCDF amplifier*

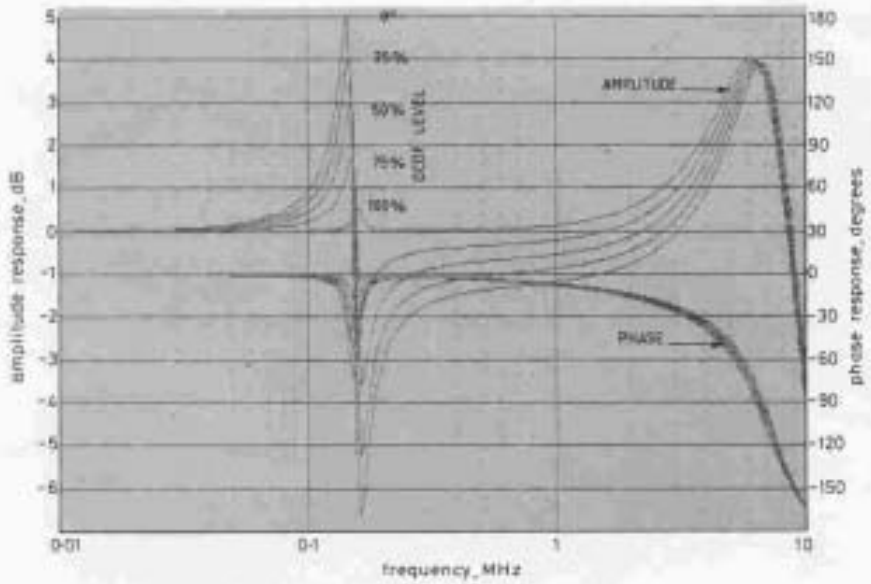


Fig. 5 *Section 2 example amplifier with 5 levels of DCDF ranging from 0 to 100%*

Although a DCDF enhanced amplifier can achieve zero output impedance with a finite forward-path gain, it is only realised if the output impedance Z_1 has a constant resistance and if the current transformer and amplifier transfer functions are optimally aligned. For high-output current amplifiers, the impedance Z_1 is nonlinear, and these conditions are not met. In Section 4, DCDF is extended to accommodate amplifiers with nonlinear output impedance.

4 Error-correction amplifiers using DCDF

4.1 Correction of voltage-transfer errors

Consider a linear amplifier of transfer function A that has a constant-value output resistance R_1, where an appropriate level of DCDF makes the closed-loop output impedance zero. If a nonlinear resistance is now connected to the amplifier output, then, because of the zero output impedance under optimal balance, the output voltage remains undistorted. This observation can be extended to a nonlinear stage N (where $N \approx 1$ and, at this stage of discussion has a very high input impedance) connected across the resistor R_1, as shown in Fig. 6. The stage N contributes to the total output current, but, because the system has a theoretical output impedance of zero, the amplifier closed-loop gain remains independent of the nonlinear gain N.

The amplifier in Fig. 6 uses three current-sensing paths (sensors L_0, L_1 and L_2) that feedback the differentials of I_1, I_2 and I_0, respectively. Initially, the forward-path amplifier is assumed to have both a zero output resistance and a voltage transfer function A, defined by eqn. 2*d*, where the closed-loop gain can be expressed as

$$\frac{V_0}{V_1} = \frac{A}{1+A} - \frac{1}{V_1} \times \left\{ \frac{sA(L_0 + L_2)I_0 + [R_1 - sA(L_1 - L_2)]I_1}{1+A} \right\} \tag{3a}$$

To make eqn. 3*a* independent of current I_1 and thus independent of amplifier nonlinearity,

$$R_1 = sA(L_1 - L_2) \tag{3b}$$

whereby eqn. 3*a* reduces to

$$V_0 = \frac{A}{1+A}[V_1 - s(L_0 + L_2)I_0] \tag{3c}$$

Hence, substituting for A from eqn. 2*d*, the closed-loop output impedance Z_{out} is

$$Z_{out} = \frac{s(L_0 + L_2)}{1 + sT_0} \tag{3d}$$

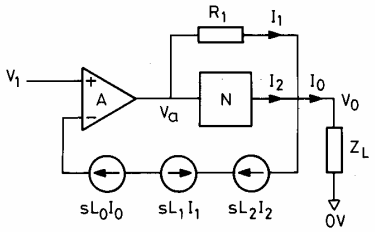


Fig. 6 *Error correcting amplifier using three current sensing loops*

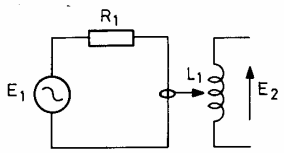


Fig. 7 *Measurement scheme for determining coil mutual inductance*

Eqns. 3*b* and *d* identify three correction scenarios:

(*a*) L_0, L_1, L_2 all finite: in general Z_{out} is finite, as defined by eqn. 3*d*; however, for the special case of $L_0 = -L_2$, then $Z_{out} = 0$. This condition can be implemented using a single sensing toroid that is simultaneously linked by I_0 and I_2, flowing in opposite directions.

(*b*) L_0 or L_2 set zero: balance remains possible, but the output impedance is finite.

(*c*) L_0 and L_2 zero, only I_1 sensed: under optimal balance $Z_{out} = 0$, and toroid magnetic flux is reduced as $I_1 \ll I_2$ for $N \approx 1$; only a single sensor is required.

An expression for closed-loop gain follows from eqn. 3*a* when the output stage is considered to have an inverse operator N^{-1}, where, if $I_1 = V_0(N^{-1} - 1)/R_1$ and $L_0 = L_2 = 0$,

$$\frac{V_0}{V_1}\left[1 + \frac{(R_1 - sAL_1)}{1+A}(N^{-1} - 1)\right] = \frac{A}{1+A} \tag{3e}$$

Defining an error function $E(s)$ where the actual gain $G(s) = V_0/V_1$ and the target gain $G_t(s) = G(s)$ for $N^{-1} = 1$, and assuming $(N^{-1} - 1)(1 - sAL_1/R_1)/(1 + A) \ll 1$, then

$$E(s) = \frac{G(s)}{G_t(s)} - 1 = (N^{-1} - 1)\left[\frac{1 - sAL_1/R_1}{1+A}\right] \tag{3f}$$

The circuit in Fig. 7 enables L_1 to be determined as $E_1 = E_2$ at the frequency where $|A| = 1$.

4.2 Error correction with finite output impedance of amplifier A

The system diagram in Fig. 8 shows a DCDF scheme where amplifier A has a finite output impedance Z_0. Both the input current I_n to the output stage and the current I_1 in impedance Z_1 can now inject nonlinear signal components into the overall feedback loop, which in turn distort the output signal. The proposed correction procedure is to use individual differential-current sensors for *both* currents I_n and I_1, where, in general,

$$I_1Z_1 + I_1Z_0 + I_nZ_0 - sA(I_nL_n + I_1L_1) = A(V_1 - V_0) - V_0$$

Error correction is achieved when the left-hand side is equated to zero, when, to desensitise system dependence on *both* I_1 and I_n, two conditions must be met where, for

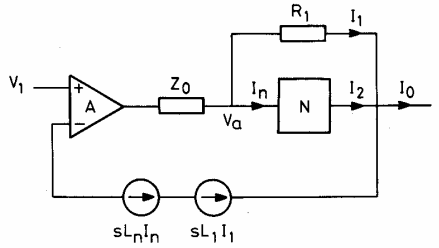


Fig. 8 *Error correction accommodating both voltage transfer errors in N and finite output impedance of A in association with nonlinear loading of output stage N*

the special case of A given by eqn. 2*d* and for resistive components $Z_0 = R_0$ and $Z_1 = R_1$,

$$L_1 = \frac{Z_0 + Z_1}{sA} = (R_0 + R_1)T_0 \tag{3g}$$

$$L_n = \frac{Z_0}{sA} = R_0 T_0 \tag{3h}$$

that is, $L_1/L_n = 1 + R_1/R_0$.

4.3 DCDF error-correction scheme using cascaded transconductance stages

Two cascaded transconductance amplifiers g_1, g_2 are shown in Fig. 9, where g_2 has local capacitive feedback via Z_0 and a local load impedance Z_2. The second stage

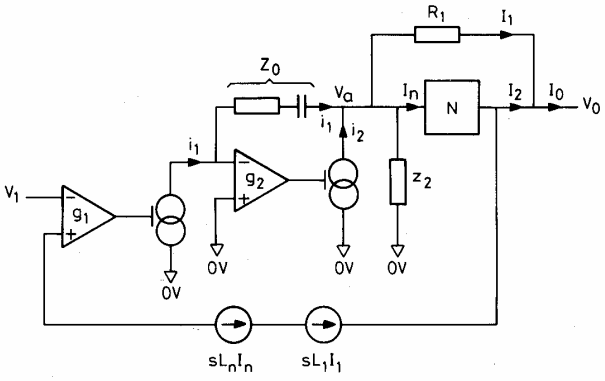


Fig. 9 *DCDF error-correction amplifier using transconductance stages*

also drives the nonlinear output stage and supplies current to the feedforward bypass impedance Z_1. As in Section 4.2, a dual DCDF scheme is used to correct both voltage-transfer errors and the effect of nonlinear loading on the frequency-dependent output impedance of the second stage.

Defining α as

$$\alpha = \frac{g_2 Z_2}{1 + g_2 Z_2} \tag{4a}$$

the closed-loop transfer function is expressed as

$$\frac{V_0}{V_1} = \frac{\alpha g_1(Z_0 - 1/g_2) - \dfrac{I_1 Z_1 - s\alpha g_1(Z_0 - 1/g_2)(L_1 I_1 + L_n I_n) + \alpha(I_1 + I_n)/g_2}{V_1}}{1 + \alpha g_1(Z_0 - 1/g_2)}$$

Equating the terms in I_1 and I_n to zero, setting $Z_0 = r_0 + 1/(sC_0)$ and $Z_1 = R_1$, $Z_2 = R_2$ and aligning $r_0 = 1/g_2$, the two balance conditions follow as

(*a*) eliminating I_1

$$L_1 = \frac{1 + g_2 Z_1 + \dfrac{Z_1}{Z_2}}{sg_1(g_2 Z_0 - 1)} = \left[\frac{1}{g_1 g_2} + \frac{R_1}{g_1} + \frac{R_1}{R_2 g_1 g_2}\right] C_0 \quad (4c)$$

(*b*) eliminating I_n

$$L_n = \frac{1}{sg_1(g_2 Z_0 - 1)} = \frac{C_0}{g_1 g_2} \quad (4d)$$

that is, $L_1/L_n = 1 + g_2 R_1 + R_1/R_2$. R_2 represents the output impedance of the second transconductance stage, which can be made large by using either a grounded base stage or an enhanced cascode topology [2], whereby, if $R_2(1/R_1 + g_2) \gg 1$, then eqn. 4*c* reduces to $L_1 = C_0(R_1 + 1/g_2)/g_1$. Hence, providing the core cells g_1, g_2 exhibit a wide bandwidth with constant transconductance and L_1, L_n are optimally aligned, then, by setting $Z_0 = 1/g_2 + 1/(sC_0)$ with $g_2 Z_2 \gg 1$, a first-order closed-loop gain follows from eqns. 4*a* and *b*

$$\frac{V_0}{V_1} = \frac{g_1 g_2 Z_2(Z_0 - 1/g_2)}{1 + g_2 Z_2[1 + g_1(Z_0 - 1/g_2)]} = \frac{g_1}{g_1 + sC_0} \quad (4e)$$

4.4 High-frequency performance limitation of current transformer

Fig. 10 shows a DCDF amplifier with feedback factor k that includes a second-order model of the HF resonance

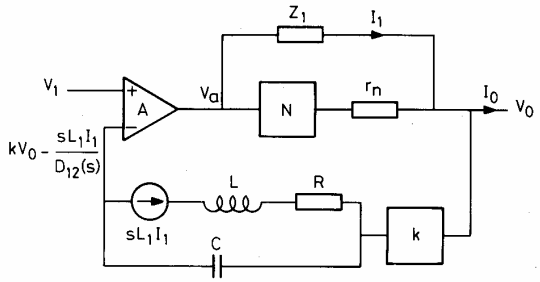


Fig. 10 *DCDF with second-order bandlimited current transformer*

of the current transformer. Using the transformation

$$sL_1 \cdots \frac{sL_1}{D_{12}(s)}$$

where

$$D_{12}(s) = 1 + \frac{1}{Q_1}\left(\frac{s}{\omega_1}\right) + \left(\frac{s}{\omega_1}\right)^2$$

enables the closed-loop gain to be expressed as

$$G_{n0}(s) = \left\{k + \frac{1}{A}\left[\frac{1}{D_{12}(s)} + \frac{1}{N_T}\left(1 - \frac{1}{D_{12}(s)}\right)\right]\right\}^{-1} \quad (4f)$$

To determine the sensitivity of $G_{n0}(s)$ to $D_{12}(s)$, an error function $E(s)$ is defined

$$E(s) = \frac{G_{n0}(s)}{G_{n0}(s)|_{D_{12}(s)=1}} - 1$$

where the advantage of making $\omega_1 >$ {amplifier gain–bandwidth product} is shown by

$$E(s) = -\left[\frac{D_{12}(s) - 1}{D_{12}(s)}\right]\left[\frac{N_T - 1}{N_T}\right]\frac{1}{1 + kA}$$
$$= -\frac{s}{\omega_1}\left[\frac{1}{Q_1} + \frac{s}{\omega_1}\right]\left[\frac{N_T - 1}{N_T}\right]\frac{1}{1 + kA} \quad (4g)$$

5 Current-dumping amplifier

The *current-dumping* amplifier illustrated in elemental form in Fig. 11*a* was introduced in 1975 by Albinson and Walker [4] and is described as a feedback amplifier with feedforward error correction [14]. The basic operation suggests that the pre-output-stage amplifier (A in the present notation) both drives the output stage N and supplies an error-correction current via a resistor R_1 ($Z_1 = R_1$) to compensate for the nonlinearity in N. An alternative thesis is proposed here that the principal distortion-correction process in the Albinson–Walker amplifier is differential-error feedback, although the supporting role of a partial bypass around N is recognised. We also show how DCDF can compensate for the finite output impedance of the class A stage, which has to supply current to both the output transistors and the feedforward resistor.

5.2 Comparison of current dumping and DCDF amplifiers

A progressive transformation from current dumping to DCDF is shown in Fig. 11*a*, *b*, *c* and *d*, where the output stage is modelled as a nonlinear transfer characteristic N with nonlinear output impedance r_n. In Fig. 11*c* and *d* this is combined with Z_2, where $Z_{2n} = r_n + Z_2$. Thus, defining $N_T^{-1} = 1 + NZ_1/Z_{2n}$, the transfer function for the unbalanced amplifier is

$$\frac{V_0}{V_1} = \frac{1}{1 + s\tau}\left[1 - \frac{I_0 Z_1}{V_1}\left(s\tau + \frac{1 - sA\tau}{AN_T^{-1}}\right)\right] \quad (5a)$$

Under optimum balance $sA\tau = 1$, where $R_1 = Z_1$ and $\tau = L_2/R_1$ are substituted,

$$\frac{V_0}{V_1} = \frac{A}{1 + A} - \frac{I_0\{(sL_2)//(R_1)\}}{V_1} \quad (5b)$$

The schematic diagram of Fig. 11*d* distinguishes between error feedforward and error feedback. Stage N has partial feedforward via Z_1 but appears within the feedback path, whereas DCDF senses the output-stage error. That is, feedforward is not fundamental to this process. However, the feedforward summing network $\{Z_1, Z_{2n}\}$ supplements output-stage performance and reduces the differentiated-error signal and, hence the dynamic range requirements of the amplifier input stage. For the *current-dumping* amplifier, increasing Z_1 requires an increase in the inductance of L_2 to maintain balance. However, with DCDF,

the output impedance need not increase as the generator $sL_2 I_0$ shown in Fig. 11*d* is independent of the balance condition. Fig. 12 shows an equivalent amplifier with output voltage derived feedback and DCDF.

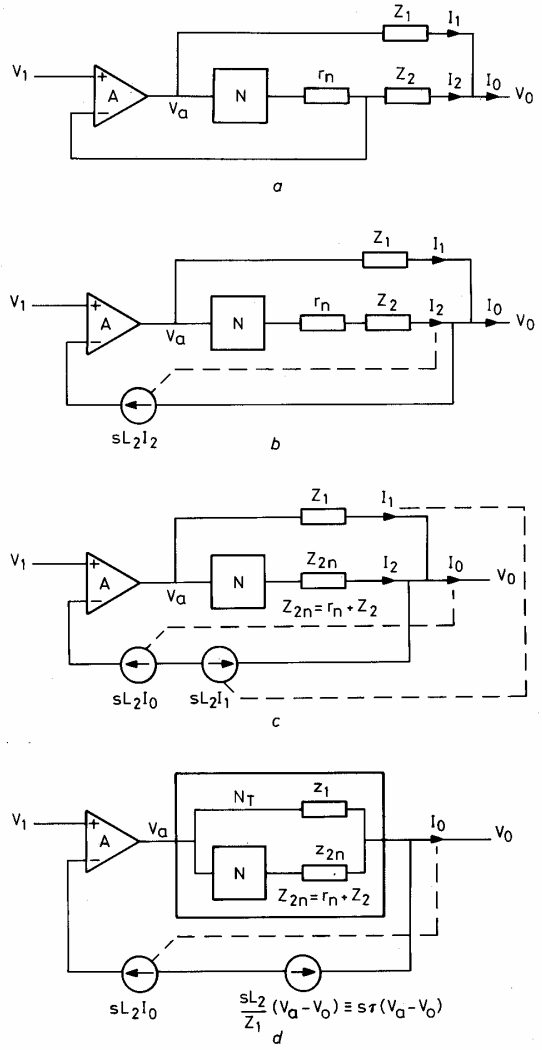


Fig. 11 *Transformation from current dumping to DCDF*
a Current dumping topology with unity-gain feedback
b Current dumping amplifier with differential-current sensor
c Current dumping amplifier with differential-current sensing of both output and feedforward currents
d Current dumping amplifier configured with DCDF

5.2 DCDF enhancement of current dumping

The class A pre-output stage of a current-dumping amplifier supplies current to both the output transistors and the feedforward impedance Z_1, where, because these currents are generally nonlinear functions of the input signal, distortion occurs when the class A amplifier has a finite output impedance Z_a. To reduce this distortion, Fig. 13 shows a DCDF enhanced current-dumping amplifier that senses both I_1 (the current in Z_1) and I_n (the nonlinear input current to the output stage N), where the closed-loop gain can be expressed as

$$\frac{V_0}{V_1} = \frac{A}{1+A} + \frac{\begin{array}{c} sA(L_n I_n + L_1 I_1) - A(I_0 - I_1)Z_2 \\ -(I_1 + I_n)Z_a - I_1 Z_1 \end{array}}{(1+A)V_1} \quad (6a)$$

where, putting $Z_2 = sL_2$ and equating terms in I_1 and I_n, the balance conditions are stated as

$$A = \frac{Z_a + Z_1}{s(L_1 + L_2)} = \frac{Z_a}{sL_n} \quad (6b)$$

that is

$$L_n = (L_1 + L_2)\frac{Z_a}{Z_a + Z_1} \quad (6c)$$

Although, in this scheme, both I_n and I_1 are sensed, in the Walker–Albinson amplifier, $L_1 = 0$ because of the

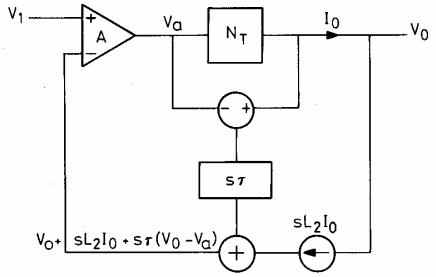


Fig. 12 *Differential-error and output-voltage derived feedback amplifier*

inclusion of Z_2; thus only a single low-current sensor for I_n is required. The technique is also applicable to the transconductance amplifier of Section 4.3, where, comparing A and Z_a with the core cells g_1, g_2 and using a Norton–Thevenin transformation,

$$A = g_1\left(r_0 - \frac{1}{g_2} + \frac{1}{sC_0}\right) \quad (6d)$$

$$Z_a = \frac{1}{g_2} \quad (6e)$$

If $g_2 r_0 = 1$, then $A = g_1/(sC_0)$, and the balance conditions follow as

$$L_1 + L_2 = \frac{C_0}{g_1}\left(Z_1 + \frac{1}{g_2}\right) \quad (6f)$$

$$L_n = \frac{C_0}{g_1 g_2} \quad (6g)$$

Providing the core of the transconductance cell g_2 has a broad bandwidth, balance accommodates both output-

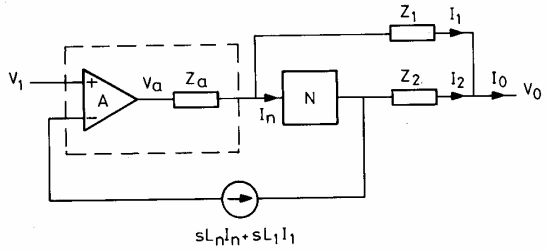


Fig. 13 *Current dumping amplifier with DCDF enhancement*

stage nonlinearity and nonlinear loading on the forward amplifier.

6 Verification of DCDF using nonlinear transient analysis

To validate the distortion reduction of a dual-sensor DCDF amplifier, a nonlinear transient analysis was per-

formed on the Fig. 9 transconductance topology using the nonlinear model described in Fig. 14, where bias voltages B_1 and B_2 determine the output-stage bias current.

class C (0 mA bias), where DCDF alignment is most sensitive; however, under optimum balance, low-level distortion cannot be observed in Fig. 15 for 100% correction.

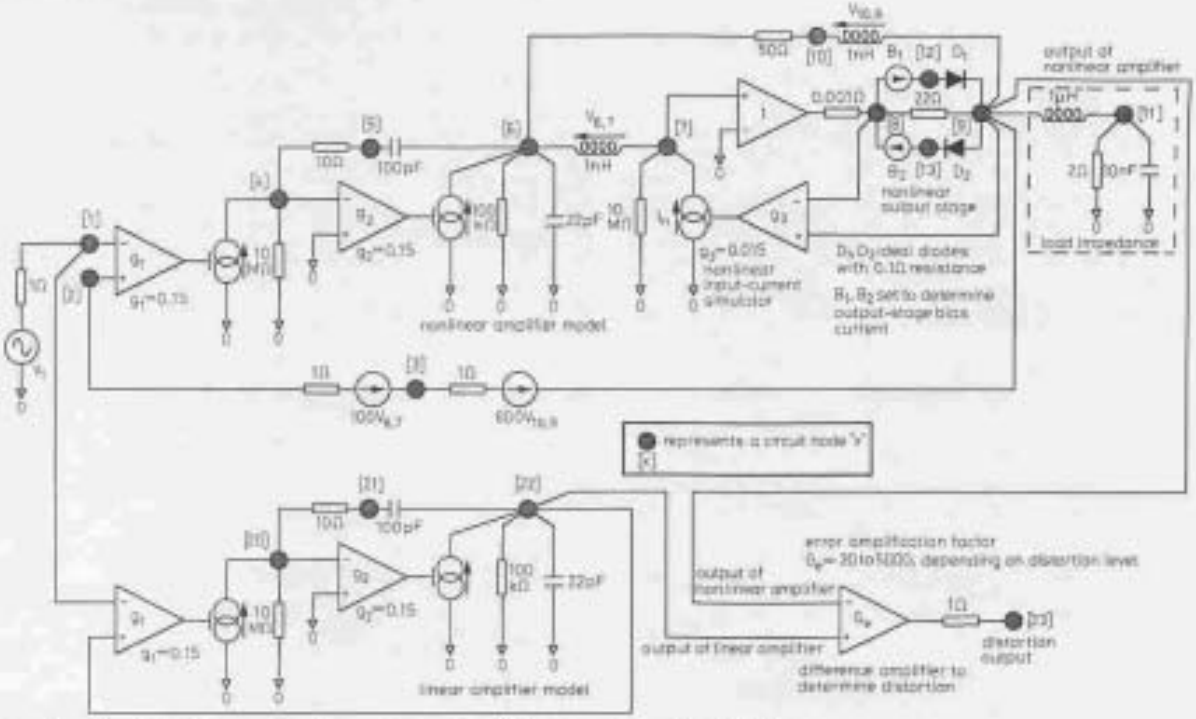


Fig. 14 *Simulation of linear and nonlinear amplifiers with a difference amplifier to determine distortion*

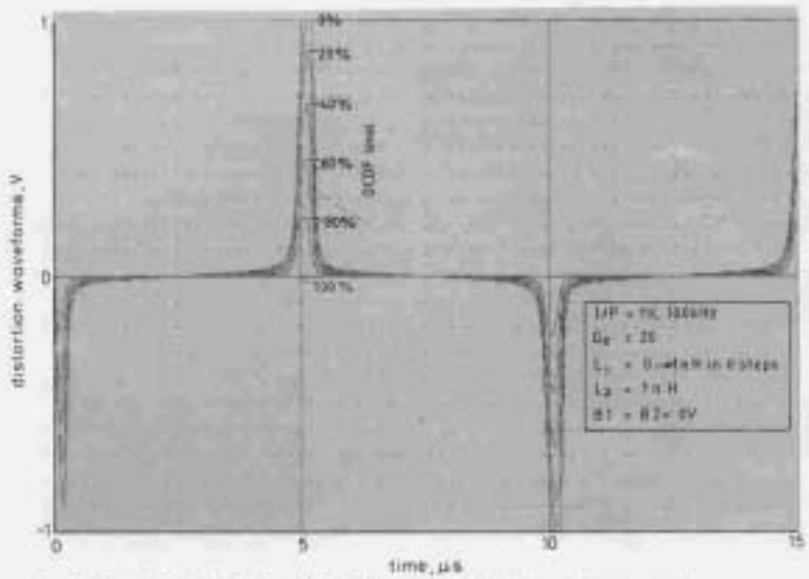


Fig. 15 *Distortion waveforms for class C biased amplifier with 6 levels of DCDF and optimum loading correction*

Three sets of distortion traces are presented in Figs. 15, 16 and 17 and relate to the data in Table 1. Component values are defined in Fig. 14, and all time-domain distortion waveforms correspond to a 1 V amplitude and a 100 kHz sinusoidal input signal (chosen to accentuate low-level crossover distortion), with a load impedance of 2 Ω//10 nF. Significant impulsive distortion results under

Under class AB (47 mA bias), distortion is reduced where the error traces in Figs. 16 and 17 confirm the balance conditions deduced in Section 4. However, for optimum distortion correction again, both DCDF loops must be aligned, where, for example, if difference amplifier gain $G_e = 5000$ is selected, no distortion is observed on the 100% trace in Fig. 17.

Table 1: Amplifier and evaluation data

	Output bias current	Time axis	Error amplifier gain	Input DCDF
Fig. 15	0 mA	0–15 μs	20	Yes
Fig. 16	47 mA	0–30 μs	420	No
Fig. 17	47 mA	0–30 μs	5000	Yes

7 *m*th-order nested DCDF error-correction amplifier

DCDF can be extended to systems of order m that incorporate m first-order amplifiers and m DCDF correction paths using m feedforward resistors. An mth-order DCDF system is shown in Fig. 18. The inner amplifier loop (designated by a gain N_1) is identical to that discussed in Sections 4.1 and 4.2 and includes dual correction for both output-stage voltage transfer and current-transfer nonlinearities. A second amplifier with a forward-path gain A_2, in association with N_1, now forms a second feedback loop, where an additional feedforward resistor R_2 and differential-current transformer realise the second correction path. It is assumed unnecessary here to incorporate current correction as the input current to amplifier A_1 is generally both linear and negligible, and hence only one transformer is required. This method of nesting correction amplifiers can then be systematically extended to a composite mth-order topology.

The closed-loop transfer function of the mth-order amplifier and its associated error function with respect to nonlinear gain N_0 can be derived as follows:

Consider the inner amplifier loop formed by A_1, N_0, R_1, Z_{01}, L_1 and L_s, which is also depicted in Fig. 8 where $A = A_1$, $N = N_0$ and $Z_0 = Z_{01}$. To obtain an expression

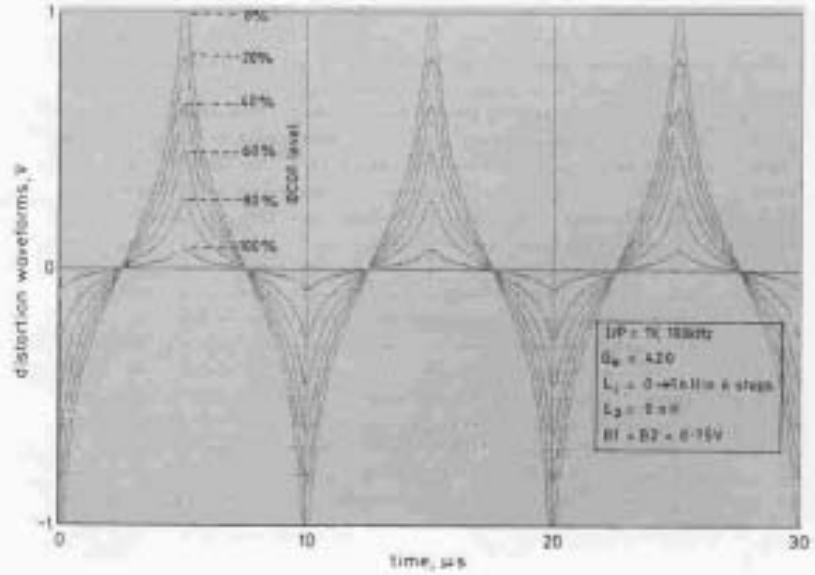


Fig. 16 *Distortion waveforms for class AB biased amplifiers with 6 levels of DCDF and no loading correction*

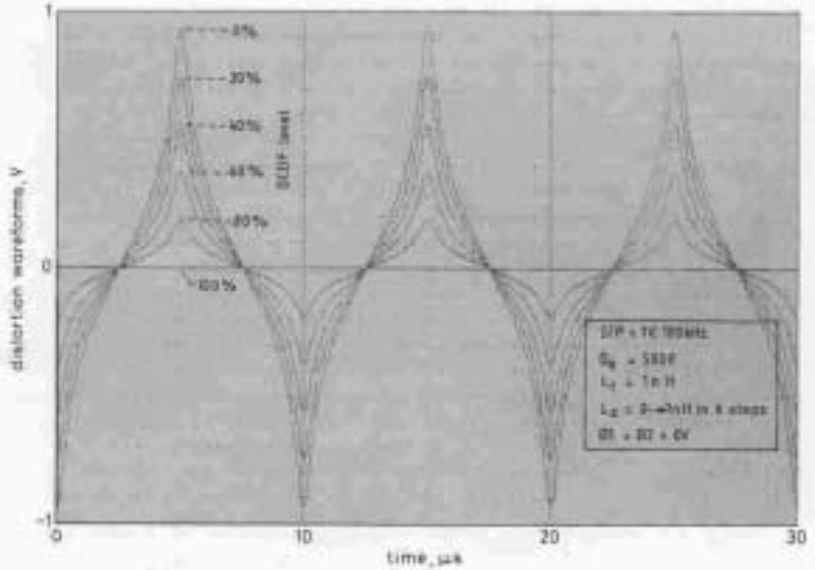


Fig. 17 *Distortion waveforms for class AB biased amplifier with optimum voltage-transfer correction and loading correction swept over 6 levels*

for I_n in terms of the output voltage V_0, the output-stage N_0 is given a nonlinear input impedance Z_{n0} (which is partially output-load-dependent), where $I_n = V_0/(N_0 Z_{n0})$. where the target transfer function is given by N_m for $N_0 = 1$ and $\lambda_n = 0$. Eqn. 7*f* reveals that the dependence upon N_0 and the input-current of the output stage are

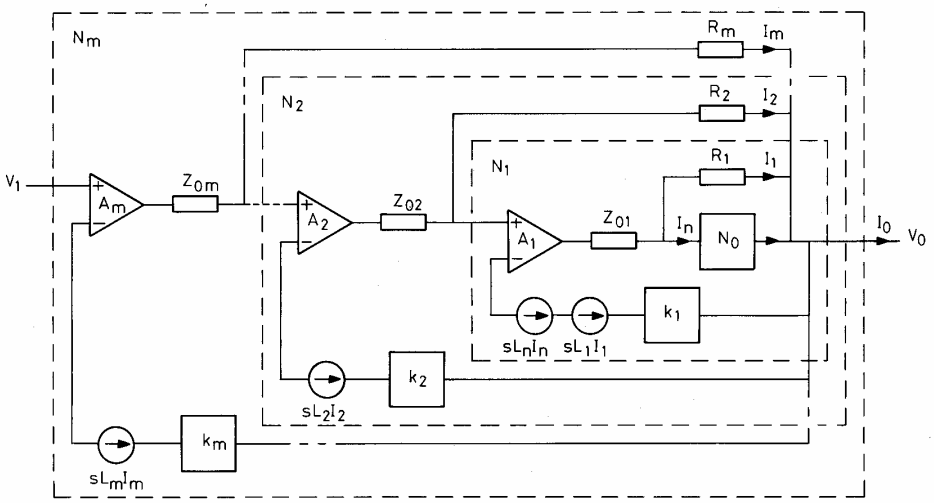


Fig. 18 *General mth-order, nested DCDF error-correction amplifier*

The transfer function of the inner amplifier loop $N_1 = V_0/V_1$ then follows as

$$\frac{V_0}{V_1} = \frac{A_1}{1 + k_1 A_1} + \frac{1}{V_1} \frac{sA_1(L_n I_n + L_1 I_1) - (I_1 + I_n)Z_{01} - I_1 R_1}{1 + k_1 A_1}$$

from which, balance equations λ_n, λ_m, and a factor Q_m can be identified as follows:

$$\lambda_n = \frac{1}{Z_{n0} N_0}(Z_{01} - sL_n A_1) \tag{7a}$$

$$\lambda_m = 1 + \frac{Z_{0m} - sL_m A_m}{R_m} \tag{7b}$$

$$Q_m = 1 + A_m(k_m - 1) \tag{7c}$$

An iterative expression for the closed-loop transfer function N_m in terms of N_{m-1} is

$$N_m = \left[1 + \frac{Q_m + \lambda_n + \lambda_m(N_{m-1}^{-1} - 1)}{A_m}\right]^{-1} \tag{7d}$$

However, except for amplifier N_1, the input impedances $Z_{n(m-1)}$ of the remaining amplifiers in the nest are large, and therefore, for $m > 1$, corresponding terms $\lambda_n \approx 0$. Hence, applying eqn. 7*d*, the closed-loop gain N_m follows as

$$N_m = \left\{1 + \frac{Q_m}{A_m} + \sum_{r=2}^{m}\left[\frac{Q_{r-1}}{A_{r-1}} \prod_{r}^{m} \frac{\lambda_r}{A_r}\right] + (N_0^{-1} - 1)\prod_{r=1}^{m}\frac{\lambda_r}{A_r} + \frac{\lambda_n}{A_1}\prod_{r=2}^{m}\frac{\lambda_r}{A_r}\right\}^{-1} \tag{7e}$$

To express the overall sensitivity to imbalance and loop gain, the error function E_m is

$$E_m = \frac{(1 - N_0^{-1}) - \lambda_n \prod_{r=2}^{m} \lambda_r}{N_m \prod_{r=1}^{m} A_r} \tag{7f}$$

reduced by both the product of the m forward-path amplifier gains $A_1 A_2 A_3 \cdots A_m$ and the product of the m transmission zeros involving the balance condition lambda functions. Also, assuming optimal alignment of the m balance conditions, eqn. 7*e* reduces to $N_m = N_{opt}$, where

$$N_{opt} = \frac{A_m}{1 + k_m A_m} \approx \frac{1}{k_m} \tag{7g}$$

where, in this idealised case, stability is dependent only upon A_m and k_m.

The advantage of this new structure over earlier nested-amplifier proposals [16] is the ability to generate m transmission zeros in the nonlinear error function, which is further desensitised by the product of the amplifier gains $A_1 A_2 A_3 \cdots A_m$. Also, the multiple DCDF loops assist in the design of a stable amplifier system where, under optimal balance, the closed-loop gain takes the form $A_m/(1 + k_m A_m)$; that is, for an ideal system, only one amplifier determines stability, although practical circuitry may lead to a more complicated result.

8 Conclusions

A method for implementing an efficient means of error correction has been described that uses a differential-current transformer where the transfer functions of transformer and forward-path amplifier are complementary. Both voltage-transfer and transconductance pre-output-stage amplifiers can be accommodated, and a second DCDF loop can correct for output-stage nonlinear current gain.

Error-correction performance and the expressions for optimum balance were confirmed using nonlinear transient analysis of a stylised, nonlinear amplifier model. However, unlike the *current-dumping* topology, the balance state is independent of the feedback factor, as the floating-current transformers can be connected directly to the input stage. This also increases their resonant frequency by allowing a reduction in the number of secondary turns.

DCDF and current-dumping feedforward error-correction were compared, and, with DCDF, the level of feedforward addition was shown to be uncritical of the balance state. However, feedforward can reduce the magnitude of DCDF sensed error signals resulting from both crossover distortion and limited output-stage high-frequency gain.

The DCDF technique was extended to an *m*th-order nested-feedback topology using *m* feedforward resistors in association with *m* differential-current transformers. This system reduces further sensitivity to output-stage nonlinearity and suggests an alternative means of controlling closed-loop stability such that, under optimal alignment, only the outermost amplifier and feedback path of the nest appear in the expression for closed-loop gain.

It is expected that the dual DCDF correction system can be introduced into existing circuits with only minor modifications to topology, although the potential for an *m*th-order loop should be of interest in very low-distortion applications. The modest-cost overhead compared with other more circuit-intensive proposals [13, 15] should also prove attractive, especially as the floating-circuit elements are passive and, in association with appropriate screening, are independent of power supply-induced distortion.

9 References

1 HAWKSFORD, M.J.: 'Distortion correction circuits for audio amplifiers', *JAES*, 1981, **29**, pp. 503–510

2 HAWKSFORD, M.O.J.: 'Reduction of transistor slope impedance dependent distortion in large-signal amplifiers', *JAES*, 1988, **36**, pp. 213–222

3 LIPSHITZ, S., and VANDERKOOY, J.: 'Is zero distortion possible with feedback?'. Presented at 76th Convention of Audio Eng. Soc., Preprint 2170 (F-4), 8th–11th October 1984, New York

4 WALKER, P.J., and ALBINSON, M.P.: 'Current dumping audio amplifier'. Presented at 50th Convention of Audio Eng. Soc., London, UK, 4th–7th March 1975

5 WALKER, P.J.: 'Current dumping audio power amplifier', *Wireless World*, 1975, **81**, pp. 560–582

6 VANDERKOOY, J., and LIPSHITZ, S.P.: 'Current dumping — does it really work?', 1978, **84**, (1510), pp. 38–40

7 VANDERKOOY, J., and LIPSHITZ, S.P.: 'Feedforward error correction in power amplifiers', *JAES*, 1980, **28**, pp. 2–16

8 McLOUGHLIN, M.: 'Current dumping review — 1', *Wireless World*, 1983, **89**, (1572), pp. 39–43

9 McLOUGHLIN, M.: 'Current dumping review — 2', *Wireless World*, 1983, **89**, (1573), pp. 35–41

10 WALKER, P.: 'Current dumping', *Wireless World*, 1983, **89**, (1575), p. 49

11 BAXANDALL, P.J.: 'Current dumping', *Wireless World*, 1983, **89**, (1575), pp. 49–50

12 VANDERKOOY, J., and LIPSHITZ, S.P.: 'Current dumping review', *Wireless World*, 1984, **90**, (1577), p. 49

13 HAWKSFORD, M.J.: 'Power amplifier output stage design incorporating error feedback correction with current dumping enhancement'. Presented at 74th Convention of AES, Preprint 1993 (B-4), 8th–12th October 1983, New York

14 ALLINSON, N.M., and WELLINGHAM, J.: 'Distortion reduction in frequency-dependent feedback-feedforward amplifiers', *Int. J. Electron.*, 1985, **59**, pp. 667–683

15 HAWKSFORD, M.J.: 'Distortion correction in audio power amplifiers', *JAES*, 1981, **29**, pp. 27–30

16 CHERRY, E.M.: 'A new result in negative-feedback theory, and its application to audio power amplifiers', *IEEE Trans.*, 1978, **CTA-6**, pp. 265–288

Quad-input current-mode asymmetric cell (CMAC) with error correction applications in single-ended and balanced audio amplifiers

M.O.Hawksford

Indexing terms: Audio amplifiers, Quad-input asymmetric cell, Cell topology

Abstract: An amplifier topology is introduced that functions principally in a current-steering mode and offers wide bandwidth and low distortion. The cell can accommodate an additional difference port and thus finds application in a range of audio amplifier incorporating error correction. The cell topology and its properties are described and methods of error correction are reviewed where it is shown that the new topology can be used in both single-ended and balanced power amplifiers.

1 Introduction

It is now widely accepted, among audio circuit designers, that simplicity in the choice of topology is desirable, providing a high standard of objective performance is maintained within the operating envelope of the system; however, this strategy requires the identification of techniques that use active devices in their most linear operating mode. Current-steering circuits [1] offer wide bandwidth as the charging of parasitic capacitance, with associated wide voltage swings on multiple circuit nodes, is reduced. For similar reasons, such cells can also yield a useful reduction in nonlinear distortion providing appropriate cascode circuitry [2] is used at circuit nodes where wide voltage swings are encountered.

A critical factor affecting audio amplifier performance relates to the signal currents that flow in ground lines and power supplies as these can both create direct errors and induce errors through interstage interaction. It is, therefore, desirable to identify circuits that minimise and localise such currents so as to enable signal currents to flow in well-defined closed paths that ideally exclude the power supply and ground line. With signals of modest level (~$2V_{RMS}$) it is straightforward to achieve low distortion. However, for higher-level signals, such as encountered in power amplifiers, this objective is more problematic because of transistor parameter modulation and output-stage nonlinearity.

IEE Proceedings online no. 19960149

Paper first received 2nd December 1994 and in revised form 20th October 1995

The author is with the Centre for Audio Research and Engineering, Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester C04 3SQ, UK

To address these problems, this paper reviews some of the principles of error correction [3–6] and shows how a new topology, defined as a current-mode asymmetric amplifier cell (CMAC), can be adapted for operation in such amplifiers by providing four input nodes, two of which function as an error-feedback port. This tutorial element describes a unified approach to error correction in both single-ended and balanced audio power amplifiers. An analysis of nonlinear distortion is presented and it is shown how a local error-correcting feedforward path within a CMAC can enhance its internal linearity.

2 Elementary CMAC topologies and their properties

Two widely adopted elementary amplifier cells [7] are shown in Figs. 1 and 2 that achieve inverting and non-inverting gain, respectively. However, in Fig. 3 a modified topology is depicted that incorporates properties of both the earlier cells and forms the kernel of a CMAC. In this basic configuration it is assumed that no loading is applied to the amplifier output port. Therefore, the signal current flows directly between collector and emitter in a closed path that can exclude the power supply. Consequently, the input current, i_{in}, must correspond to the base current, i_b, of the transistor and results in an input impedance, Z_{in}, given by

$$Z_{in} = (R_1 + r_e)(1 + \beta) \quad (1)$$

where β is the transistor collector-base current gain and $r_e = \partial V_{BE}/\partial I_E$. In applications requiring a low voltage gain, linearity improves as $R_1 >> r_e$, where in the limit of $A_v = 1$, $R_1 = \infty$ and output and input are connected only by R_2. It follows that any signal current drawn from the output must draw additional current directly from the input which lowers the input impedance, as in this simple example the input (applied at the junction of R_1 and R_2) is unbuffered and input and output are linked by R_2. Consequently, in practical circuits either the input or the output should be buffered and the use of complementary circuitry to reduce distortion and facilitate DC biasing is desirable. To illustrate how this can be achieved, a circuit example is given in Fig. 4 incorporating a complementary, compound cell formed using transistors T_1 – T_6 while T_7 and T_8 form a complementary grounded-base stage. Transistors T_5 and T_6 limit the collector–emitter voltage of transistors T_3 and T_4, allowing them enhanced bias stability and freedom from large signal swings that can cause distortion through transistor parametric modulation. The distor-

tion performance of the compound, complementary cell is studied in Section 3.

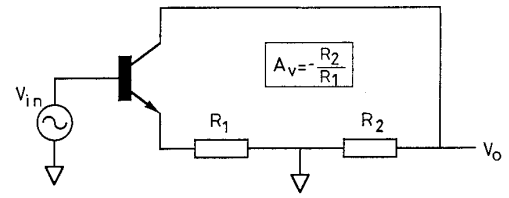


Fig. 1 *Common-emitter amplifier*

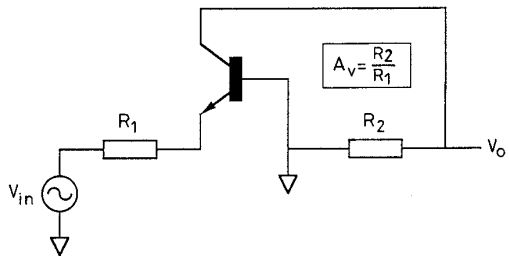


Fig. 2 *Grounded-base amplifier*

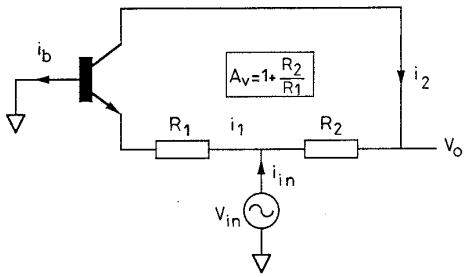


Fig. 3 *Current-mode asymetric cell*

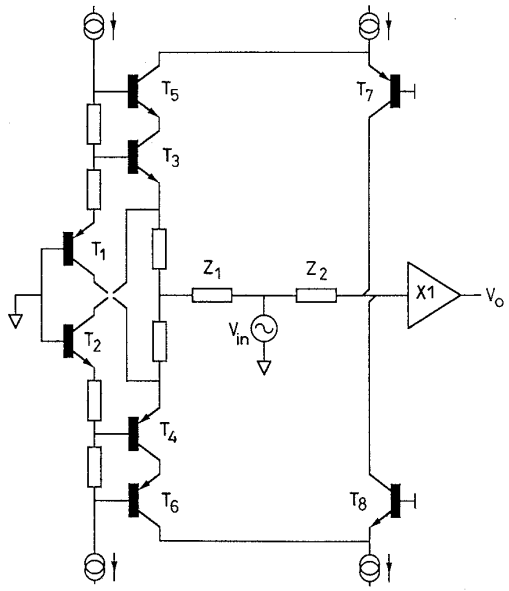


Fig. 4 *CMAC with complementary cell, output cascode and unity-gain buffer*

The circuit of Fig. 4 also can have a relatively high input impedance where to calculate the input signal current all current paths via transistor bases must be summed. However, to make the cell more tolerant to real circuit operating conditions some modifications are recommended:

(i) input-stage common-collector buffer

(ii) output-stage common-collector buffer

(iii) integral current mirror within the cell.

Recommendations (i) and (ii) are self evident; however, the incorporation of a current mirror offers a number of improvements where the basic topology is shown in Fig. 5. An input transistor T_1 acts as a unity-gain buffer stage but where the collector current i_x is mirrored with a gain m and applied to the emitter of T_2.

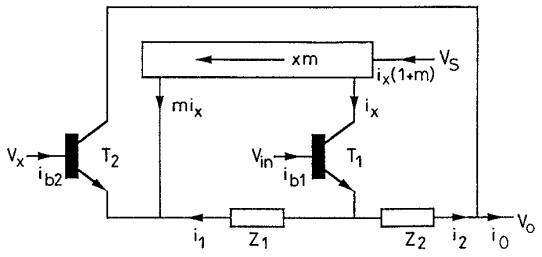


Fig. 5 *Current mirror enhanced CMAC*

The current mirror offers a useful performance tolerance to output load current i_o. In this application it is assumed that the input impedance of the current mirror is low and is referenced to the power supply voltage while the output impedance is high thus acting as a current source to the emitter of T_2. Assume initially $V_{in} = 0$V and $m = 0$, then, because of the low impedance at the emitter of T_1, most of the output current i_o flows in Z_2 yielding an output impedance $Z_o \approx Z_2$. However, for $m > 0$ and assuming V_x, $V_{in} = 0$V, then $i_x \approx i_2$ giving $i_o = i_2 (1 + m)$, whereby

$$Z_o = \frac{Z_2}{1+m} \tag{2}$$

Because the current i_x is reduced from i_o to $i_o/(1+m)$, the input impedance is increased by $(1+m)$ and any tendency for modulation of r_{e1} (i.e. for T_1, $r_{e1} = \partial V_{BE1}/\partial I_{E1}$) is correspondingly reduced, thus lowering distortion contribution from this stage.

The current mirror engenders a tolerance to loading and simultaneously aids linearisation of T_1. The value of m is uncritical in a basic CMAC (in practice being constrained by DC biasing and stability considerations) although a high value will result in both a lower output resistance and a higher input impedance. Observe that the signal current i_1 is unaffected by the choice of m although, as m is increased, a greater fraction of the output current now flows in T_2 contributing to the distortion generated by this transistor. Hence, distortion performance in a CMAC cell remains critical upon the stage represented by T_2, requiring appropriate selection of bias current and topology, where an example is given in Section 5 (see Fig. 16). However, as described later in this Section, when extra error correction input ports are included, m becomes a critical parameter that requires an accurate definition.

A current mirror results in a flow of signal current via the power supply which is contrary to the conceptual aim of a CMAC cell. However, the magnitude of this current is determined principally by the output current i_o as the only other component results from the small base current of T_2. In most CMAC applications, it is envisaged that the output current will be reduced by incorporating an output buffer stage (for example, an emitter follower), thus signal current in the current mirror is lower aiding linearity in this stage. Such a strategy is followed in Section 4.

2.1 Quad-input node CMAC

An extension to the CMAC is to use the low-impedance nodes at the emitters of T_1 and T_2 as current input ports enabling the incorporation of error correction [6] as described in Section 4. The modified cell is shown in Fig. 6.

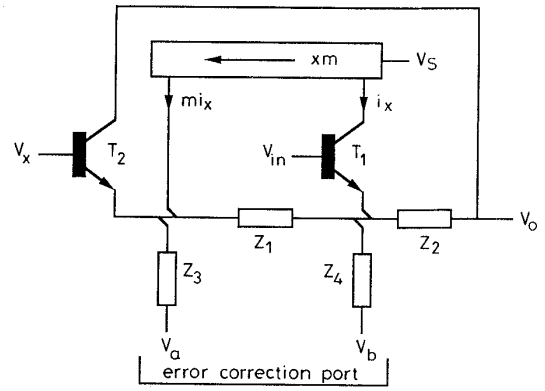


Fig. 6 *CMAC with error-correction port*

Expressing V_o as $f(V_{in}, V_x, m, Z_1, Z_2, Z_3, Z_4, V_a, V_b)$, then

$$V_o = \frac{Z_2}{1+m}\left[\frac{V_a}{Z_3} - m\frac{V_b}{Z_4}\right] - V_x\left[\frac{Z_2}{Z_1} + \left(\frac{1}{1+m}\right)\frac{Z_2}{Z_3}\right] + V_{in}\left[1 + \frac{Z_2}{Z_1} + \left(\frac{m}{1+m}\right)\frac{Z_2}{Z_4}\right]$$

To force V_o to be dependent upon $(V_a - V_b)$, make

$$\frac{Z_2}{1+m}\left[\frac{V_a}{Z_3} - m\frac{V_b}{Z_4}\right] = V_a - V_b$$

that is, expressing Z_3, Z_4 in terms of Z_2,

$$Z_3 = \left(\frac{1}{1+m}\right)Z_2 \tag{3}$$

$$Z_4 = \left(\frac{m}{1+m}\right)Z_2 \tag{4}$$

whereby

$$V_o = (V_a - V_b) - V_x\left(1 + \frac{Z_2}{Z_1}\right) + V_{in}\left(2 + \frac{Z_2}{Z_1}\right) \tag{5}$$

The CMAC with an error-correction port (with unity weighting, see '1' on symbol) is represented by the symbol shown in Fig. 7, which forms the amplifier cell used in the example power amplifier circuits incorporating error correction.

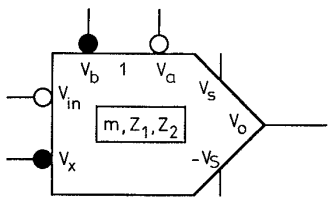


Fig. 7 *Symbol for CMAC with single error-correction port*
○—— noninverting input
•—— inverting input

3 Distortion performance of complementary amplifier cell

Consider the complementary cell shown in Fig. 8 as a subcircuit within a CMAC where the transistors are biased with two voltage generators V_B. Each emitter resistor is R_E and the input signal current, i, divides symmetrically between the two transistors such that for T_1, $I_{E1} = I_B - i$ and for T_2, $I_{E1} = I_B + i$. Because of nonlinear transistor action, the common-mode bias current I_B is also a function of input current having a quiescent value I_{BQ} when $i = 0$.

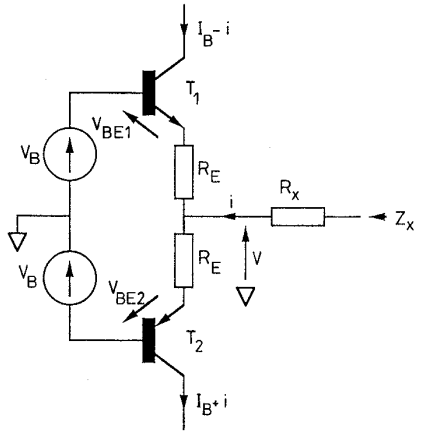


Fig. 8 *Elementary complementary emitter-coupled cell*

With reference to Fig. 8 the input impedance Z_{in} observed at the junction of the two emitter resistors R_E is established as follows:

$$v + (I_B - 0.5i)R_E + V_{BE1} - V_B = 0 \tag{6}$$

$$v - (I_B + 0.5i)R_E + V_{BE2} + V_B = 0 \tag{7}$$

Addition of eqns. 6 and 7 gives

$$v = 0.5[iR_E - V_{BE1} - V_{BE2}]$$

whereby

$$Z_{in} = \frac{\partial v}{\partial i} = 0.5\left[R_E - \left(\frac{\partial V_{BE1}}{\partial i} + \frac{\partial V_{BE2}}{\partial i}\right)\right]$$

Defining $\alpha = kT/q$, putting $V_{BE1} = \alpha \ln[(I_B - 0.5i)/I_s]$ and $V_{BE2} = -\alpha \ln[(I_B + 0.5i)/I_s]$ and differentiating V_{BE1}, V_{BE2} with respect to i, then

$$\frac{\partial V_{BE1}}{\partial i} = \frac{\alpha}{I_B - 0.5i}\left(\frac{\partial I_B}{\partial i} - 0.5\right)$$

$$\frac{\partial V_{BE2}}{\partial i} = \frac{-\alpha}{I_B + 0.5i}\left(\frac{\partial I_B}{\partial i} + 0.5\right)$$

Substituting the differentials in the expression for Z_{in} and simplifying,

$$Z_{in} = 0.5R_E + \frac{0.5\alpha}{I_B^2 - 0.25i^2}\left(I_B - i\frac{\partial I_B}{\partial i}\right) \tag{8}$$

Similarly, subtracting eqns. 6 and 7 and substituting for V_{BE1}, V_{BE2},

$$2I_BR_E + \alpha \ln\left(\frac{I_B - 0.5i}{I_s}\right) + \alpha \ln\left(\frac{I_B + 0.5i}{I_s}\right) = 2V_B$$

Under quiescent conditions $i = 0$, $I_B = I_{BQ}$, whereby

$$I_{BQ}R_E + \alpha \ln\left(\frac{I_{BQ}}{I_s}\right) = V_B$$

thus eliminating V_B and putting expression in terms of I_{BQ},

$$(I_B - I_{BQ})R_E + \frac{\alpha}{2}\ln\left(\frac{I_B^2 - 0.25i^2}{I_{BQ}^2}\right) = 0 \tag{9}$$

The differential $\partial I_B/\partial i$ can now be evaluated as,

$$\frac{\partial I_B}{\partial i} = \frac{0.25\alpha i}{R_E(I_B^2 - 0.25i^2) + \alpha I_B} \quad (10)$$

which enables the input impedance Z_{in} to be expressed as,

$$Z_{in} = \frac{R_E}{2} + \frac{\alpha}{2}\left(\frac{I_B R_E + \alpha}{R_E(I_B^2 - 0.25i^2) + \alpha I_B}\right) \quad (11)$$

For the special case where $R_E = 0$, then

$$I_B = (I_{BQ}^2 + 0.25i^2)^{0.5} \quad (12)$$

$$\left.\frac{\partial I_B}{\partial i}\right|_{R_E=0} = \frac{i}{4I_B} \quad (13)$$

$$Z_{in}|_{R_E=0} = \frac{\alpha}{2I_B} = \frac{\alpha}{2(I_{BQ}^2 + 0.25i^2)^{0.5}} \quad (14)$$

To determine the error in the general function for the total input impedance Z_x including the series resistor R_x as shown in Fig. 8, assign the nonlinear component of input resistance as $z(i)$, then

$$Z_x = R_x + Z_{in}(i = 0) + z(i) \quad (15)$$

Hence, defining a distortion factor $DZ(i) = z(i)/(R_x + Z_{in}(i = 0))$, then

$$DZ(i) = \frac{\alpha}{I_{BQ}(2R_x + R_E) + \alpha}\gamma(i) \quad (16)$$

where

$$\gamma(i) = \frac{(I_B R_E + \alpha)(I_{BQ} - I_B) + 0.25R_E i^2}{R_E(I_B^2 - 0.25i^2) + \alpha I_B} \quad (17)$$

The overall distortion level is reduced by the premultiplier of $\gamma(i)$ defined in eqn. 16 which is a function of I_{BQ}, R_x and R_E. However, for optimum selection of R_E $DZ(i)$ should be constant against i where the nonlinear function $\gamma(i)$ as defined by eqns. 17, 12 is plotted as a function of i with R_E as parameter and is shown in Fig. 9. In this example I_{BQ} = 25mA and R_E ranges from 0.02 to 1.92Ω, showing a minimum distortion factor for R_E = 0.62Ω.

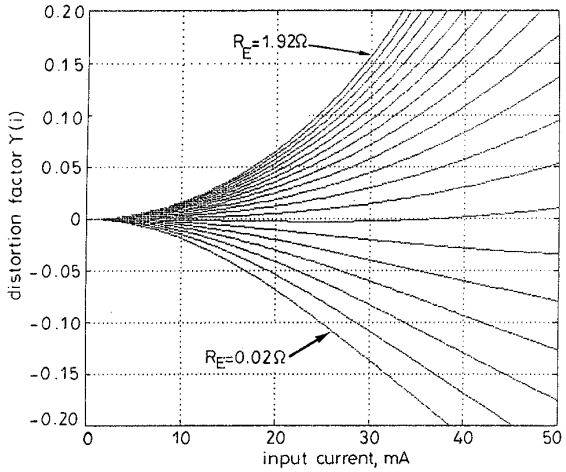


Fig. 9 *Normalised distortion characteristics for complementary emitter-coupled cell*
Circuit data: I_{BQ} = 25mA; ΔR_E = 0.1Ω; input current 0–50mA

4 Principles of distortion reduction and the application of CMAC in single-ended and balanced power amplifier

The cardinal mechanisms within an error-correction system are negative feedback, error feedback and error feedforward, where these techniques can be compounded with linearisation procedures as described in Sections 5 and 6. There has been considerable technical debate on methods of error correction [3–6] so we review only the salient features.

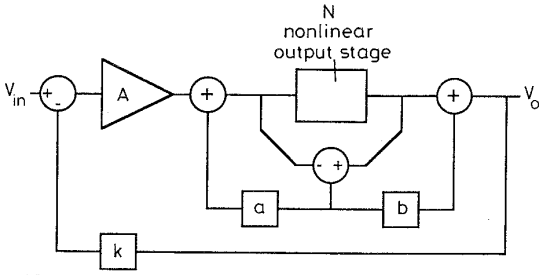


Fig. 10 *Power amplifier incorporating feedback and feedforward error correction*

For a single-ended amplifier, error correction can be applied locally to an amplifier stage as shown in Fig. 10, where N represents the nonlinear transfer function of the output stage of a power amplifier.

The closed-loop transfer function $G(f)$ of this feedback/feedforward system is given by

$$G(f) = \frac{A\left(\frac{N + bN - b}{1 - aN + a}\right)}{1 + kA\left(\frac{N + bN - b}{1 - aN + a}\right)} \quad (18)$$

To reveal the error-correction properties, define an error transfer function $E(f)$ normalised for N = 1, where

$$E(f) = \frac{G(f)}{G(f)_{N=1}} - 1$$

and

$$E(f) = \frac{(N-1)(1 + a + b)}{1 + (kbA - a)(N - 1) + kAN} \quad (19)$$

The closed-loop gain is independent of N for

$$a + b + 1 = 0 \quad (20)$$

This expression is the balance condition of the error correction amplifier where three factors contribute to reduced distortion:

(i) making $N \approx 1$

(ii) the balance condition ($a + b + 1 = 0$) introduces a zero in $E(f)$

(iii) a large value of the product $\{k\ a\ N\}$ reduces sensitivity of $E(f)$ to variation in N although we note the limitation cited by Lipshitz and Vanderkooy [8].

The topology of Fig. 10 is only one variant but is sufficient to demonstrate the principle of compound error feedback/error feedforward [3, 9] where two special cases emerge:

(i) $a = -1$, $b = 0$ pure error feedback around N

(ii) $a = 0$, $b = -1$ pure error feedforward around N.

In practice (i) is bounded by stability constraints where, if

$$a = \frac{-1}{1 + s\tau} \quad (21)$$

then, for an optimum balance condition and complete distortion cancellation,

$$b = \frac{-s\tau}{1 + s\tau} \tag{22}$$

This result demonstrates how a combination of error feedback and error feedforward can theoretically achieve total distortion cancellation even when the error feedback loop has finite bandwidth. This approach can be shown to be the basis of the Walker [10, 11] 'current-dumping' amplifier, although additional observations on the role of differential-current derived feedback (DCDF) were made in an earlier study [12].

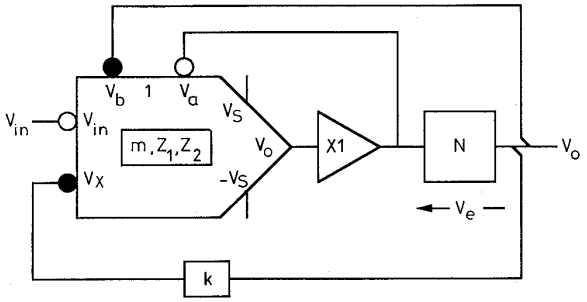


Fig. 11 *Basis error correction amplifier using CMAC*

A CMAC can be used directly within an error-feedback system, where adopting the symbology of Fig. 7, a basic single-ended amplifier is shown in Fig. 11. Alternatively, a balanced amplifier can use a CMAC with error correction where error injection can be derived either as shown in Fig. 12 or by crosscoupling between two differentially driven power amplifier, as in Fig. 13. Research by Sandman also offers additional insight into this process [13, 14].

However, observation of the crosscoupled, balanced amplifier in Fig. 13 reveals that even if N_1 and N_2 exhibit zero error, there remains a large common-mode signal applied to each error input port of the two CMACs. By combining the two configurations of error correction and introducing an additional error correction input to the CMAC as shown in Fig. 14, the common-mode signal can be cancelled contributing to lower distortion. The modified balanced power amplifier is shown in Fig. 15, where the error input ports for each CMAC have a weighting of 0.5 and the feedback factor of output to input is k.

The balance condition and operation of the power amplifier is verified as follows:

$$V_{o1} + V_{e1} = +V_{in}\left(2 + \frac{Z_2}{Z_1}\right) - kV_{o1}\left(1 + \frac{Z_2}{Z_1}\right) + 0.5(V_{e1} - V_{e2})$$

$$V_{o2} + V_{e2} = -V_{in}\left(2 + \frac{Z_2}{Z_1}\right) - kV_{o2}\left(1 + \frac{Z_2}{Z_1}\right) - 0.5(V_{e1} - V_{e2})$$

Thus, by subtraction

$$(V_{o1} - V_{o2}) + (V_{e1} - V_{e2}) = 2V_{in}\left(2 + \frac{Z_2}{Z_1}\right) - k(V_{o1} - V_{o2})\left(1 + \frac{Z_2}{Z_1}\right) + (V_{e1} - V_{e2})$$

that is, the error voltages cancel, whereby

$$\frac{V_{o1} - V_{o2}}{V_{in}} = \frac{2\left(2 + \frac{Z_2}{Z_1}\right)}{1 + k\left(1 + \frac{Z_2}{Z_1}\right)} \tag{23}$$

This Section has demonstrated how error correction can be used in both single-ended and balanced amplifiers where the underlying principle is error feedforward-feedback. Even when the power amplifiers are in a crosscoupled balanced configuration each has to function simultaneously both in a feedforward mode and recursively.

5 Error-feedforward linearisation of CMAC

A local error-feedforward correction loop [4] formed around transistor $T_{2a,b}$ can be used to sense the instantaneous base-emitter voltage and inject a correction current, as shown in Fig. 16. Effectively transistors $T_{3a,b}$ and $T_{4a,b}$ form a complementary differential amplifier which in concert with the two complementary current mirrors each of current gain m, allow a correction current $(m+1)V_e/Z_e$ to flow in Z_2.

Assuming no loss of current via the output port or transistor bases then

$$V_o - V_{in} = Z_2\left(i_1 + (m+1)\frac{V_e}{Z_e}\right)$$

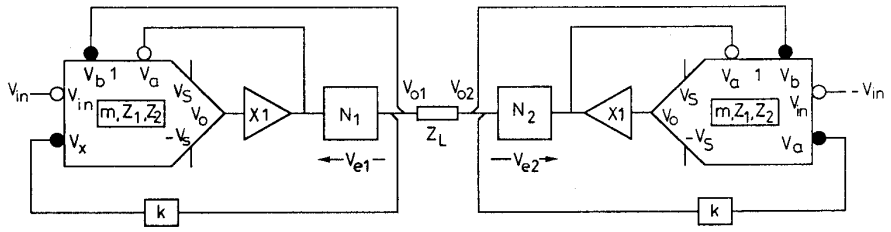


Fig. 12 *Balanced amplifier using 'normal' error correction*

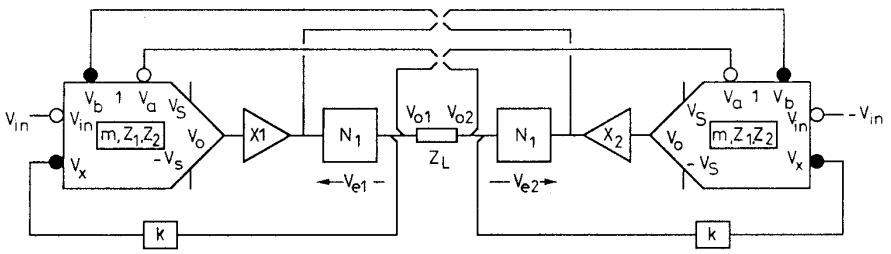


Fig. 13 *Balanced amplifier using 'crosscoupled' error correction*

and

$$V_{in} - V_x = i_1 Z_1 + V_e$$

Eliminating i_1,

$$V_o = V_{in}\left(1 + \frac{Z_2}{Z_1}\right) - V_x \frac{Z_2}{Z_1} + V_e Z_2 \left(\frac{m+1}{Z_e} - \frac{1}{Z_1}\right)$$

Hence, if

$$Z_e = (m+1)Z_1 \tag{24}$$

then dependence upon the nonlinear V_{BE} of $T_{2a,b}$ is minimised.

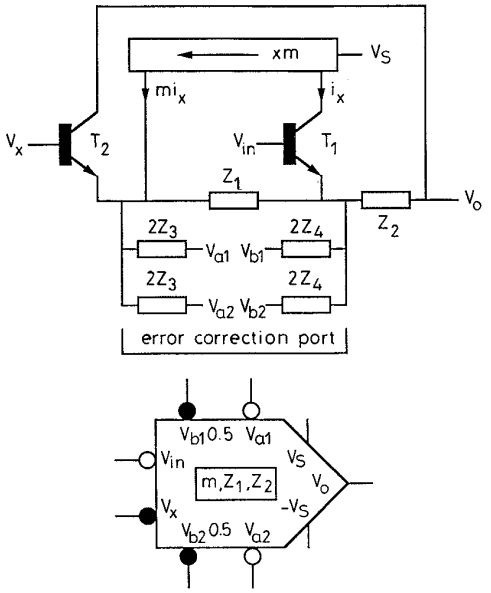


Fig. 14 *CMAC with dual error-correction ports*

In practice, a cell using the complementary configuration already exhibits low distortion as described in Section 3, however the addition of feedforward-error correction acts to further reduce distortion for applications requiring high linearity. Such correction is of course in addition to the error correction applied to the output stage.

6 Differential current-derived feedback enhanced amplifier with CMAC

Differential current-derived feedback (DCDF) [12] can also be applied to the CMAC as a means of error correction as illustrated in Fig. 17. DCDF is a method of using a current transformer to sense the time differential, of an error current that is associated with a nonlinear amplifier stage which is then used as an element within the feedback signal appearing in a first-order amplifier loop. Optimal balance of this additional feedback loop can achieve distortion cancellation. In Fig. 17 the error voltage V_e is associated with the nonlinear stage N where the derived error voltage (at the secondary of the current transformer) sL_eV_e/R_e is applied to the input of the CMAC. Analysis is straightforward and follows directly as,

$$V_o + V_e = V_{in}\left(1 + \frac{Z_2}{Z_1}\right) - \left(V_x - sL_e\frac{V_e}{R_e}\right)\frac{Z_2}{Z_1}$$

where if

$$1 = s\frac{L_e}{R_e} \times \frac{Z_2}{Z_1} \tag{25}$$

then the transfer function is independent of V_e. This is readily achieved by putting $Z_1 = R_1$ and $Z_2 = 1/sC_1$, yielding $L_e = R_e(R_1C_1)$.

7 Conclusions

This paper has considered techniques for error correction in audio power amplifiers using both single ended and balanced configurations. The principal method of error correction using feedback and feedforward was described and various power amplifier configurations considered.

An asymmetric current-steering cell was introduced designated a CMAC and its properties described. It was shown how the basic cell could be enhanced with a local feedback path using a current mirror. The CMAC offers simplicity yet through current steering and the use of unity-gain buffers, wide bandwidth and low distortion is inherent. The CMAC employs transistors in a linear configuration where the inherent properties of complementary circuitry and the operation of transistors with near constant base-collector voltages has been exploited. Where large voltage swings occur, grounded base or enhanced cascode stages [2] are recommended.

A property of the CMAC is that in the principal signal path (excluding current mirror) a single current loop is formed where effectively there is only one stage of amplification. Also it is evident that there is a direct path from input to output which is advantageous in low-gain amplifiers in terms of bandwidth and distortion. The current mirror yields a reduction in output impedance of the CMAC, but itself only has to handle signal components due to output loading and not the main signal. As such the mirror should achieve excellent linearity and can be considered as a secondary signal path operating only on demand, even if a signal is being amplified. Indeed, with appropriate design, the current mirror can function as the bias network for the main amplifier stage.

A well defined current path that does not demand signal current from the power supply can offer a similar advantage to a shunt regulator by localising the signal-current path. This enhances performance by reducing interstage coupling and the effects of signal

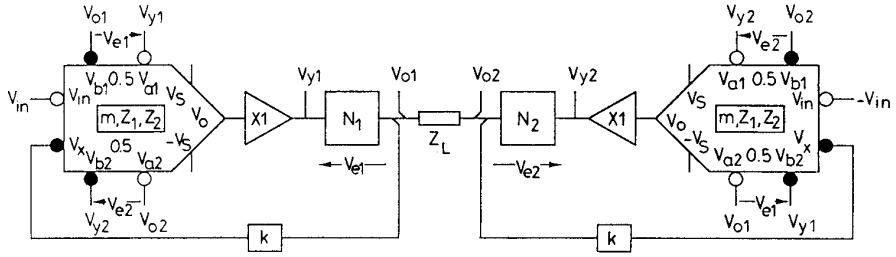


Fig. 15 *Balanced amplifier with dual error input CMAC to cancel common-mode component in error signal*

currents within the ground connection. As such the need for star grounding is less critical. The performance advantage of this stage should be compared with a two-stage amplifier that uses an inverting current mirror in the second stage, such amplifiers demand signal current from the power supply thus requiring enhanced regulation, also signal currents can circulate over wider areas of circuitry resulting in unpredictable interstage coupling.

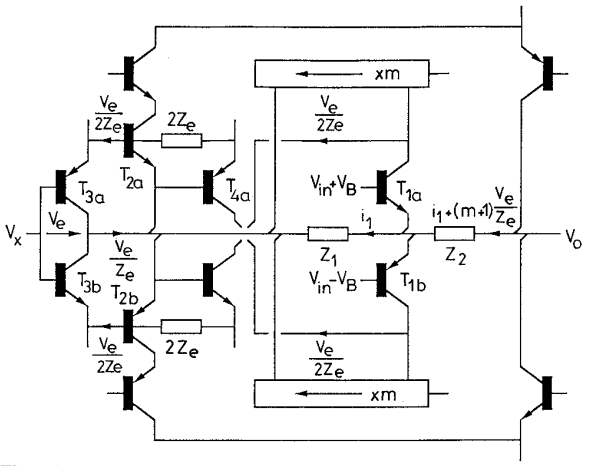


Fig. 16 *Error-feedforward linearisation of CMA*

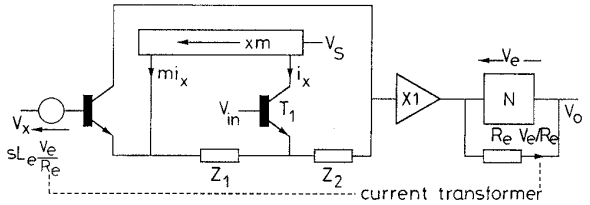


Fig. 17 *Single-ended CMAC with DCDF*

The principle of error feedforward/feedback was described and the application of the CMAC in various amplifier configurations illustrated. It was shown how the CMAC could include a difference port to accommodate the correction of output-stage distortion. Both single ended and balanced configurations were shown where the latter offered the advantage of common-mode signal cancellation in the CMAC thus improving linearity under large signal conditions.

8 References

1 ALEXANDER, M.: 'A current-feedback audio power amplifier', 88th AES convention, March 1990, preprint 290 (D-5)
2 HAWKSFORD, M.O.J.: 'Reduction of transistor slope impedance dependent distortion in large-signal amplifiers', *J. Audio Eng. Soc.*, 1988, **36**, (4), pp. 213–222
3 VANDERKOOY, J., and LIPSHITZ, S.P.: 'Feedforward error correction in power amplifiers', *J. Audio Eng. Soc.*, 1980, **28**, (1/2), pp. 2–16
4 HAWKSFORD, M.O.J.: 'Distortion correction circuits for audio amplifiers', *J. Audio Eng. Soc.*, 1981, **29**, (7/8), pp. 503–510
5 HAWKSFORD, M.J.: 'Power amplifier output stage design incorporating error feedback correction with *current dumping* enhancement', 74th convention of the AES, New York, 8–12 October 1983, preprint 1993 (B-4)
6 HAWKSFORD, M.J.: 'Distortion correction in audio power amplifiers', *J. Audio Eng. Soc.*, 1981, **29**, (1/2), pp. 27–30
7 RITCHIE, G.R.: 'Transistor circuit techniques: discrete and integrated (3rd edn.)' (Chapman & Hall, London, 1993)
8 LIPSCHITZ, S., and VANDERKOOY, J.: 'Is zero distortion possible with feedback?', 76th convention of the Audio Engineering Society, New York, 8–11 October 1984, preprint 2170 (F-4)
9 ALLINSON, N.M., and WELLINGHAM, J.: 'Distortion reduction in frequency-dependent feedback-feedforward amplifiers', *Int. J. Electron.*, 1985, **59**, (6), pp. 667–683
10 WALKER, P.J., and ALBINSON, M.P.: 'Current dumping audio amplifier', 50th convention of the Audio Engineering Society, London, 4–7 March 1975
11 VANDERKOOY, J., and LIPSHITZ, S.P.: 'Current dumping – does it really work?', *Wireless World*, 1978, **84**, (1510), pp. 38–40
12 HAWKSFORD, M.O.J.: 'Differential-current derived feedback in error-correcting audio amplifier applications', *IEE Proc., Circuits Devices Syst.*, 1994, **141**, (3), pp. 227–236
13 SANDMAN, A.M.: 'Errors – a positive approach', PhD thesis, City University, London, Nov. 1989
14 SANDMAN, A.M.: 'Who designed this?', *Electron. World*, September 1991, pp. 788–790

Relationships between Noise Shaping and Nested Differentiating Feedback Loops*

J. VANDERKOOY, *AES Fellow*

Department of Physics, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada

AND

M. O. J. HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

The application of heavy feedback is studied in two different topologies, namely, multiple-order noise shaping and nested differentiating feedback loops. Both have similar loop gain and stability considerations, although the two approaches have different implied circuit environments and areas of application. In noise shaping, emphasis is placed on the integrator characteristics of each gain stage, whereas flat-gain stages with high-frequency poles form the usual basis of the nested differentiating loop concept. This engineering report helps in understanding the application of large amounts of feedback to control noise or distortion at baseband frequencies.

0 INTRODUCTION

The concept of nesting differentiating feedback loops (NDFLs) has been introduced and promoted by Cherry [1], [2]. The basic idea with NDFLs is that for the stage that creates the most distortion (usually a power output stage), the encompassing differentiating feedback stabilizes the loop at high frequencies while allowing increased feedback at lower frequencies to reduce system distortion significantly. This engineering report makes comparisons between NDFLs and certain high-order noise-shaping loops, where it is shown that with minor topological transformations, similar loop behavior and means of stabilization are observed.

1 NOISE SHAPING

Our approach to the application of large amounts of feedback has come from a digital perspective, namely, that of noise shaping, as typically applied to requantizers in both multibit and one-bit analog-to-digital and digital-to-analog converters; for reference, see Hawksford [3]. To commence the comparative discussion, Fig. 1 illustrates a progression from a generalized error-feedback noise-shaping feedback topology to an equivalent canonic form of feedback amplifier. In the noise shaper shown in Fig. 1(a) the quantization error is filtered by a z-domain transfer function $H(z)$ and then subtracted retrospectively from the input sequence to enable partial correction for the quantizer error. This classic topology has a signal transfer of unity and a noise-shaping transfer function of $[1 - H(z)]$. The negative feedback topology of Fig. 1(d) is derived through the progression shown in Fig. 1(b) and (c), where precise equivalence in terms of both signal and noise-shaping transfer functions is achieved when

$$A(z) = \frac{H(z)}{1 - H(z)}. \tag{1}$$

Fig. 1(d) is based on a conventional digital feedback loop, although it is unusual in that a feedforward path x yields the required unity-gain signal transfer function. Since path x is outside the feedback loop, it does not modify the loop transfer function and often can be omitted at the expense of relatively benign high-frequency gain errors. It is common practice in noise shapers that are designed to maximize low-frequency performance, to implement the forward-path transfer function $A(z)$ by cascading digital integrators of the form $(1 - z^{-1})^{-1}$.

* Presented at the 93rd Convention of the Audio Engineering Society, San Francisco, CA, 1992 October 1–4; revised 1999 November 9.

However, for a loop order $N > 1$ (where N is the number of integrators in the loop), $N - 1$ transmission zeros are required in the closed-loop transfer function to enable the Bode stability criterion [4] to be satisfied. Fig. 2 shows two equivalent methods of loop synthesis based on cascaded integrators. Fig. 2(a) follows the structure presented in [3], whereas Fig. 2(b) is configured to support the discussion in Section 2. Again path x provides appropriate signals injected into the forward path to maintain a unity-gain signal transfer function. The topologies of Fig. 2 have identical characteristics when $\alpha_r = \beta_r$ for $r = 1, \ldots, N - 1$.

By way of example, both a first-order and a second-order noise shaper are considered, as shown in Figs. 3(a) and 4(a), respectively. In the first-order system the error resulting from quantizer Q is modified by just a sample delay, where $H_1(z) = z^{-1}$ (there must always be at least a one-sample delay in the loop), whereas for the second-order case $H_2(z) = z^{-1}(2 - z^{-1})$ is selected. From Eq. (1),

$$A_1(z)|_{\text{first order}} = \frac{z^{-1}}{1 - z^{-1}}$$

$$A_2(z)|_{\text{second order}} = \frac{z^{-1}(2 - z^{-1})}{1 - 2z^{-1} + z^{-2}}$$

$$= \frac{z^{-1}}{1 - z^{-1}}\left(\frac{1}{1 - z^{-1}} + 1\right).$$

$A_1(z)$ is a single integrator whereas $A_2(z)$ consists of two cascaded integrators together with a unity-gain feedfor-

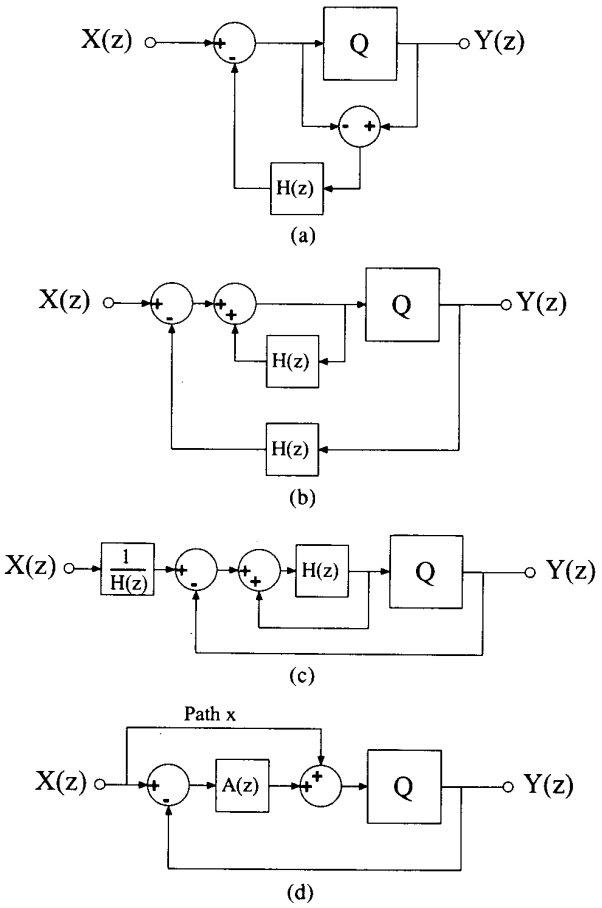


Fig. 1. Progression from classic noise-shaping configuration to exact equivalent negative-feedback loop with feedforward path x designed to maintain unity-gain signal transfer function.

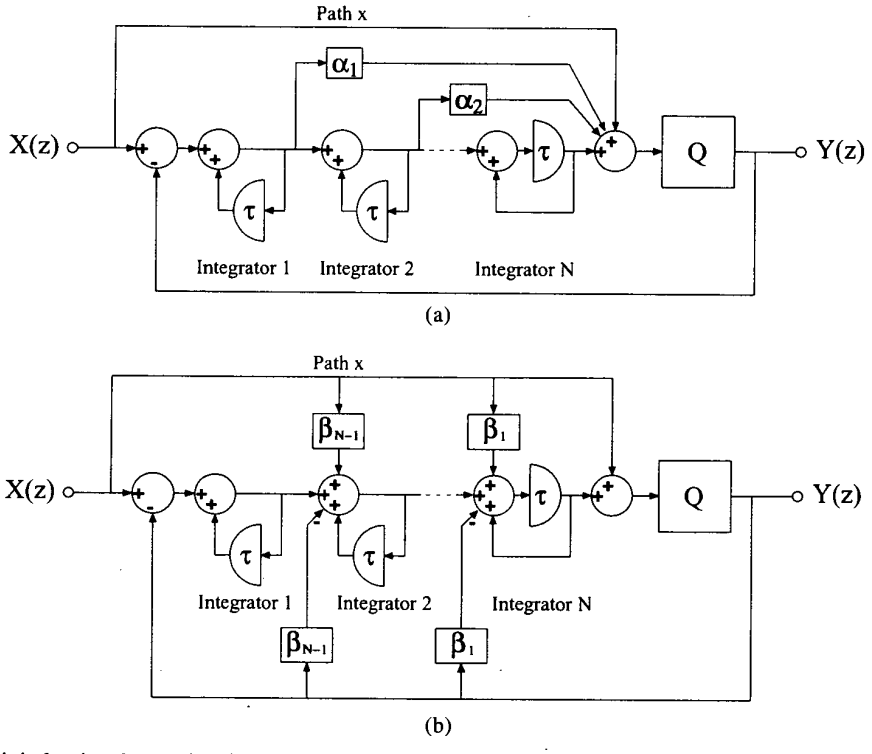


Fig. 2. Nth-order digital noise shaper showing two equivalent forms of stabilization. (a) Uses feedforward paths across integrators. (b) Uses equivalent multiple feedback paths; preferred for demonstrating equivalence with NDFL.

ward path, justifying the choice of $H_2(z)$. Figs. 3(b), 4(b), and 4(c) show the first- and second-order topologies redrawn in feedback form, which can be compared directly with a sigma–delta converter when Q is a two-level comparator. The order of the loop may be increased further by cascading additional integrators that are tailored at high frequency, for example, by using one of the techniques illustrated in Fig. 2. It follows from the earlier discussion that both these noise shapers have signal transfer functions of unity and that the quantization error (normally considered as noise) is shaped by a response $R(z)$, where

$$R(z) = (1 - z^{-1})^N . \quad (2)$$

Such noise-shaping structures have been much discussed in the literature as well [5].

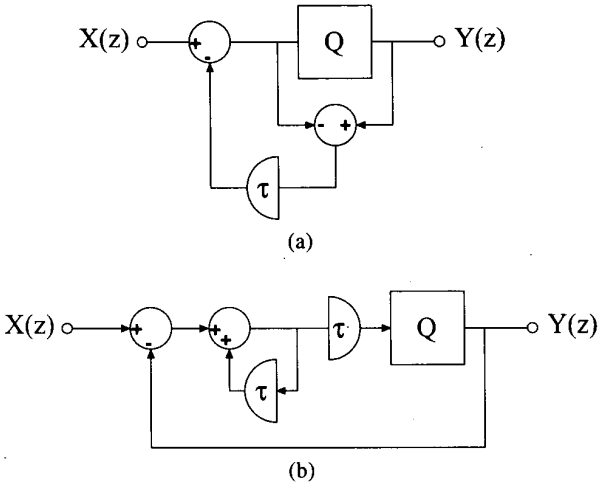


Fig. 3. Progression of equivalent circuits for first-order digital noise shaper.

2 DISTORTION SHAPING

Fig. 5 illustrates a further progression of ideas. Fig. 5(a) shows a third-order digital shaper for which the quantizer output-related error $q[z]$ is treated as an additive error, but where path x is omitted for simplicity, since with a third-order loop the signal transfer function within the audio band is virtually unity. Fig. 5(b) shows a similar system, but implemented in the analog domain that uses ideal continuous integrators, in which the quantizer is replaced by a continuous but nonlinear output stage. The analog circuit can be interpreted as a limiting digital case in which the sampling frequency has become

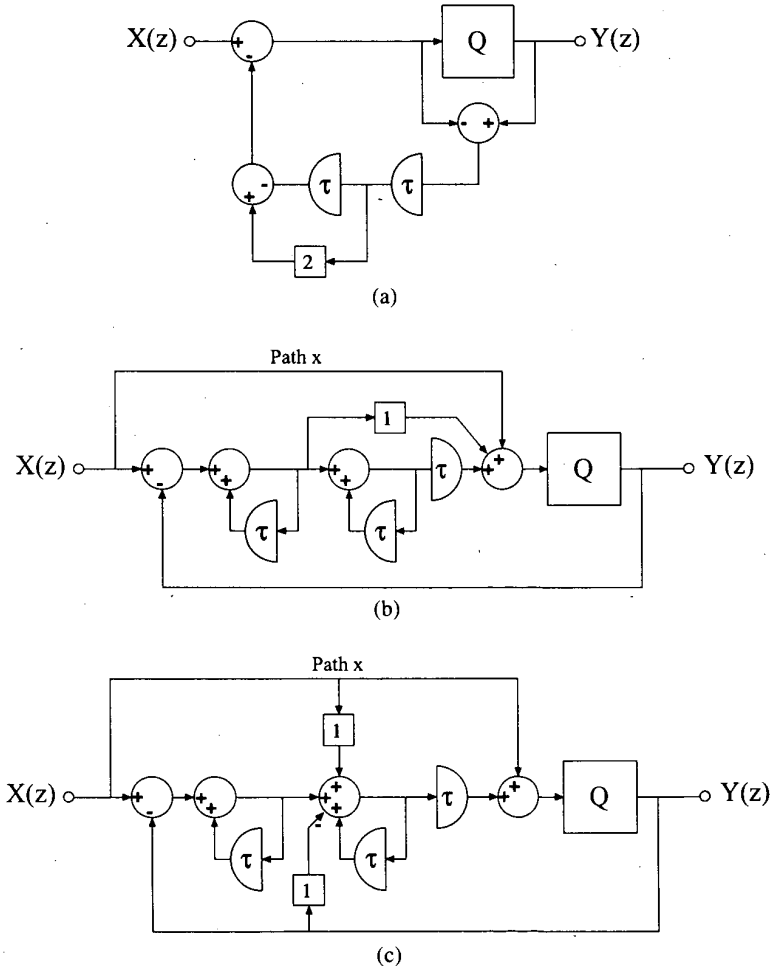


Fig. 4. (a) Progression of equivalent circuits for second-order digital noise shaper. (b) Feedforward compensation. (c) Multiple-feedback-path configuration as presented in Fig. 2.

infinite, hence removing all delays, and where the discontinuous output quantizer is replaced with a continuous but nonlinear analog output stage.

In the subsequent analysis the following notation is used to describe the function of the analog integrators. Consider integrator r, which receives an input $X_r(s)$ from the signal path and $Y_r(s)$ from the feedback path. Then the integrator output $I_r(s)$ is given by

$$I_r(s) = \frac{X_r(s)}{s\tau_{rs}} + \frac{Y_r(s)}{s\tau_{rf}}.$$

Applying this convention to the circuit in Fig. 5(c), where the integrators are shown labeled with respective time constants τ_{1f}, τ_{2f}, and τ_{3f} for the feedback paths and τ_{1s}, τ_{2s}, and τ_{3s} for the signal paths, and where the output stage has gain G, then the transfer function is

$$\left[\frac{s^3\tau_{1s}\tau_{2s}\tau_{3s}}{G} + \frac{s^2\tau_{1s}\tau_{2s}\tau_{3s}}{\tau_{3f}} + \frac{s\tau_{1s}\tau_{2s}}{\tau_{3f}} + \frac{\tau_{1s}}{\tau_{1f}}\right]Y(s)$$

$$= X(s) + \frac{s^3\tau_{1s}\tau_{2s}\tau_{3s}}{G}D(s) \qquad (3)$$

where $D(s)$ is the Laplace transform of $d(t)$, the additive error of the analog output stage.

It should be observed that in the context of a third-order loop, the τ parameters define both the stability

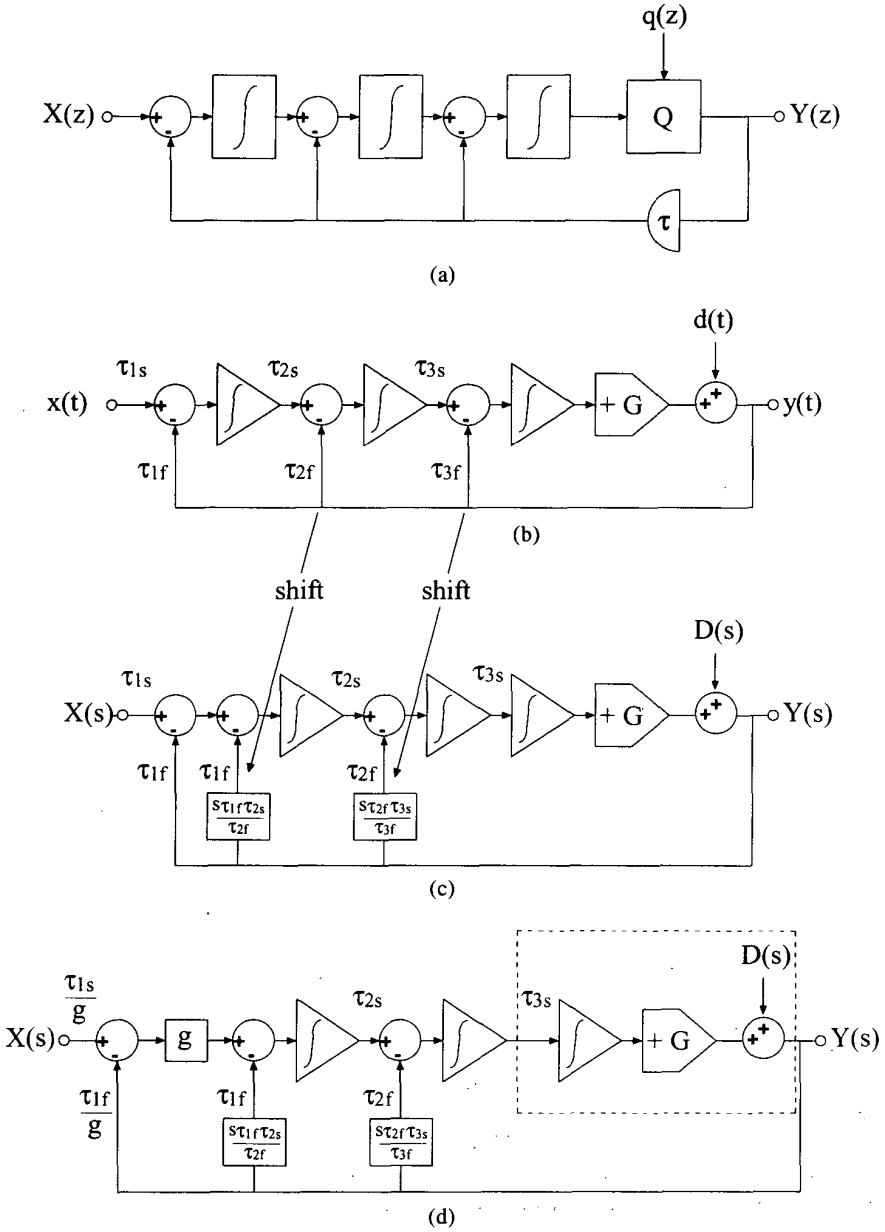


Fig. 5. Derivation of NDFL concept. (a) Digital third-order noise shaper. (b) Analog form. By moving feedback paths labeled τ_{2f} and τ_{3f} in (b) to the left, and adding compensating differentiators to maintain the same circuit transfer function, NDFL representations of (c) and (d) are achieved. (d) Dotted lines encompass what might be considered a realistic output stage of a normal amplifier.

criteria and the signal transfer function of the circuit, and that the cube of the signal frequency weights the output-stage error $D(s)$. Consequently this analog feedback circuit can be viewed as a distortion shaper, where the multiple integrators, by virtue of their large low-frequency gain, reduce the effects of nonlinearity at lower frequencies. The parallels between noise shapers and feedback amplifiers are evident.

The similarity with NDFLs is now demonstrated. In the circuit of Fig. 5(c) the two feedback paths applied to the inputs of the last two integrators are shifted to the left by one integrator stage, and compensating differentiators $s\tau_{2f}$ and $s\tau_{3f}$ are inserted to keep the transfer function unaltered. Observe that the output signal fed back directly to the input is related to $(1 + s\tau_{1f}\tau_{2s}/\tau_{2f})$. This has a similar function as the parallel resistor–capacitor phase-advance network typically used in the feedback path of an amplifier. Finally in Fig. 5(d) an extra amplifier with gain g has also been inserted to give the circuit a more usual configuration. It is this modified circuit that we wish to compare with the NDFL circuits of Cherry [1].

In this reconfiguration the output stage is modeled with frequency-independent gain G, but where the dashed box shown in Fig. 5(d) associates the third integrator with the output stage so that it can display real poles. It is also possible to think of each integrator as having a finite low-frequency gain of form $a/(1 + s\tau)$, more like the actual stages of an amplifier. The first integrator might be associated with an intermediate amplifier stage, the second integrator with the voltage-gain stage, and the third integrator and output block with a more realistic output stage.

3 DISCUSSION

The circuits of Fig. 5(c) and (d) can be identified directly with NDFL circuits, even though there may be differences of small detail, such as introducing zeros in the integrators at high frequency. The nested loops may not all take their feedback signal directly from the output, for example, although there is then a difference in implied circuit environment. In Fig. 5(b) each of the integrators (which in practice will have finite dc gain) is regarded as similar. However, in Fig. 5(c) and (d) the last integrator is considered to be part of a low-gain output stage describing, for example, a typical emitter–follower output stage of an audio amplifier. Although it is customary to include a Miller compensation capacitor across the voltage-gain stage, if this capacitor also encompasses the output stage, and feeds back to the virtual ground input of the voltage-gain stage, it becomes the $s\tau_{2f}$ differentiating feedback shown in the diagram. The effect on amplifier distortion is very beneficial, and this point has been emphasized by Cherry [2].

There are other strong parallels between distortion-shaping NDFLs and digital noise-shaping topologies. In principle by adding more stages, the order of each structure is increased, giving even more feedback at lower frequencies. Another aspect is the internal stability of the loop with respect to the output stage. It is the output stage that generally is considered the dominant source of the distortion, and a high loop gain will act to reduce it. The output-stage loop gain $A(s)$ in the circuit of Fig. 5(b) [and hence also of Fig. 5(c) and (d)] determines the distortion-shaping transfer function, where

$$\frac{Y(s)}{D(s)} = \frac{1}{1 + A(s)}.$$

Setting the input $X(s) = 0$, then $A(s)$ can be shown to be

$$A(s) = \frac{D(s)}{Y(s)} - 1 = G\left[\frac{1}{s\tau_{3f}} + \frac{1}{s^2\tau_{2f}\tau_{2s}} + \frac{1}{s^3\tau_{3f}\tau_{2s}\tau_{3s}}\right] \tag{4}$$

and this is easily generalized to higher or lower order. At the highest frequencies, only the first term survives, and for good stability the phase shift must be considerably less than $-180°$, ideally being close to $-90°$, the lag of a single integrator. It is the τ_{3f} feedback path that ensures this stability at high frequencies, but the other loops allow increasing feedback at lower frequencies, reducing distortion in the process.

As an example, let us consider a fifth-order distortion-shaper feedback circuit, for which $G = 1$. The signal transfer function $Y(s)/X(s)$ is by proper choice of τ parameters a fifth-order Butterworth unity-gain low-pass filter with cut-off frequency of, say, 50 kHz. The loop gain can be written as

$$A = \frac{s_n^5 + as_n^4 + (a + 2)s_n^3 + (a + 2)s_n^2 + as_n^1 + 1}{s_n^5} - 1$$

that is,

$$A = \frac{a}{s_n^1} + \frac{a + 2}{s_n^2} + \frac{a + 2}{s_n^3} + \frac{a}{s_n^4} + \frac{1}{s_n^5} \tag{5}$$

where

$$a = 1 + 2\cos\left(\frac{\pi}{5}\right) + 2\cos\left(\frac{2\pi}{5}\right)$$

and the normalized Laplace variable $s_n = s/\omega_0 = s/(2\pi f_0)$, and $f_0 = 50$ kHz. Fig. 6 is a plot of Eq. (5), showing the desired 6-dB per octave roll-off of the loop gain at high frequencies, but with a gain at lower frequencies proportional to f^{-5}. If the integrators in the circuit have finite dc gain, the graph is similar, but will limit near dc at some high value of gain, giving a loop gain for the output stage very similar to that indicated by Cherry [1]. In this example, unity gain occurs at 157 kHz, and the attendant phase shift is $-120°$, representing good stable behavior. Note that at 20 kHz there is already 40 dB of distortion reduction, rising in an ever-increasing way at lower frequencies.

For high-order loops such as discussed here, this system displays conditional stability, as pointed out by

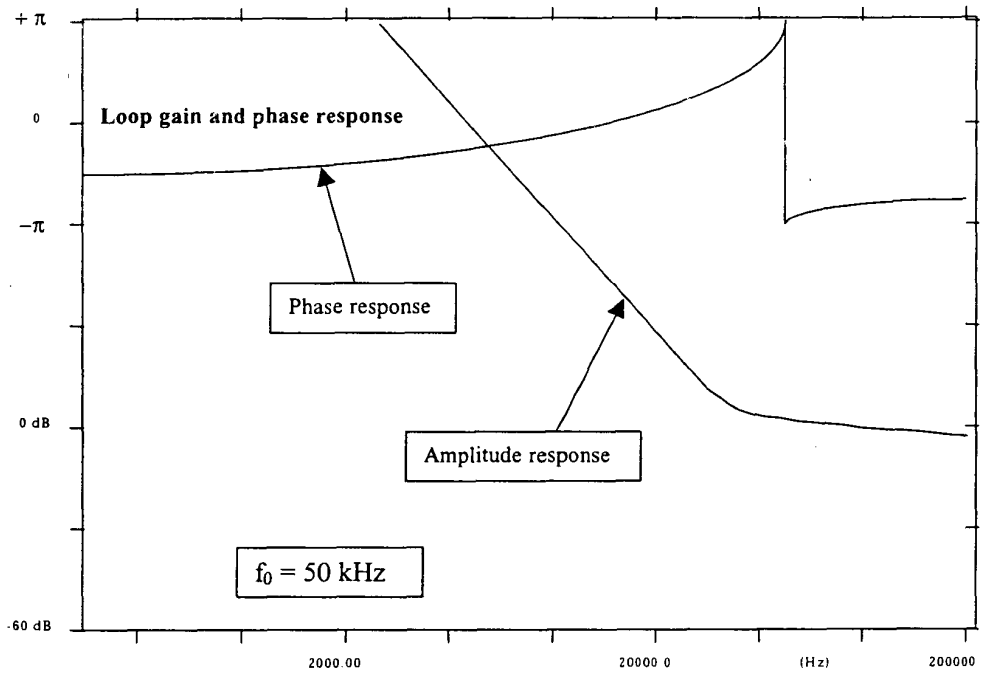


Fig. 6. Loop gain as seen by output stage of fifth-order distortion shaper having a transfer function of a fifth-order Butterworth low-pass filter with 50-kHz cut-off frequency. Rising loop gain at low frequencies vastly reduces distortion at these frequencies and effectively introduces distortion shaping.

Cherry [1] and in a later discussion of NDFLs [6]. Hence it is important to consider large-signal clipping behavior experimentally or by simulation to see whether the system can be provoked into self-oscillation or other bizarre or chaotic behavior. It is evident that there is no simple limit to the amount of distortion reduction at high orders, but increasing care must be taken in defining stable circuit parameters and behavior for large signals.

4 OVERVIEW

At first glance it almost seems trivial that the reassignment of the feedback loops around the integrators in Fig. 5(b) results in the NDFL structure of Fig. 5(c). However, there are differences in circuit realization and circuit tradition. Cherry [1] also works out a great deal of the mathematical aspects of NDFLs with sensitivities and the analysis of appropriate models. Presumably most of this work applies to the distortion-shaping topology as well, with appropriate measuring points or circuit associations in the two approaches. We do not in this engineering report work out such details or attempt a mapping between the two topologies.

In some ways the distortion-shaping approach is easier to grasp initially. But NDFLs are perhaps better if one is faced with a traditional class AB audio power amplifier and wishes to improve its performance. In fact this is suggested in the title of one of Cherry's papers [7]. Also, when the NDFL was invented, it appeared as a new result since the traditional methods of stabilizing amplifiers had limitations on the amount of simple feedback that could be applied, as Bode [4] had shown. There have been engineers who have employed selected aspects of a differentiating loop all along, but the general concept of the NDFL puts a firm footing on new aspects of feedback that relate particularly to analog amplifiers. However, the equivalence is important at a conceptual level, especially as the means of stabilizing high-order digital loops was known [8] at the time of invention of NDFL [1]. What this engineering report shows is that there is another way of looking at the application of large amounts of negative feedback, and that the two are closely related.

5 REFERENCES

[1] E. M. Cherry, "A New Result in Negative-Feedback Theory and Its Application to Audio Power Amplifiers," *IEEE J. Circuit Theory Appl.*, vol. 6, pp. 265–288 (1978 July).

[2] E. M. Cherry, "Nested Differentiating Feedback Loops in Simple Audio Power Amplifiers," *J. Audio Eng. Soc.*, vol. 30, pp. 295–305 (1982 May).

[3] M. O. J. Hawksford, "Chaos, Oversampling, and Noise Shaping in Digital-to-Analog Conversion," *J. Audio Eng. Soc.*, vol. 37, pp. 980–1001 (1989 Dec.).

[4] H. W. Bode, "Network Analysis and Feedback Amplifier Design (van Nostrand, Princeton, NJ, 1945).

[5] M. W. Hauser, "Principles of Oversampling A/D Conversion," *J. Audio Eng. Soc.*, vol. 39, pp. 3–26 (1991 Jan./Feb.).

[6] J. Scott and G. Spears, "On the Advantages of Nested Feedback Loops," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 39, pp. 140–145 (1991 Mar.); E. M. Cherry, Discussion, ibid., pp. 145–147.

[7] E. M. Cherry, "A Power Amplifier 'Improver,'"

J. Audio Eng. Soc. (Engineering Reports), vol. 29, pp. 140–147 (1981 Mar.).

[8] S. K. Tewksbury and R. W. Hallock, "Over-sampled Linear Predictive and Noise Shaping Coders of Order $N > 1$," *IEEE Trans. Circuits Sys.*, vol. CAS-25, pp. 437–447 (1978 July).

THE AUTHORS

John Vanderkooy received a B.Eng. degree in engineering physics in 1963 from McMaster University in Hamilton, Ontario, Canada. He continued studies there in low-temperature physics of metals, receiving a Ph.D. in 1967. He spent two years as a postdoctoral fellow at the University of Cambridge in England, and returned to Canada in 1969 to join the faculty at the University of Waterloo. For some years he followed his doctoral interests in magnetic properties of electrons in metals, but his research interests have slowly shifted since the late 1970s to audio and electroacoustics. He is currently a professor of physics at the University of Waterloo.

Dr. Vanderkooy is a fellow of the Audio Engineering Society, a recipient of its Silver Medal, and several Publication Awards. He has contributed a wide variety of papers at AES conventions and to the *Journal* in such areas as loudspeaker crossover design, electroacoustic measurement techniques, dithered quantization, digital signal processing, loudspeaker impedance, acoustics, and diffraction. He is a founding member of the Audio Research Group at the University of Waterloo, working together with his colleague Stanley Lipshitz and a number of graduate students. Dr. Vanderkooy's current research interests are digital audio signal processing, dithered quantization, transducers, acoustic diffraction from edges, stochastic resonance, and loudspeaker ports.

The biography for M. O. J. Hawksford was published in the 1999 September issue.

PAPERS

Low-Distortion Programmable Gain Cell Using Current-Steering Cascode Topology*

MALCOLM JOHN HAWKSFORD

University of Essex, Department of Electrical Engineering Science, Colchester, Essex, United Kingdom

A programmable gain cell is described which operates in a current-steering mode. The technique is shown to offer good linearity and a precisely defined maximum gain which exhibits near zero distortion. The cellular structure is presented together with an application circuit using error feedforward distortion correction within the input stage, while similar techniques realize a precision current mirror but configured using error feedback.

0 INTRODUCTION

The traditional circuitry that can be identified as the kernel of most analog programmable gain amplifiers (PGA) uses the translinear gain cell first described by Gilbert [1], with further derivatives reported in [2]–[6]. This basic cell is illustrated in Fig. 1.

In elementary form the cell is a differential current in (i_1), differential current out (i_2) structure that requires all transistors to have the same parameters (matched physically and thermally). If we assume

$$I_e = I_0 \, e^{qV_{be}/KT} \tag{1}$$

and since by Kirchhoff's law

$$V_{be1} - V_{be2} = V_{be3} - V_{be4}$$

then

$$\ln\left[\frac{I_1 + i_1}{I_1 - i_1}\right] = \ln\left[\frac{I_2 + i_2}{I_2 - i_2}\right]$$

which yields

$$\frac{i_1}{i_2} = \frac{I_1}{I_2} . \tag{2}$$

Eq. (2) suggests a seemingly ideal solution where a linear relationship exists between i_1 and i_2 and the gain is determined by the ratio of the bias currents I_1 and I_2.

The maximum gain setting is not specified by Eq. (2), and in fact will ultimately depend upon the device adherence to the exponential relationship specified by Eq. (1). That is, once there is deviation from Eq. (1), nonlinearity is introduced. Also unity gain ($I_1 = I_2$) is dependent upon device matching and as such is specified only as accurately as the transistors are matched.

The cell is also indirect. The signal current i_2 is derived by using T_3 and T_4 as a differential amplifier with T_1 and T_2 presented as nonlinear load resistors to the respective input currents $I_1 + i_1$ and $I_2 - i_2$. The nonlinear distortion is then minimized by matching the nonlinear device transfer characteristics.

The gain cell to be described in this paper uses a current-steering topology. It effectively eliminates a stage of amplification compared with the cell in Fig. 1 and most important, it has a well-defined upper unity gain (current in–current out) which is essentially linear and independent of device characteristics and matching—a most useful attribute for an audio channel.

This paper describes the theoretical basis of the current-steering gain cell and suggests an outline system topology that should enable high-performance PGAs to be designed. In addition to the main cell, a precision linear current mirror is described posessing wide dynamic range. The current mirror is used here within the PGA. However, it should also find application in mainstream constant-gain preamplifiers and power amplifier designs. Finally, the input circuitry proposed for the PGA includes error-correction feedforward to enhance transconductance, improve linearity, and minimize Johnson noise.

1 CURRENT-STEERING GAIN CELL

The elementary current-steering gain cell is illustrated in Fig. 2. The cell consists of two pairs of matched transistors that should adhere to the exponential I_e/V_{be} relationship defined by Eq. (1). Justification of the quiescent current distribution is given in the Appendix.

In Fig. 2, i_1 is the input signal current, $|i_1| < |I|$, i_2 and i_3 are output signal currents, I is bias current, m is a

* Presented at the 69th Convention of the Audio Engineering Society, Los Angeles, 1981 May 12–15.

 0004-7554/82/110795-05$00.75

gain-control parameter, $-1 < m < 1$, and V_k is a constant bias voltage.

1.1 Analysis

Applying Kirchhoff's law we obtain

$$V_k - V_{be1} + V_{be2} - V_{be3} + V_{be4} = V_k .$$

Therefore

$$(V_{be1} - V_{be2}) + (V_{be3} - V_{be4}) = 0 .$$

Using Eq. (1), we observe

$$V_{be1} - V_{be2} = \frac{KT}{q}\ln\left[\frac{I_{c1}}{I_0}\right] - \frac{KT}{q}\ln\left[\frac{I_{c2}}{I_0}\right] .$$

Therefore

$$V_{be1} - V_{be2} = \frac{KT}{q}\ln\left[\frac{I(1+m)/2 - i_2}{I(1+m)/2 + i_2}\right] .$$

Similarly for V_{be3} and V_{be4}:

$$V_{be3} - V_{be4} = \frac{KT}{q}\ln\left[\frac{I(1-m)/2 + i_3}{I(1-m)/2 - i_3}\right] .$$

Thus

$$\frac{I(1+m)/2 - i_2}{I(1+m)/2 + i_2} = \frac{I(1-m)/2 - i_3}{I(1-m)/2 + i_3}$$

where, after simplification,

$$\frac{i_2}{i_3} = \frac{1+m}{1-m} . \tag{3}$$

We also observe from Fig. 2 (where i_1 is the input current and i_2 and i_3 are output currents)

$$I + i_1 = \left(I\frac{1+m}{2} + i_2\right) + \left(I\frac{1-m}{2} + i_3\right) .$$

Therefore

$$i_1 = i_2 + i_3 . \tag{4}$$

Eqs. (3) and (4) reveal the operation of the current-steering PGA. The sum of the output currents is exactly equal to the input signal current (neglecting small base currents) and is independent of the I_c/V_{be} characteristics. Also the current division is linearly related to the gain parameter m. From Eqs. (3) and (4), i_2 and i_3 are derived:

$$i_2 = \frac{1+m}{2} i_1 \tag{5}$$

$$i_3 = \frac{1-m}{2} i_1 . \tag{6}$$

Thus the current gain of each cell is additionally complementary (that is, their sum is unity).

Eqs. (5) and (6) show that when $m = 1$ (or -1), then $i_2 = i_1$ (or $i_3 = i_1$), and on investigation Fig. 2 reveals that the input current is then steered through T_1 and T_2 (or T_3 and T_4) such that device nonlinearity has no effect, as the transistor pair acts as a cascode, assuming V_k is constant.

It is this latter mode of operation that is of greatest significance, since unity current gain results with excellent linearity due to the grounded base operation of transistors T_1 and T_2 or T_3 and T_4.

However, one problem still remains: the rejection of gain-control current under dynamic operation. The method by which control signal breakthrough is minimized is through the use of two techniques, a differential current mirror and two additionally complementary gain cells driven from a differential input stage. These are the subjects of Sections 2 and 4.

2 GENERALIZED PGA TOPOLOGY

This section describes the design of a basic PGA that uses two current-steering gain cells and two precision current mirrors. The circuitry is illustrated in the schematic of Fig. 3.

The input stage consists of a long-tail pair circuit with local emitter degeneration formed by R_1. Since a nonlinear fraction of the input signal $(V_1 - V_2)$ is developed across the two base–emitter junctions of the long-tail pair, two further long-tail pair stages are introduced to measure these error voltages and add corrective currents to each of the two output currents. Using feedforward error correction, greatly enhanced input stage performance with a significant reduction of nonlinearity in the transconductance transfer characteristic is achieved.

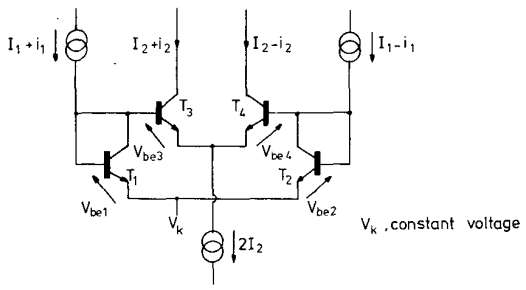


Fig. 1. Basic translinear gain cell.

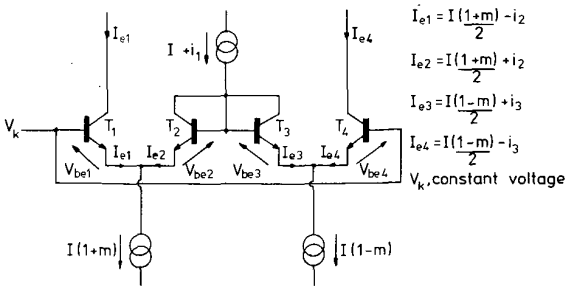


Fig. 2. Elementary current-steering gain cell.

2.1 Analysis

Referring to the input stage shown in Fig. 3,

$$V_1 - V_2 = i_1 R_1 + (V_{be1} - V_{be2})$$

where V_1 and V_2 are input signals.

Applying Eq. (1) to each error amplifier in Fig. 3,

$$V_{be1} = i_2 R_2 + \frac{KT}{q} \ln\left[\frac{I_2 + i_2}{I_2 - i_2}\right] + I_2 R_2$$

$$V_{be2} = -i_3 R_2 + \frac{KT}{q} \ln\left[\frac{I_2 - i_3}{I_2 + i_3}\right] + I_2 R_2$$

therefore,

$$V_1 - V_2 = i_1 R_1 + (i_2 + i_3) R_2 + \frac{KT}{q} \ln\left[\left(\frac{I_2 + i_2}{I_2 - i_2}\right)\left(\frac{I_2 + i_3}{I_2 - i_3}\right)\right] .$$

Assuming $i_2, i_3 < i_1$ such that $i_2, i_3 \ll I_2$,

$$V_1 - V_2 \simeq \left[i_1 + \left(\frac{i_2 + i_3}{R_1}\right)\left(R_2 + \frac{2KT}{qI_2}\right)\right] R_1 .$$

However, the output signal current i_0 is formed (see Fig. 3) by

$$i_0 = i_1 + i_2 + i_3 .$$

Hence if

$$R_1 = R_2 + 2\frac{KT}{qI_2} \tag{7}$$

then

$$i_0 = \frac{V_1 - V_2}{R_1} . \tag{8}$$

Since ΔV_{be1} and $\Delta V_{be2} < V_{in}/2$, then the error amplifiers operate well within their linear region. Thus provided that R_1 is selected according to Eq. (7), compensation for V_{be1} and V_{be2} is achieved, which results in the transconductance exhibiting excellent linearity with signal current.

The differential currents derived from the input stage provide signal currents for two current-steering gain cells, which in turn are both biased by identical gain-control currents. The outputs of the two cells are suitably crosscoupled, and when combined with two precision unity-gain current mirrors, produce additionally complementary output currents I_{01} and I_{02}. To aid understanding of this process, the circuit diagram shown in Fig. 3 should be observed with respect to the labeled currents. Under quiescent conditions $I_{x1} = I_{x2}$ and $I_{y1} = I_{y2}$, hence output currents I_{01} and I_{02} result, which are essentially independent of the gain-control currents.

The actual output currents generated when an input signal is applied are given by

$$I_{01} = I_k + \frac{V_1 - V_2}{R_1}(1 + m) \tag{9}$$

$$I_{02} = I_k - \frac{V_1 - V_2}{R_1}(1 - m) \tag{10}$$

where these currents can be converted to voltages by using suitably chosen load resistors R_L, as shown in Fig. 3.

3 LINEARIZATION USING NEGATIVE FEEDBACK

Although adequate performance can be achieved with the system shown in Fig. 3, it is possible to use the result that $I_{01} + I_{02}$ is independent of m. Thus an output voltage can be derived which is independent of the gain selected and used as a feedback signal that is returned to the input stage. Consequently the current-steering gain cells need operate only with small signal currents, which enhances linearity. The open-loop gain

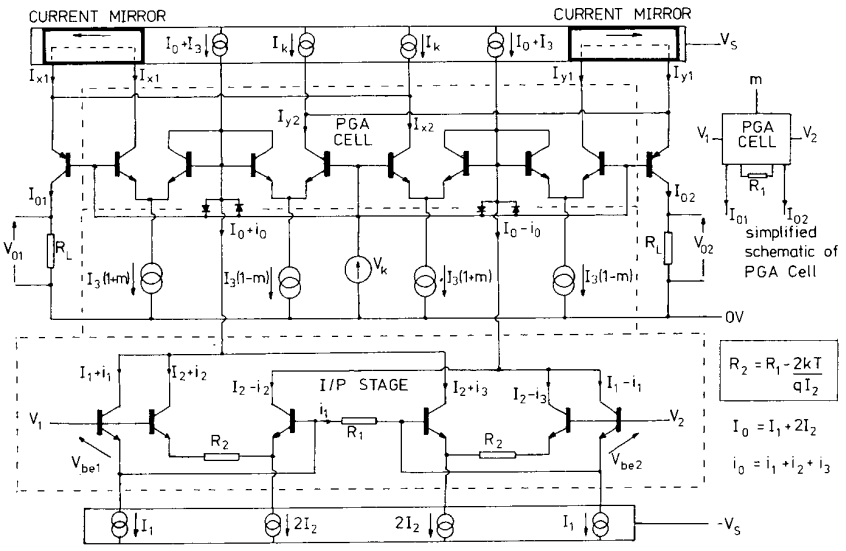


Fig. 3. Differential current-steering gain cells. (I/P stage example using error feedforward correction.)

is determined by selecting R_1 and R_5, while R_3 and R_4 determine the feedback factor. A basic feedback schematic is illustrated in Fig. 4, where the PGA cell is the circuit of Fig. 3. The output voltages V_{01} and V_{02} are then given by

$$V_{01} = V_1(1 + m)\frac{R_5(R_3 + R_4)}{R_1R_3 + R_1R_4 + R_4R_5} \quad (11)$$

$$V_{02} = V_1(1 - m)\frac{R_5(R_3 + R_4)}{R_1R_3 + R_1R_4 + R_4R_5} \quad (12)$$

4 CURRENT MIRROR

The two current mirrors in Fig. 3 can use an error correction feedback scheme to enhance linearity and minimize dependence upon the I_e/V_{be} characteristics. Fig. 5 illustrates a basic current mirror often used in discrete amplifier design, while Fig. 6 shows the enhanced design. Integrated current mirrors could be used, but they generally exhibit poor current gain linearity at currents in excess of a few milliamperes, especially at the extremes of their transfer characteristics.

4.1 Analysis

Since the bases of T_1 and T_2 are at the same potential, by Kirchhoff's law,

$$V_{be1} + I_1R + (I_1 + I_0 - i)R = V_{be2} + I_2R + (I_2 + I_0 + i)R .$$

Therefore

$$(V_{be1} - V_{be2}) + 2I_1R = 2iR + 2I_2R .$$

But within the error amplifier T_3, T_4,

$$V_{be1} - V_{be2} = (V_{be3} - V_{be4}) + iR_x$$

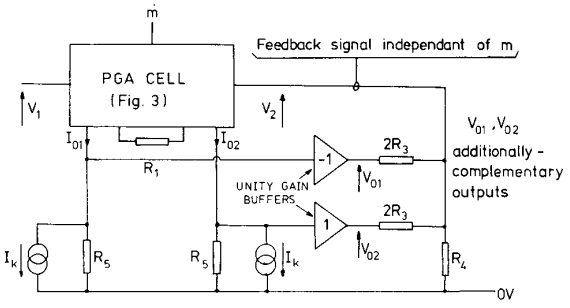


Fig. 4. Basic scheme for implementing overall negative feedback.

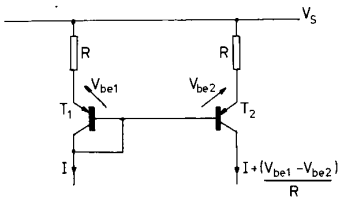


Fig. 5. Basic current mirror.

where

$$V_{be3} - V_{be4} = \frac{KT}{q}\ln\left[\frac{I_0 + i}{I_0 - i}\right] \simeq \left(\frac{2KT}{qI_0}\right)i, \quad \text{for modest } i .$$

Therefore

$$\left(\frac{2KT}{qI_0} + R_x\right)i + 2I_1R = 2iR + 2I_2R.$$

If

$$R_x = 2R - \frac{2KT}{qI_0} \quad (13)$$

then

$$I_2 = I_1 . \quad (14)$$

Since the error voltage $V_{be1} - V_{be2} \simeq 0$ due to T_1 and T_2 operating virtually with the same emitter current, the error amplifier is rendered linear. Consequently the correction system results in exceptional linearity over a wide range of the current transfer characteristic of the current mirror.

5 CONCLUSIONS

A topology for a programmable gain amplifier has been presented which can exhibit excellent gain stability and linearity. The cell offers the advantage of a precisely defined upper gain limit with the advantage that distortion is almost zero over a wide dynamic range.

The circuit requires transistor matching, but pair matching should be adequate, provided that transistors adhere closely to the logarithmic relationship between I_e and V_{be} (such as the LM394 at $I_c < 1$ mA).

An application circuit is presented which provides additionally-complementary gains, a topology which allows overall negative feedback to be used to further enhance linearity. However, the open-loop characteristics should yield adequate performance and may be preferred by designers of the low feedback school.

Additional to the basic cell, an input transconduc-

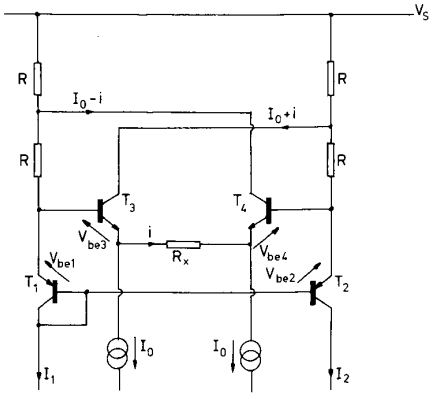


Fig. 6. Enhanced current mirror using error-correction feedback.

tance stage is described which uses error-correction feedforward to enhance linearity. Also a precision current mirror is presented which uses similar error correction, but configured within a feedback loop. These circuits offer excellent linearity, which is virtually independent of device characteristics.

6 REFERENCES

[1] B. Gilbert, "Translinear Circuits: A Proposed Classification," *IEE Electron. Lett.*, vol. 11, pp. 14–16 (1975 Jan. 9).

[2] B. Gilbert, "A New Wideband Amplifier Technique," *IEEE J. Solid-State Circuits*, vol. SC-3, pp. 353–365 (1968 Dec.).

[3] C. C. Todd, "A Monolithic Analog Compandor," *IEEE J. Solid-State Circuits*, vol. SC-11, pp. 754–762 (1976 Dec.).

[4] T. Yamaguchi, S. Takaoka, and K. Aizawa, "A New Configuration Using a Voltage Controlled Amplifier for a Dolby System Integrated Circuit," *IEEE Trans. Consumer Electron.*, vol. CE-25, pp. 723–729 (1979 Nov.).

[5] D. Baskind and H. Rubens, "Techniques for the Realization and Applications of Voltage-Controlled Amplifiers and Attenuators," presented at the 60th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 26, p. 572 (1978 July/Aug.), preprint no. 1378.

[6] D. Baskind, H. Rubens, and G. Kelson, "The Design and Integration of a High-Performance Voltage-Controlled Attenuator," presented at the 64th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 27, pp. 1018, 1020 (1979 Dec.), preprint no. 1555.

7 APPENDIX
QUIESCENT CURRENT DIVISION IN CURRENT-STEERED PGA CELLS

The current distribution shown in Fig. 2 is stated without formal justification. Here we show the current divisions to be valid. Since the circuit is considered in the quiescent state, i_1, i_2, i_3, and $i_4 = 0$. Let emitter currents of T_1, T_2, T_3, and T_4 be I_{e1}, I_{e2}, I_{e3}, and I_{e4}. Apply Kirchhoff's law to current distribution in Fig. 2:

$$I_{e1} + I_{e2} = I(1 + m) \tag{15}$$

$$I_{e3} + I_{e4} = I(1 - m) \tag{16}$$

$$I_{e2} + I_{e3} = I \; . \tag{17}$$

Apply Kirchhoff's law to the base-emitter voltages:

$$(V_{be1} - V_{be2}) + (V_{be3} - V_{be4}) = 0 \; .$$

Using Eq. (1),

$$\frac{KT}{q}\ln\left[\frac{I_{e1}}{I_{e2}}\right] + \frac{KT}{q}\ln\left[\frac{I_{e3}}{I_{e4}}\right] = 0 \; .$$

Hence

$$\frac{I_{e1}}{I_{e2}} = \frac{I_{e4}}{I_{e3}} = \lambda \; . \tag{18}$$

Substituting for λ in Eqs. (15) and (16),

$$I_{e2} = I\left(\frac{1 + m}{1 + \lambda}\right)$$

$$I_{e3} = I\left(\frac{1 - m}{1 + \lambda}\right) \; .$$

Hence from Eq. (17), $\lambda = 1$. Therefore

$$I_{e1} = I_{e2} \quad \text{and} \quad I_{e4} = I_{e3} \; .$$

We therefore derive I_{e1}, I_{e2}, I_{e3}, and I_{e4} from Eqs. (15) and (16):

$$I_{e1} = I\left(\frac{1 + m}{2}\right)$$

$$I_{e2} = I\left(\frac{1 + m}{2}\right)$$

$$I_{e3} = I\left(\frac{1 - m}{2}\right)$$

$$I_{e4} = I\left(\frac{1 - m}{2}\right)$$

The division of emitter currents is therefore as shown in Fig. 2. The output collector currents of T_1 and T_4 are almost equal to the emitter currents, provided that transistors have adequate current gain. Also near linearity between I_c and I_b (for modest signals) enhances linearity of the grounded base stage.

THE AUTHOR

Malcolm Hawksford was educated at the University of Aston in Birmingham, England, from 1965 to 1971. In 1968 he obtained a first class honors degree in electrical engineering, and that same year was awarded a BBC research scholarship to investigate the application of deltamodulation to color television. In 1972 he obtained a Ph.D. degree. In 1971 Dr. Hawksford became a lecturer at the University of Essex, England, in the Department of Electrical Engineering Science. During his time at Essex he actively pursued research projects within the field of audio engineering, where projects on power amplifier design, loudspeaker crossover design, analogue to digital conversion and music synthesis have been undertaken. He has presented papers at conventions of the Audio Engineering Society. He is currently a member of the AES, IEE and RTS.

ENGINEERING REPORTS

Topological Enhancements of Translinear Two-Quadrant Gain Cells*

MALCOLM O. J. HAWKSFORD

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, CO4 3SQ, UK

AND

P. G. L. MILLS

Tannoy Limited, Coatbridge, Strathclyde ML5 4TF, UK

A method is proposed for extending the performance regime of two-quadrant translinear gain cells that can both increase the gain-control range and lower distortion, particularly at extremes of attenuation. The general operating principles of translinear cells are reviewed and a laboratory design for an enhancement of the current-steering cell is presented. The design is supported by a range of measurements to validate the technique.

0 INTRODUCTION

The two-quadrant multiplier configured as a voltage-controlled amplifier (VCA) has found wide application as a dynamic gain element in audio systems, both for gain control and for program-controlled equalization. Consequently VCAs have been widely discussed in the literature [1]–[12]. In this paper the basic translinear, bipolar transistor gain cells are reviewed and an enhanced topology for the current-steering gain cell [12] is proposed, together with a prototype circuit and supporting performance data.

To realize gain control the audio circuit designer has available a range of methods that include motorized potentiometers, multiplying digital-to-analog converters (MDAC), FET attenuators, light-controlled resistors, pulse-modulation and switched-gain systems, and bipolar transistor arrays. However, this paper limits discussion to the bipolar array for implementing translinear gain elements as these are readily fabricated in integrated form and are compatible with large-scale analog systems.

Of particular importance when designing a VCA is the need to achieve an adequate performance regime over a wide range of the system's dynamic characteristic. A deficiency evident in several designs is the nature of distortion as a function of signal level and attenuation where, specifically at high attenuation, signal-to-distortion ratios deteriorate rapidly. This results from an inherent mechanism that yields a distortion residual at high attenuation in the output signal whose level with respect to the input remains approximately constant. Thus (at high attenuation) the distortion can be considerable. The enhanced topology presented directly addresses this problem and achieves a falling distortion with increased (high) attenuation.

To commence the study, the basic structures of translinear bipolar VCA cells are reviewed for tutorial value, and an approximate distortion analysis is presented that is directed at the VCA at high attenuation.

* Manuscript received 1988 May 4.

1 CLASSIFICATION OF TRANSLINEAR GAIN CELLS

The operation of a translinear cell is based on the logarithmic relationship between emitter current I_E and base–emitter voltage V_{BE} of a bipolar transistor, where

$$V_{BE} = \frac{kT}{q} \ln \left[\frac{I_E}{I_0}\right] \quad (1)$$

where

k = Boltzmann's constant
T = temperature, kelvin
q = charge on electron, coulomb
I_0 = saturation current, ampere.

Of particular concern is the temperature dependence of the transistor, both as a linear function of T and through the saturation current I_0, the latter being extremely temperature sensitive. To overcome this problem, the classic solution [1] is to use a transistor array of nearly identical devices operating under isothermal conditions, where a minimum of two transistors are required to compensate for I_0 and four transistors to compensate for T [see Eq. (1)].

Most of the VCA circuits are direct descendants of either the Gilbert cell or the log/antilog topologies as, for example, discussed in [6]. The dbx[1] VCA [8], which introduced the NPN/PNP array as a complementary cell, is also a direct descendant of the log/antilog configuration.

1.1 Two-Transistor Cells and Descendants

The basic two-transistor cell is illustrated in Fig. 1, where the operational amplifiers IC_1 and IC_2 suspend the transistor cell T_1, T_2 in a well-defined bias environment. Assuming IC_1 and IC_2 are ideal and ignoring base currents, then the operating conditions are

$$I_{E1} = V_i/R$$

$$I_{E2} = V_o/R$$

$$V_{CB1} = 0 \text{ V}$$

$$V_{CB2} = -V_g$$

where V_g is the gain-control voltage and V_{CB1} and V_{CB2} are the collector base voltages of T_1 and T_2, respectively.

Analysis

$$V_{BE1} - V_{BE2} = -V_g \ . \quad (2)$$

Assuming ideal behavior as stated in Eq. (1), then

$$\frac{kT}{q} \ln \left[\frac{I_{E1}}{I_{E2}}\right] = -V_g \quad (3)$$

where the dependence on saturation currents is eliminated provided T_1 and T_2 are matched and isothermal.

Hence eliminating I_{E1} and I_{E2} and rearranging,

$$\frac{V_o}{V_i} = e^{qV_g/kT} \ . \quad (4)$$

Eq. (4) shows the voltage gain of the VCA in Fig. 1 to be related to the gain-control voltage V_g through an exponential relationship; the dependence upon T should also be noted. Unfortunately, to operate correctly, the cell requires a value of $V_i > 0$ to prebias T_1 and T_2 to their active region, and this implies feedthrough of the gain-control function to the output, an undesirable characteristic for a two-quadrant cell.

To overcome this problem, two solutions have been proposed. One method consists of using two identical cells, as those in Fig. 1, but with a differential drive, as shown in Fig. 2,

$$V_{o1} = \left(V_i + \frac{v_i}{2}\right) e^{qV_g/kT}$$

$$V_{o2} = \left(V_i - \frac{v_i}{2}\right) e^{qV_g/kT} \quad (5)$$

$$V_o = V_{o1} - V_{o2} = v_i \, e^{qV_g/kT}$$

where

V_i = dc bias to bring cell into active region
v_i = signal component
V_{o1}, V_{o2} = respective outputs of two stages.

V_g also drives the noninverting inputs of the output operational amplifiers.

The second method, introduced in the dbx [8] cell, is to use a complementary four-transistor configuration, where the upper PNP transistors are also controlled by V_g, as shown in Fig. 3. Provided T_3 and T_4 have similar characteristics as T_1 and T_2 and isothermal conditions prevail, then for $V_i = 0$ V, $I_{C3} = I_{C1}$ and $I_{C4} = I_{C2}$. Thus gain-control feedthrough is compensated over the

[1] dbx is a trademark.

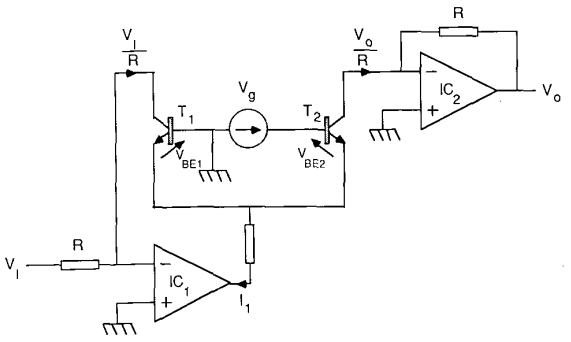


Fig. 1. Basic two-transistor log/antilog cell.

range of V_g. Of course, having decided to use a second pair of transistors, there is no reason why these cannot participate in the gain-control function, thus enhancing noise performance and dynamic range where, effectively, transistor pairs T_1/T_3 and T_2/T_4 appear in parallel. In Figs. 4 and 5 two variations are shown, one using an amplified diode while the second employs a floating supply such as a lead acid cell, a solar cell, or a switched-mode supply to eliminate bias current noise by not returning this current to ground. This latter circuit appears to offer much potential for low-noise operation, particularly where operating current levels are raised by the parallel connection of several transistor arrays. The insensitivity of gain to bias current level should also be noted.

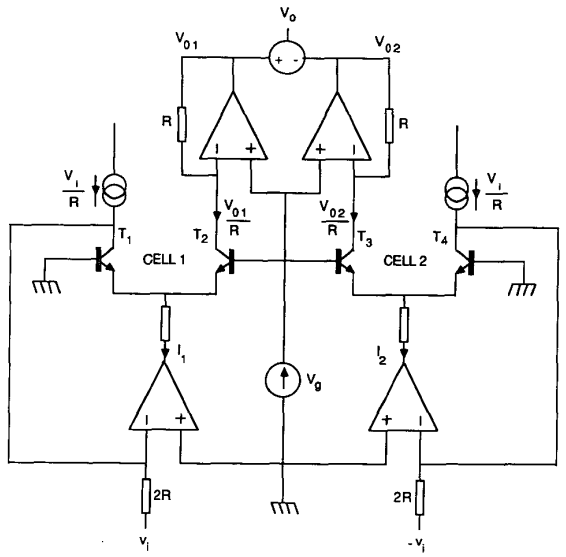


Fig. 2. Differential configuration of two-transistor log/antilog cell with bias.

Analysis of the Fig. 1 VCA revealed a linear relationship between emitter currents I_{E1} and I_{E2} where, from Eq. (3),

$$\frac{I_{E2}}{I_{E1}} = e^{qV_g/kT} .$$

Hence the tail current of T_1 and T_2 is also a linear function of I_{E1} and I_{E2} where, since $I_1 = I_{E1} + I_{E2}$, then

$$I_{E1} = I_1 \left(1 + e^{qV_g/kT}\right)^{-1} . \quad (6)$$

Noting the form of emitter currents I_{E1} and I_{E4} from Eq. 6, the voltage gain follows directly with reference to Fig. 6 as

$$\frac{v_o}{v_i} = 2g_m R \left(1 + e^{qV_g/kT}\right)^{-1} . \quad (7)$$

Consequently the Fig. 2 topology can be modified, whereby the tail currents form a differential input signal and the collector currents a differential output signal,

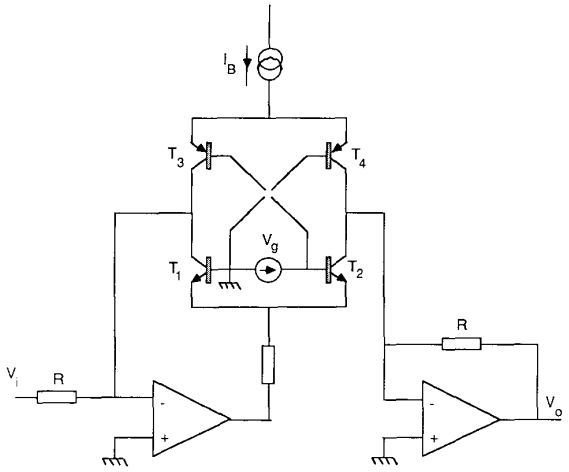


Fig. 3. Basic dbx voltage-controlled amplifier.

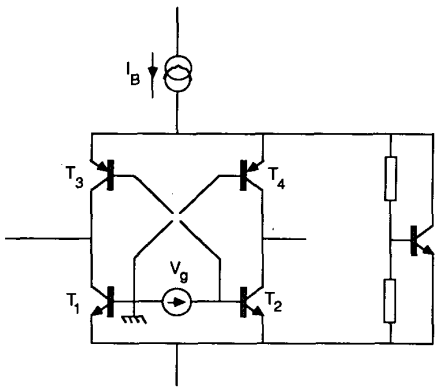


Fig. 4. dbx cell enhanced with amplified diode.

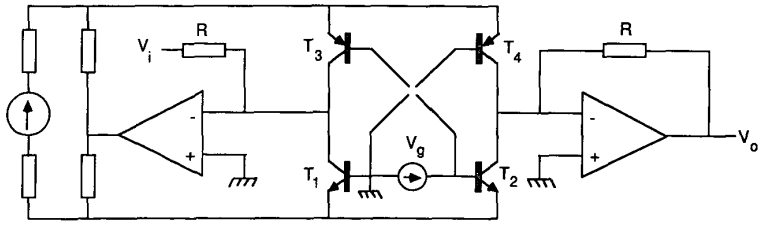


Fig. 5. Symmetrical dbx cell with floating supply.

as shown in Fig. 6 [see also Eq. (9)]. An important refinement of this reconfiguration is that for maximum gain T_1 and T_4 are now in common base (with T_2 and T_3 off); thus distortion is negligible.

1.2 Four-Transistor Cells

The two-transistor cells (and four-transistor derivatives) were shown to offer an exponential gain law and to retain a degree of temperature dependence. However, the use of four-transistor symmetrical circuits where the base–emitter voltages sum to zero can eliminate the temperature dependence and yield a linear gain law characteristic, where the desired gain law can be configured using nonlinear shaping circuits. The classic translinear circuit is the Gilbert multiplier [1] as shown in Fig. 7. Effectively, it is a differential amplifier where the base-to-base input voltage is predistorted by differentially driven diodes. However, operation is best understood by noting that the base–emitter voltages sum to zero and then applying Eq. (1) to each transistor whereby, assuming matching and isothermal operation, I_o and T are eliminated.

Analysis

$$V_{BE1} - V_{BE2} - V_{BE3} + V_{BE4} = 0 \ . \tag{8}$$

Applying Eq. (1) to each (matched) transistor, then

$$\frac{I_{E1}}{I_{E2}} = \frac{I_{E3}}{I_{E4}} \ . \tag{9}$$

Neglecting base currents, the output V_o is expressed as

$$V_o = R(I_{E3} - I_{E4})$$

$$2I_g = (I_{E3} + I_{E4})$$

where

$$V_o = \left[\frac{RI_g}{V_i}\right] v_i \tag{10}$$

that is, the gain γ is given as

$$\gamma = \frac{RI_g}{V_1} \ . \tag{11}$$

The analysis shows a gain linearly dependent on I_g. Because all base–emitter voltages were summed to zero, T cancels, thus minimizing temperature dependence.

The current-steering gain cell, which is here allowed further study, was described earlier [12] and is shown in elementary form in Fig. 8. This cell also has the property of summing the four base–emitter voltages to zero, and is thus independent of T. Hence in this respect it differs fundamentally from the similar structure of Fig. 6. An analysis of the current-steering cell was given earlier [12], where it was shown that

$$\frac{i_0}{i_1} = 1 - \frac{I_g}{2I_1} \ . \tag{12}$$

Having reviewed a range of the basic topology for translinear-gain cells, we proceed by investigating the high-attenuation distortion of the current-steering cell and then introducing a modified structure for reducing distortion and extending the attenuation range to ≈140 dB at 1 kHz.

2 DISTORTION ANALYSIS OF CURRENT-STEERING CELL AT HIGH ATTENUATION

A measurement of distortion on the current-steering class of translinear circuits reveals a distortion residual that is approximately independent of attenuation. Consequently, at high attenuation, the signal-to-distortion ratio degrades, ultimately reaching an operation regime where the distortion is greater than the signal. This distortion mechanism reduces the effectiveness of the circuit and makes acceptable operation at high attenuation especially sensitive to misalignment, inherent transistor offsets, and input-stage finite common-mode gain.

Although the four-transistor current-steering cell illustrated in Fig. 8 is considered, similar discussion applies to other translinear circuits. Investigation has revealed three principal distortion contributions: finite effective emitter bulk resistance, dc base–emitter offsets, and finite common-mode gain in the input stage. The analysis proceeds by including emitter resistors r in each transistor that is both a representation of emitter

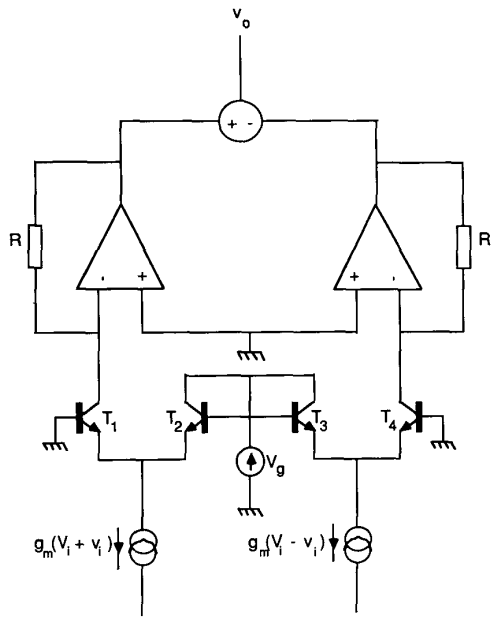


Fig. 6. Variation on log/antilog cell with input currents applied to T_1/T_2 and T_3/T_4 transistor pairs.

bulk resistance and a reflection of base bulk resistance. Also each transistor is assumed to exhibit a different saturation current, which results in an effective offset voltage V_{BEO} when the four-transistor base–emitter voltages are summed. The modified cell for analysis is shown in Fig. 9.

Analysis

$$I_g = 2I_1 - I_{01} - I_{04} . \tag{13}$$

$$\frac{kT}{q} \ln\left[\frac{I_{E2}I_{E4}}{I_{E1}I_{E3}}\right] + V_{BEO} + 2r(i_1 - i_0) = V_\varepsilon \tag{14}$$

where

$$i_0 = \frac{I_{01} - I_{04}}{2} \tag{15}$$

and V_ε is a correction voltage (see Fig. 9).

$$\lambda = -\frac{q}{2kT}\left[V_{BEO} + 2r(i_1 - i_0) - V_\varepsilon\right] \tag{16}$$

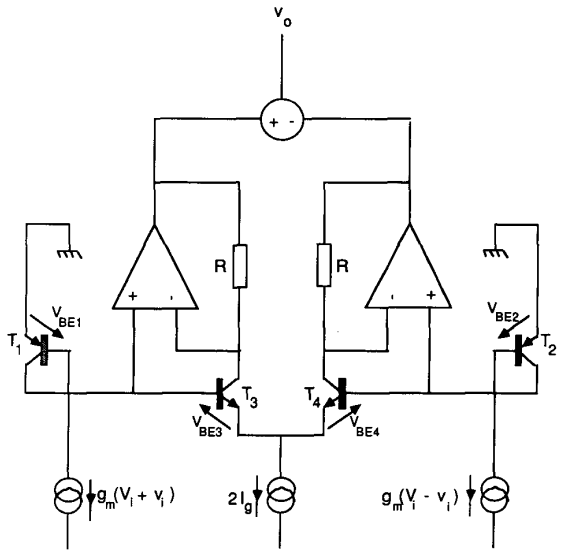


Fig. 7. Gilbert translinear gain cell [1].

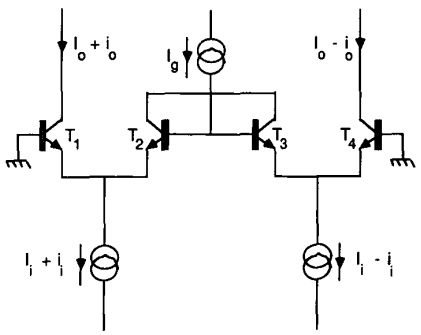


Fig. 8. Basic current-steering gain cell [12].

where

$$\ln\left[\frac{(I_1 + i_1 - I_{01})I_{04}}{(I_1 - i_1 - I_{04})I_{01}}\right] = 2\lambda . \tag{17}$$

Approximation

Let $I_{01} = \gamma(I_1 + i_1)$ and $I_{04} = \gamma(I_1 - i_1)$, where γ is the cell gain, that is,

$$\ln\left[\frac{(I_1 + i_1)(1 - \gamma)}{(I_1 - i_1)(1 - \gamma)}\frac{I_{04}}{I_{01}}\right] = 2\gamma .$$

Hence

$$\frac{(I_1 + i_1)I_{04}}{(I_1 - i_1)I_{01}} = e^{2\lambda} . \tag{18}$$

This simplification is justified for high gain ($\gamma \approx 1$) as there is minimal distortion in the output current while, for lower gains, $I_{01} << I_1 + i_1$ and $I_{04} << I_1 - i_1$.

Defining an input loading factor x,

$$x = \frac{i_1}{I_1} \tag{19}$$

and eliminating I_{01} and I_{04} using Eqs. (13), (15), (18), and (19),

$$i_0 = \frac{2I_1 - I_g}{2}\left(\frac{x - \tanh\lambda}{1 - x\tanh\lambda}\right) . \tag{20}$$

To deal with the nonlinearity and output offset current demonstrated by Eq. (20), incremental and target current gains γ_i and γ_t are introduced,

$$\gamma_i = \frac{\partial i_0}{\partial t i_1} , \qquad \gamma_t = \gamma_i\Big|_{\substack{V_{BEO} = 0 \\ r = 0}} .$$

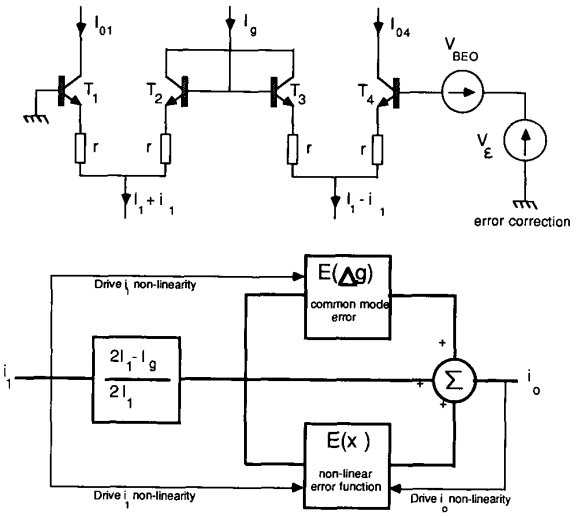


Fig. 9. Current-steering gain cell with approximate (low-level) error-function model [Eqs. (23), (27)].

Differentiating i_0 defined by Eq. (20),

$$\gamma_i = \operatorname{sech}^2 \lambda \left[\frac{1 + (1 - x^2)(1 - \gamma_i)\, qrI_1/kT}{(1 - x \tanh \lambda)^2} \right] \gamma_t \tag{21}$$

where, for $\lambda = 0$ and $r = 0\ \Omega$, the target current gain is

$$\gamma_t = \frac{2I_1 - I_g}{2I_1} \tag{22}$$

and takes the same form as the large-signal gain in Eq. (12).

To express the error in the incremental gain γ_i, an error function $E(x)$ is defined,

$$E(x) = \frac{\gamma_i}{\gamma_t} - 1$$

that is,

$$E(x) = \frac{(qrI_1/kT)(1 - x^2)(1 - \gamma_i) \operatorname{sech}^2 \lambda + \tanh \lambda\, [2x - \tanh \lambda(1 + x^2)]}{(1 - x \tanh \lambda)^2} \tag{23}$$

which, for $x \to 0$, simplifies to

$$E(0) = -\tanh^2 \lambda + \frac{qrI_1}{kT} (1 - \gamma_i) \operatorname{sech}^2 \lambda\ . \tag{24}$$

Eq. (23) describes an error function (see Fig. 9) that increases as $\gamma_i \to 0$ and is bounded by V_{BEO} and r. Even for small loading factors x, Eq. (24) shows a finite gain error that is dependent on λ. Although λ can be reduced by careful transistor matching, to minimize λ requires a corrective voltage V_ε to be added to the sum of the base-emitter voltages of the transistor array and is effectively in series with V_{BEO}, as shown in Fig. 9, where, observing Eq. (16),

$$V_\varepsilon = V_{BEO} + 2r(i_1 - i_0)\ . \tag{25}$$

This technique, which uses a constant voltage together with feedforward and feedback signals derived from input and output, respectively, is similar to that proposed by Bergstrom [13], though as this example demonstrates, it can be extended to other cell topologies.

Eq. (12) reveals the current gain to have significant sensitivity to I_1, particularly as $I_g \to 2I_g$, where small variations in either I_g or I_1 cause large fractional gain changes. At extreme attenuation, with T_1 and T_4 bias currents in the region 10^{-4} to $10^{-5} I_1$, slight distortion in the transconductance input circuitry that produces a small common-mode modulation of the bias current commensurate with T_1 and T_4 bias currents causes gain modulation of the cell. Also an imbalance between inverting and noninverting transconductance of the input stages produces a nonzero common-mode gain and has a similar effect.

To illustrate this distortion process, consider the following example where instantaneous tail currents of the T_1/T_2 and T_3/T_4 transistor pairs are i_{i1} and i_{i2}, respectively, and the noninverting and inverting transconductances are $(g_0 + \Delta_g)$ and $-(g_0 - \Delta_g)$, that is,

$$i_{i1} = g_0 v_i + (I_1 + \Delta g v_1)$$

$$i_{i2} = -g_0 v_i + (I_1 + \Delta g v_i)\ .$$

The equations for i_{i1} and i_{i2} reveal that under signal excitation a dynamic bias current $(I_1 + \Delta g v_i)$ is generated. If, for an idealized stage, $i_1 = g_0 v_i$, then the cell dynamic gain γ_d can be written following Eq. (12) as

$$\gamma_d = 1 - \frac{I_g}{2I_1 [1 + (\Delta g/g_0)\, i_1/I_1]}\ . \tag{26}$$

Hence defining an error function $E(\Delta g)$ as

$$E(\Delta g) = \frac{\gamma_d}{\gamma_t} - 1$$

where γ_t is defined by Eq. 12, then assuming $\Delta g i_1 / g_0 I_1 << 1$,

$$E(\Delta g) = \frac{1}{2I_1/I_g - 1} \cdot \frac{\Delta g i_1}{g_0 I_1}\ . \tag{27}$$

Eq. (27) shows that for high attenuation, where $I_g \to 2I_1$, the gain is extremely sensitive to Δg. Consequently measures should be taken to ensure a negligible common-mode error resulting from both linear and nonlinear distortions in the input circuitry.

The key to cell enhancement is therefore to limit the attenuation per cell and to maintain an effective operating current in all transistors. In the next section a two-stage gain cell is described which offers potential improvements in distortion performance together with a falling distortion with increased attenuation.

3 ENHANCED TWO-STAGE CURRENT-STEERING CELL

The enhanced current-steering cell employs an extended two-stage topology with attenuation divided equally between each stage. Consequently each stage now operates over a more restricted attenuation range, which offers the following principal advantages:

1) The first stage operates with the full input signal, but because its attenuation is now restricted and transistors do not operate at their limits, a relatively low distortion is returned with a low sensitivity to finite common-mode gain in the input stage [see Eq. (27)].

2) The second stage operates under identical bias conditions but accepts a much lower input signal. Consequently second-stage distortion is dramatically reduced where the first stage is the dominant distortion generator.

3) However, the distortion at the output of the first stage is attenuated by the second stage. Thus, overall, the distortion now progressively falls with increased attenuation, even though at high attenuation the first stage tends to produce a constant distortion component referred to input level.

The primitive two-stage gain cell based on the current-steering cell is shown in Fig. 10 together with two equal gain-control currents of magnitude I_g. The cell $\{T_1 \ldots T_4\}$ operates conventionally, but with collectors of T_1 and T_4 feeding into the second stage $\{T_5 \ldots T_8\}$. However, for both stages to operate under similar bias conditions at a given attenuation, a current $I_g/2$ must also be summed to compensate for the loss of current in T_1 and T_4 when $I_g > 0$. Under these conditions, the expression for current gain becomes

$$\frac{i_0}{i_1} = \left(\frac{2I_1 - I_g}{2I_1}\right)^2 . \qquad (28)$$

The circuit of Fig. 10 requires minimal extra voltage headroom where a typical value for V_B (see Fig. 10) is between 1 and 2 V. It is also relatively simple to generate the extra control current sources. Although the circuit requires an extra four transistors, this method of extension is more effective than combining arrays $\{T_1 \ldots T_4\}$ and $\{T_5 \ldots T_8\}$ in parallel.

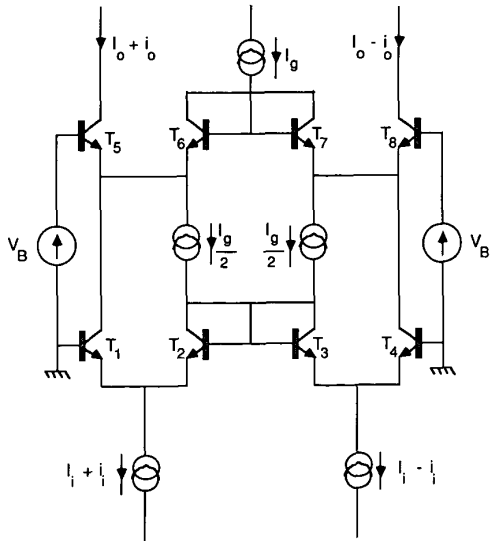


Fig. 10. Primitive two-stage current-steering gain cell.

To validate the modified current-steering cell, a prototype circuit was constructed and is shown in Fig. 11, while Fig. 12 illustrates the power supply and biasing arrangement. To minimize circuit noise through extra circuitry requirements and low bias resistor values, a dual supply rail of ± 30 V was used from which internal voltages of ± 15 V were derived.

A novel feature of the circuit was to use the gain-control voltage $V_{control}$ to drive a servo amplifier IC_6, which set the quiescent output voltages of IC_3 and thus determined the collector currents of T_5 and T_8. The servo amplifier then drove three grounded-base stages whose output currents I_{C1}, I_{C2}, and I_{C3} established the required gain-control currents and second-stage input bias compensation.

The complete VCA illustrated in Fig. 13 was assembled on a printed circuit board and metal plates attached to the LM394 arrays to aid isothermal operation. The circuit also includes switches SW_1 and SW_2 to facilitate dc alignment, hence the mimimization of gain-control feedthrough. The circuit alignment procedure is outlined in the Appendix.

4 MEASURED PERFORMANCE OF TWO-STAGE CURRENT-STEERING CELL

The enhanced circuit of Figs. 11 and 12 was assessed using a range of distortion and noise measurements, and the results are presented in Table 1. As anticipated, the distortion at maximum gain is of low level, a characteristic of the current-steering cell which effectively becomes a grounded-base stage. However, also encouraging is the distortion at 20 kHz, +15 dBV input, where for an attenuation of +100 dB a distortion of −122 dB referred to input was recorded, while for 20 kHz, +0 dBV, the distortion could not be resolved at +100 dB attenuation with available instrumentation.

5 EXTENSION OF CASCADE TO dbx ARRAY

In principle, the cascade technique can also be applied to other cells, such as the dbx array. A primitive circuit is shown in Fig. 14, which uses two complementary transistor arrays. Ideally, to minimize noise, both cells could be powered from independent floating supplies, as suggested earlier. Also, such an approach would appear to be particularly beneficial when the cells are operated in so-called class AB [8] mode (particularly for second stage), which is used to minimize noise generation from within the transistor array. However, these comments are made purely as a suggestion for further development. They are not supported at this stage with experimental verification; but, from the results obtained from the current-steering cell, improvements in distortion performance are to be anticipated.

6 CONCLUSION

This report has presented a tutorial review to outline a number of basic approaches to the translinear mul-

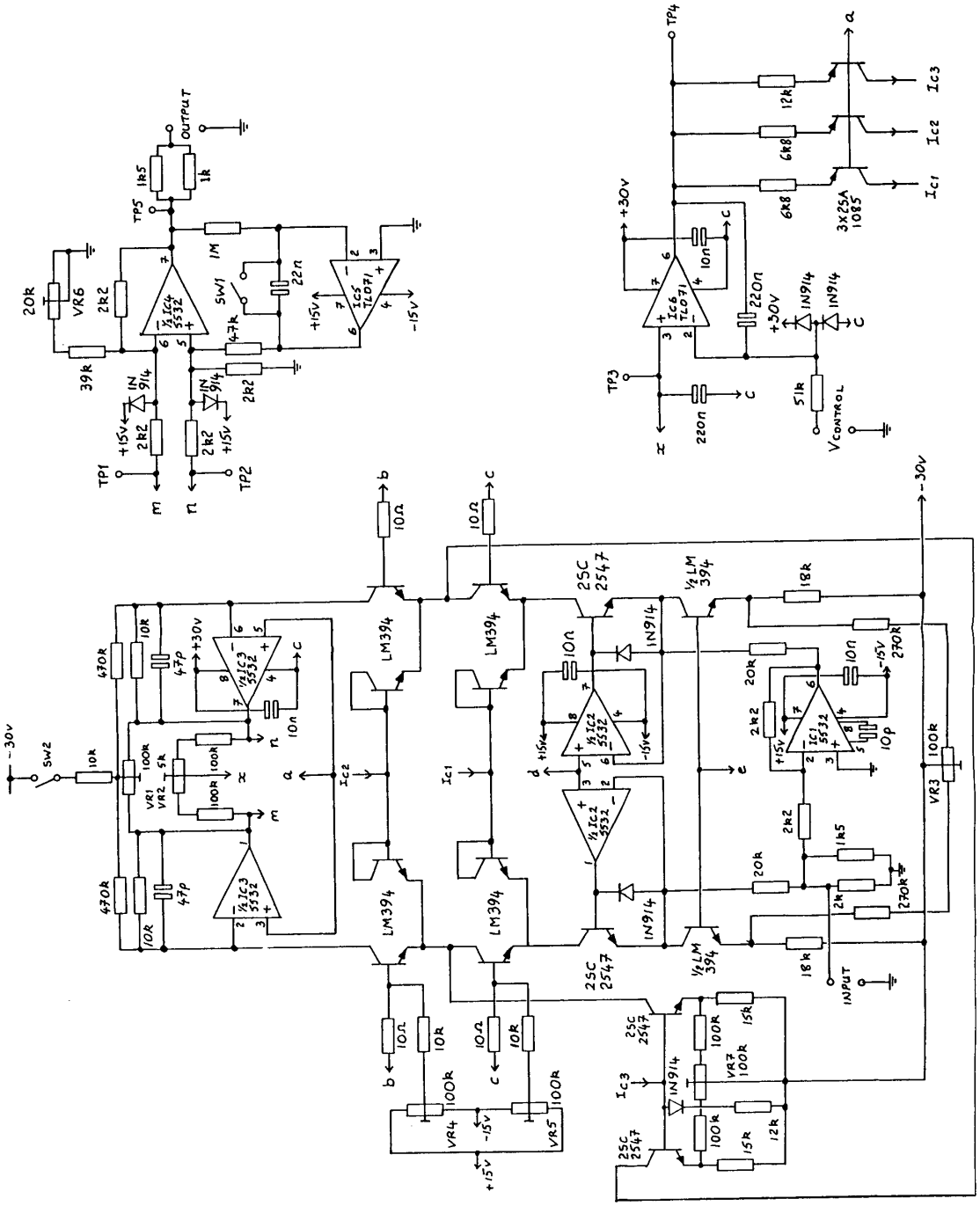


Fig. 11. Voltage-controlled amplifier, main circuit.

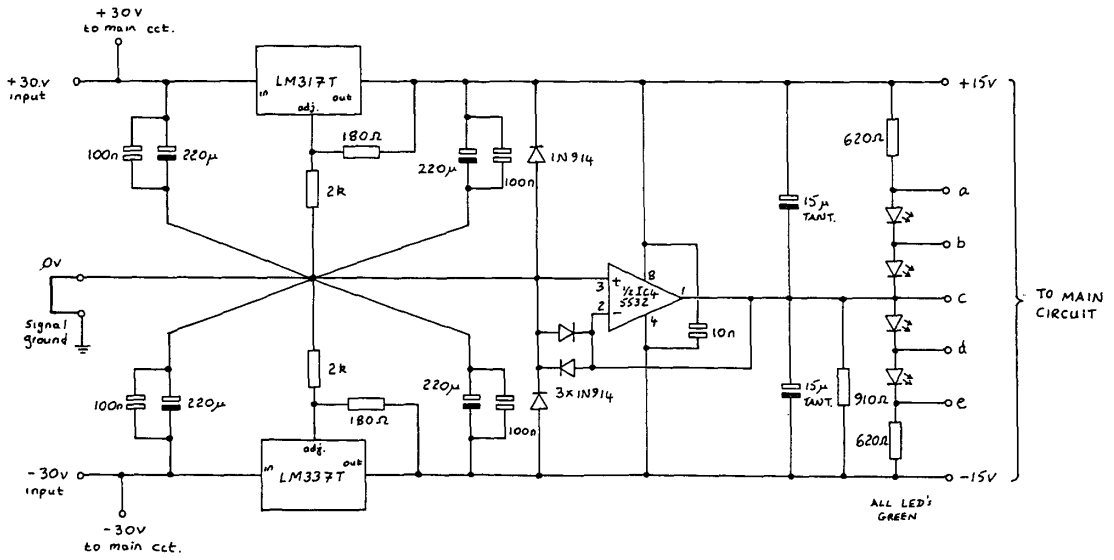


Fig. 12. Voltage-controlled amplifier, power supply.

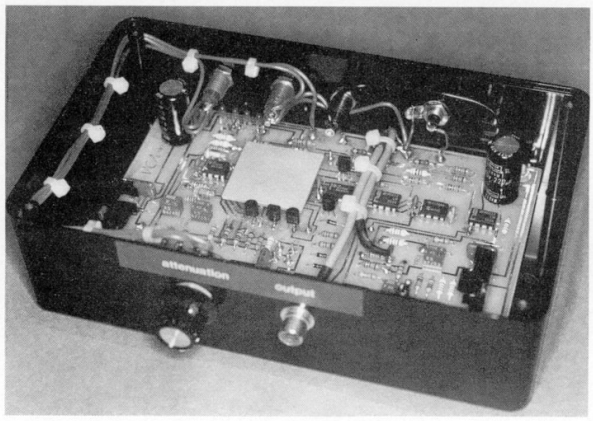


Fig. 13. Two-stage cascade prototype voltage-controlled amplifier.

Table 1. Voltage-controlled amplifier performance data.

Maximum gain		0 dB	
Maximum attenuation		140 dB at 1 kHz 115 dB at 20 kHz	
Signal-to-noise ratio (20 Hz–20 kHz unweighted)			
Maximum gain		−105 dB	
Minimum gain		−110 dB	
Distortion (THD) at maximum gain			
0-dBV input		1 kHz—not measurable; 20 kHz—0.0008%	
+15-dBV input		1 kHz—0.0004%; 20 kHz—0.0025%	
Distortion with attenuation (referred to input)			
Input level	Attenuation (dB)	1-kHz distortion (dB)	20-kHz distortion (dB)
0 dBV	+20	−94	−92
	+40	−105	−104
	+60	−115	−112
	+80	−125	−120
	+100	Not measurable	Not measurable
+15 dBV	+20	−82	−80
	+40	−89	−88
	+60	−98	−97
	+80	−108	−107
	+100	−126	−122

tiplier cell, together with an extension using two (or more) arrays connected in a cascade. To validate the technique, an experimental circuit was introduced from which a range of measurements was derived. The two-stage approach exhibited a significant improvement and showed in particular a useful reduction of distortion at high attenuation. For a given number of transistors (eight in this example) the cascade arrangement appears to offer a useful improvement over the parallel option. However, there is no fundamental reason why arrays of both serial and parallel connections should not be formed to yield further enhancement. With the experimental circuits, it proved desirable to operate from relatively high supply rails since this saved additional circuitry and the use of reduced resistor values which, in general, would otherwise have impaired the noise performance.

The report attempted an approximate distortion analysis of the current-steering cell operating at high attenuation where the need for zero common-mode bias current modulation was highlighted. The analysis also supported the desirability for using cells with more limited attenuation configured within a multistage cascade and suggested how distortion reduction may be introduced.

Although the circuits of Figs. 11 and 12 are semi-discrete prototypes, the system has potential for integrated-circuit development and for forming the basis of an effective VCA. Also, extensions to complementary NPN/PNP current-steering cells would appear feasible and allow a useful reduction in supply rail voltage, particularly if a floating gain cell supply was configured.

7 ACKNOWLEDGMENT

The authors would like to offer their appreciation to Ben Duncan of Band J Sound for his interest in this project and for drawing their attention to the Bergstrom patent [13].

8 REFERENCES

[1] B. Gilbert, "A Precise 4-Quadrant Multiplier with Subnanosecond Response," *IEEE J. Solid-State Circuits*, vol. SC-3, pp. 365–373 (1968 Dec.).

[2] Sansen and Meyer, "Distortion in Bipolar Transistor Variable Gain Amplifiers," *IEEE J. Solid-State Circuits*, vol. SC-8 (1973 Aug.).

[3] B. Gilbert, "A High-Performance Monolithic Multiplier Using Active Feedback," *IEEE J. Solid-State Circuits*, vol. SC-9, pp. 364–373 (1974 Dec.).

[4] B. Gilbert, "Translinear Circuits: A Proposed Classification," *IEE Electron. Lett.*, vol. 11, pp. 14–16 (1975 Jan. 9).

[5] B. Gilbert, "A New Technique for Analog Multiplication," *IEEE J. Solid-State Circuits*, vol. SC-10, pp. 437–447 (1975 Dec.).

[6] D. H. Sheingold, *Nonlinear Circuits Handbook* (Analog Devices, 1976).

[7] C. C. Todd, "A Monolithic Analog Compander," *IEEE J. Solid-State Circuits*, vol. SC-11, pp. 754–762 (1976 Dec.).

[8] D. Baskind and H. Rubens, "Techniques for the Realization and Application of Voltage-Controlled Amplifiers and Attenuators," presented at the 60th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 26, p. 572 (1978 July/Aug.), preprint 1378.

[9] D. Baskind, H. Rubens, and G. Kelson, "The Design and Integration of a High-Performance Voltage-Controlled Attenuator," presented at the 64th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 27, pp. 1018, 1020 (1979 Dec.), preprint 1555.

[10] T. Yamaguchi, S. Takaoka, and K. Aizawa, "A New Configuration Using a Voltage Controlled Amplifier for a Dolby System Integrated Circuit," *IEEE Trans. Consumer Electron.*, vol. CE-25, pp. 723–729 (1979 Nov.).

[11] B. Gilbert and P. Holloway, "A Wideband Two-Quadrant Analogue Multiplier," in *Proc. IEEE Int. Solid-State Circuits Conf.* (1980 Feb.), pp. 200–201.

[12] M. J. Hawksford, "Low-Distortion Programmable Gain Cell Using Current-Steering Cascode Topology," *J. Audio Eng. Soc.*, vol. 30, pp. 795–799 (1982 Nov.).

[13] G. Bergstrom, "Signal Correction for Electrical Gain Control Systems," US patient 4,234,804 (1980 Nov. 18).

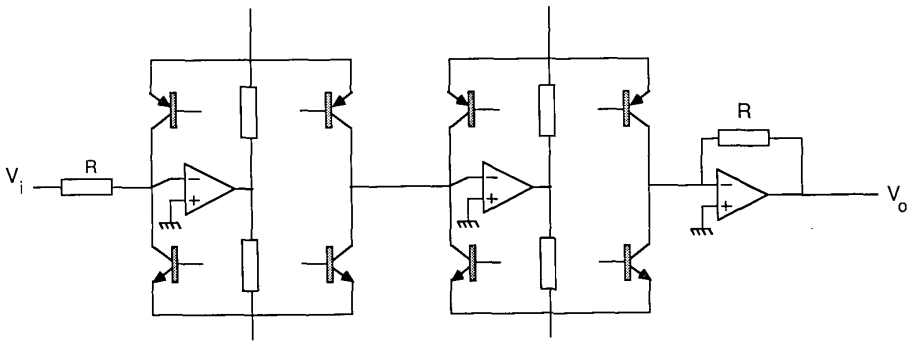


Fig. 14. Cell cascade using two dbx arrays.

APPENDIX
BASIC ADJUSTMENT PROCEDURE FOR TWO-STAGE VCA PROTOTYPE

1) Select minimum gain ($V_{control}$ = 0 V), SW_1 closed, SW_2 closed, VR_4, VR_5 midway. Adjust VR_1 to achieve $V_m = V_n$ at IC_3 outputs (T_{P1}, T_{P2}); then adjust VR_6 for zero output voltage (T_{P3}).

2) Select maximum gain, SW_1 closed, SW_2 open. Adjust VR_3 for $V_m = V_n$ at IC_3 outputs.

3) Select 12-dB attenuation. Adjust VR_7 for $V_m = V_n$ at IC_3 outputs.

4) Apply sine-wave input at maximum gain and adjust VR_2 to give zero ac output at x (T_{P3}).

Fine tuning of gain feedthrough over the range of $V_{control}$ can be achieved by iterative adjustment of VR_4, VR_5, VR_7, and the above procedure.

The circuit operates normally with both SW_1 and SW_2 open (where SW_1 selects an output servo to minimize dc offsets at the output, though it can be left closed if response down to dc is required). In practice, alignment can be minimized by the use of matched resistors (0.1% or better) and matched transistor arrays. The former was not used in prototype but is seen as an integral part of integrated circuit fabrication.

Dr. Hawksford's and Mr. Mills's biographies were published in the March issue.

3 Converter and switching amplifiers systems

3-1 Converters, noise shaping and oversampling

3-1 CHAOS, OVERSAMPLING AND NOISE SHAPING IN DIGITAL-TO-ANALOG CONVERSION, Hawksford, M.O.J., *JAES*, vol. 37, no. 12, pp 980-1001, December 1989

3-23 OVERSAMPLING FILTER DESIGN IN NOISE-SHAPING DIGITAL-TO-ANALOG CONVERSION, Hawksford, M.O.J., Wingerter, W., *JAES*, vol. 38, no. 11, pp 845-856, November 1990

3-35 OVERSAMPLED ANALOG-TO-DIGITAL CONVERSION FOR DIGITAL AUDIO SYSTEMS, Hawksford, M.O.J. and Darling, T.E., *JAES*, vol. 38, no. 12, pp 924-943, December 1990

3-55 DIGITAL-TO-ANALOG CONVERTER WITH LOW INTERSAMPLE TRANSITION DISTORTION AND LOW SENSITIVITY TO SAMPLE JITTER AND TRANSRESISTANCE AMPLIFIER SLEW RATE, Hawksford, M.O.J., *JAES*, vol. 42, no. 11, pp 901-917, November 1994

3-72 TRANSPARENT DIFFERENTIAL CODING FOR HIGH-RESOLUTION DIGITAL AUDIO, Hawksford, M. O. J., *JAES*, vol. 49, no. 6, pp 480-497, June 2001

3-2 Sigma-delta modulation (SDM)

3-90 EXACT MODEL FOR DELTAMODULATION PROCESSES, Flood, J.E. and Hawksford, M.J., *Proc. IEE*, vol.118, pp.1155-1161, 1971

3-97 UNIFIED THEORY OF DIGITAL MODULATION, Hawksford, M.J., *Proc. IEE*, vol.121, no.2, pp.109-115, February 1974

3-104 TIME-QUANTIZED FREQUENCY MODULATION, TIME-DOMAIN DITHER, DISPERSIVE CODES, AND PARAMETRICALLY CONTROLLED NOISE SHAPING IN SDM, Hawksford, M.O.J., *JAES*, vol. 52, no. 6, pp 587-617, June 2004

3-135 PARAMETRICALLY CONTROLLED NOISE SHAPING IN VARIABLE STATE-STEP-BACK PSEUDO-TRELLIS SDM, Hawksford, M.O.J., IEE Proc.-VIS. Image Signal Processing, Vol. 152, No. 1, pp 87-96, February 2005

3-3 Pulse width modulation (PWM) and switching amplifiers

3-145 DYNAMIC MODEL-BASED LINEARIZATION OF QUANTIZED PULSE-WIDTH MODULATION FOR APPLICATIONS IN DIGITAL-TO-ANALOG CONVERSION AND DIGITAL POWER AMPLIFIER SYSTEMS, Hawksford, M.O.J., *JAES*, Vol. 40, no. 4, pp 235-252, April 1992

3-163 LINEARIZATION OF MULTI-LEVEL, MULTI-WIDTH DIGITAL PWM WITH APPLICATIONS IN DIGITAL-TO-ANALOGUE CONVERSION, Hawksford, M.O.J., *JAES*, vol. 43, no. 10, pp 787-798, October 1995

3-175 AN OVERSAMPLED DIGITAL PWM LINEARIZATION TECHNIQUE FOR DIGITAL-TO-ANALOG CONVERSION, Jung, J.W. and Hawksford, M. J., Regular Papers, IEEE Transactions on [see also Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on], vol. 51, no. 9, Issue 9, September 2004-09-15, pp 1781-1789

3-184 MODULATION AND SYSTEM TECHNIQUES IN PWM AND SDM SWITCHING AMPLIFIERS, Hawksford, M.O.J. *JAES*, vol. 54, no. 3, pp. 107-139, March 2006

Chaos, Oversampling, and Noise Shaping in Digital-to-Analog Conversion*

M. O. J. Hawksford

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, CO4 3SQ, UK

A method of digital-to-analog conversion (DAC) is investigated that uses a combination of high oversampling and high-order noise shaping. Both linear analysis and computer modeling reveal many of the characteristics of a base-band noise shaper where chaotic loop behavior is shown to result in a decorrelation of DAC distortion to benign noise.

0 INTRODUCTION

Chaos [23] is about the behavior of nonlinear systems that operate within a recursive environment and exhibit nondeterministic behavior, where the slightest modification in initial conditions or system parameter can have dramatic effects on future system states. Under appropriate conditions, even simple nonlinear recursive equations can reveal complicated patterns and require the application of probability theory and topology. Many researchers have now contributed to this important subject in a number of fields, although much of the conceptual credit is due to the pioneering work of Edward Lorentz on the study of weather systems [40].

Audio engineering has had a long tradition of studying the application of feedback and distortion correction schemes to nonlinear systems, although, in general, this has not been approached within the modern framework of chaos. However, more recently, noise-shaping coders [2], [52], [53] have emerged and found applications in both analog-to-digital (ADC) and digital-to-analog conversion (DAC) systems, and since they combine both recursion and nonlinearity, their behavior can be viewed sympathetically from the perception of chaos.

In this paper we both discuss background theory to noise shaping and present results of a computer simulation of a DAC scheme that uses a combination of oversampling and noise shaping. The rationale for this approach is, first, the ability to implement more optimum linear phase reconstruction filters offering exact time synchronization; second, implementation of schemes where all the nonideal behavior of the DAC is translated to benign noise, and, third, simplification of analog circuitry, which should lead to greater signal transparency.

Present methods of oversampling in digital audio are extensions and adaptations of the work at Philips [24], where mild ratios of $\times 2$, $\times 4$, and $\times 16$ have emerged. The advantage of improved signal reconstruction using digital interpolation is evident in these schemes, where substantial filtering is performed in the digital domain. As such, the filter response can offer both a rapid attenuation with frequency as well as near-ideal group-delay characteristics. Also, where 16-bit DACs are used together with noise shaping, improvements in low-level resolution can be attained, yielding a closer approximation to the bound dictated by source data quantization.

However, there are further advantages to be gained by using even higher oversampling ratios, for example, in the region of $\times 256$. In such a scheme the number of samples per Nyquist sample is extended dramatically, allowing further enhancements in reconstruction filtering with the corresponding simplifications in DAC and analog design.

* Presented at *Reproduced Sound 3* (Windermere, Cumbria, UK, 1987 November 5–8), as "Oversampling and Noise Shaping for Digital to Analogue Conversion"; revised 1989 March 23.

To realize the full advantage of oversampling, the process must be combined with that of noise shaping [30], [31], [52], [53]. As such, it becomes feasible to both reduce the number of bits necessary in the DAC substantially and induce high-level DAC activity or chaotic behavior such that large-scale errors translate to noise. Under such operation the DAC can be viewed as a wide-band, spectrally weighted noise source and therefore takes on a more benign character.

The process of oversampling and noise shaping can be applied hierarchically to the implementation of DAC schemes. At one end of the range, there are the $\times 1$ to $\times 16$ mild oversampling ratios still retaining 14-, 16-, or 18-bit DACs, while at the other extreme a 1-bit DAC is feasible where the performance is similar to delta–sigma modulation (DSM). Indeed, the methods presented here, with the inclusion of a nonrecursive look-up table, can also be viewed as a more optimum form of PCM-to-DSM conversion [30], [32].

It may well be argued that a DAC > 16-bit resolution is not required. However, the benefits of a more natural conversion topology with predominantly digital processing should be emphasized, particularly where the DAC distortion is decorrelated and can be modeled as a simple additive noise source, virtually independent of the baseband signal. The simplification of analog circuitry is also welcome, particularly the elimination of transimpedance amplifiers as used with current-output DACs, with their extremely high bandwidth, high slew rate, and low open-loop distortion requirements. Also, there is now considerable research into digital equalization, where our own studies on loudspeaker equalization [27] have indicated the need for a dynamic range > 16 bits once equalization is performed. For each 6 dB of equalization, an increase of 1 bit in DAC resolution is required if signal truncation is to be avoided. Similarly, where a number of digital signals are mixed, if a loss of information is to be avoided, extra bits in the composite signal are required, that is, for 2 channels → +1 bit, +4 channels → +2 bits, . . . , 64 channels → +6 bits, and so on. Ideally, the composite signal is truncated in an optimal environment of digital dither. However, using a wide-range DAC, such processing can be deferred during recording to enable an appropriate artistic compromise of compression/truncation to be made.

The approach taken in this paper is to investigate the combined processes of oversampling and noise shaping in conjunction with an imperfect DAC by using computer simulation. As such, a range of DAC distortion mechanisms can be included, such as both large and small static-level errors, rise-time error, slew-rate error, and glitch error.

1 PRINCIPLES OF NOISE SHAPING AND DATA COMPACTION

The DAC system investigated in this paper requires the joint application of oversampling and noise shaping. Effectively, source data $\{M, f_{s1}\}$ are converted to output data $\{N, f_{s2}\}$, where f_{s1}, and f_{s2} are the respective input and output sampling rates ($f_{s2} >> f_{s1}$) and M and N bits are input and output word lengths. Because the input information is restricted to a Nyquist bandwidth of $f_{s1}/2$ Hz and substantial oversampling is used, information theory [49] indicates $N < M$ without loss of source data.

The basic processes are shown in Fig. 1, where a digital low-pass filter forms the sampling-rate conversion and a noise-shaper or level compaction algorithm reduces the number of quantization levels required in the output code. In this paper we shall assume ideal characteristics for the interpolation low-pass filter function and concentrate more on the performance of the noise shaper in association with a nonideal DAC, as it is the characteristic of this subsystem which enables substantial error decorrelation of output DAC nonideality. Since the results are derived by time-domain simulation, elimination of the oversampling filter is trivial as code words of length M bits at a sampling rate f_{s2} Hz can be generated from within the program. It is also important to note that the N-bit-wide output DAC is driven open loop from the output of the noise shaper. Thus the DAC dynamic performance in no way influences the noise-shaper characteristics.

1.1 Oversampling Filters

The techniques required to design oversampling filters are now well understood, and there are numerous publications describing methods for enhancing computational efficiency (see References, Sec. 7.2). Although the aim of this paper is not to present a comprehensive discussion of oversampling filters, the following brief résumé may prove useful in seeking further reference material.

1.1.1 Transposed FIR Filters

The process of oversampling requires the calculation of new samples positioned symmetrically between the Nyquist samples of the input data (such as those gen-

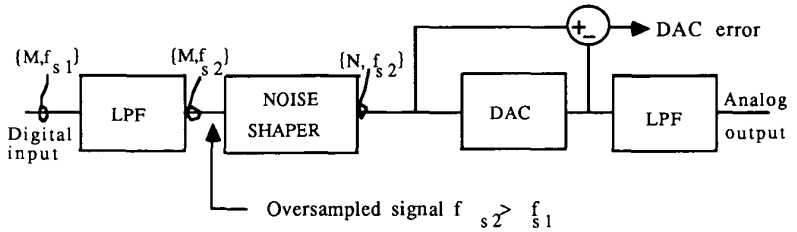


Fig. 1. Basic oversampling DAC scheme with level compaction using recursive noise shaping.

erated at 44.1 kHz). If the oversampling ratio R_N is defined,

$$R_N = \frac{\text{selected sampling rate}}{\text{Nyquist sampling rate}} \quad (1)$$

then the oversampling process requires initially the insertion of $(R_N - 1)$ zero samples per Nyquist interval followed by a low-pass filter function to band-limit the input data to one-half the Nyquist sampling rate with spectral replications about $kR_N f_{s1}$ (for integer k). The process of low-pass filtering effectively calculates the intermediate sample values, thus realizing the function of interpolation. However, if the low-pass filter is implemented by a direct realization FIR filter, then significant time is spent multiplying filter coefficients by zero-sample values. The more efficient, transposed FIR structure recognizes this redundancy and reduces the required number of multiplications by $(R_N - 1)/R_N$. Hence, for example, a 96-order symmetrical FIR filter with $R_N = 4$ now requires on average only 24 multiplications per output sample.

1.1.2 Multiband Filters

A common method of implementing high-order oversampling filters is to translate the interpolation process into a number of cascaded stages, where the interpolation factor R_N can be decomposed as

$$R_N = \prod_{i=1}^{X} R_i \ . \quad (2)$$

The intermediate factors R_i are often selected as multiples of 2 where $R_N = 2^X$ for integer X. The advantage of the technique is that, after the first stage of interpolation, "don't care" bands are generated which relax the design requirements of later stages, thus saving multiplications as shorter impulse responses can be used.

1.1.3 Half-Band Filters

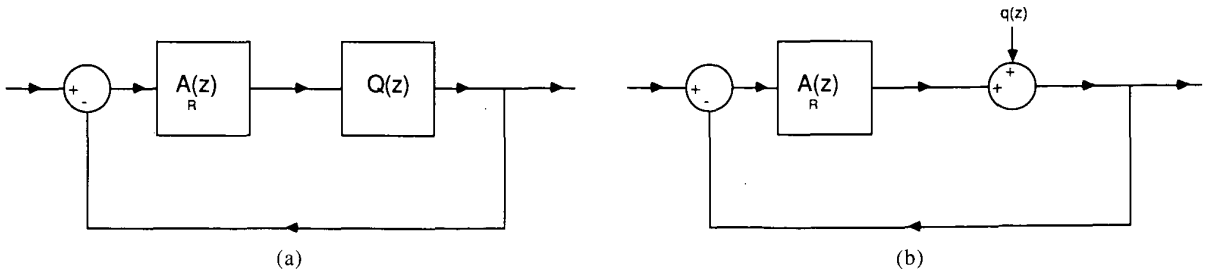
A design technique that can further save multiplications is to choose an impulse response where alternate sample values are zero. It can be shown that this condition maps into the frequency domain, where the frequency response must offer odd symmetry within the attenuation region. Hence a further constraint is placed on the interpolation, low-pass filter response which, in turn, must be accounted for in the filter optimization program. The Parks–McClellan [70] algorithm is often used to design FIR filters. However, by combining this optimization routine with the "half-band design trick" proposed by Vaidyanathan and Nguyen [90] and Ansari [59], only the nonzero coefficients need be calculated.

1.2 Noise-Shaping Coders

Information theory [49] predicts that when a band-limited signal is oversampled, the output data can tolerate a reduction in amplitude resolution, yet maintain a similar in-band signal-to-noise ratio (SNR). The process of compacting the sample resolution can be performed by either a nonrecursive or a recursive process, although the latter is the more efficient. The nonrecursive process has the advantage of simplicity and is used where a band-limited signal is oversampled and quantized, with the result that only a fraction of the total quantization noise power corrupts the signal. However, to achieve a 1-bit reduction in resolution from oversampling, a fourfold increase in the sampling rate is required, and this assumes that there is no intersample correlation.

The more efficient noise-shaping architecture uses recursion where negative feedback encloses an amplitude quantizer together with an appropriate forward-path signal processor. Such techniques both spread the overall noise power across a greater bandwidth and shape the noise spectrum so that the majority of noise is located outside of the signal bandwidth. A basic noise-shaping coder is shown in Fig. 2, where $Q(z)$ is the quantizer and $A_R(z)$ the forward-path processor of order R. This scheme also includes the delta–sigma modulator [33], [34] class of coder, where $Q(z)$ is limited to a two-level comparator and $A_R(z)$, selected to be either a first-order ($R = 1$) or second-order ($R = 2$) integrator.

The delta modulator [16], [17], [51] and later the delta–sigma modulator [33], [34] are both examples of early forms of noise-shaping coder. However, the two-level restriction on $Q(z)$ is a formidable limitation, even though the system is well suited to DAC. Simulation studies, however, reveal that for loop orders ≥ 3, a two-level quantizer sets a performance bound that causes the coder to enter a nonlinear regime with minimal coding performance. A solution is to introduce a multilevel quantizer [30] and to prevent amplitude saturation, whereby loop orders ≥ 3 become feasible with the potential for enhanced dynamic range and distortion



(a) (b)

Fig. 2. Noise-shaping coder. (a) Recursive model. (b) Error addition, recursive model.

reduction.

In designing noise-shaping coders where the quantizer $Q(z)$ has a linear quantization law, the form of the noise-shaping characteristic, together with the need for closed-loop stability, must be addressed. In this paper we tackle this problem in two stages. First, we derive by linear analysis the best form of noise-shaping processor that can also meet the need of closed-loop stability when $Q(z) = 1$. Second, nonlinearity is introduced where the loop is tested for stability and performance is investigated through exact time-domain simulation. It is in this latter stage that the combination of nonlinearity (quantizer and computer truncation errors) and recursion exhibit elements of chaos and prevent a deterministic analysis from describing output sequences, allowing only a statistical estimate of SNR.

The theoretical discussion of noise shaping commences by assuming $Q(z)$ to be a linear network with additive distortion $q(z)$, as shown in Fig. 2(b), where $q(z)$ is noise shaped by the recursive noise shaper to form an output noise spectrum $N(z)$.

Let $V_o(z)$ be the output sequence of an Rth-order noise shaper when $V_i(z)$ is the input sequence. Referring to Fig. 2(b),

$$V_o(z) = \frac{V_i(z)A_R(z)}{1 + A_R(z)} + \frac{q(z)}{1 + A_R(z)} . \tag{3}$$

For $A_R(z) >> 1$,

$$V_o(z) = V_i(z) + N(z) \tag{4}$$

that is,

$$N(z) = \frac{q(z)}{1 + A_R(z)} = q(z)D_R(z) . \tag{5}$$

Tewksbury and Hallock [53] have shown that an optimal function for the noise-shaper characteristic $D_R(z)$ is

$$D_R(z) = \left(\frac{z - 1}{z}\right)^R \tag{6}$$

which is effectively R cascaded digital differentiators. Thus for a given order R, the slope of the shaping function against frequency is maximum and gives the best suppression of low-frequency distortion.

The frequency-domain representation of $D_R(z)$ is determined by substituting

$$z = e^{j2\pi f/f_{s2}}$$

where f_{s2} Hz is the noise-shaper sampling frequency, giving

$$|D_R(f)| = \left[2 \sin\left(\frac{\pi f}{f_{s2}}\right)\right]^R . \tag{7}$$

Fig. 3 shows a family of $|D_R(f)|$ for $R = 1$ to 6 and illustrates the predicted noise-shaping functions based on linear analysis. The curves reveal two frequencies of interest,

$$\left|D_R\left(\frac{f_{s2}}{6}\right)\right| = 1 \tag{8}$$

$$D_R\left(\frac{f_{s2}}{2}\right) = 2^R . \tag{9}$$

Eq. (8) reveals that all curves take unit value at $f_{s2}/6$ Hz and that distortion reduction is achieved only for $f < f_{s2}/6$ Hz, while for $f_{s2}/6 < f < f_{s2}/2$ the noise spectrum is actually amplified, reaching a maximum at $f_{s2}/2$ Hz.

From Eqs. (5) and (6) the optimal z-domain description of $A_R(z)$ is

$$A_R(z) = \left(\frac{z}{z - 1}\right)^R - 1 \tag{10}$$

which also takes the expanded form

$$A_R(z) = \left[\frac{z^{-1}}{(1 - z^{-1})^1} + \frac{z^{-1}}{(1 - z^{-1})^2} + \cdots + \frac{z^{-1}}{(1 - z^{-1})^R}\right] . \tag{11}$$

Using this expansion, together with a quantizer $Q(z)$, the topological form shown in Fig. 4 is generated for an Rth-order noise shaper, while Fig. 5 illustrates a family of frequency-domain plots of $A_R(z)$ for $R = 1$ to 6.

1.3 Properties of Noise Shaper

1) The topology uses no multiplications, only additions. Hence the processor speed potential is high.

2) For $f = f_{s2}/2$ Hz, $A_R(z)$ is minimum where, noting

$$z = e^{j2\pi f/f_{s2}} = e^{j\pi} = -1$$

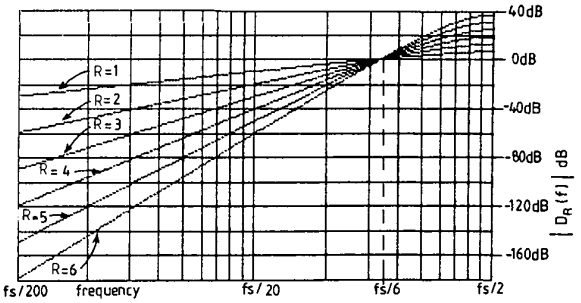


Fig. 3. Family of $|D_R(f)|$ for $R = 1$ to 6 noise shaper.

then

$$A_R(-1) = 0.5^R - 1, \text{ that is, } |A_R(-1)| < 1 . \quad (12)$$

3) For $f << f_{s2}/2$ Hz, the modulus of the gain of $A_R(z)$ follows by substituting

$$z = 1 + j2\pi f/f_{s2}$$

namely,

$$|A_R(f)|_{f<<f_{s2/2}} = \left[\frac{j2\pi f}{f_{s2}}\right]^R \quad (13)$$

that is, the transfer function approximates to R cascaded integrators.

4) The closed-loop gain $G(z)$ follows from Eq. (3), where assuming a linear quantizer, namely, $Q(z) = 1$ and $q(z) = 0$, then

$$G(z) = \frac{A_R(z)}{1 + A_R(z)} . \quad (14)$$

Substituting $A_R(z)$ from Eq. (10),

$$G(z) = 1 - \left(\frac{z - 1}{z}\right)^R . \quad (15)$$

For $Q(z) = 1$, the closed-loop gain $G(z)$ has the pole–zero distribution shown in Fig. 6 and exhibits closed-loop stability (in the linear sense) with R coincident poles located at $z = 0$.

5) Transforming the noise-shaper topology to a form similar to delta modulation [16], [30], [51], as shown in Fig. 7, then

$$Y = \sum_{r=0}^{R} H(r) = H(0) + \sum_{r=1}^{R} H(r)$$

and

$$H = \sum_{r=1}^{R} H(r) .$$

Hence if $Q(z)$ is uniformly quantized and transparent to integer data, then for integer input data $H(0) = X - H$, whereby $X = Y$, that is, in the modified configuration, the coder is transparent to integer data and independent of loop activity, provided $Q(z)$ does not saturate. Although this arrangement offers no useful noise shaping or resolution compaction, it nevertheless demonstrates an important characteristic of the optimal coder topology.

6) Over the band 0 to f_{s2} Hz, the noise-shaping function is symmetrical about $f_{s2}/2$ Hz.

So far, the investigation of the noise shaper has assumed the quantizer $Q(z)$ to be a unity-gain cell with additive noise. Clearly, such an assumption can only give a partial insight into coder behavior, where a more detailed study requires the inclusion of a discontinuous quantizer. It is in this nonlinear arena that elements of chaos are encountered where high orders of nonlinearity are combined within a recursive topology. A deterministic study would attempt to predict the exact output sequence based on a knowledge of the input excitation

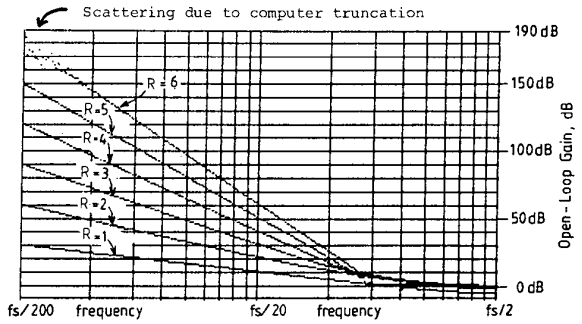


Fig. 5. Family of open-loop frequency response curves for $R_N = 1$ to 6 noise shaper.

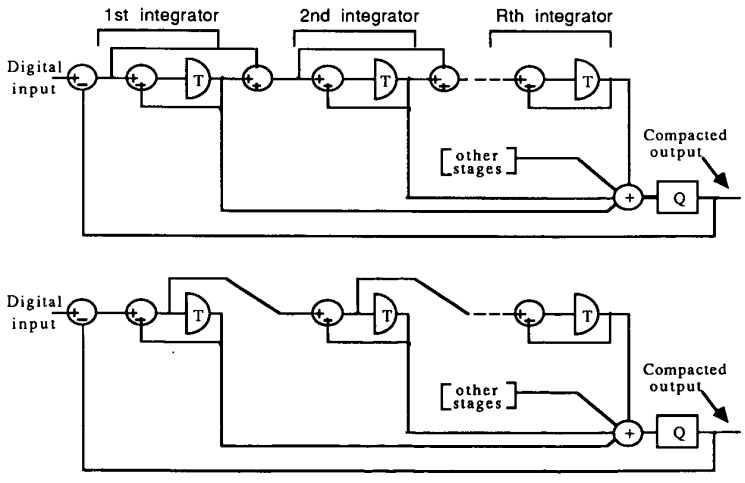


Fig. 4. Two equivalent topologies for Rth-order noise shapers.

and transfer characteristic of the coder. However, as computer simulation reveals, such an analytic approach is unrealistic as the smallest change in input signal, system initial conditions, or arithmetic truncation can propagate through to large-scale events and totally transform the output data sequences. However, as will be demonstrated, although the output sequence exhibits significant chaotic behavior, the embedded input signal is still preserved and chaos is instrumental in decorrelating hardware-related distortions. In the next section a computer model is described which simulates the noise-shaping DAC and discusses how DAC nonidealities are included in the model.

2 COMPUTER SIMULATION

The DAC system to be investigated is based on the noise-shaping coder illustrated in Figs. 1 and 4. The linear quantizer within the loop is implemented by using an integer function where the quantizer has a midtread characteristic with a unit-step quantum.

Eq. (4) reveals that because the feedback path has unity gain, the overall input–output transfer function excluding noise $N(z)$ is also unity gain for frequency $<<f_{s2}/2$ Hz. It is convenient, therefore, to use the unit step of the quantizer as a reference level for input signals, where we define a sine wave of amplitude $\sqrt{2}$ quantum as an input signal of 0 dB. However, the input data to the noise shaper are samples with full system resolution, where, for an M-bit input sample with 2^M linear quantization steps of amplitude q_m, it follows that an input quantum with respect to a unit-output quantum is defined as

$$q_m = \frac{1}{(2^M - 1)} . \tag{16}$$

The system modeling proceeds by calculating output data samples for a predetermined input sequence (such as a sine-wave sequence) using the model of Fig. 4.

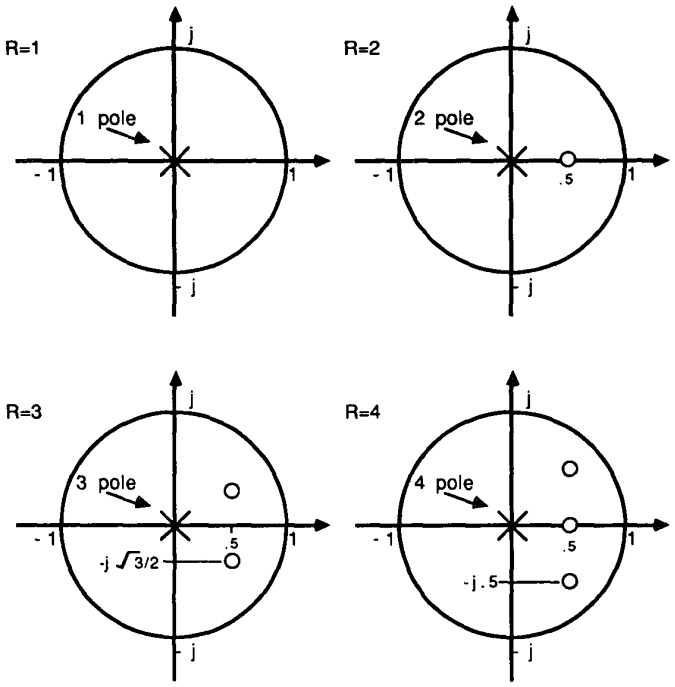


Fig. 6. Noise-shaper pole-zero plots of $A_R(z)$ for $Q(z) = 1$ and $R = 1$ to 4.

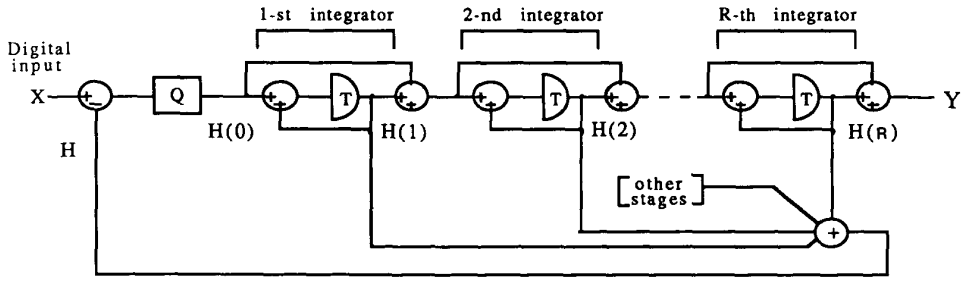


Fig. 7. Nth-order multilevel delta modulator.

The calculation is accomplished by first setting initial conditions to zero and then computing the state of the machine at each subsequent sample of period $1/f_{s2}$ second. The R cascaded, digital integrators in the forward-path processor are each formed using unity-gain positive feedback, enclosing a delay element of $1/f_{s2}$ second, and act as simple arithmetic accumulators. Hence the only distortion introduced in modeling is the fundamental quantizer nonlinearity (formed taking integer values) and the truncation errors inherent in computer arithmetic. If a sinusoidal excitation is used, then it is convenient to set the input frequency so that when the number of output samples is a power of 2 (such as 512), then an integer number of input cycles are processed. This enables a fast Fourier transform (FFT) algorithm to compute the distortion spectra. It is also good practice to compute at least one cycle of input data before capturing a segment of output data for FFT analysis to ensure that startup transients have dispersed (as is evident in Fig. 15).

If the rigor of an FFT analysis is not required, then the in-band SNR can be computed by using an error-difference technique by running two identical models side by side with a common input sequence. One model includes a linear quantizer, while the second sets $Q(z) = 1$. Hence by taking the output sequence difference, the true error can be estimated, while to extract the in-band noise, the coder outputs are band-limited using an appropriate FIR or IIR filter. A running mean-square-error summation is then formed over a known data sequence from which the SNR is calculated. This approach ensures that linear errors do not distort the very small in-band errors encountered in high-order noise shapers.

The simulation is completed by including a model of three classes of DAC error mechanisms, which can modify the noise-shaper output to introduce DAC distortion. Each output pulse then receives an additive error component, where the only approximation is that subtle changes in pulse shape are assumed not to have spectral significance in the band 0 to $f_{s2}/2$ H as $f_{s2} >> f_{s1}$.

The three primary distortions of static, glitch, and slew-rate limit are shown in Fig. 8 and modeled as follows.

2.1 Static DAC Nonlinearity

A static error has no memory and is calculated on a sample-by-sample basis. Such DAC errors result from resistor ladder networks or current sources which deviate from their optimum values. Consequently a DAC can exhibit large-scale curvature in its overall transfer characteristic together with superimposed small-scale random displacements.

The simulation used a composite error function consisting of a sine function nonlinearity spanning the full quantizer range, together with randomly generated deviations about each reconstruction level. An example static error function is shown in Fig. 9 for an $N = 4$-bit DAC, where the error quantum q_m is referred to the noise-shaper input data with $M = 16$ bits [Eq. (16)].

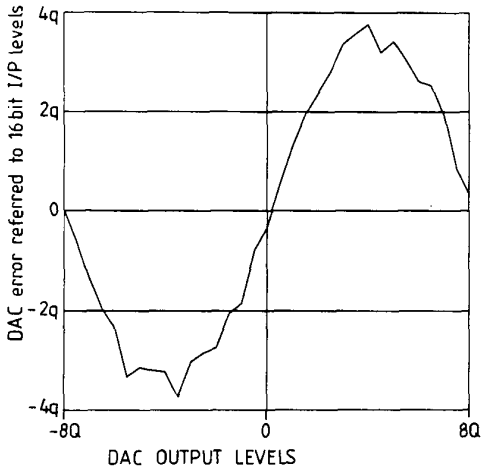


Fig. 9. DAC static nonlinearity.

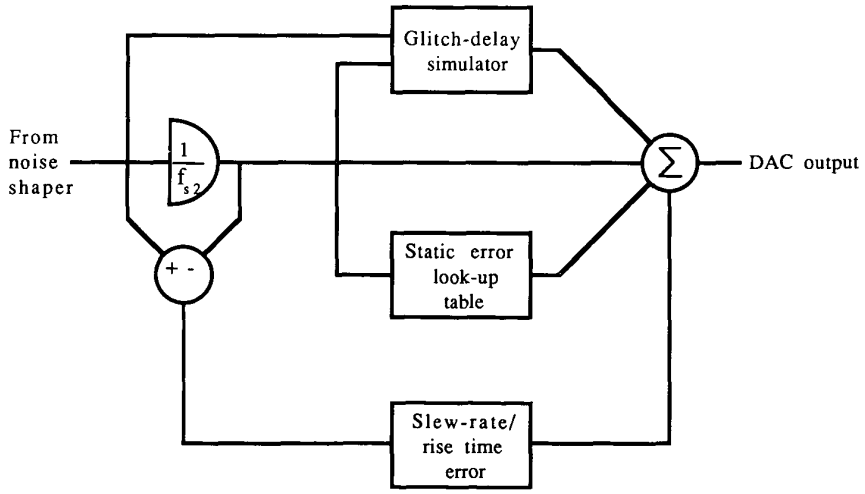


Fig. 8. Overall model of DAC error simulator.

Using this approach, the output DAC accuracy can easily be varied from, say, 12 to 16 bits.

The static error function of Fig. 9 is used as a look-up table where the noise-shaper output forms the address, whereby the output error is added to the output sample, as shown in Fig. 8.

2.2 Glitch Distortion for Binary Weighted and Parallel DAC Architectures

Glitch distortion in DACs is a dynamic distortion that depends on signal transitions from one level to another. Indeed, the characteristic of glitch will depend on the actual levels involved in the transition. Therefore it can be said to have a dynamic characteristic. Simplistically, glitch arises due to the small time-response differences that occur within the DAC on each of the effective data lines. For example, when the most significant bit settles, this may not be synchronized with, say, the least significant bit. Thus the output is momentarily in error until both events have settled. Hence we can relate one source of glitch to internal logic timing; consequently, the resulting error is a function of the type of DAC and its internal functionality.

In this study we include the basic models, as shown in Fig. 10, for two types of DAC:

1) Binary weighted DAC, where each binary input is weighted 2^0, 2^1, 2^2, . . . , and differential timing errors are introduced on a per-bit basis

2) A parallel DAC, where a coded word is loaded to a parallel register and each output is given equal weight.

Because the sampling rate is high, the model need only predict the error in sample area due to glitch and subtract the appropriate error from the sample value. The exact shape of the glitch is of little consequence, because the glitch duration is only a small fraction of a Nyquist sample period as $f_{s2} >> f_{s1}$.

For each of the two DAC structures, the program selected random pulse delays up to a maximum of 150 ps. For the parallel DAC the maximum glitch area recorded was 1.240 pV · s, while for the binary weighted DAC 0.770 pV · s resulted. The program then interrogated output corresponding glitch error which, effectively, represents a loss of pulse area.

2.3 Rise Time and Slew-Rate Distortion

Possibly the most significant distortion is that due to slew rate [21], where either finite logic rise time or slewing in subsequent stages results in an error in the

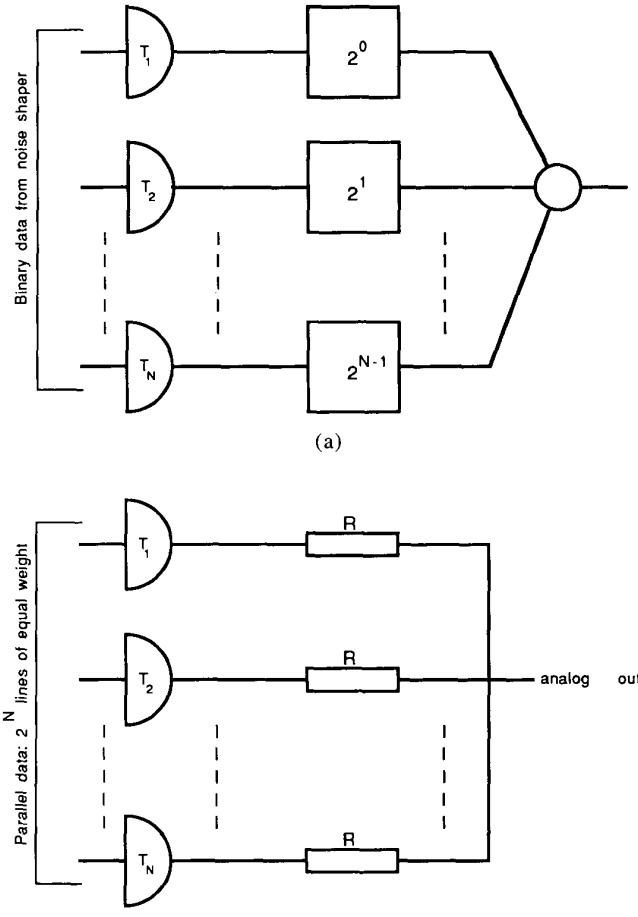


Fig. 10. Binary and parallel output glitch models. T_1, T_2, . . ., T_N are logic timing errors chosen randomly from a range of 0–150 ps.

area of each sample. This distortion is a function of the logic speed, the sampling rate, and intersample differences, where the more frequent the data transitions, the more significant the distortion. Thus the distortion degrades with an increase in sample rate. In Fig. 11 a pulse rise time is shown that has both a slew-limited region and a linear region (exponential in this example), where the slew rise time t_x and the linear rise time t_r are also defined.

It is clear that the greater ΔV_0, the greater is the loss of pulse area SA due to slew rate. The interrelationship between slew-rate-limited and linear rise time then produces an error that is nonlinearly related to ΔV_0, where the number of error areas per second also increases with the number of transitions. In the computer program the error area SA was calculated and the pulse value correspondingly modified, as follows. Let

ΔV_0 = pulse change in amplitude
x = slew rate, volt per second
τ_1 = linear region exponential time constant
V_x, t_x = nonlinear/linear transition coordinates
t_y = virtual time commencement of linear region
t_r = time duration of linear region to $0.9\Delta V_0$.

For $t > t_y$,

$$v = \Delta V_0\left[1 - e^{-(t-t_y)/\tau_1}\right]$$

but at $t = t_x$, $dv/dt = x$. Hence

$$\left.\frac{dv}{dt}\right|_{t=t_x} = \frac{\Delta V_0}{\tau_1} e^{-(t_x-t_y)/\tau_1} = x$$

whereby

$$V_x = \Delta V_0 - x\tau_1 ,$$

$$t_x = \frac{\Delta V_0}{x} - \tau_1$$

$$\text{area}_1 = \left(\Delta V_0 - \frac{V_x}{2}\right)t_x .$$

Hence substituting for V_x and t_x,

$$\text{area}_1 = \frac{1}{2}\left(\frac{\Delta V_0^2}{x} - x\tau_1^2\right) .$$

For the linear region $t > t_x$,

$$\text{area}_2 = \int_{t_x}^{\infty} (\Delta V_0 - v)\, dt$$

where, substituting for v and integrating,

$$\text{area}_2 = \Delta V_0\tau_1 e^{-(t_x-t_y)/\tau_1} .$$

But, $\Delta V_0 e^{-(t_x-t_y)/\tau_1} = x\tau_1$. Hence,

$$\text{area}_2 = x\tau_1^2 .$$

Since $\text{area}_1 \geq 0$ for $\Delta V_0 \geq x\tau_1$, then the total area SA is either

$$\text{SA}\Big|_{\Delta V_0 \geq x\tau_1} = \frac{\Delta V_0^2}{2x} + \frac{x\tau_1^2}{2} \qquad (17)$$

or

$$\text{SA}\Big|_{\Delta V_0 < x\tau_1} = \Delta V_0\tau_1 . \qquad (18)$$

The expressions for total area enable the distortion to be estimated as a function of intersample pulse difference ΔV_0, slew rate x and linear rise time constant τ_1. The distortion is therefore dynamic and requires a one-sample delay to estimate the sample difference ΔV_0. Since for $\Delta V_0 > x\tau_1$ the area is related to ΔV_0^2, this error is nonlinear, while for $\Delta V_0 < x\tau_1$ the area is proportional to ΔV_0 and is linear, hence is of little effect.

However, the time spent slewing compared with a linear rise is an important characteristic in determining degradation due to rise-time limitation. A parameter $\eta_{\%}$ is therefore introduced to represent the ratio of slew-rate-limited rise time t_x to overall rise time t_x +

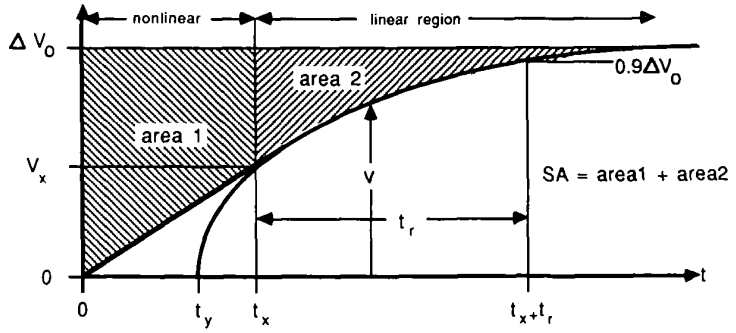


Fig. 11. Representation of rectangular pulse with both slew-rate and linear rise-time limitations.

t_r (that is, time to $0.9\Delta V_0$),

$$\eta_{\%} = \frac{100t_x}{t_x + t_r} . \quad (19)$$

The overall program flowchart prepared by McCrea [42] is shown in Fig. 12 and outlines the system strategy. Input data include sampling ratio, noise-shaper order, and nonlinear DAC errors, which can be included either individually or in composite form. Output data are presented as time-domain output sequences, frequency-domain plots, SNR, histogram of noise-shaper output, and DAC error data.

In Sec. 3 the program is used to generate a number of example output data in order to quantify the noise-shaper/DAC performance for a range of conditions applicable to digital audio systems.

3 COMPUTER-GENERATED RESULTS

The results in this section offer an image of the performance potential of a noise-shaped DAC with high oversampling. To commence this presentation, the time-domain behavior is illustrated where the following basic

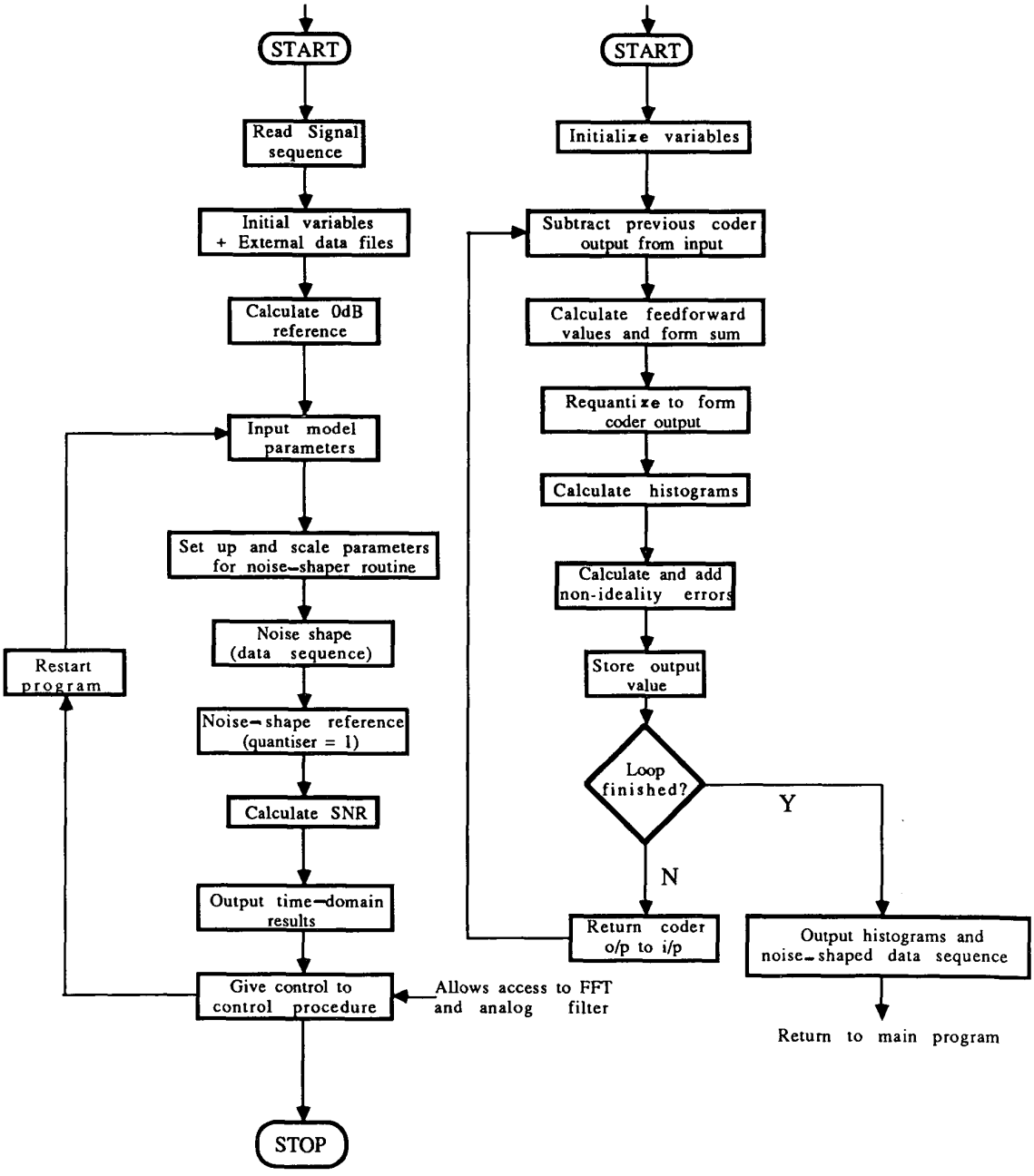


Fig. 12. Flowchart of computer program (after McCrea [42]). (a) Main program. (b) Noise-shaping procedure.

data are used:

Input sine-wave frequency = 8 kHz

O-dB input corresponds to amplitude of $\sqrt{2}$ (where output quantum is normalized to 1)

Noise-shaper order $R = \{1, 2, 3, 4\}$

Output sampling rate $f_{s2} = 5.12$ MHz (corresponding to $R_N = 128$ ref $f_{s1} = 40$ kHz).

In Fig. 13 four time-domain sequences are shown, corresponding to $R = \{1, 2, 3, 4\}$. The key factor to observe is the increase in quantizer activity where for $R = 4$ the quantizer range spans about -8 to $+8$, corresponding to a 4-bit DAC. The $R = 4$ raster exhibits chaotic behavior with an apparent random or noiselike structure and is a by-product of the noise shaper with both recursion and discontinuous nonlinearity. An essential characteristic is that the quantizer range is now much greater than the peak input signal level, such that all levels over a range of -8 to $+8$ on average participate in the conversion process.

Also, by restricting the peak input signal to be close to an output quantum, the chaotic activity in the large-scale sense exhibits only minor changes with the input signal level. Hence even for low-level signals, virtually all DAC levels participate in the conversion process.

To demonstrate the statistical distribution of output levels, Fig. 14 shows a histogram of quantizer activity for $R = \{1, 2, 3, 4\}$, where the progression of activity with increasing order can be observed.

The chaotic activity acts as an intelligent dither [30], [31] or self-dither [2]. The range of activity can be estimated approximately from the linear analysis of Sec. 1, where Eq. (9) reveals an effective gain of 2^R at $f = f_{s2}/2$ Hz. Hence if we assume that the quantization distortion of the noise-shaper quantizer $Q(z)$ spans ± 0.5 (that is, a quantum is normalized to 1), then the approximate DAC activity spans a range $\pm Q_R$, where

$$Q_R = 2^{R-1} . \tag{20}$$

For the case $R = 4$ then $Q_R = 8$, that is, the range of chaos $\cong \pm 8$. Eq. (20) is useful as it estimates the size of DAC required and relates it to the noise-shaper order, hence providing the input-signal peak value $\cong < 1.5$. Then

$$\text{minimum DAC size} = R \text{ bits} . \tag{21}$$

Although the DAC output reveals high-level activity,

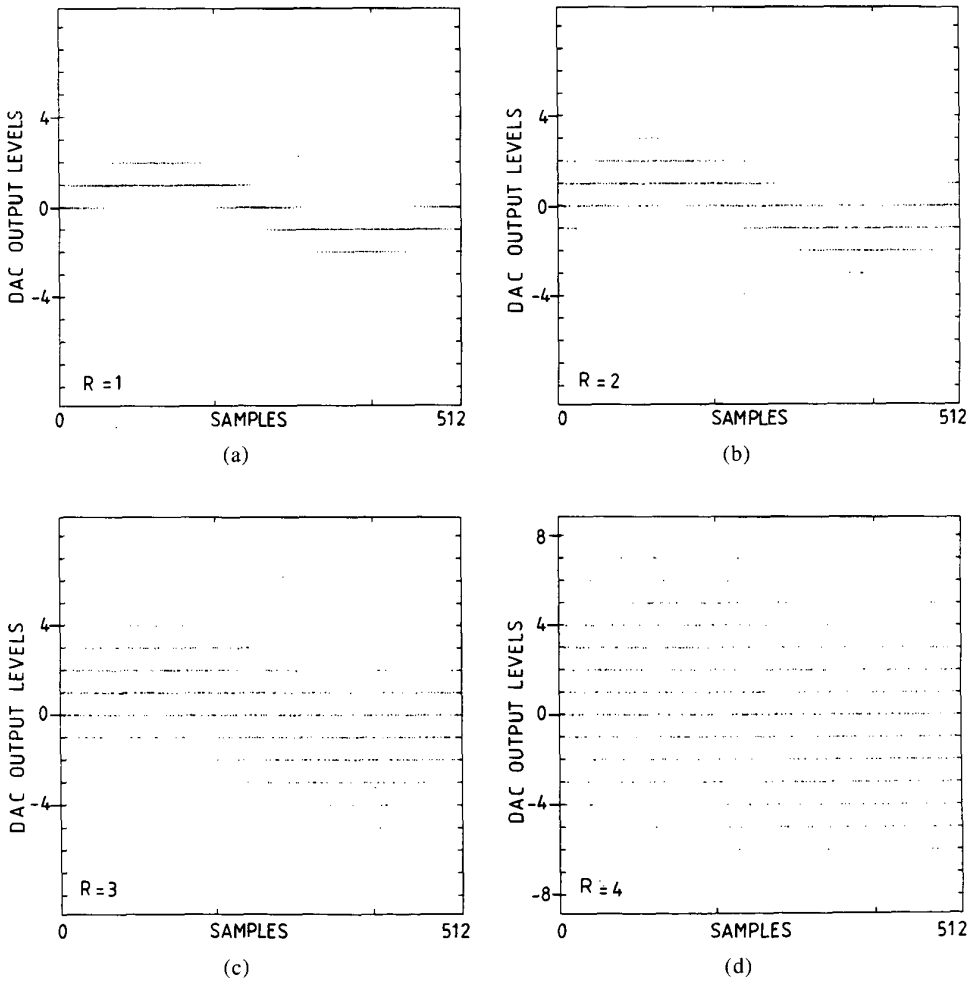
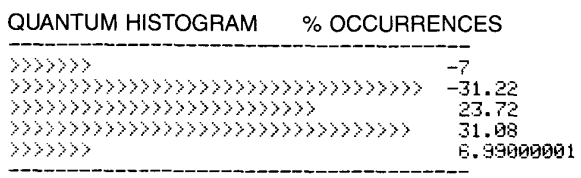
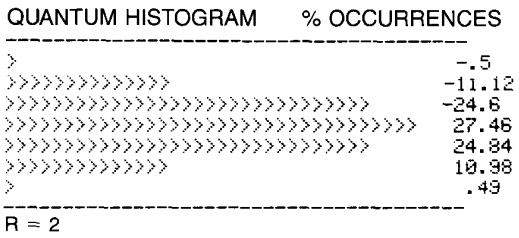
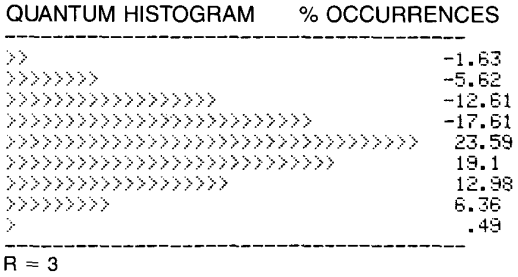


Fig. 13. Noise-shaper time-domain outputs for $R = \{1, 2, 3, 4\}$. Input 0 dB (corresponding to $\pm\sqrt{2}Q$); $f_{s2} = 5 \cdot 12$ MHz.

when this signal is band-limited, excellent low-level coding accuracy of the baseband signal 0 to $f_{s1}/2$ Hz is achieved (for $f_{s2} >> f_{s1}$). As an example, Fig. 15 shows two time-domain examples where a 2-kHz input signal set to −126 dB and specified to full computer precision (that is, not quantized to 16 bits) is applied to both an $R = 3$ and an $R = 4$ noise shaper, with $f_{s2} = 8$ MHz. The output sequence of an ideal DAC is band-limited by a 20-kHz, coincident-pole filter of order 10, where for $R = 3$, SNR = 132 dB and for N = 4, SNR = 157 dB (re. 0-dB input signal level). However, $R = 3$ shows the noise shaper to exhibit significant HF noise, while for $R = 4$ reconstruction is excellent and shows the potential for converting signals to > 20 bits.







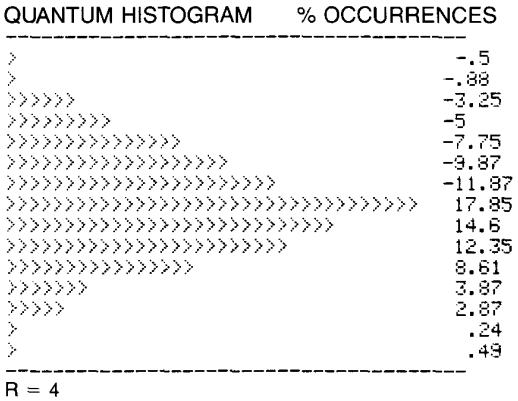


Fig. 14. Noise-shaper histogram distributions. Input −20 dB, 20 kHz, $R_N = 200$, $R = 1, 2, 3, 4$.

However, at this juncture the DAC is assumed perfect. The next stage is to introduce DAC nonlinearity and to assess the effect upon overall performance, where time-domain modeling allows an exact estimation of output distortion.

In observing the following results, it is emphasized that the DAC operates both open loop and noninteractively with the noise-shaper coder. Thus by using an FFT routine performed on the output data sequence that is derived using a dedicated transient analysis procedure, it is possible to assess the magnitude and spectral distribution of the DAC distortion.

3.1 Ideal DAC

Fig. 16 shows four spectral plots for $R = \{1, 2, 3, 4\}$, where the input is a 0-dB 8-kHz sine wave and $f_{s2} = 5.12$ MHz. The noise-shaping characteristic can be observed, together with enhancements for higher order noise-shaping filters. Since the slope of the noise-shaping characteristic at low frequency approximates $6R$ dB per octave, it follows that an $(R + 1)$th-order coincident-pole low-pass filter will cause the output spectrum to fall for a frequency greater than the filter cutoff, where the rate of attenuation approximates −6 dB per octave. Fig. 17 shows the same spectra as Fig. 16, but with the addition of the $(R + 1)$th-order low-pass reconstruction filter with a 3-dB cutoff of 30 kHz.

3.2 DAC with Static Error

Again, using a 0-dB 8-kHz sine-wave input the two sets of four spectral plots of Figs. 18 and 19 were calculated using the static error example of Fig. 9. The

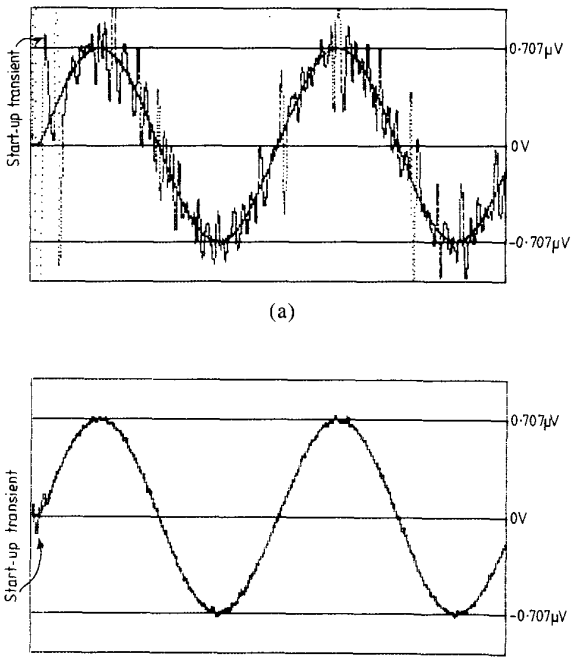


Fig. 15. Time-domain examples of low-level coding performance. Input −126 dB, 2 kHz, $R_N = 200$. (a) $R = 3$. (b) $R = 4$.

results show that for $R > 3$, all traces of second-harmonic distortion have been translated to a benign, noiselike residue, provided the input was limited in level to less than 1.5 quanta peak. If the input had exceeded this level, then the chaotic loop behavior would ride with the signal over a greater quantizer range, whereby some high-level artifacts of DAC nonlinearity are not decorrelated. It is a condition for decorrelation that the input signal is restricted in amplitude. It also eliminates the need for a DAC with $N > 4$ for the $R = 4$ application. In this sense, the results should be contrasted against the work of Adams [2] in ADC systems, where loop chaos is limited by a more conservative design of loop forward-path processor and where high-level signals are expected to traverse a significant range of the 4-bit (or, later, 6-bit) quantizer. Of course, in ADC applications, the DAC is enclosed within the noise-shaper feedback path, which sets more demanding stability criteria, possibly preventing an optimum noise shaper, as proposed by Tewksbury and Hallock [53], from being realized.

To further explore the static DAC distortion, Fig. 20 shows the spectrum of the DAC error for $R = 4$ over a band 0 to $f_{s2}/2$ Hz. An interesting feature of this result is the noise shaping of the open-loop DAC error waveform that, in part, follows the noise-shaper output spectrum. However, a separate study has revealed that this shaping relates only to the sine-function component of the nonlinearity of Fig. 9, while if the spectrum due to random level perturbations is analyzed in isolation, a near-flat DAC error spectrum emerges. The DAC error spectrum can therefore discriminate against large-scale and small-scale DAC nonlinearity.

3.3 Glitch Distortion

Fig. 21 presents spectra for a binary weighted and a parallel DAC, showing both DAC error spectra and overall noise. Results are presented only for $R = 4$, where the peak delay is set at 150 ps. The main conclusion is that no harmonic distortion is evident and that distortion levels are low and decorrelated.

3.4 Rise Time and Slew Rate

Example results for linear rise time and slew rate are shown in Fig. 22, where each plot includes the spectrum of the nondistorted noise-shaper output for $R = 4$ as reference. For the simulation, the slew rate was fixed and made equal to the maximum slope of an

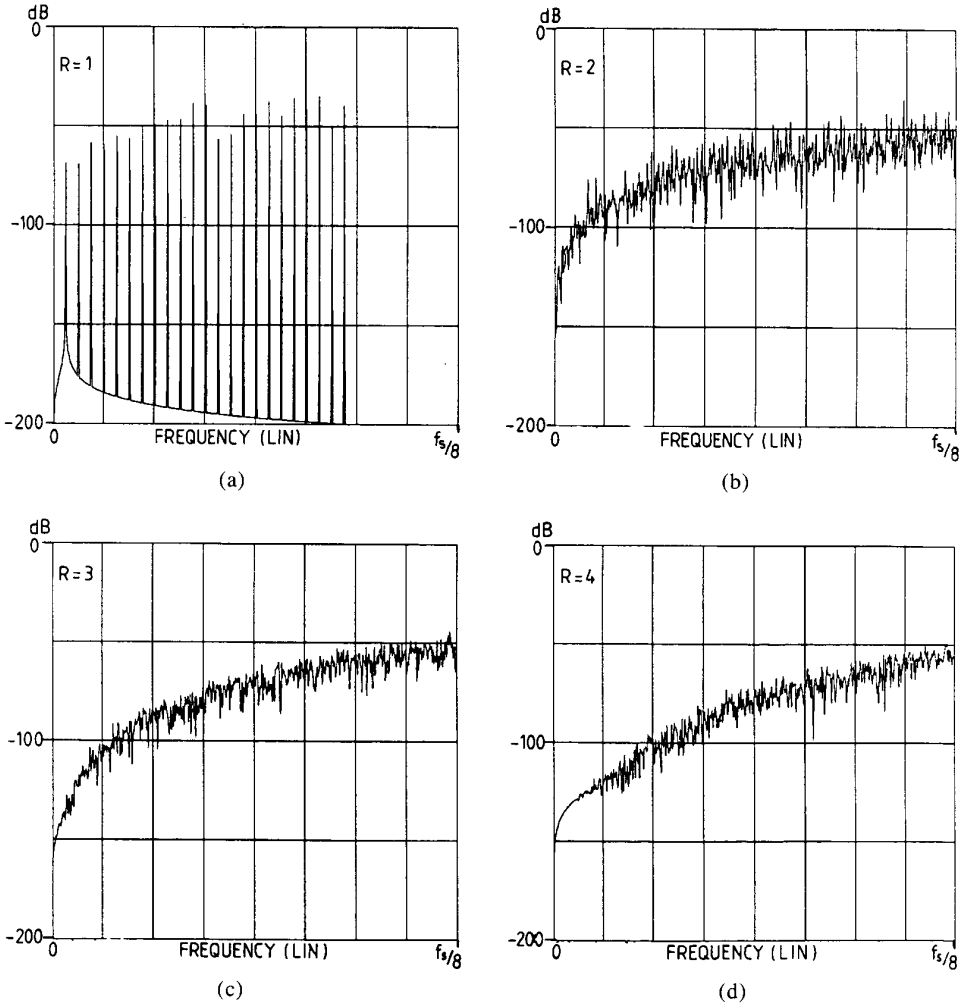


Fig. 16. Output noise spectrum for ideal DAC. Input 0 dB, 8 kHz, $f_{s2} = 5 \cdot 12$ MHz, $R = 1, 2, 3, 4$.

exponential function where the rise time is 170 ns (in a 195-ns period) and the amplitude transition is 16 V (assuming $N = 4$ bits and a noise-shaper quantum = 1 V). Hence if the corresponding time constant is τ_{lm}, then the slew rate is $16/\tau_{lm}$. For the linear DAC error spectrum of Fig. 22(a) the time constant is set to τ_{lm} to prevent slew limiting. Consequently, only a mild error occurs, whose spectrum follows closely that of the noise shaper. However, by setting the time constant $\tau_l = 0$, yet maintaining the slew rate at $16/\tau_{lm}$, each data transition exhibits slew rate and produces the DAC error spectrum of Fig. 22(b). The DAC error spectrum now reaches a greater level, $\cong -70$ dB, and takes a noise shape similar to that of the static sine function. The spectrum is replicated in Fig. 22(d), though here the frequency band is restricted to 0–100 kHz.

To explore the sensitivity of DAC errors as a function of the blend of linear rise time and slew-rate limit, as described in the model of Sec. 2.3, the 0–20-kHz integrated DAC error is plotted in Fig. 22(c) against $\eta_{\%}$, defined by Eq. (19), where the slew rate is maintained at $16/\tau_{lm}$, but the time constant τ_l is progressively adjusted over a range $0 \leqslant \tau_l \leqslant \tau_{lm}$. The curve reveals a rapid transition in DAC error for $15\% < \eta_{\%} < 22\%$ and is a result of the statistics of ΔV_0, which has a greater probability of lower levels.

Although significant degradation can result from slew-induced distortion, the DAC error spectra show no trace of harmonic distortion. It is therefore concluded that decorrelation is complete and that hardware nonlinearities produce only noiselike residues. We also observe a common factor that, where modest perturbations are introduced to the noise-shaper output, whether the source is static or dynamic, the chaotic activity of the noise-shaper loop, together with the restriction of the peak input signal being smaller than 1.5 quanta, result in substantial error decorrelation that appears complete for $R \geqslant 4$.

4 THEORETICAL ESTIMATE OF NOISE-SHAPER IN-BAND NOISE CONTRIBUTION

The SNR for an Rth-order oversampled noise shaper is here estimated as a function of R_N and R. The analysis

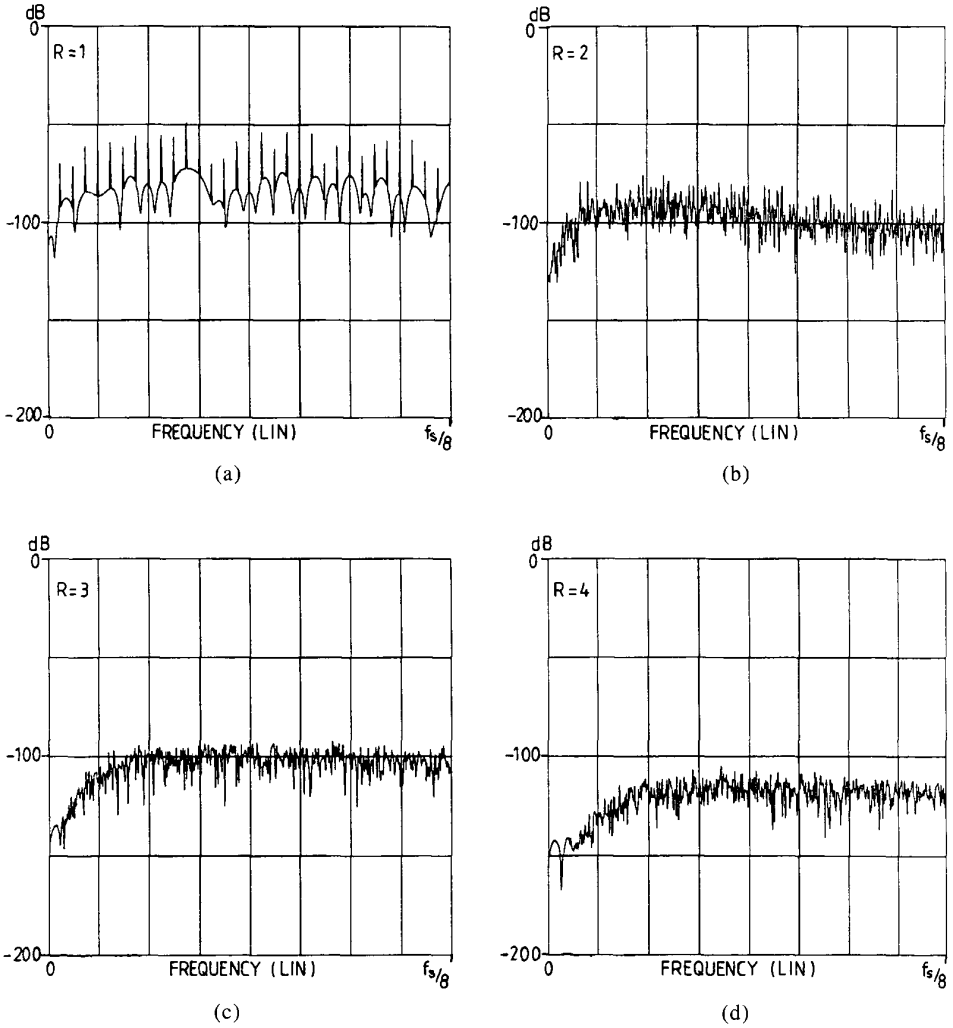


Fig. 17. As Fig. 16, but with coincident pole filter of order $R + 1$, 30-kHz bandwidth, Input 0 dB, 8 kHz, $f_{s2} = 5 \cdot 12$ MHz, $R = 1, 2, 3, 4$.

assumes that the only noise source is the noise-shaper quantizer and that the noise is integrated over a band 0 to $f_{s2}/2R_N$ Hz. Examination of the histogram distribution of Fig. 14 reveals approximately a linear function that spans a range closely matched to Q_R given by Eq. (20). The analysis therefore assumes a linear distribution, where Fig. 23 illustrates the idealized histogram for $R = 3$.

From this linear histogram, the total noise power N_t is determined for a general loop order R as

$$N_t = 2 \sum_{r=1}^{2^{R-1}} r^2 \left[\frac{2^{R-1} + 1 - r}{(2^{R-1} + 1)^2} \right] . \quad (22)$$

The in-band noise N_a is estimated from a knowledge of the distortion reduction factor $|D_R(f)|$ defined by Eq. (7), where if k_R is a spectral weighting factor for an Rth-order noise shaper, then

$$N_a = k_R \int_0^{f_{s2}/2R_N} \left[2 \sin \left(\frac{\pi f}{f_{s2}} \right) \right]^{2R} df . \quad (23)$$

For large R_N, $\sin(\pi f/f_{s2}) \approx \pi f/f_{s2}$, where by integrating,

$$N_a = \frac{\{k_R f_{s2}\}}{2(2R + 1)R_N} \left(\frac{\pi}{R_N} \right)^{2R} . \quad (24)$$

The factor k_R is here estimated by equating the area $\{k_R f_{s2}\}$ to the area under the curve $\{N_t/f_{s2}\}|D_R(f)|^2$ taken over $0 < f < f_{s2}$, that is,

$$\{k_R f_{s2}\} = \frac{N_t}{f_{s2}} \int_0^{f_{s2}} \left[2 \sin \left(\frac{\pi f}{f_{s2}} \right) \right]^{2R} df$$

where, observing spectral symmetry about $f_{s2}/2$ Hz and using the recurrence formula for $\int_0^{\pi/2} \sin^n x \, dx$ for even n, then

$$\{k_R f_{s2}\} = 2^{2R} N_t \prod_{r=1}^{2R} \left(\frac{2r - 1}{2r} \right) . \quad (25)$$

Hence from Eqs. (22), (24), and (25), eliminating N_t and $\{k_R f_{s2}\}$, the in-band noise power N_a approximates

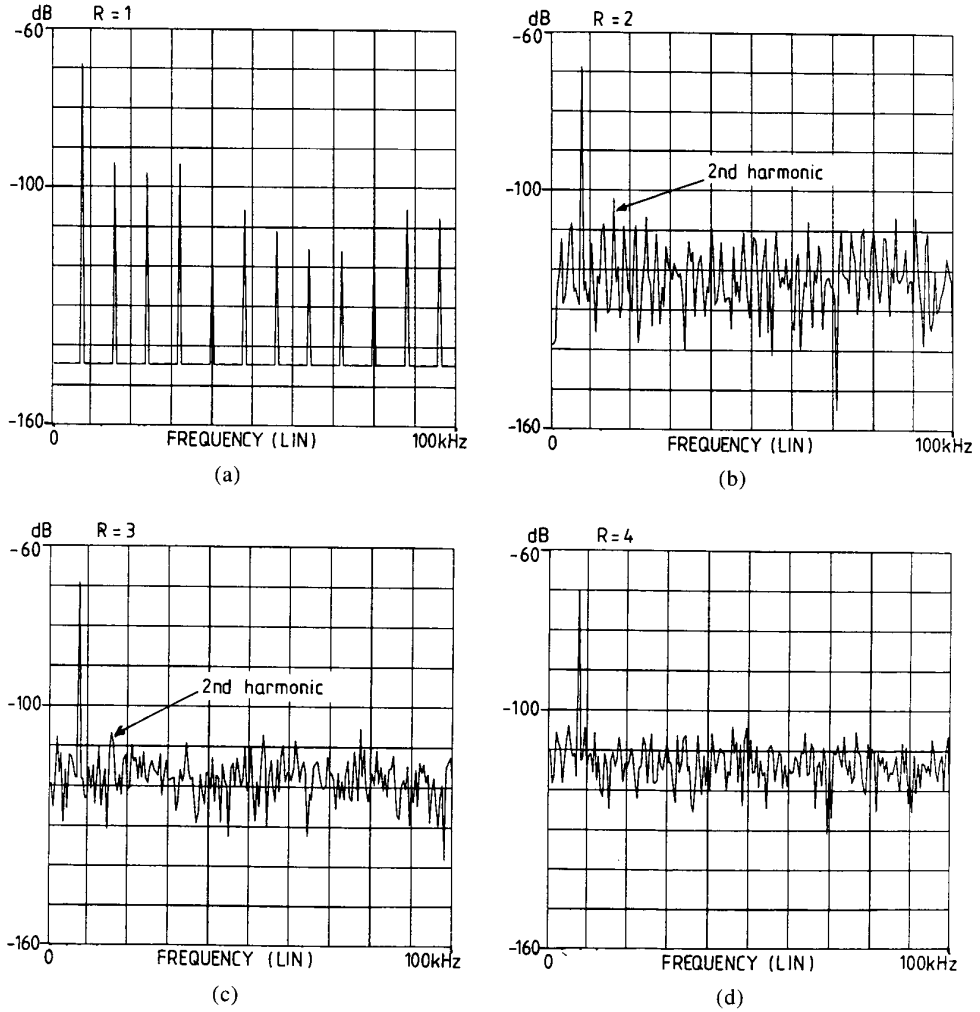


Fig. 18. DAC error spectrum using DAC error from Fig. 9. Input 0 dB, 8 kHz, $f_{s2} = 5 \cdot 12$ MHz, $R = 1, 2, 3, 4$.

to

$$N_a = \frac{1}{(2R + 1)R_N}\left\{\frac{2\pi}{R_N}\right\}^{2R}\sum_{r=1}^{2^{R-1}} r^2\left[\frac{2^{R-1} + 1 - r}{(2^{R-1} + 1)^2}\right]\prod_{r=1}^{2^{R-1}}\left(\frac{2r - 1}{2r}\right). \quad (26)$$

As a measure of the accuracy of Eq. (26), a range of results was estimated both by evaluating Eq. (26) and by computer simulation using the error-difference technique described in Sec. 2. The data in Table 1 were generated using a sinusoidal input level of −20 dB with a 20-kHz frequency, where 0 dB corresponds to a sine-wave amplitude of $\sqrt{2}$ (re the output quantum of unity) and SNRs are presented with reference to 0 dB. The results demonstrate reasonable agreement, particularly for $R = 4$, where the histogram distributions are more representative of the idealized form shown in Fig. 23.

To further explore the fundamental distortion of the idealized noise shaper that excludes the effect of source data quantization and DAC errors, a range of in-band SNR was computed, and the results are shown in Fig. 24. The SNRs, although optimistic for a practical system, do reveal the potential enhancements in resolution as R_N and R are increased simultaneously. Interestingly, for relatively low orders of oversampling (that is, $R_N = 25$), the SNR curves show a clustering as R is increased. It is not until $R_N \geq 50$ that significant enhancements in target SNR are achieved for increases in R, indicating that high-order noise shapers should only be used with high oversampling ratios.

5 CONCLUSION

This paper has attempted to demonstrate the potential performance of a high-order noise shaper (for example, $R = 3, 4$) used in association with high oversampling (for example, $R_N \cong 200$). The principal advantages have been shown to be a reduction in the number of

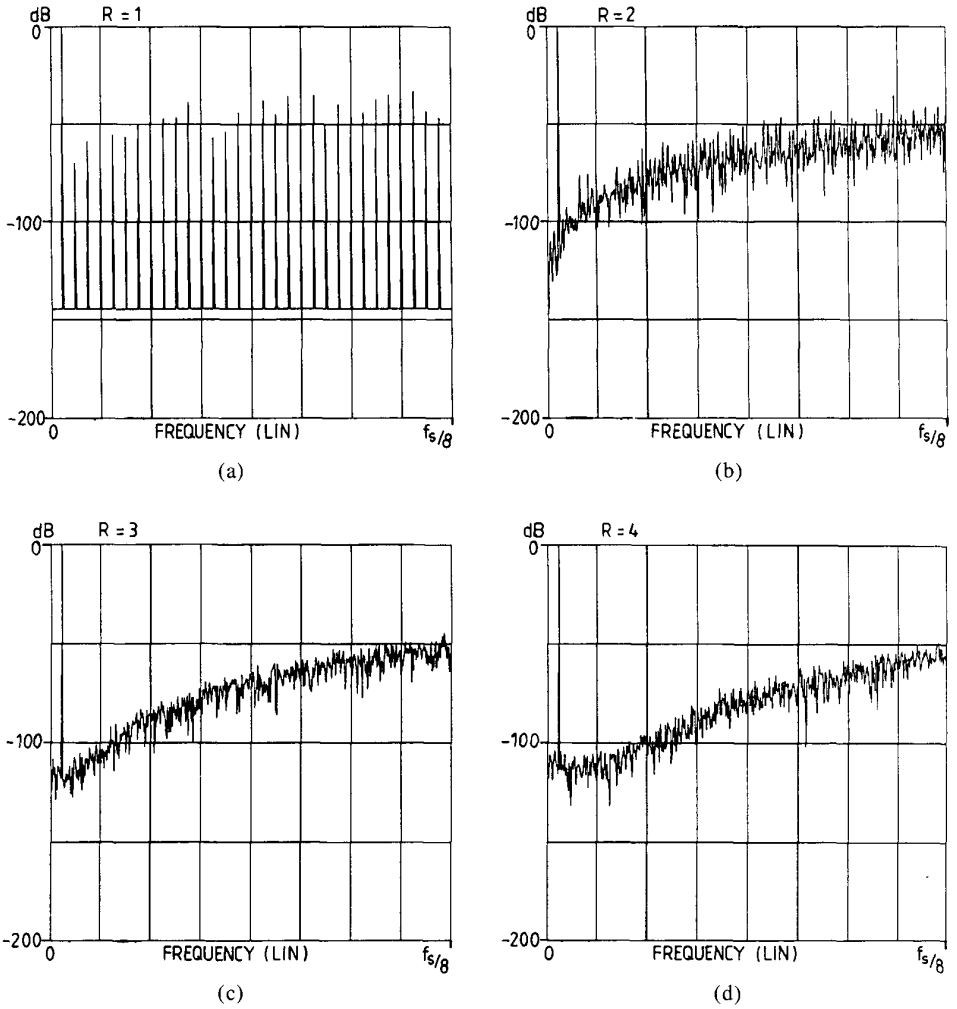


Fig. 19. As Fig. 18, but results are for overall noise at output of DAC/noise shaper. Input 0 dB, 8 kHz, $f_{s2} = 5 \cdot 12$ MHz, $R = 1, 2, 3, 4$.

levels required by the DAC, where typically 16 levels result for $R = 4$, together with a decorrelation of DAC nonidealities to a noiselike error residue. The decorrelation was shown to be a consequence of chaotic loop behavior where the output raster of the DAC spanned a range of levels so that, on average, all levels participated in the conversion process with large-scale statistics virtually independent of input signal, provided the peak signal level did not exceed about 1.5 output quanta.

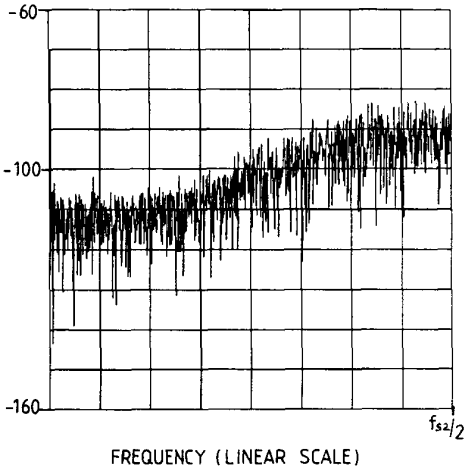


Fig. 20. DAC error spectrum (using Fig. 9 nonlinearity). $R = 4$.

A noise-shaper topology was developed that required no internal multiplications and could match the requirements of the optimal noise shaper proposed by Tewksbury and Hallock [53]. Since the machine is digital, it does not suffer the response limitations inherent in oversampled/noise-shaped ADC structures, which require a more modest design of loop filter and thus show greater nonlinear dependence on DAC distortion, which is an intrinsic part of the feedback loop.

The results are representative of a hardware system as a number of imperfections were accurately modeled, including errors both with and without memory. The results indicated that, although only 16 DAC levels are required for an $R = 4$ system, these should be specified to high accuracy (>12 bits) if low in-band noise is required. However, the limited number of levels allows the use of parallel-architecture DACs, using, for ex-

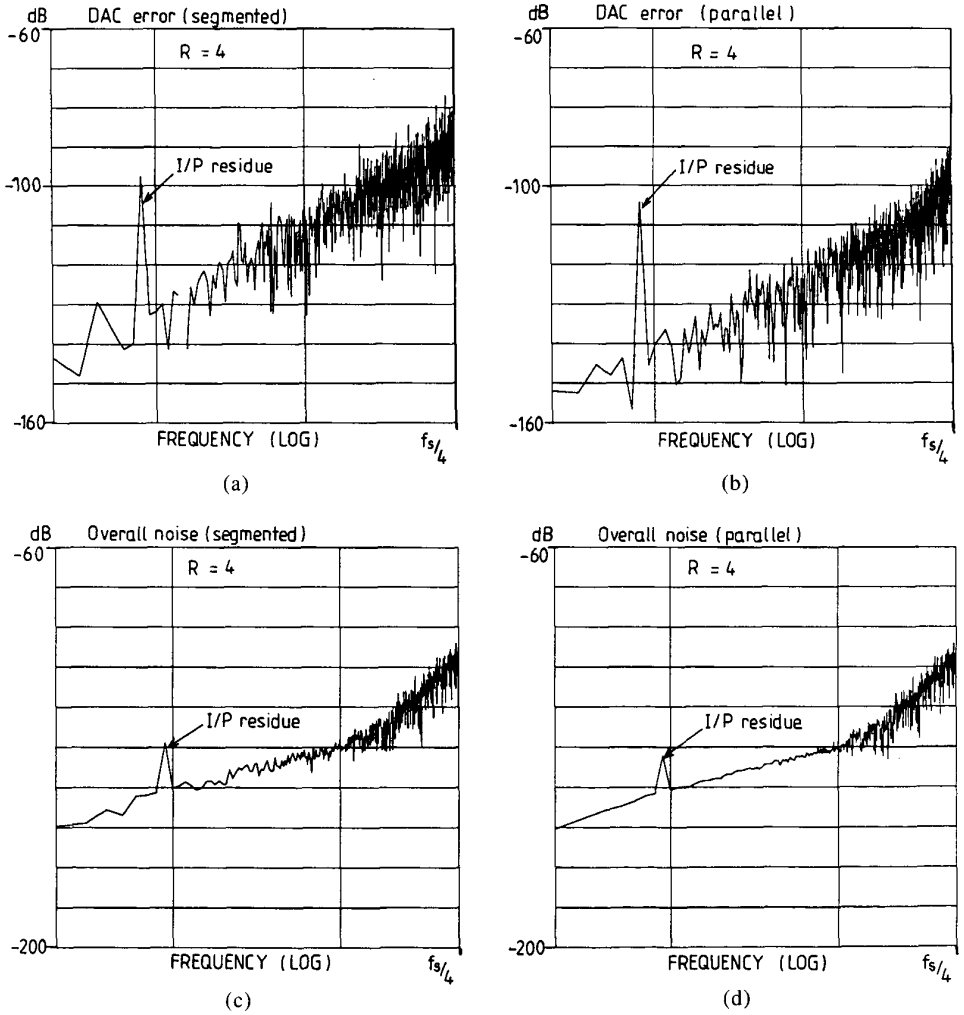


Fig. 21. Error spectra for glitch distortion for both parallel and segmented DACs. Actual maximum glitch area—parallel 1240 pV · s, segmented 770 pV · s. Maximum delay 150 ps, input 0 dB, 8 kHz, $f_{s2} = 5 \cdot 12$ MHz.

ample, either summing resistors of equal value or precision-switched current sources.

A principal feature of the technique is the minimal analog processing required in signal recovery, where computer results reveal that a coincident pole filter of order $R + 1$ will achieve an out-of-band attenuation of -6 dB per octave when mapped against the noise-shaped spectrum, which rises approximately as $6R$ dB per octave for an Rth-order system.

Simulations suggest that with an appropriate design of the DAC to minimize both static and dynamic errors (such as slew rate), a system resolution greater than 16 bits is achievable. Indeed, since the noise-shaper architecture is digital, its dynamic range can readily exceed 20 bits, although, in practice, the overall performance will be degraded by DAC errors. However, due to chaos, the DAC distortion translates to benign noise, and since the DAC is external to the noise shaper, its error is simply modeled as a low-level additive noise source.

The study did not include oversampling filters, as the aim was to investigate the distortions due only to the noise shaper and the DAC. It is intended, however, to extend the work to include oversampling and to incorporate a number of filter architectures, although it is anticipated that provided high-frequency spuriae are

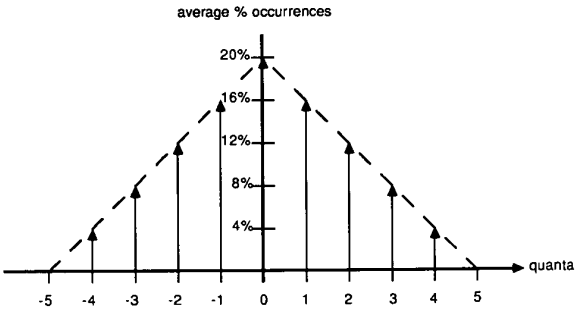


Fig. 23. Idealized histogram of quantizer activity for $R = 3$.

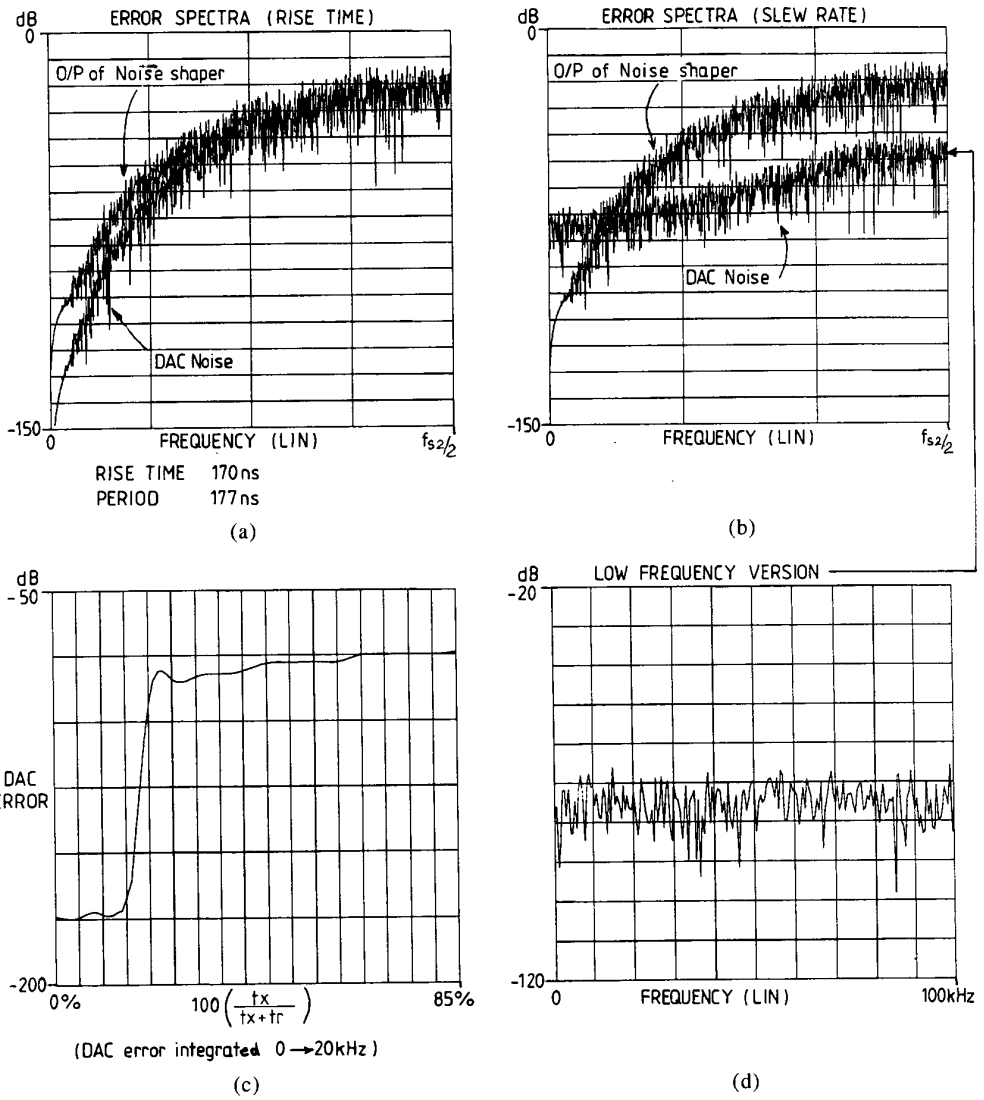


Fig. 22. Slew-rate and rise-time distortions. Input 0 dB, 8 kHz, $f_{s2} = 5 \cdot 12$ MHz.

minimized, overall degradation should be small. Evidence for the effect of out-of-band input transients is shown in Fig. 15, where the rapid change of the first derivative at the commencement of the sinusoidal input reveals a mild transient in the noise-shaper response. However, this soon decays and is a consequence of an input-signal condition that does not occur in practice, as an oversampling filter would smooth the input sequence.

Simulations also revealed the necessity for the noise shaper to have an adequate quantizer range, where quantizer saturation could instigate entry into a non-recoverable, nonlinear oscillatory mode. In practice, the input signal to the noise shaper is bounded in both amplitude and bandwidth. Consequently there is no danger of entry into such a state. However, for switch-on or some other form of transient misbehavior/interference, the system should incorporate an overload/oscillation detector so that reinitialization can be achieved. Such a scheme could incorporate a noise-shaper quantizer with a range in excess of that required by normal chaotic loop activity, whereby if an out-of-range signal is detected, all noise-shaper integrators are momentarily reset to zero, possibly in association with a short-duration (circa 0.1 ms) soft output mute.

Table 1. Comparison of theoretical and computed SNR (re 0-dB input level) for idealized noise-shaping filter computed with 20-kHz input signal at −20 dB (re $\sqrt{2}$ amplitude).

R_N	R	SNR, Eq. (26), (dB)	Computer Simulation (dB)
50	1	50	49
50	2	67	68
50	3	83	87
50	4	97	103
100	1	59	53
100	2	82	82
100	3	104	115
100	4	124	134
200	1	68	62
200	2	98	98
200	3	125	130
200	4	152	156

6 ACKNOWLEDGMENT

The author wishes to offer his appreciation to the discussions with and contributions by Timothy Darling, Li Mu, Basil McCrea, and Wolfgang Wingerter during their study periods within the Audio Research Group at Essex University.

7 REFERENCES AND COMPLEMENTARY READING

7.1 Deltamodulation, Delta–Sigma Modulation, Noise-Shaping ADC/DAC

[1] R. W. Adams, "Companded Predictive Delta Modulation: A Low-Cost Conversion Technique for Digital Recording," *J. Audio Eng. Soc.*, vol. 32, pp.

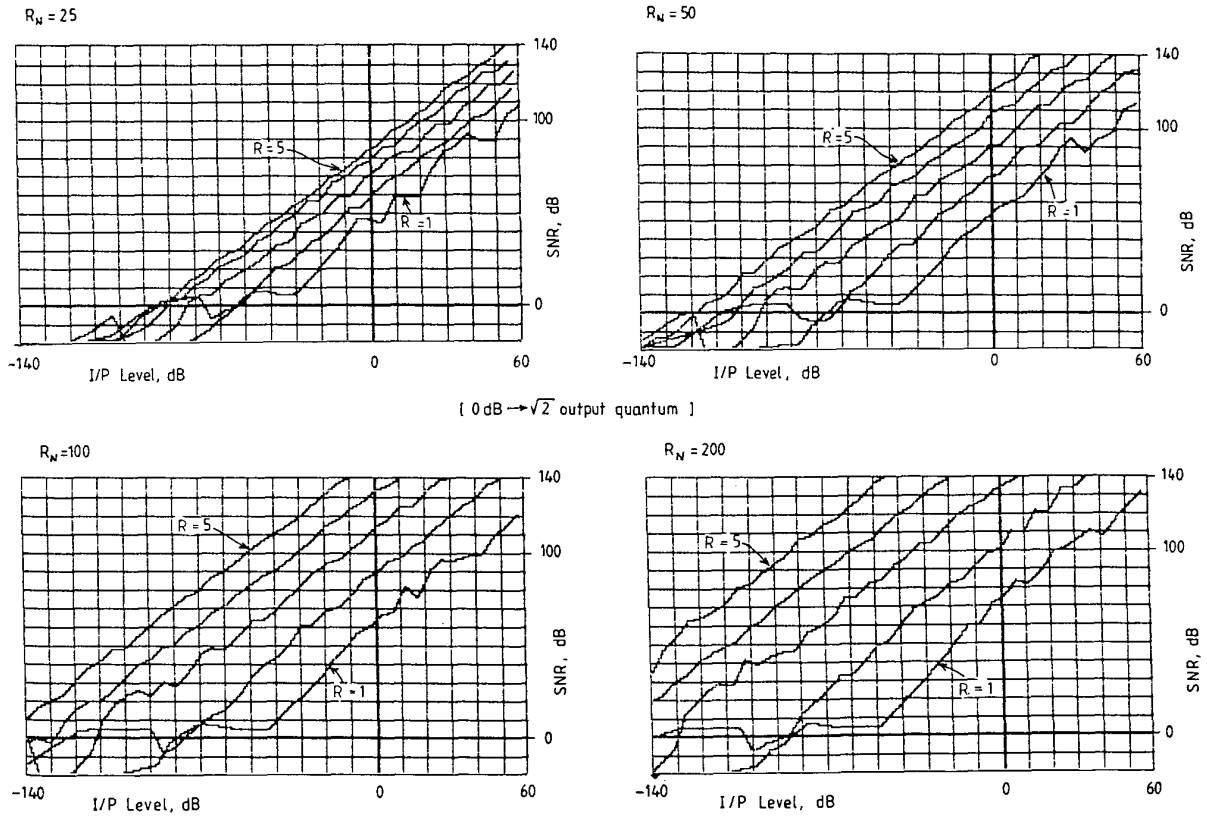


Fig. 24. SNR versus input level for noise shaper with perfect DAC. R = 1 to 5, R_N = 25, 50, 100, 200.

659–672 (1984 Sept.).

[2] R. W. Adams, "Design and Implementation of an Audio 18-Bit Analog-to-Digital Converter Using Oversampling Techniques," *J. Audio Eng. Soc.*, vol. 34, pp. 153–166 (1986 Mar.).

[3] B. P. Agrawal and K. Sheno, "Design Methology for Sigma Delta Modulation," *IEEE Trans. Commun.*, vol. COM-31 (1983 Mar.).

[4] G. Bars and J. P. Petit, "High Quality Sound Decoder Using Noise Shaping Techniques and Digital Filters," presented at the IEEE ASSP Workshop, Monk Inn, 1986 Sept.

[5] R. A. Belcher et al., "Digital Sound; an Investigation of Delta-Modulation/Pulse-Code-Modulation Analogue-to-Digital Conversion," BBC Res. Rep. RD 1980/3.

[6] B. Blesser, "Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice," *J. Audio Eng. Soc.*, vol. 26, pp. 739–771 (1978 Oct.).

[7] B. Blesser, "Advanced Analog-to-Digital Conversion and Filtering: Data Conversion," in *Digital Audio, Collected Paper from the AES Premiere Conf.* (Rye, NY, 1982 June), pp. 37–53.

[8] B. Blesser, B. Locanthi, and T. G. Stockham (Eds.), *Digital Audio, Collected Papers from the AES Premier Conf.* (Rye, NY, 1982 June 3–6).

[9] B. E. Boser, K. P. Karmann, H. Martin, and B. A. Wooley, "Simulating and Testing Oversampled Analog-to-Digital Converters," *IEEE Trans. CAD*, vol. 7, p. 668 (1988 June).

[10] J. C. Candy, "A Use of Double Integration in Sigma Deltamodulation," *IEEE Trans. Commun.*, vol. COM-33, pp. 249–258 (1985 Mar.).

[11] J. C. Candy, "Decimation for Sigma Delta Modulation," *IEEE Trans. Commun.*, vol. COM-34, pp. 72–76 (1986 Jan.).

[12] K. W. Cattermole, *Principles of Pulse Code Modulation* (Iliffe, 1969).

[13] T. A. C. M. Claasen, W. F. G. Mecklenbrauker, et al., "Signal Processing Method for Improving the Dynamic Range of A/D and D/A Converters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28 (1980 Oct.).

[14] T. F. Darling, "Oversampled Analogue–Digital Conversion for Digital Audio Systems," M.Phil. thesis, submitted to University of Essex, UK, 1987.

[15] T. F. Darling and M. O. J. Hawksford, "Oversampled Analog-to-Digital Conversion Systems for Digital Audio," presented at the 85th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 36, p. 1034 (1988 Dec.), preprint 2740.

[16] F. de Jager, "Deltamodulation, a Method of PCM Transmission Using 1-Unit Code," *Philips Res. Rep.*, vol. 7, pp. 442–466 (1952).

[17] E. M. Deloraine, Van Mierlos, and Derjavitch, "Méthodes et système de transmission par impulsions," French patent 932.140, 1947/48.

[18] J. D. Everard, "Improvements to Delta-Sigma Modulators When Used for PCM Encoding," *Electron. Lett.*, vol. 12, no. 15 (1976 July).

[19] L. D. Fielder, "Evaluation of the Audible Distortion and Noise Produced by Digital Audio Converters," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, pp. 379, 380 (1987 May), preprint 2424.

[20] J. E. Flood and M. J. Hawksford, "Exact Model for Deltamodulation Processes," *Proc. IEE (London)*, vol. 118, p. 115 (1971).

[21] D. M. Freeman, "Slewing Distortion in Digital-to-Analog Conversion," *J. Audio Eng. Soc.*, vol. 25, pp. 178–183 (1987 Apr.).

[22] N. H. C. Gilchrist, "Analog-to-Digital and Digital-to-Analog Converters for High-Quality Sound," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 372 (1980 May), preprint 1583.

[23] J. Gleick, *Chaos: Making a New Science* (William Heinemann Ltd. ISBN 0434 29554X, 1988).

[24] D. Goedhart, R. J., Van de Plassche, and E. F. Stikvoort, "Digital to Analogue Conversion in Playing a Compact Disc," *Philips Tech. Rev.*, vol. 40, pp. 174–179 (1982).

[25] D. J. Goodman, "The Application of Delta Modulation to Analogue-to-PCM Encoding," *Bell Sys. Tech. J.*, vol. 48, pp. 321–343 (1969 Feb.).

[26] J. A. Greefkes and F. de Jager, "Continuous Deltamodulation," *Philips Res. Rep.*, p. 233 (1968).

[27] R. Greenfield and M. O. J. Hawksford, "Efficient Filter Design for Loudspeaker Equalization," presented at the 86th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 37, p. 394 (1989 May), preprint 2764.

[28] M. J. Hawksford, "Unified Theory of Digital Modulation," *Proc. IEE (London)*, vol. 121, pp. 109–115 (1974 Feb.).

[29] M. J. Hawksford, "Deltamodulation Coder Using a Parallel Realisation," in *IERE Conf. Proc.*, vol. 37, pp. 547–557 (1977 Sept.).

[30] M. J. Hawksford, "*N*th-Order Recursive Sigma–ADC Machinery at the Analog–Digital Gateway," presented at the 78th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, pp. 586, 588 (1985 July/Aug.), preprint 2248.

[31] M. J. Hawksford, "Oversampling and Noise Shaping for Digital to Analogue Conversion," *Reproduced Sound*, Inst. of Acoust., vol. 3, pp. 151–175 (1987).

[32] M. J. Hawksford, "Multi-Level to 1 Bit Transformations for Applications in Digital-to-Analogue Converters Using Oversampling and Noise Shaping," *Proc. Inst. Acoustic.*, vol. 10, pp. 129–143 (1988 Nov.).

[33] H. Inose and Y. Yasuda, "A Unity Bit Coding Method by Negative Feedback, *Proc. IEEE*, vol. 51, p. 1524 (1963 Nov.).

[34] H. Inose, Y. Yasuda, and J. Murakami, "A Telemetering System by Code Modulation–Delta Sigma Modulation," *IRE Trans. Space Electron. Telem.*, vol.

SET-8, p. 204 (1962 Sept.).

[35] J. E. Iwersen, "Calculated Quantisation Noise of Single-Integration Deltamodulation Coders," *Bell Sys. Tech. J.*, vol. 48, p. 2359 (1969 Sept.).

[36] N. S. Jayant and P. Noll, *Digital Coding of Waveforms* (Prentice-Hall, Signal Processing Series, Englewood Cliffs, NJ, 1984).

[37] A. Jongepier, "A D/A Converter Which Improves Dynamic Range by Oversampling, Slope Modulation and Slope Demodulation," in *Proc. Int. Conf. on Circuits and Systems (ISCAS)*, 1980.

[38] M. Kasug, "An Approach to High Resolution D/A Converters Utilizing Linear Predictive Coding," in *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing* (Tokyo, 1986 Apr.).

[39] R. R. Laane, "Measured Quantisation Noise Spectrum for Single-Integration Deltamodulation Coders, *Bell Sys. Tech. J.*, vol. 49, p. 159 (1970 Feb.).

[40] E. N. Lorentz, 'Deterministic Nonperiodic Flow," *J. Atmos. Sci. (Boston)*, vol. 20, pp. 130–141 (1963 Mar.).

[41] Y. Matsuya, "A 16-Bit Oversampling A-to-D Conversion Technology Using Triple-Integration Noise Shaping," *IEEE J. Solid-State Circuits*, vol. SC-22 (1987 Dec.).

[42] B. A. McCrea, "Simulation of Audio Digital-to-Analogue Conversion Using Noise Shaping and Oversampling," M.Sc. dissertation, Dept. of ESE, University of Essex, UK, 1987.

[43] D. Mitra, "Large Amplitude, Self-Sustained Oscillations in Difference Equations which Describe Digital Filter Sections Using Saturation Arithmetic," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-25, pp. 134–143 (1977 Apr.).

[44] P. J. Naus, E. C. Dijkmans, E. C., E. F. Stikvoort, A. J. McKnight, D. J. Holland, and W. Bradinal, "A CMOS Stereo 16-Bit D/A Converter for Digital Audio," *IEEE J. Solid-State Circuits*, vol. SC-22 (1987 June).

[45] P. T. Nielsen, "On the Stability of a Double Integration Delta Modulator," *IEEE Trans. Commun. Technol.*, pp. 364–366 (1971 June).

[46] M. Sandler, "Investigation by Simulation of a Digitally Addressed Audio Power Amplifier," Ph.D. dissertation, University of Essex, UK, 1983.

[47] M. Sandler, "Techniques for Digital Power Amplification," *Reproduced Sound* (Inst. of Acoust.), vol. 3, pp. 177–186 (1987).

[48] J. F. Schouten, F. de Jager, and J. A. Greefkes, "Deltamodulation, a New Modulation System for Telecommunications," (in Dutch), *Philips Tech. Tidjdschr.* vol. 13, p. 249 (1951 Sept.); (in English), *Philips Tech. Rev.*, vol. 13, p. 237 (1952 Mar.).

[49] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656 (1948).

[50] H. A. Spang and P. M. Schultheiss, "Reduction of Quantizing Noise by Use of Feedback," *IRE Trans. Commun. Sys.*, pp. 373–380 (1962 Dec.).

[51] R. Steele, *Deltamodulation Systems* (Pentech Press, ISBN 0727304011, 1975).

[52] P. Stritek, "Prospective Conversion Techniques for Improved Signal-to-Noise Ratio in Digital Audio Systems," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, p. 394 (1987 May), preprint 2477.

[53] S. K. Tewksbury and R. W. Hallock, "Oversampled Linear Predictive and Noise-Shaping Coders of Order $N > 1$," *IEEE Trans. Circuits Sys.*, vol. CAS-25, pp. 437–447 (1978 July).

[54] R. J. Van de Plassche, "Dynamic Element Matching for High Accuracy Monolithic D/A Converters," *IEEE J. Solid-State Circuits*, vol. SC-11, pp. 795–800 (1976 Dec.).

[55] A. Van De Plassche, "Sigma–Delta Modulator as an A/D Converter," *IEEE Trans. Circuits Sys.*, vol. CAS-25, pp. 510–514 (1978 July).

[56] R. J. Van de Plassche and E. C. Dijkmans, "A Monolithic 16-Bit D/A Conversion System for Digital Audio," in *Digital Audio, Collected Papers from the AES Premiere Conf.* (Rye, NY, 1982 June), pp. 53–60.

[57] H. van de Weg, "Quantising Noise of a Single-Integration Deltamodulation System with an N-Digit Code," in *Philips Res. Rep.*, vol. 8, p. 367 (1953).

[58] P. P. Wang, "An Absolute Stability Criterion for Delta Modulation," *IEEE Trans. Commun. Technol.*, pp. 186–188 (1968 Feb.).

7.2 Digital Filter Design

[59] R. Ansari, "Satisfying the Haar Condition in Halfband FIR Filter Design," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-36, pp. 123–124 (1988 Jan.).

[60] A. Antoniou, *Digital Filters: Analysis and Design* (McGraw-Hill, New York, ISBN0-07-002117-1, 1979).

[61] M. G. Bellanger, J. L. Daguet, and G. P. Lepagnol, "Interpolation, Extrapolation and Reduction of Computation Speed in Digital Filters," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-22, pp. 231–235 (1974 Aug.).

[62] R. E., Crochiere and L. R. Rabiner, "Optimum FIR Filter Implementation for Decimation, Interpolation and Narrow Band Filtering," *IEEE. Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, pp. 444–456 (1975 Oct.).

[63] R. E. Crochiere and L. R. Rabiner, "Interpolation and Decimation of Digital Signals—A Tutorial Review," *Proc. IEEE*, vol. 69, pp. 300–331 (1981 Mar.).

[64] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[65] A. J. Gibbs and L. R. Rabiner, "Techniques for Designing Finite-Duration Impulse-Response Digital Filters," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 188–195 (1971 Apr.).

[66] D. J. Goodman and M. J. Carey, "Nine Digital Filters for Decimation and Interpolation," *IEEE Trans.*

Acoust. Speech, Signal Process., vol. ASSP-25, pp. 121–126 (1977 Apr.).

[67] V. Hansen, "Design of a Multistage Decimation–Interpolation Filter," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing* (ICASSP), pp. 21.9.1–4, 1987.

[68] L. B. Jackson, *Digital Filters and Signal Processing* (Kluwer Academic Publ., Boston, 1986).

[69] E. G. Kimme and F. F. Kuo, "Synthesis of Optimal Filters for a Feedback-Quantisation system," *IEEE Trans. Circuit Theory*, pp. 405–413 (1963 Sept.).

[70] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," *IEEE Trans. Audio Electroacoust.* (Dec. 1973), see also *Programs for Digital Signal Processing* (IEEE Press, 1979) New York, pp. 5.1.1–13.

[71] F. Mintzer, "On Half-Band, Third-Band and *N*th-Band FIR Filters and Their Design," *IEEE Trans. Acoust., Speech, Signal Process*, vol. ASSP-30, pp. 734–738 (1982 Oct.).

[72] G. Oetken, T. W. Parks, and H. W. Schuessler, "New Results in the Design of Digital Interpolators," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23, pp. 301–309 (1975 June).

[73] G. Oetken, T. W. Parks, and H. W. Schuessler, "A Computer Program for Digital Interpolator Design," in *Programs for Digital Signal Processing* (IEEE Press, New York, 1979) pp. 8.1.1–6.

[74] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[75] A. V. Oppenheim et al., *Selected Papers in Digital Signal Processing*, vol. II (IEE Press, Reprint Series, New York 1975).

[76] T. W. Parks and J. H. McClellan, "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 189–194 (1972 Mar.).

[77] L. R. Rabiner, "Techniques for Designing Finite-Duration Impulse Response Digital Filters," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 188–195 (1971 Apr.).

[78] L. R. Rabiner, "The Design of Finite Impulse Response Digital Filters Using Linear Programming Techniques," *Bell Sys. Tech. J.*, vol. 51, pp. 1177–1198 (1972 July–Aug.).

[79] L. B. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[80] L. R. Rabiner and Ch. M. Rader, *Digital Signal Processing* (IEEE Press, Selected Reprint Series, New York, 1972).

[81] L. R. Rabiner and R. W. Schafer, "Recursive and Nonrecursive Realizations of Digital Filters Designed by Frequency Sampling Techniques," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 200–207 (1971 Sept.).

[82] D. W. Rorabacher, "Efficient FIR Filter Design for Sample Rate Reduction or Interpolation," in *Proc 1975 Int. Symp. on Circuits and Systems* (1975 Apr.).

[83] T. Saramaki and Y. Neuvo, "A Class of FIR Nyquist (*N*th-Band) Filters with Zero Intersymbol Interference," *IEEE Trans. Circuits Sys.*, vol. CAS-34, pp. 1182–1190 (1987 Oct.).

[84] R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," *Proc. IEEE*, vol. 61, pp. 692–702 (1973 June).

[85] K. S. Steiglitz, "Optimal Design of FIR Filters with Monotone Passband Response," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-27, pp. 643–649 (1979 Dec.)

[86] M. T. Sun and L. Wu, "On the Design of Digital Oversampling Filters," presented at the 81st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34 (1986 Dec.), preprint 2378.

[87] W. Sung and S. K. Mitra, "Implementation of Digital Filtering Algorithms Using Pipelined Vector Processors," *Proc. of IEEE*, vol. 75, pp. 1293–1302 (1987 Sept.).

[88] H. Urkowit, "Parallel Realizations of Digital Interpolation Filters for Increasing the Sampling Rate," *IEEE Trans. Circuits Sys.*, vol. CAS-22, pp. 146–154 (1975 Feb.)

[89] P. P. Vaidyanathan, "New Cascaded Lattice Structure for FIR Filters Having Extremely Low Coefficient Sensitivity," in *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing* (Tokyo, Japan, 1986 Apr.) pp. 497–450.

[90] P. P. Vaidyanathn and T. Q. Nguyen, "A 'Trick' for the Design of FIR Half-Band Filters," *IEEE Trans. Circuits Sys.*, vol. CAS-34, pp. 297–300 (1987 Mar.).

Dr. Hawksford's biography was published in the March issue.

PAPERS

Oversampling Filter Design in Noise-Shaping Digital-to-Analog Conversion*

M. O. J. HAWKSFORD AND W. WINGERTER**

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex CO4 3SQ, UK

A number of oversampling filters in association with a heavily oversampled and noise-shaped digital-to-analog converter (DAC) are presented for high-resolution applications in professional and consumer digital audio environments. Computer simulations of oversampling filter, noise shaper, and DAC errors are described, together with results to demonstrate interactive performance dependence on filter nonideality.

0 INTRODUCTION

The performance of a digital audio system that is bounded by uniform source quantization and sampling should result only in band limitation of the input signal, together with an additive noise component. The realization of this target is dependent on identifying analog-to-digital (ADC) and digital-to-analog conversion (DAC) systems that can attain near ideal anti-aliasing and signal recovery filters in association with a virtually perfect quantization characteristic and the means of decorrelating signal and quantization distortion so as to mimic the characteristics of purely additive noise.

In conventional ADC and DAC systems a principal limitation is low-level linearity of the quantization characteristics [1], where the least significant bit can reveal both relative and absolute errors in the converter. This level displacement prevents the resolution potential of the digital channel from being achieved, where, as signals approach and enter the quantization noise floor, the nonlinearity produces waveform distortion, rather than the ideal, where a signal sinks into the quantization noise without observable correlation.

To achieve this performance target requires the identification of hardward that can attain the linearity potential of the digital format, implement effective signal quantization decorrelation, and decorrelate the distortion arising from hardware nonlinearities so as to form a noiselike residue.

In an earlier paper [2], [3] a DAC was presented that used a high oversampling ratio in association with a noise shaper where, for order $R_N = 4$, a wide dynamic range was demonstrated that in the absence of DAC nonidealities, extended well beyond the requirement of current digital formats. The fourth-order noise shaper was shown to exhibit a chaotic behavior as the output signal spanned approximately 16 quantization levels (that is, 4 bit), which resulted in the desirable quality of translating DAC imperfections to a noiselike residue.

To maximize the performance of this system, the 16 DAC reconstruction levels should adhere to an effective 16-bit precision, although adequate performance was still demonstrated if this figure was reduced to 12 bit by introducing a combination of random and systematic displacements of the DAC levels. However, since the DAC has only a limited amplitude range and each reconstruction level requires equal weight, the desired accuracy is relatively simple to achieve.

As an alternative to the parallel DAC architecture, a second study [4] has shown how the 16 output level code of an $R_N = 4$ noise shaper can be translated to a serial bit stream using codes of optimized low-frequency spectral form. This signal format offers a similar advantage to that of delta–sigma modulation [5]–[7], whereby the DAC reduces to a 1-bit gateway and exhibits an excellent tolerance of hardware imperfection. The advantage of combining a recursive noise shaper and a nonrecursive code converter is that a high-order loop

* Manuscript received 1989 August 10.
** Currently with the Radio Frequency Group, SL Division, Cern/Geneva, Switzerland.

can be achieved using equally weighted loop integrators of optimal gain–bandwidth product, whereas a similar 1-bit noise shaper requires special attention in the loop design to achieve a stable performance. Also, where serial bit rates greater than 100 MHz are anticipated, the two-stage system appears more practical in terms of hardware realization.

Irrespective of whether a parallel or a serial port is used for the noise shaper, both systems require an interpolation filter to upconvert audio data at 44.1 kHz (or 48 kHz) to about 5.64 MHz sampling, where the filter also achieves the signal recovery filter function used in more conventional systems. In this paper a number of interpolation filters are presented and, by means of computer simulation, the relationship between filter design and overall system performance is investigated. The filters can be modeled as either a single-stage or a multistage process. The two methods are compared with respect to computational rate and coefficient storage requirements. The results of the computer simulation, which include interpolation filter, noise shaper, and DAC errors, are presented using spectral analysis of computed data sequences.

1 OVERSAMPLING FILTER, NOISE SHAPER, AND DAC

The technique of noise shaping illustrated in Fig. 1 was described in an earlier paper [2], where an analysis of coding behavior was given. Noise shaping has also been the subject of numerous other studies [8]–[10]. In the present study the oversampling filter is considered in relationship to the noise shaper and DAC, as shown in Fig. 1(a). The function of the oversampling filter is to remove the spectral replications about integer multiples of the Nyquist sampling frequency, which become redundant when the sampling rate is increased by a factor L. The process is illustrated in Fig. 2 in both the time and the frequency domain. Once a signal is oversampled, the amplitude resolution of the output samples can be reduced and the requantization distortion located in the now redundant signal space created by oversampling. It is the function of the recursive noise shaper in association with the requantizer to shape the noise spectrum to fit the available signal space, yet to inflict minimum linear and nonlinear distortion on the 0–22.5-kHz audio band.

The oversampling and noise-shaping system is cascaded with a high-speed low-amplitude resolution DAC and low-order analog reconstruction filter to complete the conversion. Earlier studies [2], [3], [10] have already discussed both the inclusion of a nonideal DAC and the translation of DAC errors to a random process, though the oversampling filters, at this stage, were assumed perfect.

2 FILTER STRUCTURES

The approach taken in this paper is to review several filter design methods that are applicable to interpolation and then present simulated results based on a number of filter examples. However, detailed design and theory of each filter are not given as they are adequately described in the cited references.

The choice of filter considered is limited to nonrecursive structures as exact linear-phase and reduced-word-length effects are desirable characteristics for oversampling. Consequently an output sequence $y(n)$ is computed from the input sequence $x(n)$ and the discrete impulse response $h(k)$ using the discrete con-

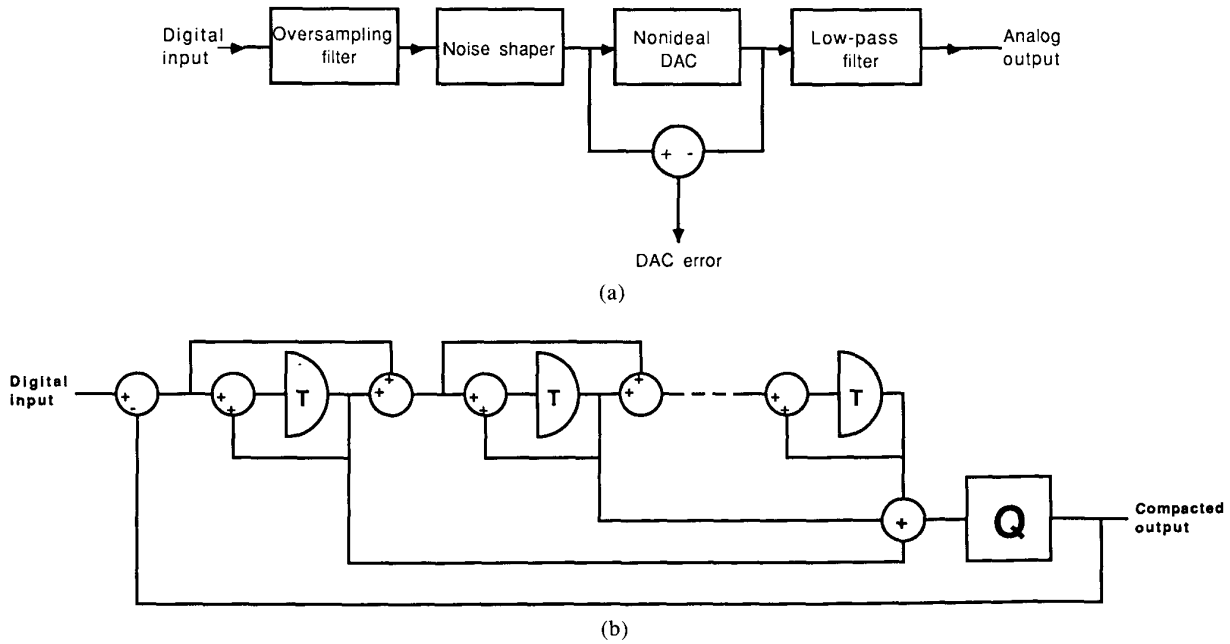


Fig. 1. (a) Oversampling and noise-shaping DAC with level compaction. (b) Nth-order linear-feedback noise-shaping coder.

volution operation

$$y(n) = \sum_{k=1}^{m} h(k) \cdot x(n - k) \ . \quad (1)$$

In the case of the oversampling filter, the arithmetic operations are performed at a rate Lf_s, where f_s is the Nyquist sampling rate of the input sequence in hertz. However, the nonrecursive structure enables the computation rate to be reduced by a factor L by identifying the redundant, zero multiplications, as shown in the signal graphs of Fig. 3 [12].

A range of design techniques are available to address filter design problems. The most prominent are as follows:

1) Equiripple FIR design [13]
2) FIR design with or without "don't care bands" [14]
3) Half-band filters [15]
4) Minimum-mean-square error design and the linear/Lagrange interpolator [16].

The widely selected optimal FIR design uses the Chebyshev approximation for the desired frequency response. The program of McClellan et al. [13] cal-

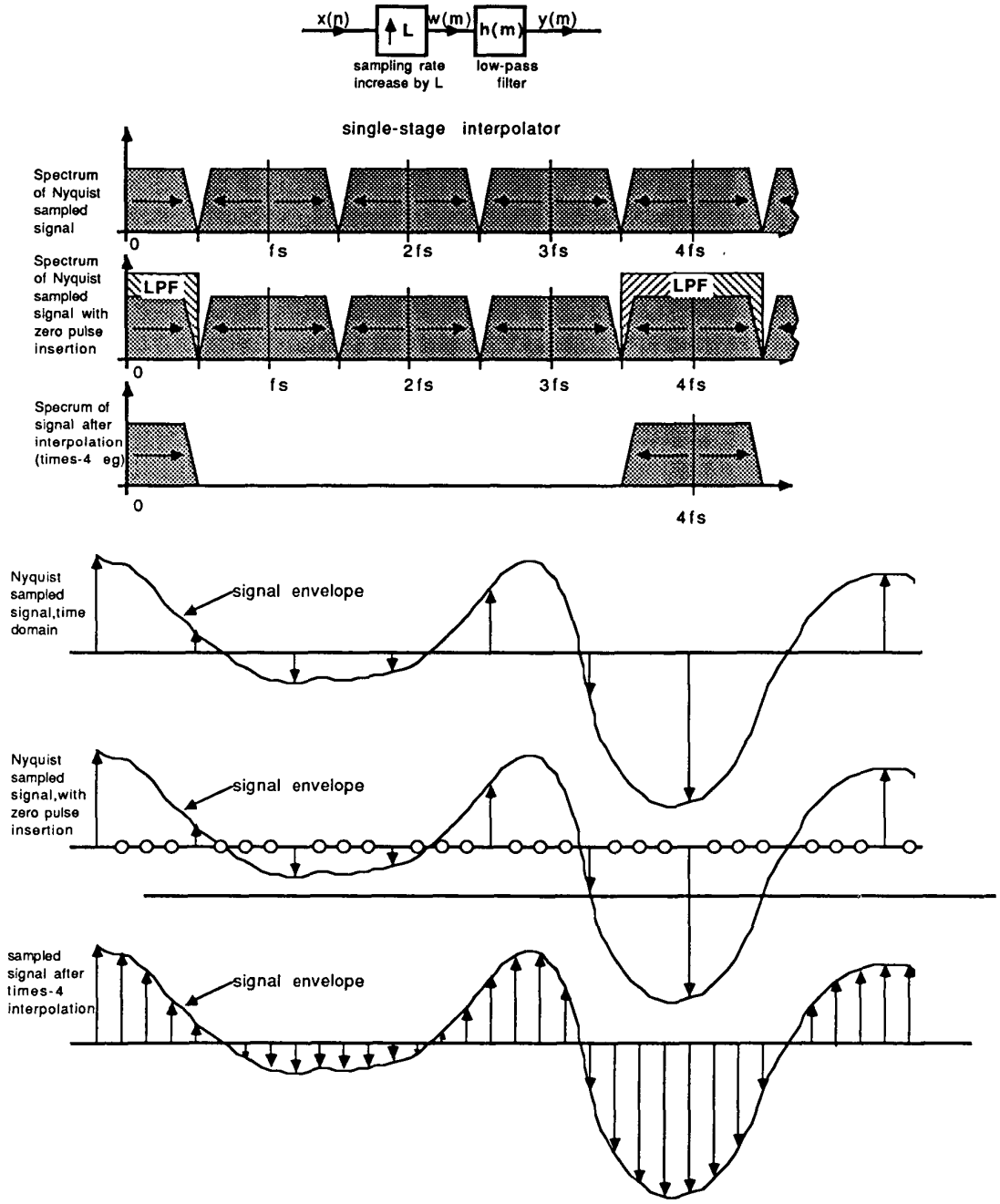


Fig. 2. Example of single-stage interpolation for $L = 4$; in both frequency and time domains.

culates the filter coefficients. A tolerance scheme, shown in Fig. 4, sets limits for the design that bounds passband amplitude ripple, stopband amplitude ripple attenuation, and transition region, together with the number of filter coefficients, which then form the program input parameters.

A significant reduction in the multiplication rate is achieved by partitioning the interpolation process into a number of intermediate interpolation stages of ratio L_i, where

$$L = \prod_{i=1}^{L} L_i \ . \tag{2}$$

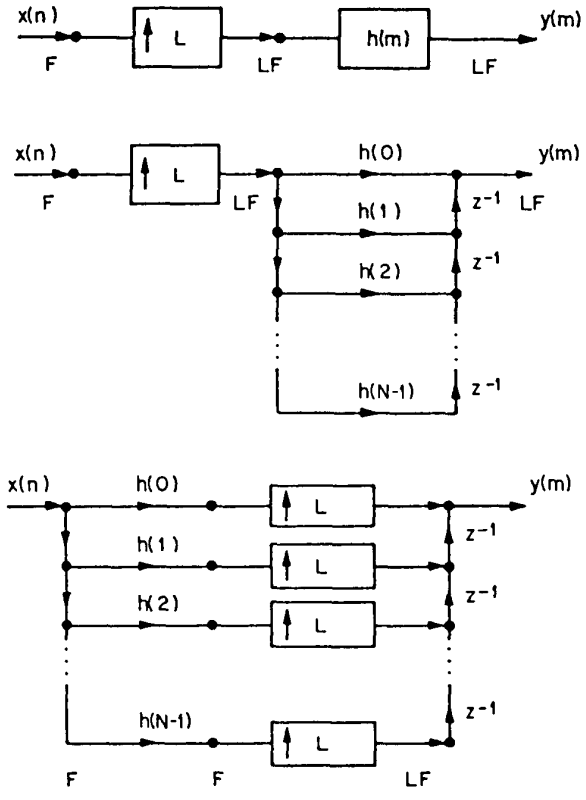


Fig. 3. Generation of efficient structure of 1-to-L interpolator.

Fig. 5 presents an illustrative example of ×4 interpolation implemented as two cascaded stages of ×2 interpolation. Inspection of the second interpolation filter reveals that, because of "don't care bands," the design can be relaxed and the broader transition band allows the number of filter coefficients to be reduced if, for example, the McClellan program is used. Multistage filters reduce both the overall computation and the storage requirements, where each stage is now an independent interpolation stage with simplified design. A method by Crochière and Rabiner [14] is to choose all stages with $L_i = 2$, which has the advantge of combining half-band filters, while another approach by the same authors [14] uses an optimization procedure which minimizes the total computation rate.

The half-band filter exploits the possible symmetry in the transition region where, for a transfer function of the form shown in Fig. 6, the impulse response can be shown to have every other coefficient equal to zero, which represents a 50% saving in multiplications. Half-band filters can be designed using the McClellan procedure, but the resulting coefficients do not exactly match the required constraint of symmetry. However, a new "design trick" [17], [18] facilitates the filter design by enabling only the nonzero $h(k)$ coefficients to be calculated.

A further design procedure makes use of time-domain optimization [14], where the signal error $y(m) - \hat{y}(m)$ of the target interpolated signal $\hat{y}(m)$ and the actual interpolated signal $y(m)$ is minimized. A computer program [16] uses this algorithm, which is based on a solution of linear equations. Finally, linear/Lagrange interpolators are filters with coefficients calculated by mathematical approximation, where, effectively, new

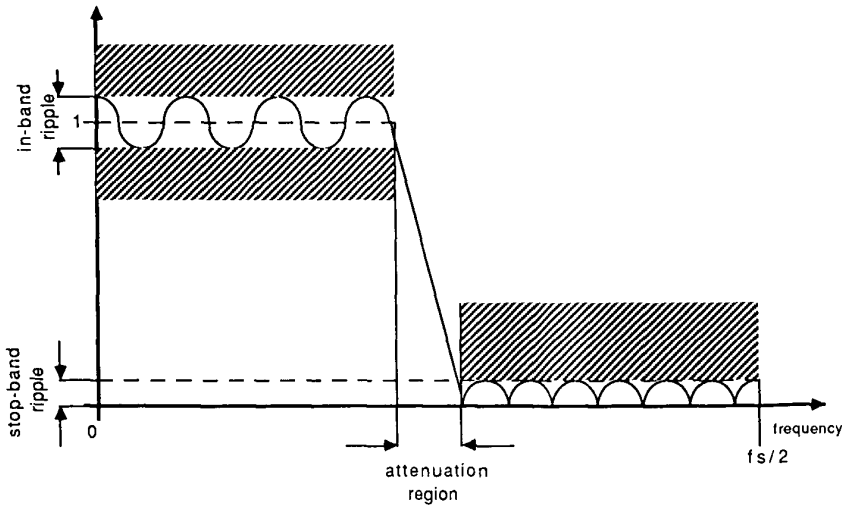


Fig. 4. Tolerance scheme for low-pass filter design.

samples are created by interpolating between two or more of the originals.

3 COMPUTER SIMULATION

The computer simulations follow earlier procedures [2], [3], [10] but now include the interpolation filter to enable the effect of nonideal interpolation to be calculated for a range of filter designs. The simulation uses 4096 sample points of a sine wave signal to calculate the operations of filtering, noise shaping, and DAC nonidealities (both static and dynamic errors). The following data and filter specification were selected:

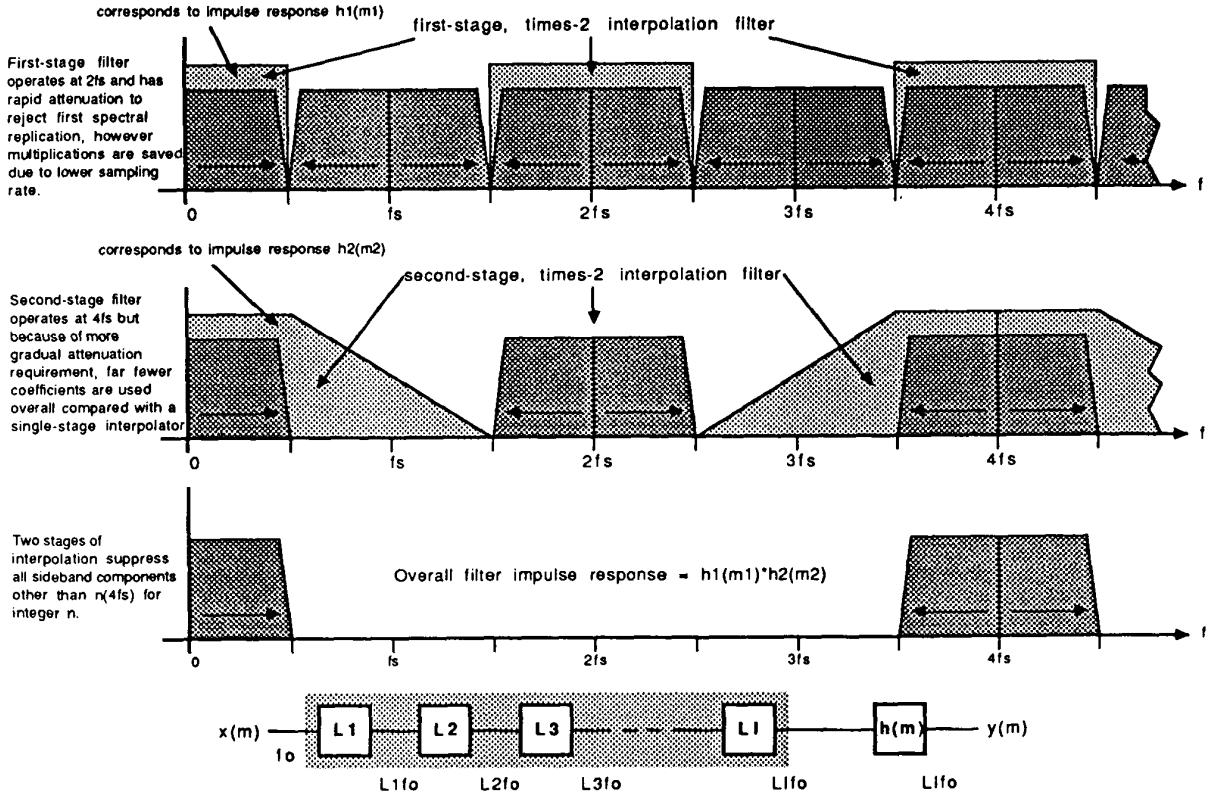
Data:

Input signal frequency	11.025 kHz
Input signal amplitude	$\sqrt{2}$ units
Noise shaper requantization interval	1 unit
Noise shaper sampling frequency	5.6448 MHz
Number of sample points	4096
Number of requantization bits	4 bit
Noise shaper, loop order	4 integrators
Oversampling ratio	128
Number of input cycles shown by computation	8
Filter:	
Passband ripple (0–20 kHz)	0.0001 dB
Transition band	20–24.1 kHz
Stopband (24.1 kHz)	−100 dB
Number of stages	2–6

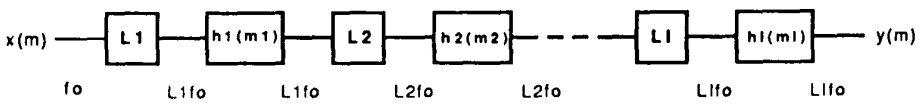
4 RESULTS

4.1 FILTER PERFORMANCE

The multistage structure of an oversampling filter enables a decrease in the required memory and computation rate. A single-stage interpolator where L = 128 needs about 8223 coefficients to achieve an attenuation of −100 dB with a multiplication rate of 181 ×



Single-stage interpolator showing decomposition of oversampling factors to L1,L2, ... Li



Multi-stage interpolator showing decomposition of oversampling factors to L1,L2, ... Li with correspondig impulse responses h1(m1), h2(m2), ... hI(mI).

Fig. 5. Multistage interpolation with illustrative example of two-stage interpolation process.

10^6 multiplications per second. A two-stage filter reduces the coefficients to 508 with a corresponding 28 × 10^6 multiplications per second, which results from a widening of the transition band. Most of the gain for optimized computation has been achieved taking L_i = 2 for the first subfilter, although a half-band filter reduces the number of coefficients by 50%.

Further design shows that the largest decrease of computation rate occurs when going from one to two or three stages; further partitioning does not result in significant reductions. The digital interpolator design [16] creates a multistopband filter for $L > 2$ and achieves better results than the Lagrange interpolator. For L_i = 2, both techniques approach the result of half-band filters with even better results than the Remez algorithm for short filters and large transition bands (N = 7, . . ., 11), such as for the last stages.

Tables 1 and 2 present a comparative overview of some design examples using multistage structures (2 to 6 stages) and Fig. 7 shows the example of ×4 interpolation presented as a series of frequency response plots.

4.2 Quantization of Filter Coefficients

The quantization of the filter coefficients $h(k)$ is simulated with both fixed-point and floating-point arithmetic, where the more optimum floating-point reveals frequency responses enhanced by up to 10 dB, as shown in Fig. 8. The reason lies in the greater dynamic range of floating-point arithmetic, which is particularly important for the first filter stage when the number of filter coefficients exceeds 100.

4.3 DAC Nonideality

4.3.1 Ideal DAC

The overall error spectrum for a fourth-order noise shaper using the simulated oversampled filters does not deviate significantly from the simulation for ideal

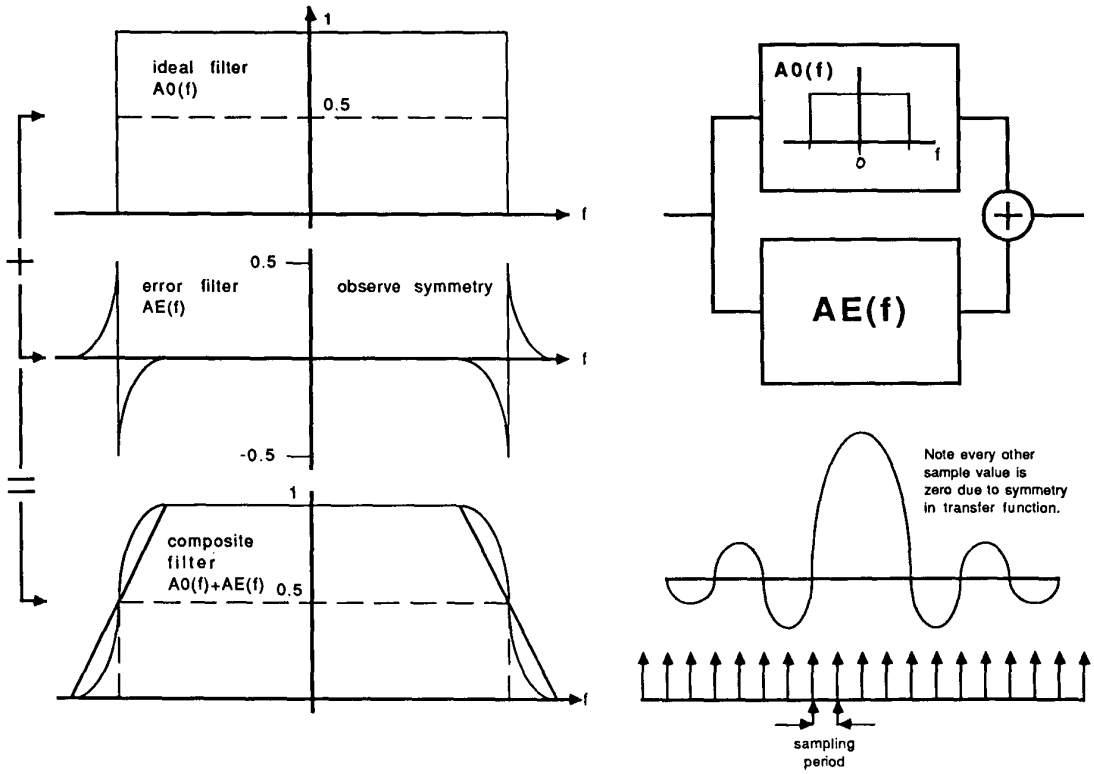


Fig. 6. Half-band filter showing symmetry in frequency response and corresponding alternate zeros in impulse response.

Table 1. Filters designed with multistage optimization structure.

Stages (L_j)	Coefficients	Memory	$R/10^6$ s	Passband (dB)	Stopband (dB)
2 (4, 32)	255, 253	255	28.05	−0.000162	−97.95
3 (2, 4, 16)	143, 49, 113	119	23.94	−0.000105	−108.47
3 (2, 8, 8)	143, 97, 53	113	25.00	0.000116	−105.61
4 (2, 2, 2, 16)	143, 25, 15, 113	115	24.30	0.000065	−108.46
4 (2, 2, 4, 8)	143, 25, 31, 53	93	24.65	−0.000110	−105.61
5 (2, 2, 2, 2, 8)	143, 25, 15, 15, 53	90	25.00	0.000066	−105.61
5 (2, 2, 2, 4, 4)	143, 25, 15, 31, 23	86	26.77	0.000113	−107.58
6 (2, 2, 2, 2, 2, 4)	135, 25, 15, 15, 15, 23	76	26.33	0.000097	−103.37
6 (2, 2, 2, 2, 2, 4)	143, 25, 15, 15, 15, 23	78	26.42	0.000067	−107.58

oversampling, where on average an in-band noise of about −130 dB is achieved.

4.3.2 Nonideal DAC

The inclusion of errors in the DAC has already been discussed [2], [3], where a range of errors including both static and dynamic mechanisms were modeled. Again, the example simulations shown in Figs. 9–11 demonstrate that the type of oversampling filter does not significantly affect the in-band noise level (−106 to −109 dB) sine degradation from the ideal DAC case is primarily attributable to DAC errors.

4.3.3 Filter Quantization Noise

A further noise source inherent to the oversampled system is associated with quantization or truncation of the signals within the FIR filters. To illustrate this mechanism, a four-stage filter was quantized with different resolutions for each stage as follows:

Stage 1 16 bit

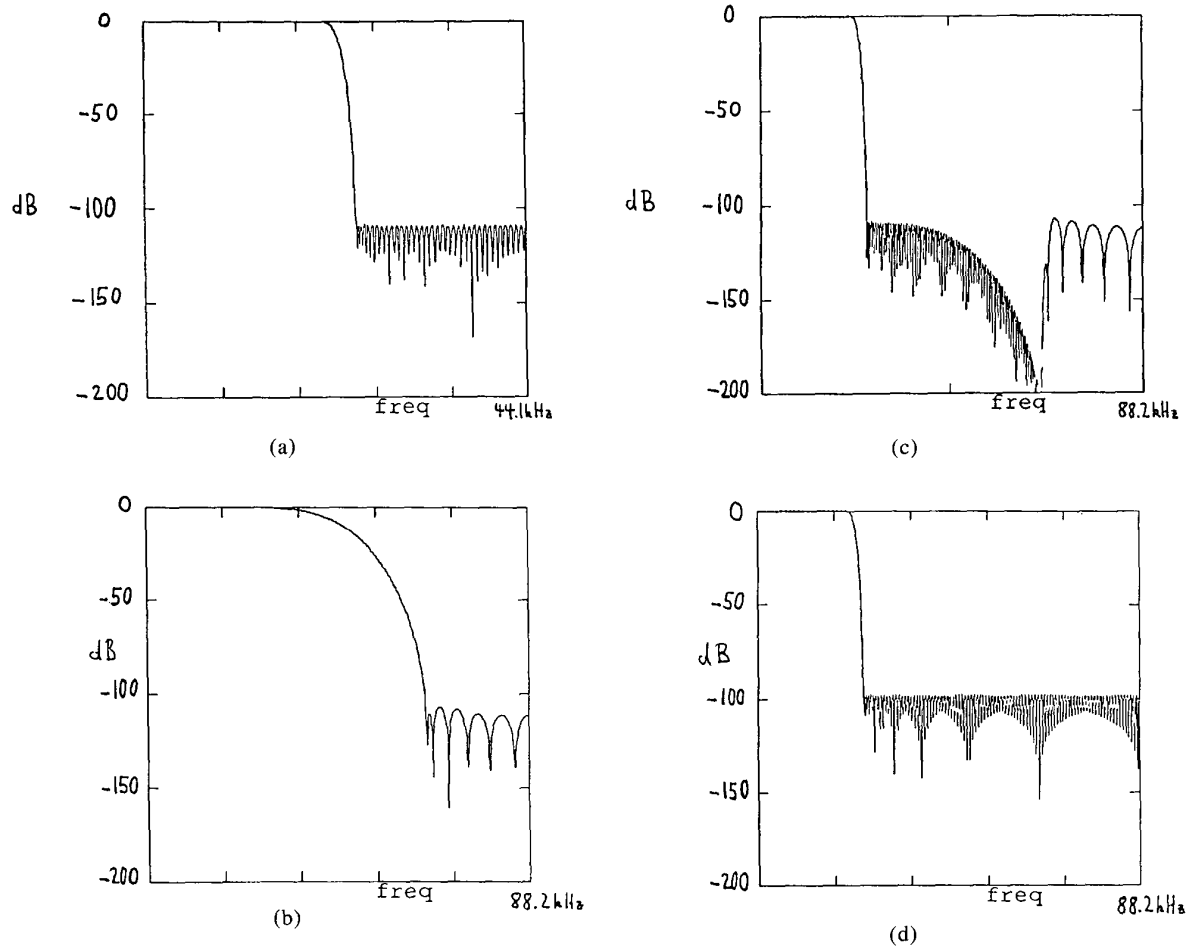


Fig. 7. Comparison of single-stage and two-stage interpolation structures. (a) 1-to-2 interpolation filter (143 coefficients). (b) 2-to-4 interpolation filter (23 coefficients). (c) Overall frequency response. (d) 1-to-4 interpolation filter (255 coefficients).

Table 2. Filters designed with halfband design procedure.

Stages (L_j)	Coefficients	Memory	$R/10^6$ s	Passband (dB)	Stopband (dB)
3 (2, 2, 32)	127, 19, 253	166	24.30	0.000259	−98.45
3 (2, 2, 32)	131, 23, 253	167	24.52	−0.000170	−100.92
3 (2, 2, 32)	139, 23, 253	170	24.61	−0.000101	−104.51
3 (2, 2, 32)	143, 27, 253	171	24.65	−0.000084	−104.51
4 (2, 2, 2, 16)	143, 27, 11, 113	106	23.15	0.000122	−108.44
4 (2, 2, 2, 16)	143, 27, 15, 113	107	23.33	0.000044	−108.44
4 (2, 2, 2, 16)	131, 23, 11, 113	103	22.93	0.000208	−100.92
4 (2, 2, 2, 16)	131, 23, 15, 113	103	23.14	0.000149	−100.92
5 (2, 2, 2, 2, 8)	143, 27, 11, 11, 53	80	23.51	0.000120	−105.61
5 (2, 2, 2, 2, 8)	143, 27, 15, 11, 53	81	23.68	0.000045	−105.61
5 (2, 2, 2, 2, 8)	143, 27, 11, 15, 53	80	23.86	0.000118	−105.61
5 (2, 2, 2, 2, 8)	143, 27, 15, 15, 53	81	24.03	−0.000043	−105.61
6 (2, 2, 2, 2, 2, 4)	143, 27, 15, 11, 7, 23	69	23.68	−0.000054	−107.59

Stage 2 14 bit
Stage 3 14 bit
Stage 4 12 bit.

The input signal $x(n)$ and the output signal $y(n)$ are quantized with the same bit number whereon the filters increase the noise level of the noise shaper (ideal DAC) from −130 dB average in-band noise to about −112 dB. For floating-point arithmetic the increase is less, as shown in Tables 3 and 4. However, when DAC errors are included, the difference in noise level with and without quantization is small; DAC nonidealities are again the dominant noise source.

5 CONCLUSION

This paper has demonstrated the feasibility of a digital oversampling filter that is compatible with the requirements of a fourth-order noise-shaping and 4-bit DAC in terms of inherent noise generation and signal recovery filtering.

Multistage structures reduce significantly both the computation rate and the required filter coefficient storage. Half-band filters are also efficient in this application, where attenuations of −100 dB can be achieved with less than 100 words of coefficient memory.

The filter simulations illustrate that filter-induced nonidealities overlap those of DAC errors but do not significantly degrade the coder performance, where the overall error spectrum of the system is virtually identical to that measured with an ideal oversampled input signal. The output noise level of the complete system lies well below −100 dB and can readily attain −110 dB, so enabling a signal coded to 18 bit to maintain its signal-to-noise ratio.

A useful characteristic of the process is the virtual elimination of low-level nonlinearity compared with a conventional DAC, where nonlinearity results from suboptimal weights, particularly with the least significant bits. The system imperfections instead manifest themselves as a noise residue, whereby signals, provided they are appropriately coded, can enter well into the system noise without nonlinear impairment. A measure of the effectiveness of this technique is to observe the noise-shaper noise floor in association with an ideal DAC, where for ×128 oversampling and a fourth-order noise shaper, the output noise level is about −130 dB.

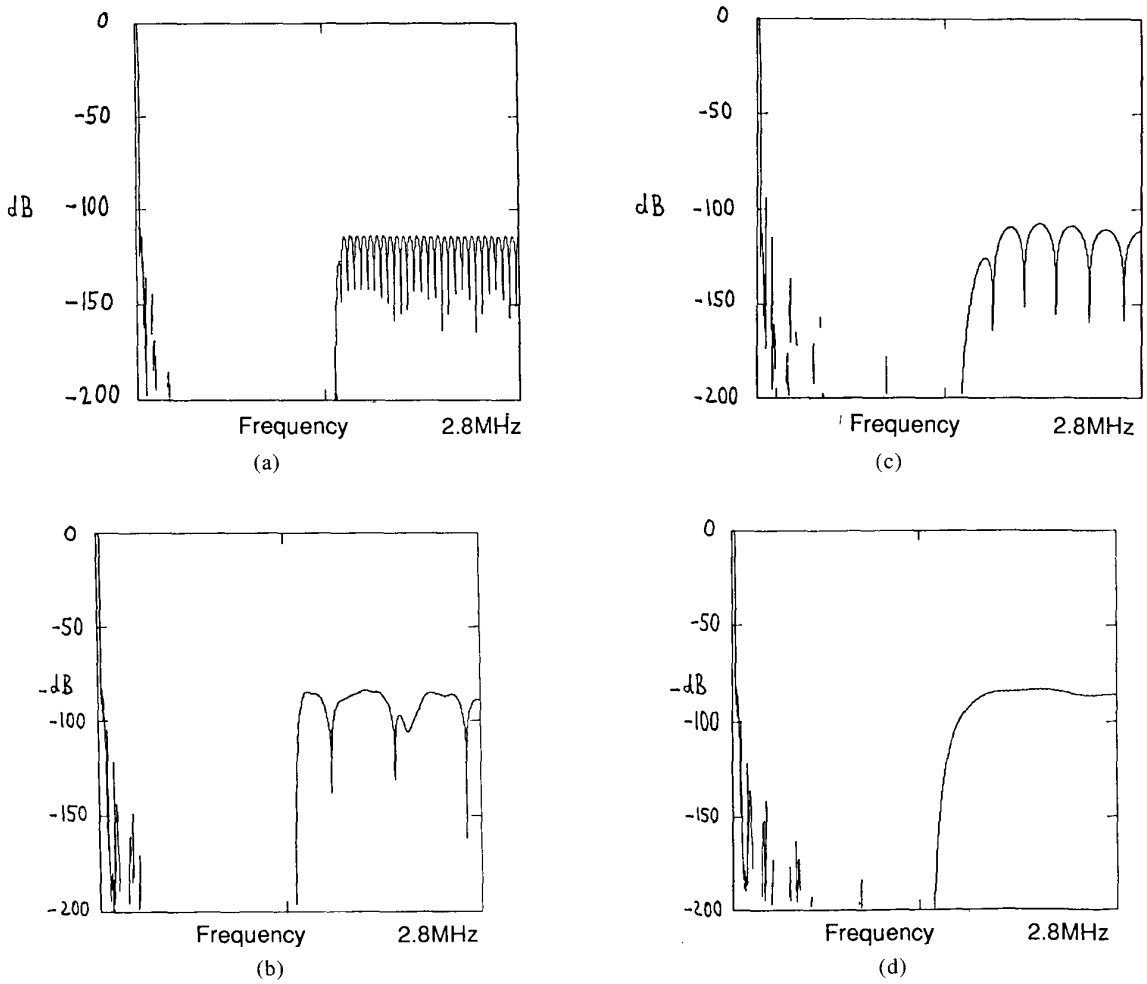


Fig. 8. Four-stage and six-stage ×128 interpolation filters with and without quantization. (a) Four-stage filter. (b) Four-stage filter quantized with 16 bit. (c) Six-stage filter. (d) six-stage filter quantized with 16 bit.

Hence with an appropriate 4-bit DAC design using a precision-weighted network to achieve 16-bit accuracy, or with code conversion to a 1-bit serial format, resolution approaching 19–20 bit should be attainable together with a clean low-level signal characteristic, commensurate, of course, with optimal ADC and signal-distortion decorrelation.

The study of filter quantization effects supports the use of floating-point arithmetic for high-resolution systems, particularly as the coefficient range is high, where quantization can cause more noise in the higher frequencies.

The technique of high oversampling and noise shaping therefore addresses most of the requirements of a DAC system for high-performance digital audio systems. The oversampling filters, when partitioned to three or four cascaded stages, are practical and can be designed for exemplary in-band amplitude ripple and out-of-band attenuation and, when using floating-point arithmetic, can achieve a high dynamic range. The heavily oversampled noise shaper of order 4 does not appear to suffer significant impairment from slight ultrasonic components as a consequence of nonideal interpolation and can also achieve a wide dynamic range, particularly as the topology does not require multiplications. The main source of impairment is again the DAC, though this is aided by errors translating to noise and the relative ease of designing a converter of high linearity.

The heavily oversampled and noise-shaped DAC is seen as complementary to the ADC system of Adams [19], where oversampling and noise shaping simplifies analog circuitry, eliminates problems associated with the least significant bits of conventional DACs, and reduces sensitivity to sampling jitter significantly when allied with switched capacitor or similar jitter reduction techniques [7]. The data rate at 44.1 kHz then becomes only an information channel, where distortions related to the direct conversion of Nyquist samples are eliminated. Also the techniques are equally applicable where data-reduction algorithms are implemented. Indeed, with near-transparent conversion the true performance potential of such algorithms can be investigated without masking effects from suboptimal ADC and DAC systems.

6 REFERENCES

[1] D. Seitzer, G. Pretzl, and N. A. Hamdy, *Electronic Analog-to-Digital Converters* (Wiley, New York, 1984).

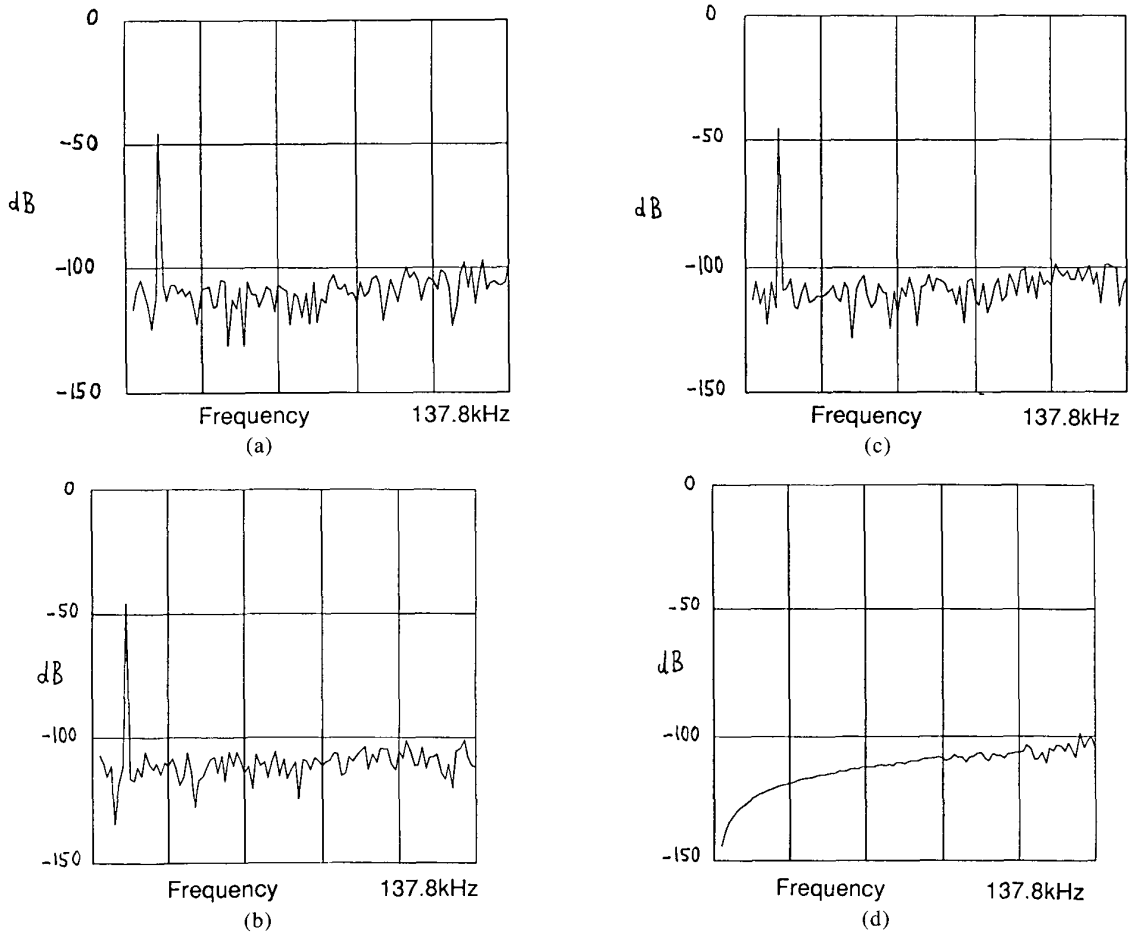


Fig. 9. Overall error (nonideal DAC), 0–137.8 kHz. (a) Filter 1. (b) Filter 2. (c) Filter 3. (d) Ideal input.

[2] M. O. J. Hawksford, "Oversampling and Noise Shaping for Digital to Analogue Conversion," *Reproduced Sound 3, Instit. of Acoustics*, pp. 151–175 (1987 Nov.).

[3] B. A. McCrea, "Simulation of Audio Digital-to-Analogue Conversion Using Noise Shaping and Oversampling," M.Sc. dissertation, Department of Electronic Systems Engineering, University of Essex, Colchester (1987).

[4] M. O. J. Hawksford, "Multi-Level to 1-Bit Transformations for Applications in Digital-to-Analogue Converters Using Oversampling and Noise Shaping," *Reproduced Sound 4, Proc. Instit. of Acoustics*, vol. 10, no. 7, pp. 129–150 (1988).

[5] H. Inose, Y. Yasuda, and J. Murakami, "A Telemetering System by Code Modulation—Delta Sigma Modulation," *IRE Trans.*, vol. 8, p. 204 (1962 Sept.).

[6] H. Inose and Y. Yasuda, "A Unity Bit Coding Method by Negative Feedback," *Proc. IEEE*, vol. 51, p. 1524 (1963 Nov.).

[7] P. J. A. Naus, E. C. Dijkmans, E. F. Stikvoort, A. J. McKnight, D. J. Holland, and W. Bradinal, "A CMOS Stereo 16-bit D/A Converter for Digital Audio," *IEEE J. Syst. Sci. Cybern.*, vol. SC-22, pp. 390–394 (1987 June).

[8] S. K. Tewksbury and R. W. Hallock, "Oversampled Linear Predictive and Noise-Shaping Coder of Order $N > 1$," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 436–447 (1978 July).

[9] P. Skritek, "Prospective Converter Techniques for Improved Signal-to-Noise Ratio in Digital Audio Systems," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, p. 394 (1987 May), preprint 2477.

[10] M. J. Hawksford, "*N*th-Order Recursive Sigma–ADC Machinery at the Analog–Digital Gateway," presented at the 78th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, pp. 586, 588 (1985 July/Aug.), preprint 2248.

[11] M. T. Sun, "Design of Digital Oversampling Filters," presented at the 81st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 1030 (1986 Dec.), preprint 2378.

[12] R. E. Crochière and L. R. Rabiner, "Interpolation and Decimation of Digital Signals—A Tutorial Review," *Proc. IEEE*, vol. 69, pp. 300–331 (1981 Mar.).

[13] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "Computer Program for Designing Optimum FIR Linear

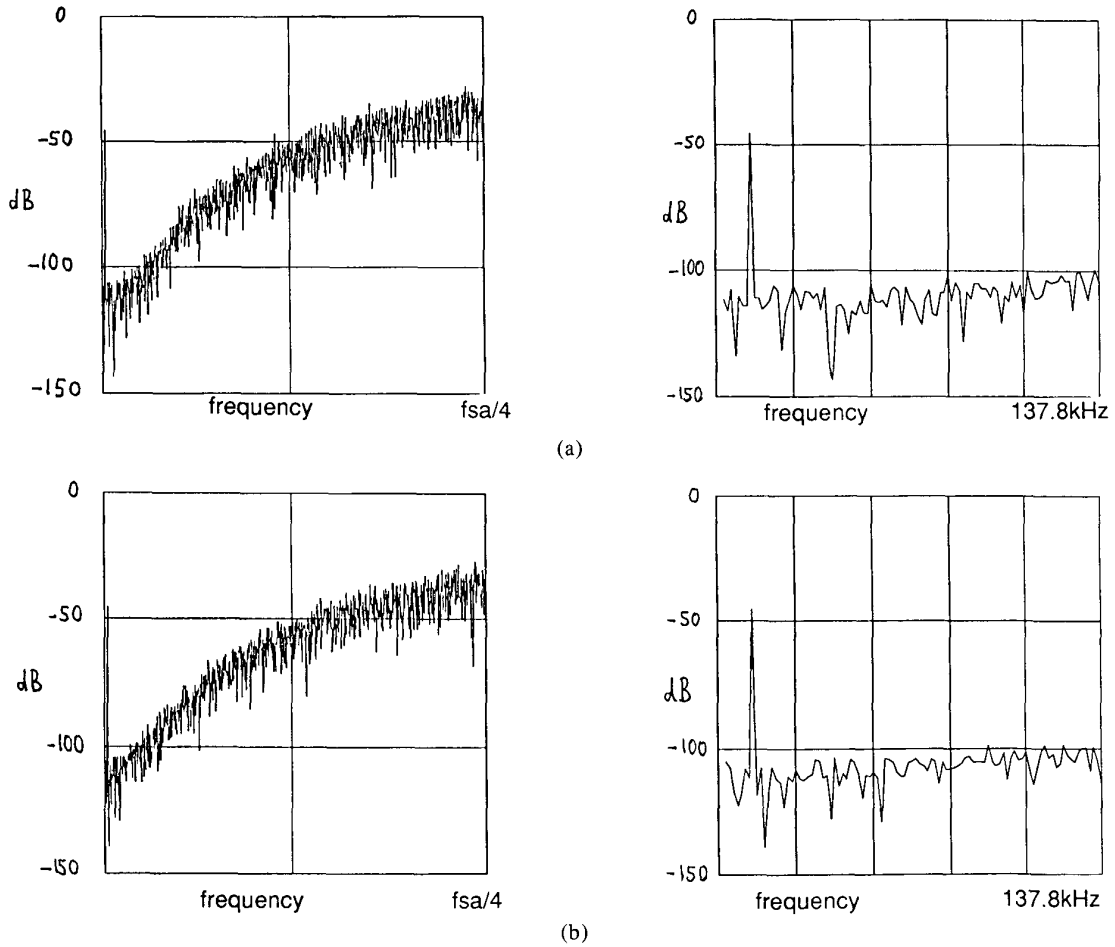


Fig. 10. Overall error, filter 1, with quantized coefficients. (a) Signal 16 bit. (b) Signal 18 bit.

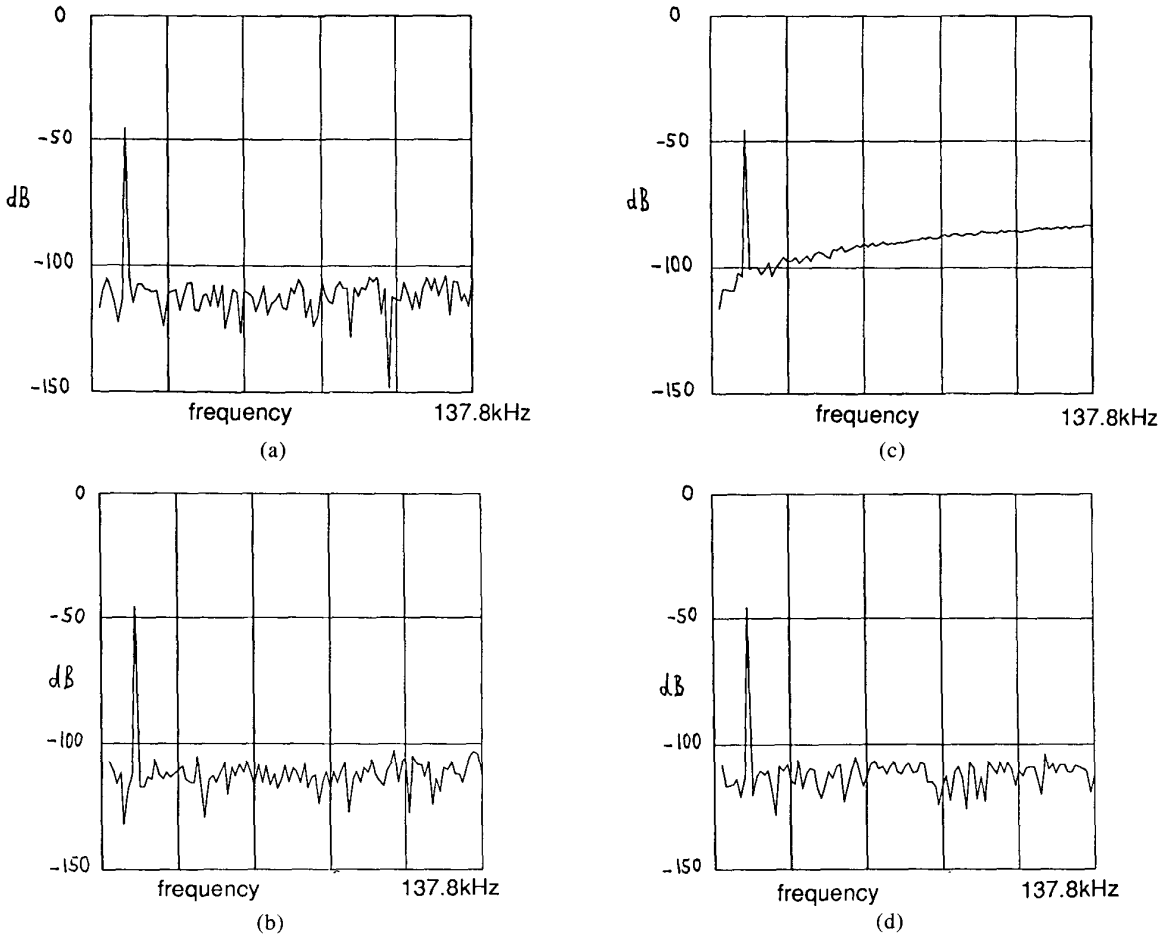


Fig. 11. DAC error, 0–137.8 kHz. (a) Filter 1. (b) Filter 2. (c) Filter 3. (d) Ideal input.

Table 3. Fixed-point arithmetic, average overall error, in decibels.

	Ideal DAC		Nonideal DAC	
Signal bit	0–137 kHz	0–20 kHz	0–137 kHz	0–20 kHz
16	−97.166	−112.652	−105.609	−106.231
17	−99.362	−114.778	−109.092	−109.367
18	−95.566	−111.028	−110.875	−109.152
19	−96.086	−111.536	−109.500	−108.465

Table 4. Floating-point arithmetic, average overall error, in decibels.

	Ideal DAC		Nonideal DAC	
Signal bit	0–137 kHz	0–20 kHz	0–137 kHz	0–20 kHz
16	−99.913	−115.381	−110.718	−108.645
17	−95.392	−110.836	−111.303	−108.370
18	−110.879	−126.401	−107.598	−109.872
19	−108.420	−123.901	−106.009	−110.188

Phase Digital Filters," *IEEE Trans. Audio Electroacoust.* (1973 Dec.).

[14] R. E. Crochière and L. R. Rabiner, *Multirate Digital Signal Processing*. (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[15] F. Mintzer, "On Half-Band, Third-Band and *N*th-Band FIR Filters and their Design," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, pp. 734–738 (1982 Oct.).

[16] G. Oetken, T. W. Parks, and H. W. Schuessler, "A Computer Program for Digital Interpolator Design," in *Programs for Digital Signal Processing*. (IEEE Press, New York, 1979), pp. 8.1.1–8.1.6.

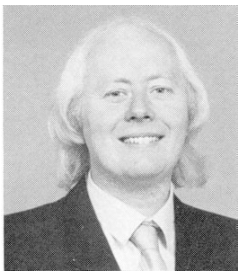
[17] R. Ansari, "Satisfying the Haar Condition in

Halfband FIR Filter Design," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, pp. 123–124, (1988 Jan.).

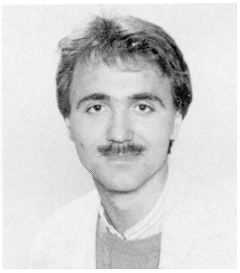
[18] P. P. Vaidyanathan and T. Q. Nguyen, "A Trick for the Design of FIR Half-Band Filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 297–300 (1987 Mar.).

[19] R. W. Adams, "Design and Implementation of an Audio 18-Bit Analog-to-Digital Converter Using Oversampling Techniques," *J. Audio Eng. Soc.*, vol. 34, pp. 153–166 (1986 Mar.).

THE AUTHORS



M. O. J. Hawksford



W. Wingerter

Malcolm Omar Hawksford is a reader in the Department of Electronic Systems Engineering at the University of Essex, where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honors B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship, where the field of study was the application of delta modulation to color television and the development of a time compression/time multiplex system for combining luminance and chrominance signals.

Since his employment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing, and loudspeaker systems has been undertaken. Since 1982 research into digital crossover systems has begun within the group and, more recently, oversampling and noise shaping investigated as a means of analog-to-digital/digital-to-analog conversion.

Dr. Hawksford has had several AES publications that include topics on error correction in amplifiers and oversampling techniques. His supplementary activities include writing articles for *Hi-Fi News* and designing commercial audio equipment. He is a member of the IEE, a chartered engineer, a fellow of the AES and of the Institute of Acoustics, and a member of the review board of the *AES Journal*. He is also a technical adviser for *HFN* and *RR*.

•

Wolfgang Wingerter was born in Landau, FRG, in 1962. He studied electrical engineering from 1982 to 1988 at the Technical University of Karlsruhe in West Germany and took part in the Joint Electrical Engineering Degree Scheme within Europe at the University of Karlsruhe, ESIEE Paris (France), and the University of Essex, U.K. He was a member of the Audio Research Group at the Department of Electronics Systems Engineering, Essex University, during 1987/1988.

He is now a fellow of CERN/Geneva where he has worked in the Radio Frequency Group of the SL Division since 1988. His current research interests lie particularly in the areas of analog-to-digital/digital-to-analog conversion, signal processing, and fast data acquisition.

Oversampled Analog-to-Digital Conversion for Digital Audio Systems*

TIMOTHY F. DARLING

Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106, USA

AND

MALCOLM O. J. HAWKSFORD

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, CO4 3SQ, UK

An alternative method of digitizing music signals is presented. It relies on the techniques of spectral noise shaping and oversampling, which achieve theoretical and computed signal-to-quantization-noise ratios of 110–120 dB. This encoding scheme is based on higher order delta–sigma modulation and is referred to as HLDSM A/D conversion. A hardware version of this system was developed measuring 80-dB SQNR and 0.007% total harmonic distortion. Potentially the HLDSM A/D converter offers noise and distortion performance superior to pulse-code modulation, a simpler hardware realization, and reduced cost.

0 INTRODUCTION

This paper analyzes the errors and limitations regarding Nyquist-sampled linear pulse-code modulation (PCM) as it applies to present-day digital audio systems. Alternative methods of digitizing band-limited analog signals, which have found application in telephony, speech, and low-quality audio, are explored. Investigations into certain structures reveal the theoretical possibility of extending delta–sigma modulation (DSM) to achieve increased (≥16-bit) resolution. The techniques of oversampling and spectral noise shaping are applied to the primitive DSM encoder with an emphasis on improved signal-to-noise ratio (SNR), stability, and simplistic hardware realizations. The expanded encoder successfully uses spectral noise shaping, oversampling, and a multilevel inner loop quantizer to form a topologically efficient encoder, referred to as HLDSM A/D converter. These techniques have evolved a hardware conversion system that does not require troublesome sample-and-hold circuits (SHCs) or complex anti-aliasing filters (AAFs).

1 PULSE-CODE MODULATION—OPERATION, ERRORS, AND LIMITATIONS

The most common approach to encoding low-level audio analog signals for consumer and most professional applications is binary-coded PCM [1]–[5], whose principles were developed in the late 1930s. The PCM system (Fig. 1) consists of a high-order analog antialiasing filter (AAF), a sample-and-hold amplifier, and a linear analog-to-digital (A/D) converter. It divides

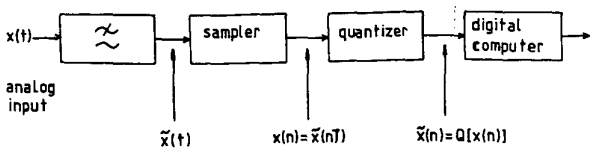


Fig. 1. Linear Nyquist-sampled PCM encoder.

* Presented at the 85th Convention of the Audio Engineering Society, Los Angeles, 1988 November 3–6.

the incoming signal into 2^B equal intervals, which are assigned digital words in monotonic order. The parameters of this initial conversion establish the maximum signal quality provided by the total system. Moreover any degradation introduced at this stage will remain with the digitized signal.

There are several errors indigenous to the PCM quantization of analog signals and subsequent decoding [3]–[8]. These errors, which are produced by the analog AAF, SHC, and A/D conversion process, place a practical limit on the resolution and transparency of the recording process.

1.1 Quantizer Errors

Waveform coding by PCM requires the input signal to be band-limited and sampled at the Nyquist frequency. There are only a finite number of quantization levels. Therefore input voltages in the range of $-\Delta/2 < x(n) < +\Delta/2$ will be assigned to $\pm\Delta$ (Fig. 2). This mapping of analog voltages into discrete words constitutes a quantization error which leads to a signal-to-quantization-noise ratio (SQNR) expression given by [9]

$$\text{SQNR} = 6.02B + 10.79 + 10 \log_{10} \left(\frac{v_{\text{rms}}}{v_{\text{peak}}} \right)^2 \quad (1)$$

where it is apparent that SQNR increases approximately 6 dB for each bit.

1.2 Anti-Aliasing Filters

The A/D conversion process requires extremely high-performance analog filters to ensure ≥16-bit accuracy. Ideally the filters should possess an SNR >96 dB, which demands both low-noise and low-distortion performance, low ripple in the passband, linear phase response, and extremely steep out-of-band attenuation.

The input AAF is required to remove all spectral components above the Nyquist frequency prior to the sampling and quantization process. If these components were not removed, they would fold back into the passband and corrupt the original spectrum. The AAF, therefore, requires very steep "brickwall" filtering outside the passband. To obtain this high rate of signal loss one can use common filter structures such as Bessel, Butterworth, Chebyshev, or elliptical polynomials. Each of these structures offers certain amplitude and phase characteristics [10]–[12].

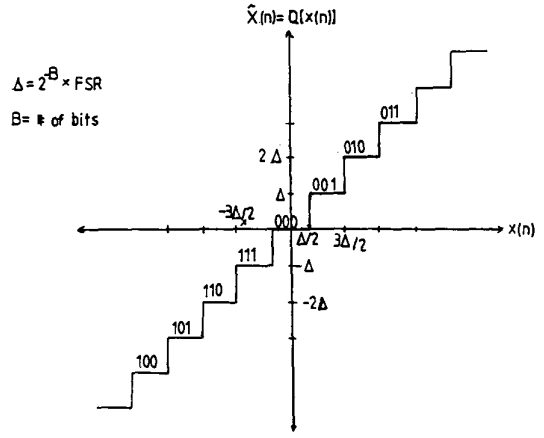


Fig. 2. Uniform quantizer transfer function.

The Butterworth function is known for being maximally flat at direct current, with minimum ripple in the passband and an attenuation rate of $6n$ dB per octave, where n is an integer representing the order of the filter polynomial.

The Chebyshev gives increased out-of-band attenuation. If we compare the difference in attenuation rates α between Chebyshev and Butterworth structures we arrive at [10]

$$\alpha_{\text{CH}} - \alpha_{\text{BW}} = 6(n - 1) \quad \text{dB} \quad (2)$$

which demonstrates a significant difference. For example, if the order of the filter is eighth, there is a 42-dB difference in their respective outputs at a certain frequency. Unfortunately the price one has to pay for this increased attenuation is a phase characteristic of increasing nonlinearity, which gives rise to group-delay distortion, and ripple in the pass- and stopbands, while its transient response clearly shows more ringing than the Butterworth structure [13].

The elliptical approximation is the most commonly used function in the design of filters. If our main concern is one of attenuation, this function will, in general, require a lower order than the Butterworth or the Chebyshev.

The Butterworth, Chebyshev, and elliptical approximations were mainly concerned with obtaining the highest rate of attenuation, while ignoring the phase or delay characteristics of the filter. The Bessel approximation is a structure that concentrates on the phase and delay characteristics, and it provides the user with constant group delay over the passband region. In addition it has a step response that displays insignificant amounts of ringing, providing the user with excellent time-domain properties [12]. The Bessel approximation does provide us with excellent time delay properties, but unfortunately its filtering action in the stopband is much worse than that of the Butterworth. The poor stopband characteristics of the Bessel approximation make it impractical for certain filtering applications and would not be used for Nyquist-sampled digital audio, which requires a very high rate of attenuation. Thus we are left with filters that provide us with high rates of attenuation and poor time-domain responses or vice versa.

If we are to choose a filter structure based on its ability to attenuate signals outside the passband, the obvious choices are Chebyshev or elliptical approximations of reasonable order. A ninth-order Chebyshev filter with +0.1-dB passband ripple, f_b = 20 kHz, would provide us with 70 dB of attenuation at 25 kHz. To increase the stopband filtering action one must increase

the amount of ripple or the order of the structure. Interestingly, as the filter order increases, the phase characteristic becomes more nonlinear [13]. This nonlinear phase will lead to different group delays over the 20-kHz bandwidth. Meyer [14] has developed an all-pass filter that corrects for the time misalignment of conventional AAFs and provides constant group delay up to 18 kHz.

Not only do high-order analog filters introduce nonlinear time delay, which results in phase distortion, they also possess significant step response overshoot and an impulse response that contains excessive ringing as the order becomes ≥ 3 [13]. This ringing leads to the phenomenon of time dispersion. The concept of time dispersion can be considered analogous to striking a bell with an impulse and hearing the original pulse and its associated decay components. These decay "echo" components are the dispersed sound of the initial pulse and as such are a source of error, which may be audible. The source of these echoes can be linked to a filter's passband ripple. The only type of analog filter that will not contribute significant time dispersivity to the A/D–D/A conversion chain is the Bessel filter. Unfortunately, as stated earlier, because of their out-of-band attenuation these filters are not adequate for use as AAFs. The design of high-quality analog AAFs that minimize group delay, ripple, and time dispersion appears to be a very formidable task.

A solution that eliminates analog AAF errors, simplifies the system output low-pass filter, and promises to maintain consistent high-quality performance throughout coding and decoding activities is an all-digital filter implementation using the techniques of oversampling (Fig. 3). If we use recursive filter structures, it is possible to obtain passband ripples ≤ 0.001 dB [15]–[17] and constant group delay (linear phase) over the entire passband, ensuring a respectable system transient response.

Through the use of oversampling techniques (Sec. 3) on both the encode and the decode processes we can take advantage of inherently accurate digital filters, whose characteristics can be specified exactly and will yield consistent results.

1.3 Sample-and-Hold Circuits

SHCs [18]–[21] are used in all consumer and most professional PCM audio systems to store an analog voltage accurately over an approximate 15-μs period while the A/D converter converts the held voltage into an n-bit digital word, and are responsible for introducing both amplitude and timing errors.

Fig. 4 shows an SHC in conceptual form. A switch is connected to a capacitor and when the switch closes, the capacitor charges to the input value. A practical implementation of the SHC (Fig. 5) consists of an input buffer to provide adequate drive current, an output buffer that isolates the hold capacitor from the load, and a very fast FET-type switch driver to obtain fast acquisition of the signals. Of the two major classes of errors (amplitude and timing) the timing errors can be the most significant. The acquisition time, which is the time required, after the sample command is given, for the hold capacitor to charge to a full-scale voltage change and remain in a specified error band about its final value [22], is an important specification, for it tells us how fast we can sample the PCM system to the required error.

Fig. 6 graphically illustrates the entire acquisition time as being comprised of a finite switching delay (gain–bandwidth product limitations of the FET switch) Δt_1, charging the hold capacitor at a maximum rate determined by the available buffer charging current and capacitance value Δt_2, until it reaches almost final value. At this point it is charged exponentially due to finite amplifier output and switch resistance Δt_3.

There are several system criteria that establish the acquisition time: speed, stability of the input amplifier when driving a capacitive load, slew-rate limitations, finite amplifier and switch resistance, numerical value

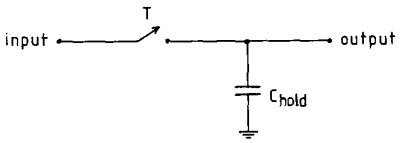


Fig. 4. Sample-and-hold circuit in conceptual form.

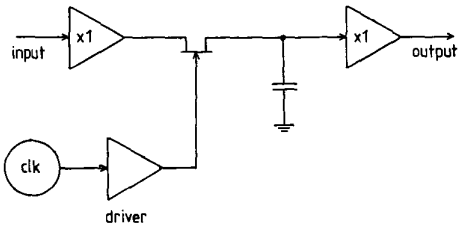


Fig. 5. Practical realization of sample-and-hold circuit.

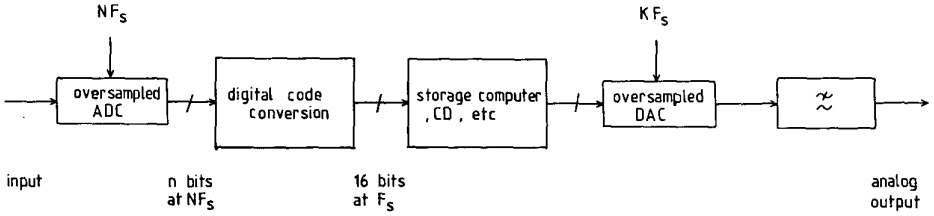


Fig. 3. Encode–decode process using techniques of oversampling.

of hold capacitor, and available driving currents.

Let us consider other timing-related errors that affect SHC performance. One such error is the aperture time, which is the delay between the hold command and an input voltage transition such that the transition does not affect the held output. The aperture time is variable and occurs because FET switch capacitance (C_{gs}, C_{gd}) is nonlinear, having a strong dependence on signal amplitude [23]. This timing error implies that the analog voltage, which is held, differs from the one that occurred at the "true" sampling time. "We can therefore view this error as an additive voltage which is equal to the signal derivative multiplied by the signal aperture time" [3].

Proceeding with numerical examples to illustrate the severity of this nonlinearity we shall assume that the input signal is a sine function, the sampling rate F_s = 44.1 kHz, and $\Delta\varepsilon$ represents the aperture error uncertainty. The error in decibels below the fundamental can be expressed as

$$\text{error}|_{\text{dB}} = -20 \log_{10} \left[\frac{dv_{\text{in}}}{dt} \Delta\varepsilon \right] \quad (3)$$

The error signal $\Delta\varepsilon$ is given in Table 1 for various input frequencies, f_{in}.

Therefore to reduce the worst-case peak error to that of 1Δ LSB, a 16-bit converter requires that the aperture uncertainty be approximately 200 ps for a full-level maximum frequency signal, while for 18 bit $\leq$100 ps is required.

Sampling errors can also originate from phase jitter in the clock signals. Clock jitter produces additive noise, which is proportional to the signal amplitude and frequency. A general expression for the noise produced by a Gaussian time jitter is given by [3]

$$\text{SNR}_{\text{ws}} = -20 \log (2\pi f \Delta t) \quad (4)$$

and for sine-wave approximated time jitter it can be expressed as [24]

$$\text{SNR}_{\text{jitter}} = -20 \log \left(\frac{\pi}{\sqrt{3} f \Delta t} \right) \quad (5)$$

Even though the two equations were derived using different input signal approximations, they yield similar results. By solving Eq. (4) for Δt we find that time jitter components of 250 ps for 16 bit and 100 ps for 18 bit are required.

There are two main sources of amplitude-related SHC errors that occur during the hold mode and are referred to as hold-mode droop and hold-mode feedthrough. Hold-mode droop is defined as output voltage change per unit of time. The five leakage current components that contribute to droop are capacitor insulation leakage I_{CL}, switch leakage currents I_{OS}, amplifier bias current I_{B}, and stray leakage from the common terminal connection (Fig. 7). The change in output voltage during the hold mode can be expressed as the sum of the leakage currents to hold capacitance. 1Δ LSB of error for an 18-bit system requires a total leakage current $\leq$500 pA for a 100-pF hold capacitor.

Hold-mode feedthrough is simply the percentage of an input signal that can be measured at the output of a SHC while in the hold mode. For state-of-the-art SHCs this can be on the order of 0.05–0.005%.

Capacitors in the SHC must use high-quality dielectric material such as polypropylene to minimize losses. It is a well-known fact, and one that can be measured

Table 1. Error in decibels below fundamental.

f_{in} (kHz)	$\Delta\varepsilon$ 50 ns	200 ps	10 ps
1	70	120	145
10	50	100	125
20	45	95	120

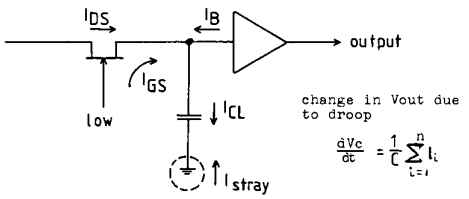


Fig. 7. Droop in output voltage due to leakage and stray currents.

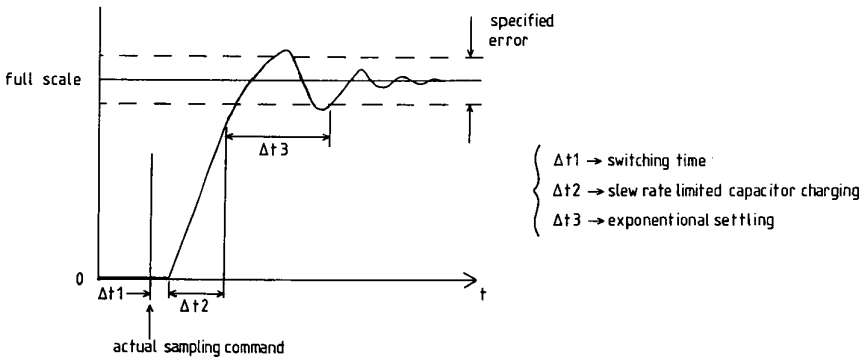


Fig. 6. Entire acquisition time of capacitor charging over one sampling period.

[25], that capacitors exhibit dielectric absorption (DA). This phenomenon is essentially a reluctance on the dielectric material to give up electrons that it has stored within itself whenever the capacitor is asked to discharge. The effect of this change in voltage is similar to that produced by droop, that is, the held voltage is changing during the n-bit test and compare process. Therefore high-quality polypropylene capacitors with low DA must be used.

1.4 Pulse-Code-Modulated Analog-to-Digital Conversion

We have briefly investigated certain common errors inherent in analog AAFs and SHCs, whose circuit functions are essential in a Nyquist-sampled audio system in reducing aliasing distortion and maintaining a constant band-limited analog voltage while performing A/D conversion.

The most common form of A/D conversion found in commercial PCM audio systems is coined a successive approximation (SA) type [1]–[5], [26]. The basic topology, shown in Fig. 8, consists of a comparator, a successive approximation register (SAR), and a D/A converter arranged within the feedback loop. The principle of operation is straightforward. The SAR seeks to find the digital word which, when driving the D/A converter, will produce the best approximation to the held input signal. If we are aspiring toward $\geq$16-bit performance, the comparator must make 16 tests in succession over a 22-μs period for a 44.1-kHz sampling rate. If we allow $\cong$5 μs to acquire the signal to the specified linear error, we are left with 17 μs for the n-bit digitization process, thus yielding $\cong$ 1 μs for 16 bits or 900 ns for 18 bits to test each bit accurately.

When a bit is switched on by the SAR, a certain waiting time must elapse until the SAR can test the sign of the comparator output. If we assume that the settling process can be modeled as an RC time constant, then 11 (for 16 bits) or 13 (for 18 bits) time constants are required to elapse for the MSB to settle to an accuracy of $\frac{1}{2}$ Δ LSB. This translates into 90- or 70-ns accumulated time constants for the SAR, D/A converter, and comparator. Even high-performance operational amplifiers (such as SN5532, OP-27, and OP-37), with GBP of 5–70 MHz will generate (if driven from a zero-impedance source) 10-μv broad-band rms noise voltage. This implies that the peak signal must be scaled to at least 2.5 V (16-bits) or 10 V (18 bits) to produce the required resolution.

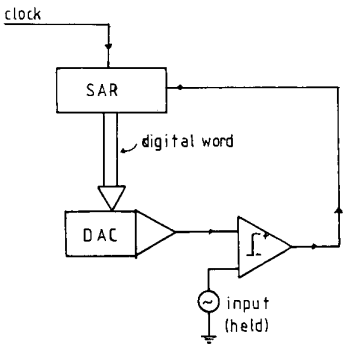


Fig. 8. Successive-approximation analog-to-digital conversion.

1.5 Summary

In this section we have outlined some of the major errors inherent in Nyquist-sampled PCM to generate digital audio signals. SHCs produce amplitude and timing errors. Analog AAFs possess nonlinear phase, ripple in the passband, insufficient attenuation in the stopband, and invariably have very poor (that is, ringing) transient responses. Successive-approximation A/D conversion requires both high-bandwidth and low-noise operation to obtain the required resolution.

Most of these errors described illuminate the difficult task of obtaining >16-bit performance in a Nyquist-sampled PCM system. If may be possible with sophisticated circuit techniques and topologies to minimize SHC-related inaccuracies. It becomes more difficult to constrain the clock jitter to $\leq$100 ps. The requirements for high-bandwidth and low-noise comparators are in theoretical opposition.

Therefore if we are interested in achieving >16-bit-resolution A/D conversion capable of interfacing directly to transducer signals, we must investigate alternative schemes of converting analog audio signals into digital words which eliminate the various errors described.

2 ALTERNATIVE METHODS OF DIGITIZING ANALOG SIGNALS

The consumer audio industry has adopted 16-bit linear PCM quantization for most consumer and professional digital audio systems. There is an increasing need to develop high-resolution A/D conversion (20–24 bits), such that microphone and other analog transducer signals can be digitized directly, eliminating the need for low-level preamplification. Even with today's advanced integrated-circuit technology and clever circuit techniques, it becomes a formidable design challenge (as was shown in Sec. 1) to successfully implement >16-bit PCM digital audio systems. Therefore if we are journeying toward increased resolution in our conversion systems, we must investigate alternative encoding techniques which offer a more transparent method of encoding (negating the need for SHC and AAF).

This section encapsulates a review of various encoding techniques that have found applications in audio, speech, and telephony and could provide us with different approaches to the direct digitization of analog (music) signals.

2.1 Flash Conversion

A simple, fast approach to A/D conversion, which has found application in video, radar, instrumentation, and low-resolution audio, is the flash converter method shown in Fig. 9. This method is fast because there is

only one step in determining the input voltage. A series of comparators test the input signal against a set of voltage thresholds established by the resistor ladder. Digital logic is then used to convert the comparator outputs to a binary number in the classical 2's complement format, which allows for easy interfacing to digital signal-processing equipment.

To operate in such a fast manner requires $2^B - 1$ comparators. Therefore, for 16-bit resolution we would need 65 535 comparators and a significant amount of priority encoding logic to achieve current PCM-compatible performance. Even though flash conversion eliminates the need for SHC and steep AAF, it currently would be stretching the capabilities of integrated-circuit technology to make it a directly viable alternative to PCM.

2.2 Residual Expansion

In Sec. 2.1 we discussed an extremely fast method of conversion that would require a significant amount of circuitry to maintain equivalence with present digital audio precision. A practical compromise is to imbed two flash converters (reducing chip density) and process the digital word in two parts (Fig. 10).

The incoming held signal is converted into an n-bit digital word (comprising the MSBs) and stored in counterregister A. This same word is converted back to analog using a high-precision D/A converter and substracted from the input signal, thereby generating the error signal. This error signal is the difference between the first approximation of n bits and the original signal. The error is then scaled up by 2^n, converted to digital, and loaded into register B. The first and second conversions are thus the equivalent of two tests of n bits, resulting in $(2n - 1)$-bit accuracy. One bit of resolution is lost in the conversion process due to quantizer nonlinearity within the flash converters. The residual expansion conversion scheme is obviously not as fast as the simple parallel flash converter, yet the advantage of this system is that each converter requires only 50% of the final output accuracy.

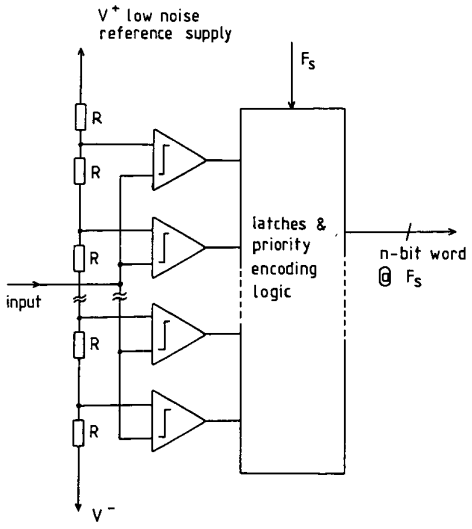


Fig. 9. Flash conversion, a high-speed approach to digitizing analog signals.

2.3 Differential Conversion

Differential A/D conversion can provide the user with a reduction in the number of bits for a given quality by digitizing the difference between two adjacent samples of the analog data. Proponents of this method [27] have observed that low-frequency energy (most musical signals obtain a substantial percentage of low-frequency information), which can have a large amplitude, has a very small derivative. Therefore the dynamic range of the difference signal is small and a lesser number of bits can be used than for conventional PCM. McDonald [28] has shown through computer simulation that differential PCM (DPCM) offers a 6–10-dB advantage in SQNR over PCM for speech signals at the same clock rate.

The block diagram of Fig. 11 shows a differential coder which digitizes the difference between two samples of the analog signal. At the decoding end of the system the quantized signal is converted back to analog and added to the previous value to form the reconstructed signal. The signal Y_r is a series of samples related to the input by

$$Y_r = X_r - X_{r-1} \tag{6}$$

r being the rth sample of X. A time-domain representation of the signal to the quantizer is

$$y(t) = x(nt) - x(nT - T) \tag{7}$$

where T is the sampling period and n is an integer. The error signal, which is the difference between input and output, is given by

$$\Delta e_r = X_r - X_r' \ . \tag{8}$$

For purposes of noise analysis the modified block diagram in Fig. 12 is useful. We shall compute the noise in X_r' based on the assumption that only the quantization process produces noise N_q. The Laplace transform for transmission from the input to the output port is given by

$$\frac{X_r'}{X_r} = H(s) = e^{-st} \ . \tag{9}$$

Except for a phase shift, the gain is unity. Therefore the noise appearing at the output of a DPCM system is analogous to that of a conventional PCM system. Only when we consider the statistical properties of the input signal can we evaluate the SQNR difference between DPCM and PCM. With speech signals and other low-frequency signals an improvement of 3–11 dB in SQNR using DCPM has been reported [27], [28]. If we are willing to make the assumption that audio program material has dominant low frequencies, then such systems have a certain appeal.

2.4 Delta Modulation

The number of bits required to digitize a differential PCM signal is less than that required by linear PCM over a certain signal bandwidth and sampling rate. By increasing the sampling rate, the prediction method becomes more accurate, since the input signal cannot change significantly during such a short interval. Thus the summed difference signals will make a better approximation to the input signal. In the limit with very high sampling rates (oversampling) the error signal becomes exceedingly small. We can take advantage of this fact to develop a simple quantizing strategy which displays efficient coding activities.

Delta modulation (DM) [29]–[33] is such a scheme, whereby an analog signal $x(t)$ is encoded into a very fast serial bit stream $L(t)$ (Fig. 13). These pulses occur at a sampling rate that greatly exceeds the Nyquist rate. At each sample time the sign of error $e(t)$ is determined. If the error is positive, an incrementing pulse is applied to the integrator to increase the signal. Negative errors result in a corresponding negative pulse. The integrator output is a past approximation to the input, while the correcting pulse is the differential correcting error required to bring the integrator output closer to the present-value input.

Fig. 14 illustrates how the voltage waveform of a DM varies with time when the input signal is a sinusoid. When the binary pulses $L(t)$ are integrated by the loop filter in the feedback loop, the resulting output waveform $y(t)$ consists of steps having magnitude $\pm\Delta$ of duration T seconds, which oscillate about the analog input signal $x(t)$. The difference between $x(t)$ and $y(t)$ is the error signal $e(t)$. The error signal will be quantized to $\pm\Delta$ if the system is not slope overloaded (that is, the input signal is changing too rapidly for the encoder to track it).

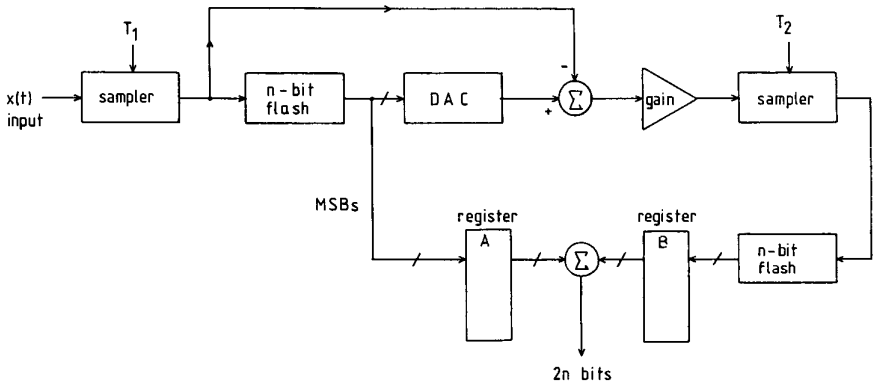


Fig. 10. Two-pass analog-to-digital conversion using flash convertors.

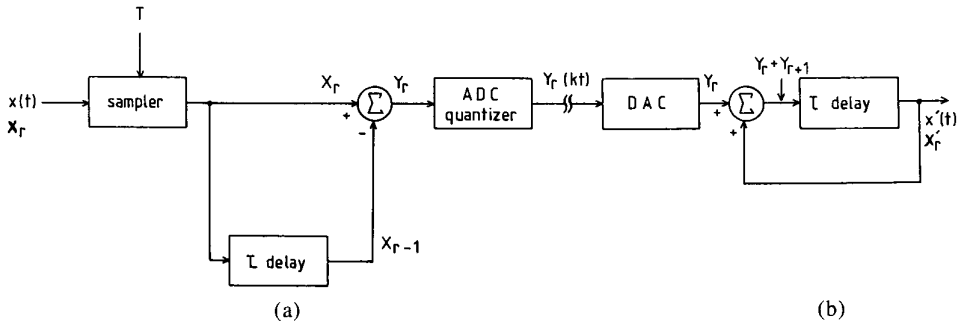


Fig. 11. Differential conversion system. (a) Encoder. (b) Decoder.

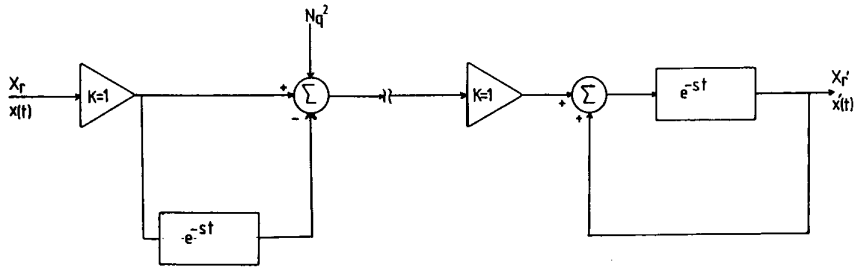


Fig. 12. Quantization noise model of differential encoder.

2.5 Noise Analysis of Delta Modulation

Quantization noise occurs in DM due to the time and level quantization of the input signal. In the encoding process two types of quantization errors are generated, granular noise and slope overload noise (Fig. 15). The two types of noise are not equally noticeable to the human hearing process. Since slope overload [35] is correlated to the input signal, it is perceptually less disturbing then granular noise at equal power levels [36].

To calculate the quantization noise we will assume that the system is not being overloaded and the noise random [31]. Steele states that for a wide variety of input signals, particularly band-limited random signals (certainly music is random), the quantization noise is substantially uncorrelated with the input signal. This will enable the noise to be specified without reference to the input signal.

To ensure that the system is tracking correctly, the integrated signal $y(t)$ must be $\leq\Delta$ (step size) of the input signal. Therefore the maximum signal slope shall not exceed Δ/T. The maximum slope of a sine wave of amplitude A and frequency f_i is $2\pi f_i A$. Thus equating

$$\frac{\Delta}{T} = 2\pi f_i A \tag{10}$$

ensures proper tracking. The step size is easily given as

$$\Delta = \frac{2\pi f_i}{f_s A} \tag{11}$$

and the quantizing noise power spans the range $\pm\Delta$ with an audio noise bandwidth of f_b. The total noise power is given as

$$N_q^2 = \int_0^{f_b/f_s} \Delta^2 \, df = \Delta^2 \frac{f_b}{f_s} . \tag{12}$$

Substituting Eq. (11) into Eq. (12) yields

$$N_q^2 = A^2 4\pi^2 \frac{f_i^2 f_b}{f_s^3} \tag{13}$$

which shows that for simple DM the noise power depends on the cube of the sampling rate and the square of the signal frequency. This result shows good correlation with Steele et al. [35], [37], [38], who have analyzed the quantization noise in DM for Gaussian-type input signals. The mean square value of the sine wave is $A^2/2$. Hence the dynamic range of the system

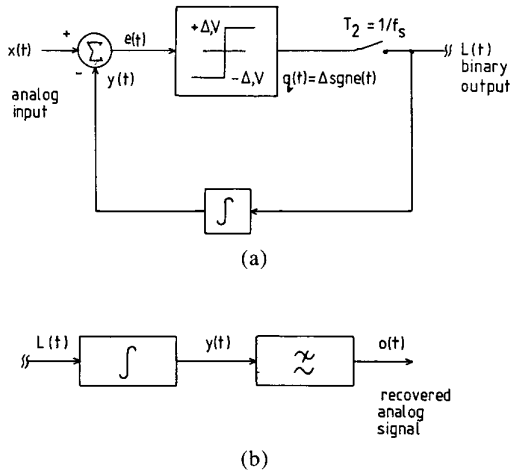


Fig. 13. Delta modulation. (a) Encoder. (b) Decoder.

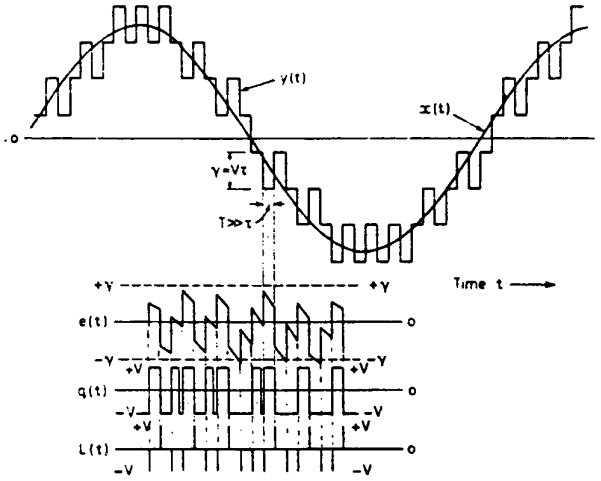


Fig. 14. Delta-modulation signal waveforms. (After Steele [34])

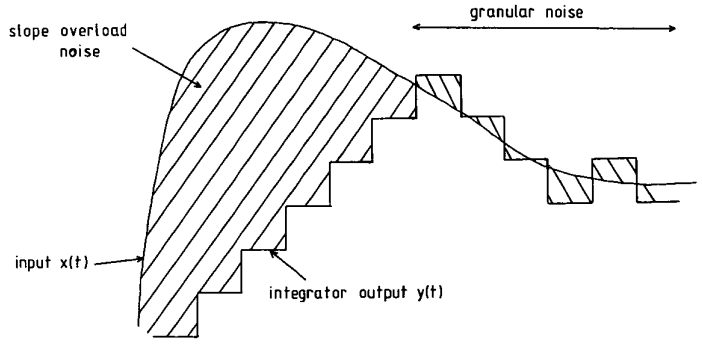


Fig. 15. Delta-modulation waveforms, demonstrating two types of noise: granular and slope overload.

in decibels is given by

$$\text{SQNR} = 10 \log \frac{A^2/2}{A^2 4\pi^2 f_i^2 f_b / f_s^3} = \left[-16 + 10 \log \left(\frac{f_s^3}{f_b f_i^2} \right) \right] \text{ dB} . \quad (14)$$

To achieve 96-dB SQNR, which makes DM compatible with 16-bit PCM, requires a clock rate of 200 MHz, choosing $f_i = f_b = 20$ kHz.

DM offers the user an alternative (much simpler) scheme for encoding audio signals with an increase in quantization noise as compared to PCM or DPCM.

2.6 Delta–Sigma Modulation

Linear DSM [31] is similar in operation to DM. It is only the rearrangement of the integrator into the forward path that is different. Through rearrangement of the loop structure the SQNR is determined to be

$$\text{SQNR} = \left[-14 + 10 \log \left(\frac{f_s}{f_b} \right)^3 \right] \text{ dB} \quad (15)$$

for a single integrator system. Comparing this result with Eq. (14) we observe that in both cases the SQNR is proportional to the clock rate cubed, while in contrast to the DM encoding scheme, the SQNR in a DSM system is independent of the input signal frequency.

DM and DSM convert analog voltages into very fast serial bit stream with a minimum of circuit architecture. The attractiveness of these conversion techniques is its inherent simplicity, lack of problematic SHCs, and high-order analog AAFs (due to oversampling). For these techniques to be applicable toward high-resolution digital audio we need to observe an increase in SQNR for realistic clock rates. Another requirement for compatibility with existing digital audio system, is that the high-speed digital code be decimated with no loss in SQNR to $\geq$16 bit $\times$ 44.1 or 50 kHz.

2.7 High-Order Linear Delta or Delta–Sigma Modulation

One must ask the question, can we enhance the resolution of DM- or DSM-type coders such that their performance would begin to mimic state-of-the-art PCM-type coding schemes, which currently achieve 96-dB SQNR? We have already shown that the SQNR of DSM is increased by the cube of the clock rate [Eq. (15)] for a first-order system. Therefore what is the effect of increasing the number of integrators N in the forward path on the SQNR?

Inose et al. [39] have developed a general expression for the quantization noise power N_q^2 of a unity-bit (DSM)-type coder, where

$$N_q^2 = \left(\frac{\Delta^2 T_s}{6\pi} \right) \int_{-\omega_c}^{\omega_c} \left| \frac{1}{H(\omega)} \right|^2 \left(\frac{\omega T_s}{2} \right) S_a(\omega)^2 \, d\omega . \quad (16)$$

Here $H(\omega)$ is the transfer function of the forward path loop filter, T_s is the sampling rate, Δ is the peak output of $H(\omega)$, ω_c is the signal bandwith, and $S_a(\omega) = \sin \omega/\omega$.

From Eq. (16) we note that the quantization noise is weighted by $H(\omega)$. Therefore by increasing the number of cascaded integrators we gain a significant reduction of in-band quantization noise power. Computing the SQNR for a second-order $H(\omega)$ we find that the SQNR is proportional to the fifth power of the sampling frequency, whereas for first order it goes as the third power. Thus the theory predicts an enhancement in performance as the order of the loop increases.

However, when more than two integrators are embedded within the structure, severe stability constraints result, whereby only second-order filters are deemed practical [40], [41]. Several attempts have been made using adaptive and predictive techniques to guarantee the stability performance of high-order DSM or DM systems [42]–[46] with varying degrees of success.

If we are to apply DSM successfully toward high-resolution digital audio, we must sample at a rate >> Nyquist frequency (oversampling) and develop a loop filter $H(\omega)$ that will provide sufficient noise shaping and be stable under all input conditions.

2.8 Summary

In this section we have discussed various techniques for converting analog signals into digital bit streams. We have shown that certain conversion systems offer a simpler technique of conversion (as compared to PCM) and in some instances similar noise performance.

Through our development of SQNR expressions for DM systems it was concluded that the in-band quantization noise power can be reduced by increasing the clock rate and the number of cascaded integrators (order of loop filter).

Therefore by utilizing the techniques of oversampling and high-order filter structures (if one can maintain closed-loop stability) it should be possible to develop a high-resolution $\geq$16-bit coder which employs topologically simple circuit structures to convert complex analog music signals into accurate digital bit streams. Section 3 describes one such system, which is capable of approximately 20-bit theoretical and computed resolution and has been implemented in hardware.

3 HIGH-RESOLUTION, LINEAR ANALOG-TO-DIGITAL CONVERSION USING OVERSAMPLING AND SPECTRAL NOISE SHAPING TECHNIQUES

In this section we describe a linear method of digitizing music signals based upon higher order, linear delta–sigma modulation (HLDSM) (Fig. 16). The encoder incorporates an Nth-order feedback–feedforward integrating filtering structure $H(s)$, which has theoretical gains in excess of 250 dB at 20 Hz and 30 dB at 20 kHz. By placing this loop filter prior to the 4-bit quantizer, we substantially decrease the in-band (20 Hz to 20 kHz) quantization noise power by $|1 + H(f)|^{-2}$

while increasing the out-of-band noise power (this is what is meant by spectral noise shaping) (Fig. 17). As the loop gain of the encoder approaches 0 dB, this unique loop filter topology exhibits single-pole attenuation, making the system stable over a wide range of operating conditions.

The HLDSM A/D converter operates at clock rates many times greater than the Nyquist rate, which reduces in-band noise power by spreading the original noise spectrum over a much larger bandwidth (Fig. 18). This drastically relaxes the analog AAF requirements and eliminates the SHC, while the importance of each new digital word is lessened (since we have lots of samples representing the input signal) as compared to Nyquist-sampled PCM.

These virtues have allowed us to construct a topologically simple A/D converter, which rivals state-of-the-art PCM audio quantizer performance.

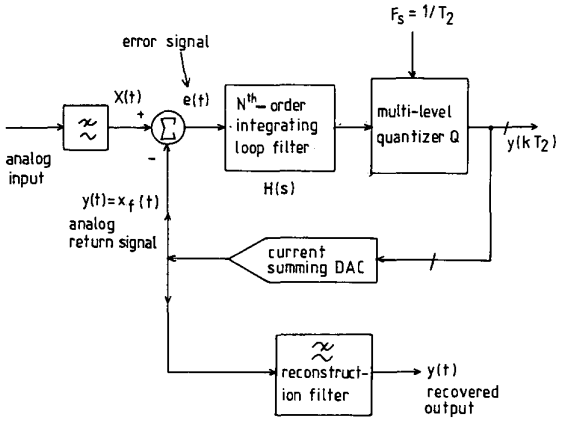


Fig. 16. Block diagram of Nth-order noise-shaped oversampled analog-to-digital converter.

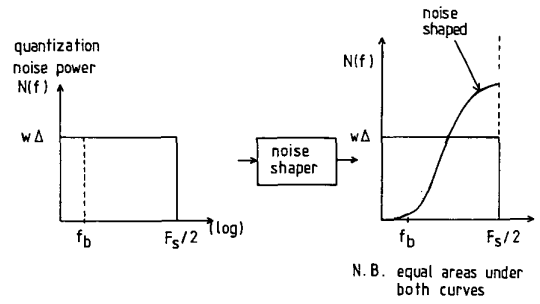


Fig. 17. Spectral noise shaping of quantization noise when f_b represents upper frequency of interest.

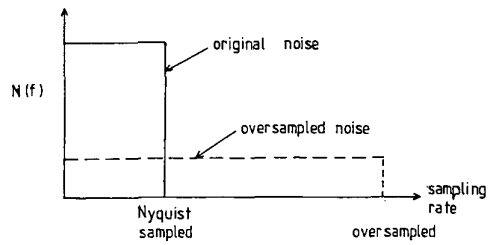


Fig. 18. Reduction of quantization noise through oversampling of analog input.

3.1 Operation of the HLDSM Analog-to-Digital Converter

The HLDSM A/D converter accepts two analog inputs (Fig. 16), the input signal $x(t)$ and the feedback signal $x_f(t)$. The front-end differencing-integrating loop filter (DLF) with transfer function $H(s)$ subtracts and integrates the two signals to produce the integrated error signal $e(t)$. This error signal for $N = 1$ is given by

$$\int_{k_i T_2}^{(k_i+n)T_2} e(t)\,dt = \int_0^T \{x(t) - x_f(t)\}\,dt \qquad (17)$$

where for the first clock cycle $x_f(t) \cong 0$. Therefore from Eq. (17), $e(t) \cong x(t)$ or, equivalently, the output sequence $y[kT_2]$ represents the input signal $x(t)$. Over k clock cycles the feedback signal tracks the input signal, providing a difference signal which is added to past values, ensuring that our oversampled output is a concise representation of the input signal.

3.2 Computer Simulation of the HLDSM Analog-to-Digital Converter

A computer program was written to evaluate the SQNR, quantizer activity, and stability of the HLDSM A/D conversion system. The programmable variables are the number of integrators N, the sampling-rate to input-signal-frequency ratio R, and the form of the input signal. The computer simulation shall provide us with sufficient information to design a hardware version of the encoder.

An appropriate computer model of the HLDSM A/D converter is constructed after Hawksford [47] and consists of cascaded integrators with feedback and feedforward paths. This approach allows us to obtain a loop filter with transfer function $H(s)$ which has single-pole behavior at high frequencies and N-pole behavior at audio frequencies. Achieving an $H(s)$ of this form helps provide closed-loop stability, while providing excellent noise shaping within the audio band. Also included in the model is a nonsaturating quantizer Q and a 10-pole analog recovery filter (ARF). This model forms the basis for the computer algorithm.

Since we are interested in computing the SQNR of our A/D converter, two simulation coders are placed in parallel, one with a quantizer, the other without (Fig. 19). This allows us to compute the error due to quantization noise only, through taking the difference between the two outputs at each sampling instant and summing these error and signal values over the total sampling period. The number of time samples used to compute the SQNR was 2000.

Table 2 summarizes the results of the SQNR simulation as a function of oversampling ratio R, number of integrators N, and clock rate F_s for a sine-wave input of the form $v(t) = A \sin(\omega t + \phi)$.

The computed SQNR results clearly indicate the ad-

vantage of oversampling and increasing the number of integrators. For example, if we choose $R = 200$, we gain a 25-dB advantage in SQNR by going from two to three integrators.

Another important area to look at in the simulation is for a given value of N and R, how many quantum levels ($\Delta = A/\sqrt{2}$) are being exercised at the quantizer input. Fig. 20 gives histograms for $N = 1$ to 4 and $R = 200$, showing the percent occurrence of each quantum over the entire sampling period. In this instance the number of quantum levels excited increases from 5 to 15 over this range of N. Fig. 21 plots the quantizer activity for $N = 6$ and $R = 400$, and we see distinctly that as N becomes excessive, the activity, hence, complexity of the quantizer (from a hardware point of view) increases substantially.

Since the quantizer would be of the flash conversion type, which is required for high-speed operation, we need to trade off quantizer activity versus SQNR. By minimizing the number of quantum levels, the hardware

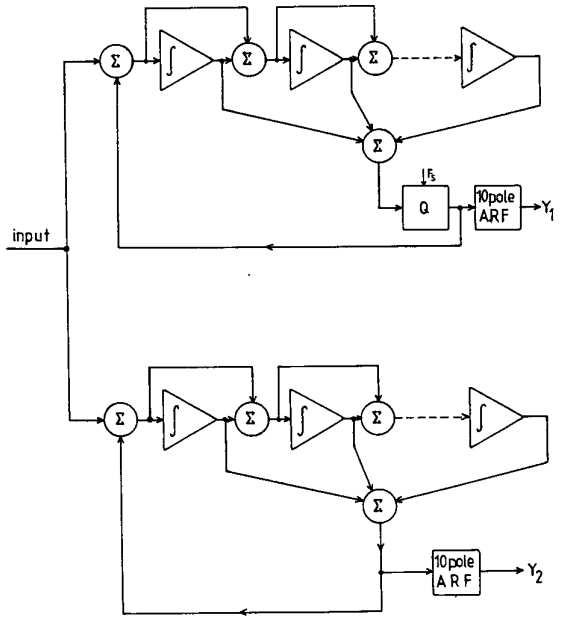


Fig. 19. Computer simulation model which computes SQNR due to quantizer noise.

Table 2. Computed SQNR results.

R	F_s	N	Computed SQNR (dB)
100	2 MHz	1	48
100	2 MHz	2	66
100	2 MHz	3	85
100	2 MHz	4	100
200	4 MHz	1	53
200	4 MHz	2	79
200	4 MHz	3	105
200	4 MHz	4	125
400	8 MHz	1	57
400	8 MHz	2	92
400	8 MHz	3	122
400	8 MHz	4	151

realization is greatly simplified, whereas if we had chosen $N = 6$ and $R = 400$, we would require 50 high-speed comparators and latches.

Another reason to keep N reasonably bounded is to maintain inner loop stability. Since each integrator produces an inner loop time delay due to finite gain–bandwidth product, instability could result for a given clock rate T_2, integration time constant T_1, and gain K.

The computed sine- and square-wave responses of the HLDSM A/D converter are shown in Fig. 22 for an oversampling ratio of 200 and a third-order loop filter. The sine-wave output allows us to observe that the loop produces a phase shift, while the square-wave output exhibits a slightly underdamped response, demonstrating excellent inner loop stability.

3.3 HLDSM Analog-to-Digital Converter Noise Mechanisms

The encoder in Fig. 16 consists of a differencing loop filter with transfer function $H(f)$, a uniform quantizer Q, and a D/A converter. The two main sources of noise are introduced by the loop filter and quantizer. A noise model is proposed, where E_{ni}^2 is the noise due to the loop filter and N_q^2 accounts for quantizer noise (Fig. 23).

If we assume that the noise sources are uncorrelated and additive, we can determine how the output noise power is affected by each internal noise source separately. Applying the principle of linear system theory, the output noise voltage squared E_{01}^2 due to E_{ni}^2 is given by

$$E_{01}^2 = E_{ni}^2 \left[\frac{|H(f)|^2}{|1 + H(f)|^2} \right] . \tag{18}$$

For our applications, $|H(f)| >> 1$ within the information bandwidth (dc → 20 kHz), Eq. (18) becomes

$$E_{01}^2 \cong E_{ni}^2 \tag{19}$$

making it apparent that the loop filter must be of low-noise construct if we are to realize an A/D converter with ≥16-bit performance.

The output noise due to N_q^2 is easily found to be

$$E_{oq}^2 = N_q^2 \left[\frac{1}{|1 + H(f)|^2} \right] \tag{20}$$

which shows that the quantizer noise power is weighted by the inverse magnitude of the transfer function. The results of Eqs. (18)–(20) indicate that the noise performance of the loop filter will place an upper bound on the SQNR capabilities of the encoder.

3.4 Differencing Loop Filter

The design of the DLF is of paramount importance to the successful implementation of the oversampled

noise-shaped HLDSM A/D converter. The filter must maintain high open-loop gain at audio frequencies to ensure adequate loop gain for reducing system distortion and noise, difference and integrate the error signal accurately at speeds up to 10 MHz, while maintaining low-noise operation and high-frequency stability. A high-gain, feedback–forward third-order loop filter using low-noise Signetics 5532 operational amplifiers was implemented successfully (Fig. 24).

The general form of an Nth-order DLF transfer function for identical time constants using the principle of superposition is given by

$$V_0 = K(V_f - V_i)\left[\frac{(1 + j\omega\tau)^{N-1}}{(j\omega\tau)^N}\right] \quad (21)$$

where N represents the number of cascaded integrators and K is the gain.

By choosing the filter time constant judiciously, the DLF shall maintain substantial open-loop gain at frequencies less than the clock rate, which ensures substantial noise shaping over the audio bandwidth. In addition, at frequencies when the system loop gain approaches 0 dB, the filter must exhibit single-pole rolloff to maintain overall closed-loop system stability.

```
A=1.41     N=1

        QUANTUM HISTOGRAM % OCCURRENCES
*****************************************
>>>>>>>                                -6.75
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>  -31.42
>>>>>>>>>>>>>>>>>>>>>>>>>               23.67
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>   31.41
>>>>>>>                                  6.74
*****************************************
NUMBER OF QUANTA =           5
QUANTISED SIGNAL RMS = 1.08074973
Q-OUTPUT SNR = -0.700420447 DB


A=1.41     N=3        R=200

SNR=105.234145

        QUANTUM HISTOGRAM % OCCURRENCES
*****************************************
>                                      -8E-2
>                                     -1.22
>>>>>>                                -4.5
>>>>>>>>>>>>>>>>>                      -12
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>  -20.7
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>   24.19
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>   19.05
>>>>>>>>>>>>>>>>>>                     11.92
>>>>>>>>                                4.92
>>                                      1.42
>                                        0
*****************************************
NUMBER OF QUANTA =          11
QUANTISED SIGNAL RMS = 1.62554554
Q-OUTPUT SNR = -4.24590049 DB
```

The theoretical open-loop gain of the DLF for τ = 0.1, 10, and 100 μs is found to display significant gain over the audio bandwidth and reduced gain slope attenuation above 20 kHz (Fig. 25). This assists in providing appropriate gain and phase margins of the HLDSM A/D converter under closed-loop conditions.

3.5 Noise Analysis of Differencing Loop Filter

To analyze and evaluate the circuit noise of the Nth-stage loop filter (DLF) directly (Fig. 24) would require identifying each noise source and determining its effect on the output noise voltage. An alternative method of analysis is to recognize that the DLF is essentially a high-gain multiple-stage feedback–forward amplifier topology. By choosing such a filter structure, certain generalizations about the expected noise performance are made, which shall ease the amount of computational analysis. Negative feedback–forward does not change

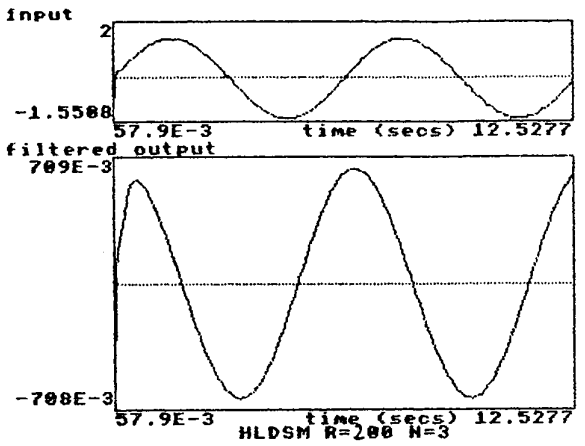
```
A=1.41     N=2

   QUANTUM HISTOGRAM % OCCURRENCES
*************************************
>
>>>>>>>>>>
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>>>>>>>>
>
*************************************
NUMBER OF QUANTA =           7
QUANTISED SIGNAL RMS = 1.94746749
Q-OUTPUT SNR = -5.81532201 DB


  1.41     N=4

   QUANTUM HISTOGRAM % OCCURRENCES
*************************************
>
>
>
>>
>>>>>
>>>>>>>>>>>>>>>>
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
>>>>>>>>>>>>>>>>>
>>>>>>
>>
>
>
>
*************************************
NUMBER OF QUANTA =          15
QUANTISED SIGNAL RMS = 5.82115074
Q-OUTPUT SNR = -15.3260946 DB
```

Fig. 20. Plots of quantizer activity, showing percent occurrence of each quantum for R = 200, N = 1–4.

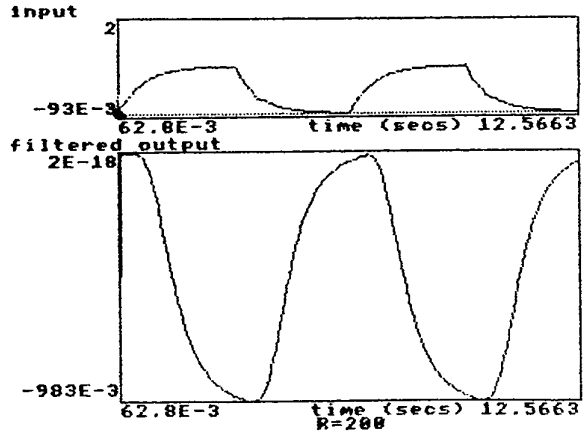
```
Q=1.41     N=6        R=400

      QUANTUM HISTOGRAM % OCCURRENCES
*****************************************
>                                  -7E-2
>                                  -0.13
>                                  -7E-2
>                                  -0.25
>                                  -0.32
>                                  -0.44
>                                  -0.13
>                                 -0.69
>>                                -1.13
>                                 -0.94
>                                 -0.88
>>                                -1.38
>>                                -1.57
>>>>                              -2.25
>>>                               -1.94
>>>>                              -2.57
>>>>>>                            -3.5
>>>>>>                            -3.38
>>>>>>                            -3.57
>>>>>>>>                          -4.5
>>>>>>>>>>>>                      -6.81
>>>>>>>>>>>>>>>>                  -8.62
>>>>>>>>>>>>>>>>>>>>>>>>         -13.5
>>>>>>>>>>>>>>>>>>>>>>>>>>>      -15.62
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>   -17.24
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>  18.61
>>>>>>>>>>>>>>>>>>>>>>>>>>>>     16.98
>>>>>>>>>>>>>>>>>>>>>>>>>        14.61
>>>>>>>>>>>>>>>>>>>>>>           12.67
>>>>>>>>>>>>>>>>>              9.24
>>>>>>>>>>>>>                  6.68
>>>>>>>>                       3.81
>>>>>>>>>                     4.99
>>>>                           2.37
>>>>>                          2.99
>>>>                           2.37
>>>>>                          2.74
>>>                            2.06
>>>                            1.93
>>>                            1.68
>>                             1.12
>                              0.87
>                              1.06
>                               0.43
>                               0.18
>                               0.49
>                               0.24
>                               6E-2
>                               0.18
>                               0.12
>                               0
*****************************************
NUMBER OF QUANTA =          51
QUANTISED SIGNAL RMS = 9.64643542
Q-OUTPUT SNR = -19.7132549 DB
```

Fig. 21. Plot of quantizer activity for $R = 400$, $N = 6$, demonstrating excessive activity.

the equivalent input noise voltage of an amplifier at frequencies for which the internal feedback capacitance of the active device provides isolation [48]–[50]. This does not imply that the output noise of an amplifier is unaffected by feedback–forward paths. We know that the output noise is linearly related to gain. Only the SNR remains unchanged by negative feedback. Although feedback does not add noise, the resistors of the feedback network can contribute to E_{ni}^2 because of their excess and thermal noise mechanisms. By analyzing the theoretical noise performance of multiple-



(a)



(b)

Fig. 22. Time-domain response for HLDSM analog-to-digital converter. (a) Sine-wave response for $R = 200$, $N = 3$. (b) Underdamped square-wave response.

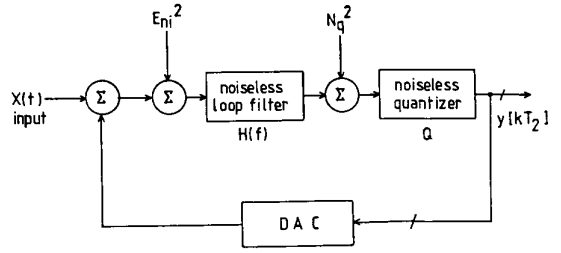


Fig. 23. Complete HLDSM analog-to-digital converter noise model. E_{ni}^2—filter noise, N_q^2—quantizer noise.

stage cascaded amplifiers we show that the noise of the loop filter is confined to the first-stage integrator.

In an earlier paper [51] the theoretical noise of an N-stage cascaded amplifier was evaluated mathematically. Using these results and the noise-equivalent model (Fig. 26), we find that the input noise for the DLF can be approximated by

$$E_{ni}^2 = E_1^2 + [1 + (\omega\tau)^2](E_2 + E_n)^2 + (2RI_n)^2 \quad (22)$$

where we assume that the loop filter noise is dominated by the first-stage integrator. Even though the reactive components do not contribute thermal noise, they do accentuate the noise voltage term of the amplifier, causing an increase in the effective noise voltage over the desired bandwidth.

The SNR is calculated for loop filter time constants of 0.1, 1, 10, and 100 μs over a 20-kHz noise bandwidth and an input signal of 1 V_{rms}. These results show that the SNR contributed by the DLF is 115–116 dB, depending on the time constant. To increase the SNR, we would need to consider lower noise DLF constructs that do not depend on the operational amplifier, which is inherently noisy for the first-stage integrator. A discrete operational amplifier first-stage integrator is proposed (Fig. 27), which offers a 3–6-dB improvement in noise performance, depending on transisitors and biasing conditions.

Since the final-stage integrator has $N - 1$ direct feedforward paths, it must be capable of extremely fast operation. To achieve this speed, the final-stage integrator can be realized using transconductance gain cells driving capacitance loads (Fig. 28). The advantage to this scheme is that the integration capacitor combines with the output transistor's nonlinear collector–base capacitance to extend the cells bandwidth. The hybrid approach offers a more elegant solution toward the implementation of the DLF, increasing both SNR and speed of operation. For ultimate performance of the oversampled A/D converter, the first- and final-stage integrators should be discrete.

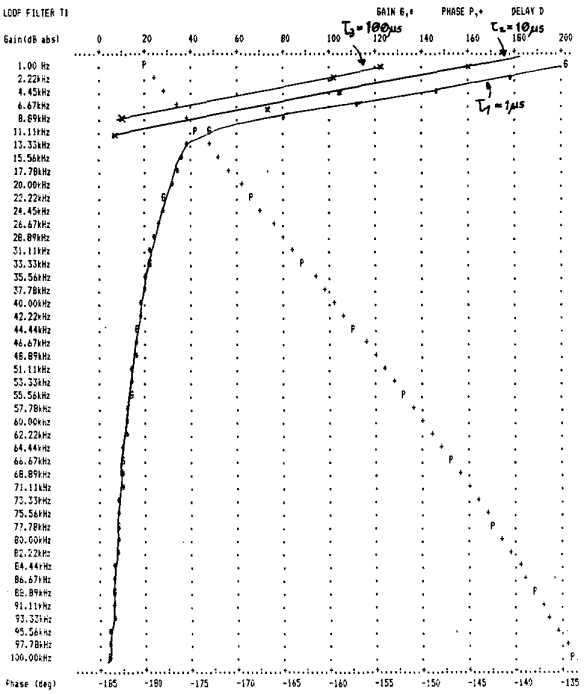


Fig. 25. Differencing loop filter for τ = 1, 10, and 100 μs displays single-pole behavior at frequencies of 100 kHz for T = 1 μs.

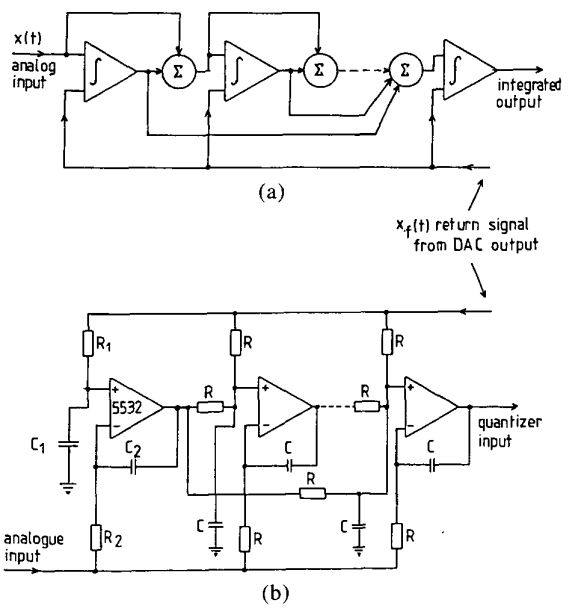


Fig. 24. (a) Block diagram of differencing–integrating loop filter (DLF). (b) Actual implementation of Nth-order DLF using 5532 operational amplifiers.

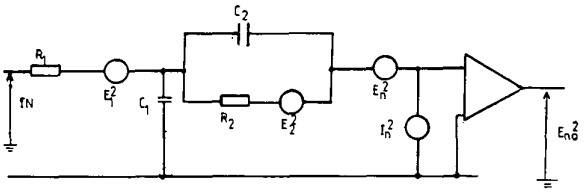


Fig. 26. Noise-equivalent model of single-stage integrator.

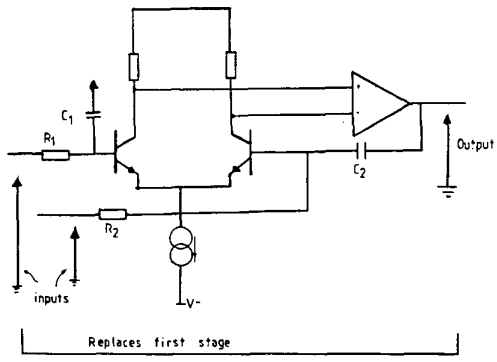


Fig. 27. Discrete operational amplifier topology that will improve noise performance by 6 dB.

3.6 Theoretical Noise Analysis Due to Quantization Noise Only

To evaluate the output noise power due to the inner loop n-level quantizer we integrate Eq. (20) over the information bandwidth, which gives

$$E_{oq}^2 = \frac{N_q^2}{f_s} \int_0^{f_s/R} \frac{1}{|1 + H(f)|^2} \, df \tag{23}$$

where N_q^2/f_s is the quantization noise per unit bandwidth and $f_s/R = 20$ kHz. To make the analysis reasonable, we assume that N_q^2 has a uniform spectral density up to 20 kHz. By observing quantizer activity (Figs. 20 and 21) over the entire sampling period, we observe that total quantizer noise is approximated by

$$N_q^2 \cong (w\Delta)^2 \tag{24}$$

where w is the span of the quantizer.

To evaluate the integral of Eq. (23), we need to obtain an expression for $|1 + H(f)|^2$. For reasonable values of frequency (<100 kHz) Eq. (21) becomes

$$\frac{1}{|1 + H(f)|^2} \cong \left(\frac{1}{K}\right)^2 (2\pi f \tau_1)^{2N} \tag{25}$$

where N represents the number of integrators with time constant τ_1 and gain K. Substituting Eqs. (24) and (25) into Eq. (23) yields

$$E_{oq}^2 = \left\{\left(\frac{w\Delta}{K}\right)^2 (2\pi\tau_1)^{2N} \left(\frac{1}{f_s}\right)\right\} \int_0^{f_s/R} f^{2N} \, df \, . \tag{26}$$

Evaluating the integral, we find that the quantization noise power over a 20-kHz noise bandwidth is given by

$$E_{oq}^2 = \left(\frac{w\Delta}{K}\right)^2 \left(\frac{1}{2N + 1}\right) (2\pi\tau_1 f_s)^{2N} \left(\frac{1}{R}\right)^{2N+1} \, . \tag{27}$$

Eq. (27) clearly indicates that the output noise power (due to N_q^2 only) depends on loop filter gain K, integrator time constant τ_1, number of integrators, span of the quantizer, and system sampling rate (or equivalently oversampling ratio).

To compare the theoretical noise performance, Eq. (26), with the SQNR values obtained via computer simulation (Table 2), the clock rate should be equivalent to τ_1, $K = 1$, and the input signal magnitude will be equal to a quantum level Δ. Under these conditions the theoretical $(SQNR)_T$ becomes

$$(SQNR)_T = 10 \log \left[\frac{1}{w^2\{1/R(2N + 1)\}(2/R)^{2N}}\right] \, . \tag{28}$$

Table 3 compares theoretically evaluated and computer-generated SQNRs (from Table 2) directly, based on quantization noise only.

The tabulated results show reasonable correlation between theoretical and computer-generated SQNRs, particularly for the higher order loop ($N = 3, 4$) filter topologies.

Significant discrepancies are revealed for $N = 1$ and 2, which is largely attributed to the assumption that the quantization noise power has uniform spectral density. This is clearly not the case for the lower order loop filters since we have limited quantizer activity, leading to a lower theoretical SQNR.

Table 3 and Eq. (22) indicate that our target SQNR performance based on loop filter constraints and noise-

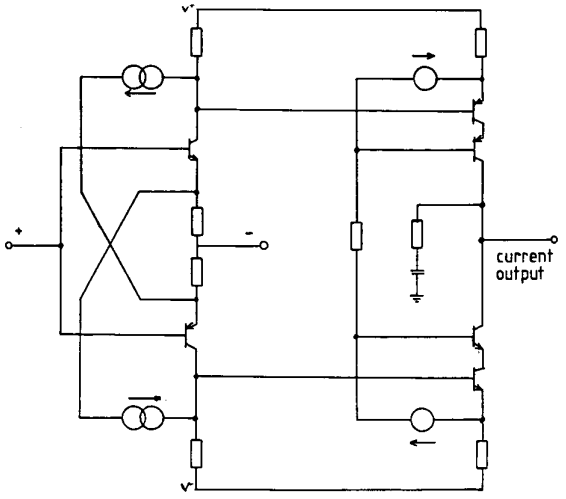


Fig. 28. Final-stage high-speed integrator utilizing transconductance gain cells for extended bandwidth.

Table 3. Theoretical versus computed SQNR.

Sampling rate F_s (MHz)	Oversampling ratio R	Integrators N	Theoretical SQNR (dB)	Computed SQNR (dB)
2	100	1	36	48
2	100	2	59	66
2	100	3	83	85
2	100	4	102	105
4	200	1	44	53
4	200	2	74	79
4	200	3	103	105
4	200	4	129	125

shaped quantization noise is approximately 115 dB. This target SQNR will allow us to determine the operating conditions for the HLDSM A/D converter, which is to be realized in hardware.

3.7 Stability Analysis of the HLDSM Analog-to-Digital Converter

To perform the stability analysis we shall construct a dicrete-time model of the encoder (Fig. 29). The D/A converter is represented by a zero-order hold (ZOH) with transfer function $Z_0(s)$, and the loop filter has an integration time constant τ_1 with transfer function $H(s)$. The D/A converter and the perfect sampler are to be clocked at the same rate T_2. For this analysis we shall assume that the inner loop delays are small compared with the clock speed.

If the assumption is made that the input signal and internal noise is bounded, it is then sufficient for the proof of system stability to show that a bounded input inserted into the loop produces a bounded output. By obtaining the Z transforms of the input and output we are able to ascertain the open-loop transfer function $G(z)$, based on the discrete-time open-loop model (Fig. 30). By assessing the roots (via a root-locus plot) of the loop gain the overall stability of the encoder is known as a function of the system variables τ_1, K, and T_2.

In order to analyze the system in the z plane one is required to construct an accurate linear model of the D/A conversion (ZOH) process. The D/A converter over one clock cycle produces an output of height $x[kT_2]$ and width T_2, having an impulse as its input $x[k]$. The impulse response of the ZOH is a unit pulse of T_2 seconds. Therefore the Laplace transform is given by

$$\begin{aligned} Z_0(s) &= \mathcal{L}\{h(t)\} \\ &= \int_0^\infty [1(t) - 1(t - T_2)]\, e^{-st}\, dt \\ &= \frac{1 - e^{-sT_2}}{s} . \end{aligned} \tag{29}$$

Knowing $Z_0(s)$ and $H(s)$ we can write an expression

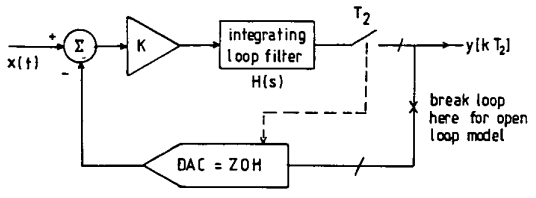


Fig. 29. Discrete-time model of HLDSM analog-to-digital converter.

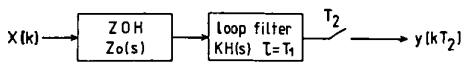


Fig. 30. Discrete-time open-loop system model.

for the open-loop transfer function,

$$G(z) = \frac{Y(z)}{X(z)} = Z\left\{\left(\frac{1 - e^{-T_2 s}}{s}\right)[KH(s)]\right\} \tag{30}$$

where the loop filter transfer function $H(s)$ is

$$H(s) = \frac{(1 + sT_1)^{N-1}}{(sT_1)^N} . \tag{31}$$

Therefore Eq. (30) can be rewritten as

$$G(z) = Z\left\{\left(\frac{1 - e^{-T_2 s}}{s}\right)\left[K\frac{(1 + s\tau_1)^{N-1}}{(s\tau_1)^N}\right]\right\} . \tag{32}$$

Using Z-transform tables and letting $N = 3$,[1] Eq. (32) becomes

$$G(z) = K\left[\frac{a_1 z^2 + a_2 z + a_3}{(z - 1)^3}\right] \tag{33}$$

where $a_1 = T^3 + 3T^2 + T_1$, $a_2 = 4T^3 - 2T$, and $a_3 = T^3 - 3T^2 + T$. T is equivalent to the clock-to-integration-time ratio ($T = T_2/\tau_1$).

The overall closed-loop transfer function is given by

$$A_{CL}(z) = \frac{G(z)}{1 + G(z)\beta} \tag{34}$$

where $\beta = 1$ because we have 100% feedback.

The stability of the system was evaluated [9] by obtaining root-locus plots in the z plane for various values of K and T. To summarize these results, we note that as K is increased, one root will move toward the closest zero inside the unit circle, while the other two roots will initially travel at right angles to the real axis in opposite directions. Over the range of K these roots complete a semicircle, terminating inside the unit circle on the real positive axis. The two roots then split toward the remaining zero and negative infinity. Through keeping track of the root, which tends to negative infinity as a function of K and T (Fig. 31), we observe that as $T \geq 0.35$, the encoder is theoretically unstable for all values of K.

To explain this phenomenon we need to look at how the DLF frequency response is altered by increasing T and how this affects stability. Increasing T implies that the integration time constant is decreasing for a given clock speed. By allowing the filter time constant to decrease we significantly alter the frequency response of the loop filter, as can be seen from Fig. 25. We no longer have single-pole behavior at frequencies when

[1] $N \leq 3$ was always stable in hardware.

the loop gain of the system approaches 0 dB. The increased slope of the loop filter in conjunction with the ZOH sacrifices the phase and gain margin of the HLDSM A/D converter, tending toward instability. The encoder, as realized in hardware, was stable for all values of $N \leq 3$ and for $0.1 < T < 1.0$.

The stability analysis is in reasonable agreement with the actual performance results of the HLDSM A/D converter. Therefore the proposed linear discrete-time model supports the results of the hardware system.

3.8 Hardware Development and Measurements of the HLDSM Analog-to-Digital Converter

Referring to Fig. 16 we observe that the HLDSM A/D converter consists of a differencing-integrating loop filter, a quantizer that performs both level and time quantization, and a D/A converter in the feedback loop to reconstruct the analog signal. A target SQNR based on loop filter and quantization noise and hardware constraints is 115 dB. Therefore from Table 3 we desire an oversampling ratio of ≥ 200 and a fourth-order loop filter.

A hardware system was implemented successfully using a third-order loop filter, a uniform 16-level flash quantizer which performed time and level quantization, and a current-summing nonclocked D/A converter. The 16 levels were chosen based on computer simulation of the expected quantizer activity as a function of the number of integrators and the oversampling ratio.

The total harmonic distortion was measured (Fig. 32) as a function of the input signal amplitude and the loop filter order. It varies from -47 to -82 dB for $N = 1$ to 3.

The noise of the HLDSM encoder with signal was also measured by taking the output of the feedback loop D/A converter, passing the signal through a passive low-pass filter and measuring the resultant product via a spectrum analyzer (Fig. 33). The SQNR is approximately 90–100 dB over a 100-Hz bandwidth, or 67–77 dB over a 20-kHz noise bandwidth.

The disadvantage to this measurement scheme is that we are measuring the analog output of the simple D/A converter in the feedback loop. Therefore we obtain the noise and distortion of this simple decoder. A more accurate method of measurement and one the authors are currently pursuing is to measure the unique high-speed digital output code of the HLDSM A/D converter using buffer memory and a high-speed computer to compute the error spectrum via an FFT. The results of these tests should indicate a 20–30-dB improvement in measured noise performance.

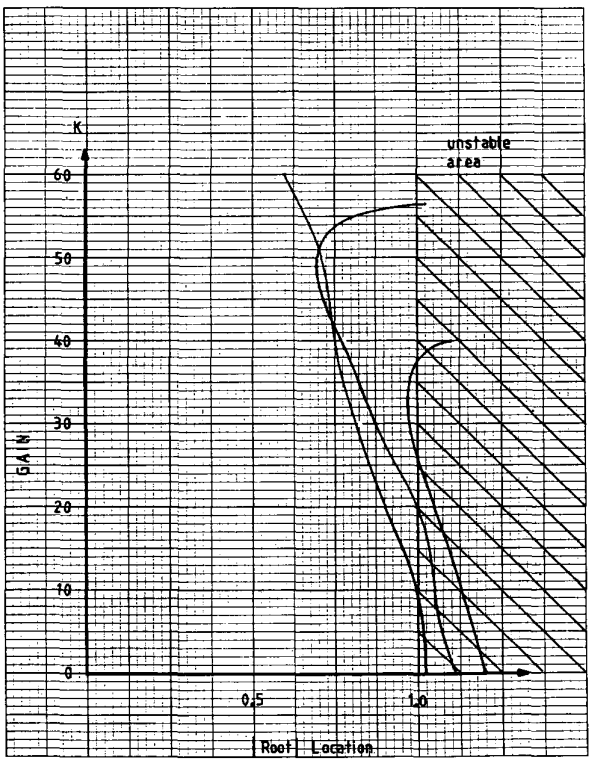


Fig. 31. Magnitude of runaway root in z plane as a function of K and T.

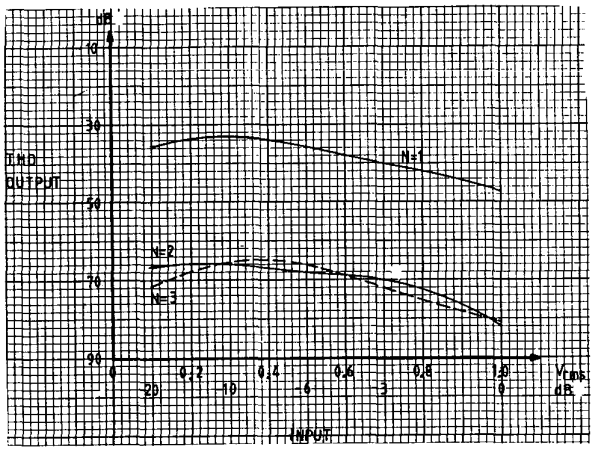


Fig. 32. Total harmonic distortion of HLDSM analog-to-digital converter as N and v_{in} are varied.

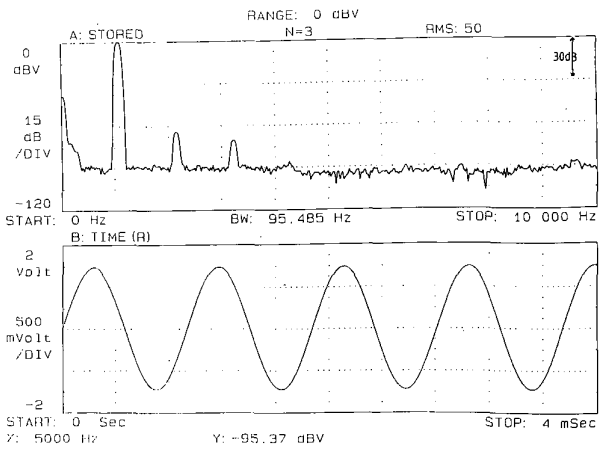


Fig. 33. Measured noise and output signal of HLDSM analog-to-digital converter.

3.9 Summary

The oversampled noise-shaped encoder offers very reasonable measured performance and leads to topologically simple hardware structures. Such structures do not require problematic SHCs, high-order analog AAFs, or complex quantizers to achieve results comparable to Nyquist-sampled PCM.

4 CONCLUSIONS

This paper has investigated the limitations and errors of Nyquist-sampled PCM. It was demonstrated that it is difficult to achieve greater than 16-bit performance working within the constraints of current technology.

An oversampled noise-shaped encoder based on DSM was investigated. This technique exploited oversampling to eliminate problematic SHCs and high-order analog AAFs. Inner loop noise shaping greatly simplified the design of the loop quantizer.

The theoretical and computed signal-to-quantization-noise analysis suggest 110–120 dB as a practical limit for this encoding system. The measured SQNR was limited by the measurement approach.

The virtues of the HLDSM A/D converter are the following:

1) SHCs with their problematic amplitude and timing errors are eliminated.

2) Analog AAF requirements are substantially relaxed.

3) Theoretically increased resolution.

4) Less complex circuit building blocks, reducing A/D converter cost.

5) Easily implemented in integrated circuit form.

6) By employing FIR decimation filters the HLDSM A/D converter output can be resequenced to any digital code format with minimum loss in SQNR.

5 ACKNOWLEDGMENT

The ideas presented in this paper are dedicated to the loving memory of Jeffery Reiter.

6 REFERENCES

[1] K. W. Cattermole, *Principles of Pulse Code Modulation* (Iliffe, 1969).

[2] G. M. Gordon, "Linear Electronic Analog/Digital Conversion Architectures—Their Origins, Parameters, Limitations and Applications," *IEEE Trans. Circuits Sys.*, vol. CAS-25, pp. 391–418 (1978 July).

[3] B. Blesser, "Digitization of Audio: A Comprehensive Examination of Theory, Implementation, and Current Practice," *J. Audio Eng. Soc.*, vol. 26, pp. 739–771 (1978 Oct.).

[4] B. Blesser, B. Locanthi, and T. G. Stockham (Eds.), *Digital Audio, Collected Papers from the AES Premier Conf.* (Rye, NY, 1982 June 3–6).

[5] P. J. Bloom, "High Quality Digital Audio in the Entertainment Industry," *IEEE ASSP Mag.*, pp. 2–25 (1985 Oct.).

[6] R. P. Talambiras, "Some Considerations in the Design of Wide Dynamic Range Audio Digitizing Systems," presented at the 57th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 25, p. 514 (1977 July/Aug.), preprint 1226.

[7] R. P. Talambiras, "Digital–Analogue Converters: Some Problems in Producing High Fidelity Signals," *Comput. Des.*, pp. 63–69 (1976 Jan.).

[8] J. Schenkel, "From Analog to Digital in Data Acquisition," *Electron. Prod.*, pp. 74–78 (1984 July 10).

[9] T. F. Darling, "Oversampled Analogue–Digital Conversion for Digital Audio Systems," M. Phil. thesis, University of Essex, UK (1987).

[10] M. E. Valkenburg, *Analog Filter Design* (CBS College Pub., New York, 1982).

[11] G. Daryanani, *Principles of Active Network Synthesis and Design* (Wiley, New York, 1982).

[12] R. Lagadec, D. Weiss, and R. Greutmann, "High-Quality Analog Filters for Digital Audio," presented at the 67th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 923 (1980 Dec.), preprint 1707.

[13] K. W. Henderson and W. H. Kautz, "Transient Response of Conventional Filters," *IRE Trans. Circuit Theory*, pp. 333–347 (1958 Dec.).

[14] J. Meyer, "Time Correction of Anti-Aliasing Filters in Digital Audio Systems," *J. Audio Eng. Soc. (Engineering Reports)*, pp. 132–137 (1984 Mar.).

[15] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1975).

[16] A. Antoniou, *Digital Filters: Analysis and Design* (McGraw-Hill, New York, 1979).

[17] R. Lagadec and T. G. Stockham, "Dispersive Models for A-to-D and D-to-A Conversion Systems," presented at the 75th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 32, p. 469 (1984 June), preprint 2097.

[18] E. L. Zuch, "Designing with a Sample-Hold Won't Be a Problem if You Use the Right Circuit," *Electron. Des.*, no. 23, pp. 84–89 (1978 Nov. 8).

[19] E. L. Zuch, "Keep Track of a Sample-Hold from Mode to Mode to Locate Error Sources," *Electron. Des.*, no. 25, pp. 80–87 (1978 Dec. 6).

[20] E. L. Zuck, "Pick Sample-Holds by Accuracy and Speed and Keep Hold Capacitors in Mind," *Electron. Des.*, no. 26, pp. 84–90 (1978 Dec. 20).

[21] H. Pichler and P. Skritek, "Design Principles of Sample-and-Hold Circuits for Digital Audio Systems," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 372 (1980 May), preprint 1584.

[22] *Linear Databook* (National Semiconductor Corp., San Jose, CA, 1982), sec. 7-4.

[23] *Small Signal FET Design Catalog* (Siliconix Inc., 1982 Nov.).

[24] R. Lagadec, "Digital Sampling Frequency Conversion," in B. Blesser, B. Locanthi, and T. G.

Stockham (Eds.), *Digial Audio, Collected Papers from the AES Premier Conf.* (Rye, NY, 1982, June 3–6), pp. 90–96.

[25] N. W. Jung and R. Marsh, "Picking Capacitors: Selection of Capacitors for Optimum Performance, pts. 1 and 2," *Audio* (1980 Feb., Mar.).

[26] N. H. C. Gilchrist, "Analog-to-Digital and Digital-to-Analog Converters for High-Quality Sound," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 372 (1980 May), preprint 1583.

[27] R. Karwoski, "Predictive Coding for Greater Accuracy in Successive Approximation A/D Converters," presented at the 57th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 77, p. 514 (1977 July/Aug.), preprint 1228.

[28] R. A. McDonald, "Signal to Noise and Idle Channel Performance of Differential Pulse Code Modulation Systems—Particular Applications to Voice Signals," *Bell Sys. Tech. J.*, pp. 1123–1151 (1966 Sept.).

[29] E. M. Deloraine, S. Van Mierlo, and B. Derjavitch, French patent 932,140 (1946 Aug. 10); U.S. patent 2,629,857 (1953, Feb. 24).

[30] F. de Jager, "Deltamodulation: A Method of PCM Transmission Using 1 Unit Code," Philips Res. Rep. 7, pp. 442–466 (1952).

[31] R. Steele, *Delta Modulation Systems* (Pentech Press, London, 1975).

[32] G. L. Baldwin and S. K. Tewksbury, "Linear Delta Modulator Integrated Circuit with 17-Mbit/s Sampling Rate," *IEEE Trans. Commun.*, vol. COM-22, pp. 977–985 (1974 July).

[33] A. Van De Plassche, "A Sigma–Delta Modulator as an A/D Converter," *IEEE Trans. Circuits Sys.*, vol. CAS-25, pp. 510–514 (1978 July).

[34] R. Steele, "Peak Signal-Noise Ratio Formulas for Multistage Delta Modulation with *RC* Shaped Gaussian Input Signals," *Bell Sys. Tech. J.*, vol. 61, pp. 347–362 (1982 Mar.).

[35] H. Levitt et al., "Perception of Slope Overload Distortion in Delta-Modulated Speech Signals," *IEEE Trans. Audio Electroacoust.*, pp. 240–247 (1970 Sept.).

[36] N. S. Jayant, and A. E. Rosenberg, "The Preference of Slope Overload to Granularity in the Delta Modulation of Speech," *Bell Sys. Tech. J.*, vol. 50, pp. 3117–3125 (1971 Dec.).

[37] R. Steele, "SNR Formula for Linear Delta Modulation with Band Limited Flat and *RC* Shaped Gaussian Signals," *IEEE Trans. Commun.*, vol. COM-28, pp. 1978–1984 (1980 Dec.).

[38] J. B. O'Neal, "Delta Modulation Quantizing Noise Analytical and Computer Simulation Results for Gaussian and Television Signals," *Bell Sys. Tech. J.*, vol. 45, pp. 117–141 (1966 Jan.).

[39] H. Inose and Y. Yasuda, "A Unity Bit Coding Method by Negative Feedback," *Proc. IEEE*, vol. 51, pp. 1524–1535 (1963 Nov.).

[40] P. P. Wang, "An Absolute Stability Criterion for Delta Modulation," *IEEE Trans. Commun. Technol.*, vol. COM-16, pp. 186–188 (1968 Feb.).

[41] P. T. Nielsen, "On the Stability of a Double Integration Delta Modulator," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 364–366 (1971 June).

[42] R. W. Adams, "Companded Predictive Delta Modulation: A Low-Cost Conversion Technique for Digital Recording," *J. Audio Eng. Soc.*, vol. 32, pp. 659–672 (1984 Sept.).

[43] R. Elen, "dBX Digital—An Overview," *Studio Sound*, pp. 50–52 (1983 Feb.).

[44] R. O. Carter, "Theory of Syllabic Compandors," *Proc. IEE* (London), vol. 111, pp. 503–511 (1964 Mar.).

[45] B. Smith, "Instantaneous Companding of Quantized Signals," *Bell Sys. Tech. J.*, vol. 36, pp. 654–709 (1957 May).

[46] P. Cummiskey, "Single-Integration Adaptive Delta Modulation," *Bell Sys. Tech. J.*, pp. 1463–1473 (1975 Oct.).

[47] M. J. Hawksford, "*N*th-Order Recursive Sigma–ADC Machinery at the Analog–Digital Gateway," presented at the 78th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, pp. 586, 588 (1985 July/Aug.), preprint 2248.

[48] J. D. Meindl, *Micropower Circuits* (Wiley, New York, 1979), p. 145.

[49] E. G. Nielsen, "Behavior of Noise Figure in Junction Transistors," *Proc. IRE*, vol. 45, pp. 957–963 (1957 July).

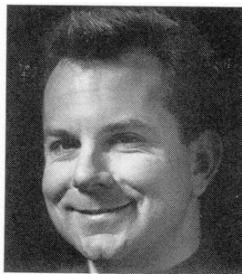
[50] R. D. Middlebrook, "Optimum Noise Performance of Transistor Input Circuits," *Semiconductor Prod.*, pp. 14–20 (1958 July/Aug.).

[51] T. F. Darling, "Mathematical Noise Modeling and Analysis of Some Popular Preamplifier Circuit Topologies," *J. Audio Eng. Soc.*, vol. 35, pp. 15–23 (1987 Jan./Feb.).

THE AUTHOR

Timothy Darling received the B.Sc. degree in electronics and the M.Sc. degree in device physics from the University of California, Santa Barbara, and the M. Phil. degree from the Electronics Systems Engineering Department at the University of Essex, UK.

At present he is a Senior Design Engineer with Science Applications International Corporation (SAIC/MariPro), Goleta, CA, where he is involved in the design of analog and digital acquisition and data transmission systems for sonar applications. He is currently developing a multichannel, fiberoptic underwater transmission link which has telephone and range-tracking applications. In addition, he maintains a visiting lectureship in the Department of Electrical and



Computer Engineering at the University of California, Santa Barbara. His primary lecture responsibilities include courses in circuit design, audio engineering, and device physics and processing. The ideas formulated in this paper evolved during a 2-year post-graduate study at the University of Essex audio research group. The author's main research areas are high-resolution analog-to-digital conversion (using oversampling and noise-shaping techniques), linear digital filtering for sample rate reduction, and circuit design.

PAPERS

Digital-to-Analog Converter with Low Intersample Transition Distortion and Low Sensitivity to Sample Jitter and Transresistance Amplifier Slew Rate*

MALCOLM OMAR HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

Multibit digital-to-analog converter technology now claims a sample amplitude accuracy of about 20 bit. However, to achieve commensurate performance in digital audio applications, both sample timing and the complete sample waveform must also have corresponding accuracies. The errors due to jitter and slew rate are analyzed, as they are treated as a unified process in the presence of a correlation between the audio signal and sample timing. The concept of jitter-equivalent slew-rate-induced distortion is introduced and an enhanced multibit topology proposed, which offers low sensitivity to both jitter and slew-rate distortion and improves upon waveform reconstruction by exhibiting no waveform discontinuities.

0 INTRODUCTION

The fundamentals of sampling theory, uniform amplitude quantization, and dither are well documented [1]. If these processes are implemented correctly, the only errors at the digital-to-analog gateway output are a band limitation of the input signal and a predictable increase in noise level. However, there is evidence [2] that errors resulting from imperfect electronics do have a deleterious effect on sonic performance, even though the perceptual correlations are not completely understood. For example, although multibit digital-to-analog converters (DACs) using error-correction techniques can achieve precise level reconstruction, nonlinearity within the sample transition region resulting from slew-rate and induced jitter can produce impairment [3]–[6]. Jitter on the DAC conversion clock can be nonnoiselike and arises within digital circuits [7] from EMC-related interference and from imperfect phase-locked-loop (PLL) performance responding to correlation between the digital audio data and pulse timing in the digital data stream [8]–[10]. Although these problems can be corrected [11], [12] by retiming, this is not always achieved within the constraints of practical circuitry.

A technique is presented that lowers the sensitivity of the DAC to sample timing errors and enables the converter to operate with band-limited signals, which eliminates rapid signal transitions and discontinuities. Jitter and slew-rate-induced distortion are analyzed and unified, as similar errors result from correlation with the intersample values of the audio data. The importance of controlling both pulse shape and timing is also emphasized in sample reconstruction.

1 JITTER IN DIGITAL-TO-ANALOG CONVERSION

In this paper we define two forms of jitter and use the terminology *random jitter* and *correlated jitter*, the latter describing a sample time displacement that is correlated with the state of the system. Jitter is strictly a random event. However, the foregoing definitions are now achieving common usage in this subject area.

In general a reconstructed sample can undergo both amplitude and time displacement, which together constitute an error vector $\bar{E}$, as shown in Fig. 1. However, considering only jitter, two classes of sample format are identified:

1) Samples that are impulsive, of uniform shape, and

* Presented at the 93rd Convention of the Audio Engineering Society, San Francisco, CA, 1992 October 1–4; revised 1994 August 20.

noninteracting, such as switched-capacitor converters

2) 100% duration pulses, allowing nonlinear intersample interaction within the sample sequence

1.1 Uniform Sampling with Jitter and Noninteracting Pulses

Consider a uniform and impulsive data sequence of sampling rate f_s Hz. The jitter model for the rth sample of weight A_r with jitter ΔT_r is shown in Fig. 2, where the error is the difference between a sample of nominal location and a time-displaced version, and the impulse weight $\{A_r/f_s\}$ of a sample is defined with respect to a nominal rectangular pulse of amplitude A_r and width $1/f_s$.

The Fourier transform $E_r(f)$ of the error for the rth sample is

$$E_r(f) = \frac{A_r}{f_s}(1 - e^{-j2\pi f \Delta T_r}) \tag{1a}$$

which, for $2\pi f \Delta T_r << 1$, approximates

$$E_r(f) = j\frac{A_r}{f_s} 2\pi f \Delta T_r . \tag{1b}$$

Hence for an N-sample cyclic sequence, the error $E_N(Lf_s/N)$ is given by

$$E_N\left(\frac{Lf_s}{N}\right) = j\,2\pi \frac{L}{N^2} f_s \sum_{r=0}^{N-1} A_r \Delta T_r e^{-j2\pi Lr/N} \tag{2}$$

Eq. (2) shows that the error spectrum is proportional to the harmonic number L of the sequence repetition frequency f_s/N Hz, but that the microstructure of the spectrum depends on intermodulation between the pulse weighting sequence $\{A_r\}_N$ and the pulse jitter sequence $\{\Delta T_r\}_N$.

1.2 Uniform Sampling with Jitter and Samples with Nonlinear Interaction

Although sample timing errors give rise to the error spectra described in Section 1.1, it is more common for a DAC to use 100% duration sample reconstruction. This may arise directly from the DAC output or via a sample-and-hold circuit used to eliminate glitches during DAC sample transitions. Although this strategy maximizes signal energy and improves immunity to system noise, the effect of sample jitter now not only modifies sample timing but also affects the sample weight, as the area under a reconstructed pulse changes as it interacts with the two adjacent samples in the sequence.

Consider a sample of amplitude A_r, nominal width $1/f_s$, and with leading- and trailing-edge jitter ΔT_r and ΔT_{r+1}, respectively. The sample construction is shown in Fig. 3. The Fourier transform $P_r(f)$ of the rectangular pulse shown in Fig. 3 can be expressed as

$$P_r(f) = \frac{A_r}{j2\pi f}\left\{\exp\left[-j2\pi f\left(-\frac{1}{2f_s} + \Delta T_r\right)\right] - \exp\left[-j2\pi f\left(\frac{1}{2f_s} + \Delta T_{r+1}\right)\right]\right\}$$

which, for $\{2\pi f \Delta T_r\} << 1$ and $\{2\pi f \Delta T_{r+1}\} << 1$, simplifies to

$$P_r(f) = \frac{A_r}{f_s}\frac{\sin(\pi f/f_s)}{\pi f/f_s} - A_r\left[\cos\left(\frac{\pi f}{f_s}\right)(\Delta T_r - \Delta T_{r+1}) + j\sin\left(\frac{\pi f}{f_s}\right)(\Delta T_r + \Delta T_{r+1})\right].$$

However, for $\Delta T_r = 0$ and $\Delta T_{r+1} = 0$, the target Fourier transform of a rectangular sample $P_T(f)$ is expressed as

$$P_T(f) = \frac{A_r}{f_s}\frac{\sin(\pi f/f_s)}{\pi f/f_s} .$$

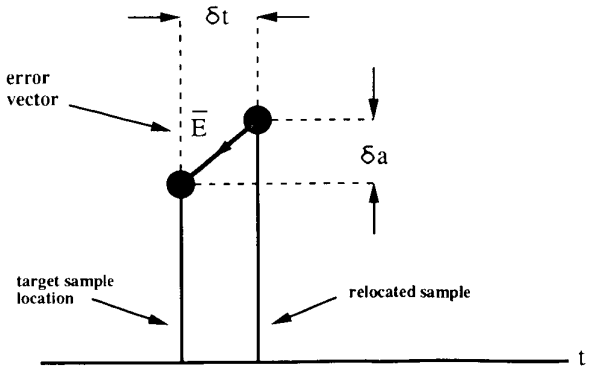


Fig. 1. Error vector resulting from simultaneous amplitude and time errors of a sample.

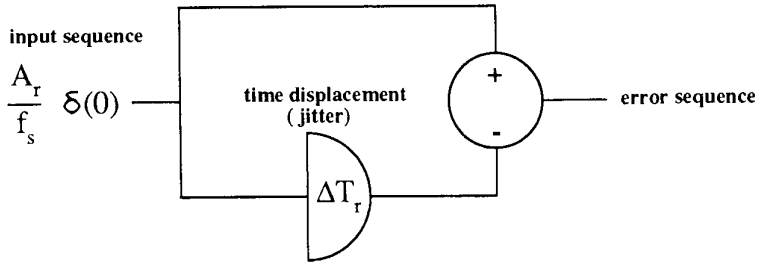


Fig. 2. Elementary model of sampling jitter.

Thus the error spectrum $E_{Rr}(f) = P_T(f) - P_r(f)$, that is,

$$E_{Rr}(f) = A_r\left[\cos\left(\frac{\pi f}{f_s}\right)(\Delta T_r - \Delta T_{r+1}) + j\sin\left(\frac{\pi f}{f_s}\right)(\Delta T_r + \Delta T_{r+1})\right]. \quad (3)$$

Summing the terms over N samples of a cyclic sequence, the Fourier transform $E_{RN}(Lf_s/N)$ is

$$E_{RN}\left(\frac{Lf_s}{N}\right) = \frac{f_s}{N}\left[\cos\left(\frac{\pi L}{N}\right)\sum_{r=0}^{N-1} A_r(\Delta T_r - \Delta T_{r+1})\,e^{-j2\pi Lr/N} + j\sin\left(\frac{\pi L}{N}\right)\sum_{r=0}^{N-1} A_r(\Delta T_r + \Delta T_{r+1})\,e^{-j2\pi Lr/N}\right]. \quad (4)$$

Comparing Eqs. (2) and (4) there is now an in-phase component weighted by a $\cos(\pi L/N)$ multiplier that extends the spectrum to dc, where correlation between timing error and signal results in a complicated error spectrum that may not be masked by program material.

Alternatively, an impulsive error sequence can be located at the interface between adjacent samples where the impulse weight is proportional to the pulse-area error resulting from jitter, whereas the impulse timing corresponds to the jitter. This error impulse sequence has simultaneous amplitude and timing modulation, where

$$\text{pulse-area error} = \{A_{r+1} - A_r\}\{f_s\,\Delta T_{r+1}\}.$$

An error pulse is assumed rectangular with the leading edge located at $t = (r + 0.5)/f_s$ and the trailing edge at $t = (r + 0.5)/f_s + \Delta T_{r+1}$. Thus with reference to the error pulse center the pulse timing error equals $\Delta T_{r+1}/2$. Hence the Fourier transform of the rth error pulse in the sequence is

$$E_{Rr}(f) = [A_{r+1} - A_r][\Delta T_{r+1} f_s]\exp\left[-j2\pi\left(\frac{r + 0.5}{f_s} + 0.5\,\Delta T_{r+1}\right)\right].$$

By forming a summation over an N-sample cyclic sequence, a discrete transform follows as

$$E_{RN}\left(\frac{Lf_s}{N}\right) = \frac{f_s}{N}\sum_{r=0}^{N-1}(A_{r+1} - A_r)\,\Delta T_{r+1}\exp\left[-j\frac{\pi L}{N}(2r + f_s\,\Delta T_{r+1})\right]. \quad (5)$$

To illustrate example error characterizations of jitter when mapped onto and correlated with the audio data sequence, sets of error spectra are presented. The first set uses the data and jitter sequences $\{A_r\}_N$ and $\{\Delta T_r\}_N$ tabulated below and computed over $N = 4096$ samples, where the sampling frequency $f_s = 44.1$ kHz,

$$f_0 = \frac{f_s}{N}\text{ Hz} \qquad f_1 = 752 f_0\text{ Hz}$$

$$f_2 = 1760 f_0\text{ Hz} \qquad d = 10\text{ ns}$$

$$A_r = \left\{\sin\left(2\pi\frac{rf_1}{f_s}\right) + \sin\left(2\pi\frac{rf_2}{f_s}\right)\right\}$$

$$\Delta T_r = d\left\{\sin\left(2\pi\frac{rf_1}{f_s}\right) + \sin\left(2\pi\frac{rf_2}{f_s}\right)\right\}.$$

Here f_0 is the sequence repetition frequency, f_1 and f_2 are the selected signal frequencies, and d is the jitter noise. In this example $f_0 \approx 10.8$ Hz, $f_1 \approx 8.096$ kHz, and $f_2 \approx 18.949$ kHz.

In the following simulated results all spectra are referenced to the input sequence $\{A_r\}_N$, which is designated 0

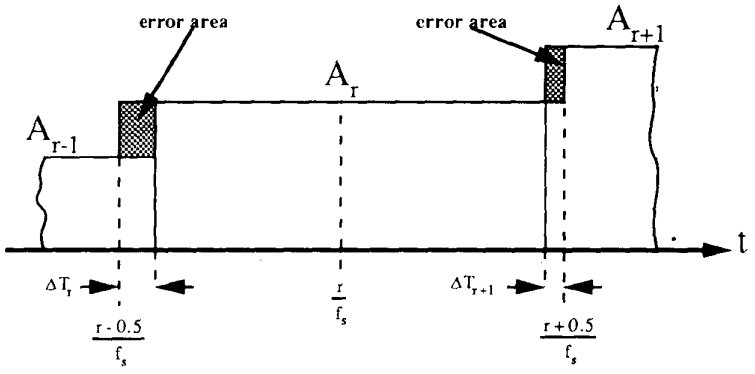


Fig. 3. Construction of 100% rectangular samples with jitter.

dB. The results shown in Fig. 4 correspond to impulsive samples as described in Section 1.1, those in Fig. 5 to the 100% duration samples described in Section 2.1. The lower frequency components are now predictably of higher level.

The second set of results uses similar correlated sequences $\{A_r\}_N$ and $\{\Delta T_r\}_N$, but now a weighted random noise sequence is added to the jitter function so that

$$\Delta T_r = d\left\{\sin\left(2\pi\frac{rf_1}{f_s}\right) + \sin\left(2\pi\frac{rf_2}{f_s}\right)\right\} + J_n\,\mathrm{Rand}(r)$$

where Rand(r) is a random function with a triangular probability distribution function spanning -1 to $+1$, and J_n is the noise weighting factor.

Figs. 6 and 7 show the corresponding results with and without 100% sample reconstruction, where $J_n = 10$ ns, meaning that the jitter probability distribution function is triangular and spans -10 ns to 10 ns. In Fig. 8 a three-dimensional plot illustrates the effects of varying levels of random jitter noise together with the correlated displacements of the sample locations for the case of 100% samples, where J_n is defined as

$$J_n = d\,10^{0.25(1-x)}$$

for $1 \leqslant x \leqslant 16$ in unit steps of x and $d = 10$ ns.

Finally, in the third set a modified jitter sequence is simulated to demonstrate the effect of incorporating a slowly varying frequency-modulated jitter component that is superimposed upon the correlated components already described. The frequency modulation is sinusoidal with a frequency equal to the sequence repetition frequency f_0 Hz, and

$$\Delta T_r = d\left\{\sin\left(2\pi\frac{r\alpha_{fm}f_1}{f_s}\right) + \sin\left(2\pi\frac{r\alpha_{fm}f_2}{f_s}\right)\right\} + J_n\,\mathrm{Rand}(r)$$

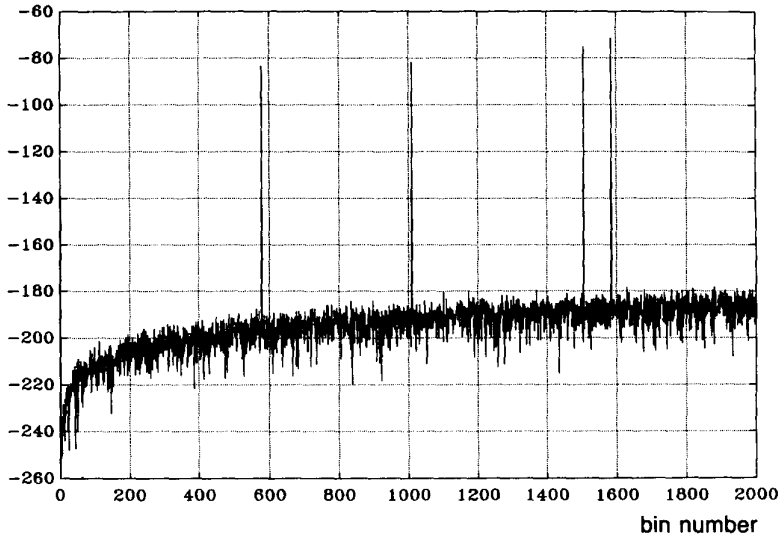


Fig. 4. Output jitter spectrum; no random jitter, impulsive samples.

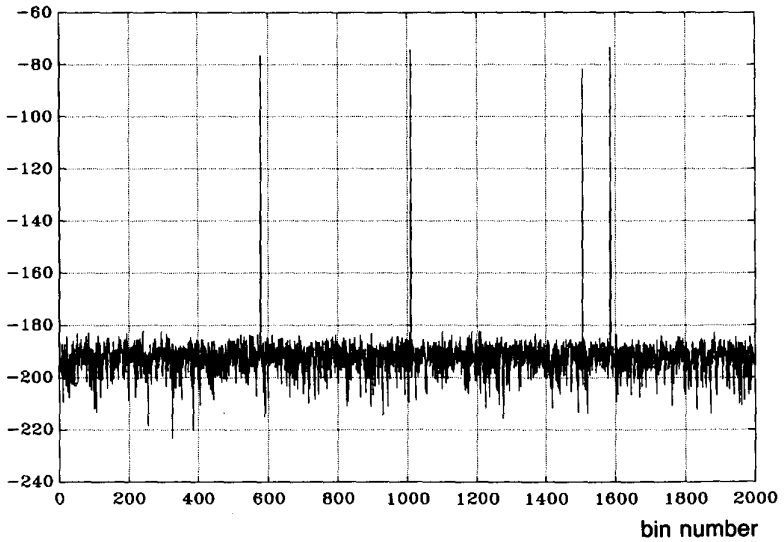


Fig. 5. Output jitter spectrum; no random jitter, 100% samples.

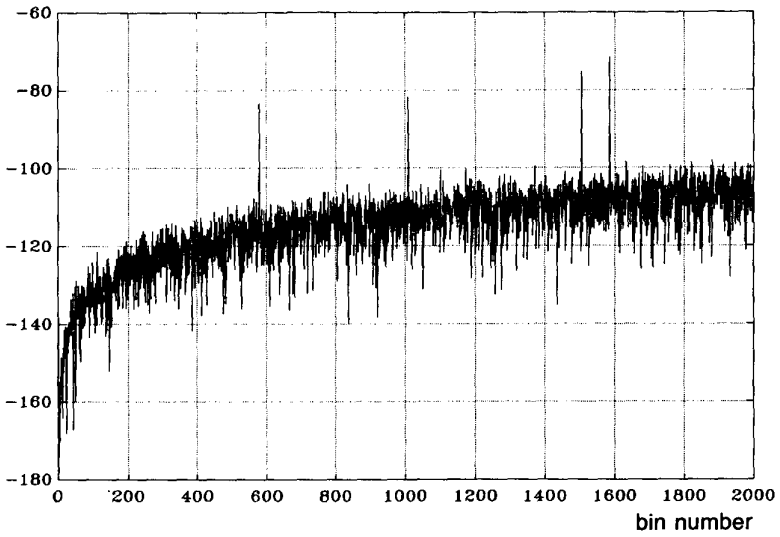


Fig. 6. Output jitter spectrum; including random jitter, impulsive samples.

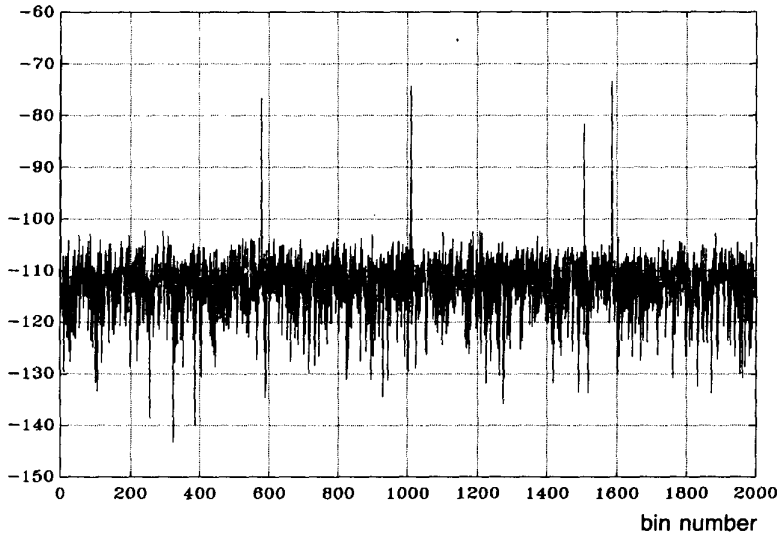


Fig. 7. Output jitter spectrum; including random jitter, 100% samples.

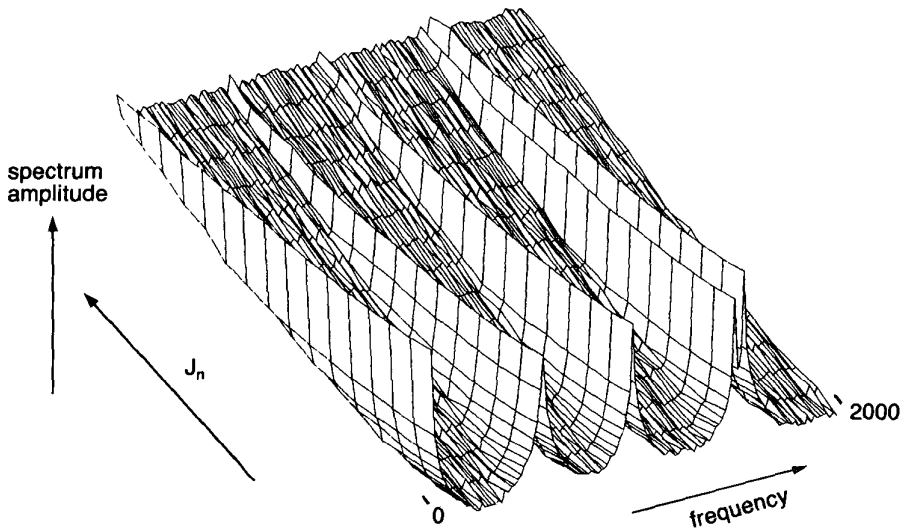


Fig. 8. Jitter spectral family with varying levels of random jitter noise.

where

$$\alpha_{fm} = 1 + J_{fm} \sin\left(2\pi \frac{r}{N}\right)$$

J_{fm} being the modulation depth, $0 \leq J_{fm} \leq 1$.

This additional jitter component is introduced as a frequency modulation of the periodic jitter sequence, where the modulation depth is J_{fm}. The results for $J_{fm} = 1.8 \times 10^{-4}$ and $J_{fm} = 1$ for $J_n = 0$ ns are shown in Fig. 9, whereas the three-dimensional plots in Figs. 10 and 11 illustrate the modification in spectral form for a range of modulation depths over $1.8 \times 10^{-4} \leq J_{fm} \leq 1$ for $J_n = 0$ ns and $J_n = 1$ ns, respectively. Finally Fig. 12 shows a family of spectra to demonstrate that the total level of noise and distortion varies as a function of signal amplitude. Here $J_n = 0.1$ ns, $J_{fm} = 0$, $d = 10$ ns, and the modified signal function is

$$A_r = \sin^2\left(\frac{h\pi}{17}\right)\left\{\sin\left(2\pi \frac{rf_1}{f_s}\right) + \sin\left(2\pi \frac{rf_2}{f_s}\right)\right\}$$

where the integer parameter h is scanned over a range of $1 \leq h \leq 16$, which in turn addresses the sine-squared amplitude weighting.

2 SLEW-RATE-INDUCED DISTORTION

The performance requirements of multibit DAC electronics for digital audio systems are stringent. First the reconstruction levels must be accurately specified. Edge jitter must be minimized. While it is a primary function of clock performance, it can result from internal circuitry exhibiting variability on propagation delay and response time as well as electromagnetic interaction between system modules. Finally the trajectory of the signal between adjacent samples should be determined by a linear network, forming, for example, an exponential curve.

However, because of the rapid response times encountered, even when sample-and-hold circuitry is used as a deglitcher, momentary nonlinearity can result in a small

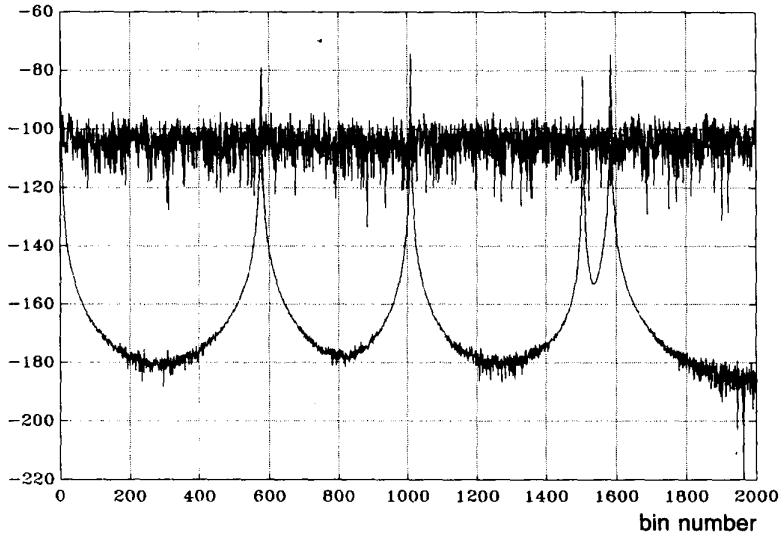


Fig. 9. Jitter spectrum with FM jitter component, used to calibrate Fig. 10.

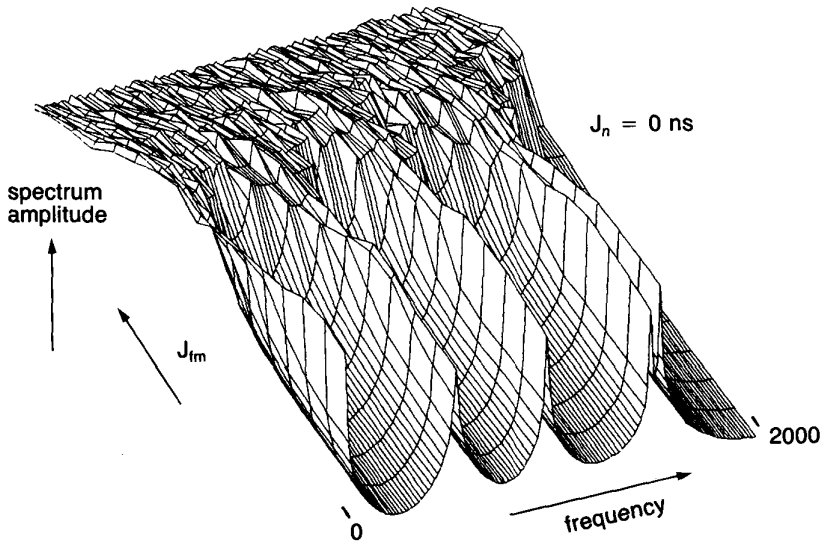


Fig. 10. Family of jitter spectra for varying FM depths; no random jitter.

change in the pulse area, which is related to intersample values. Because of the small time duration of each nonlinear event, ideal rectangular transitions are assumed with the error modeled using an equivalent sample jitter component ΔT_{nr}.

Fig. 13 shows a nonlinear transition between two 100% rectangular pulses of weight A_r and A_{r+1}, constrained by a constant slew rate S V/s, which results in a loss of pulse area ΔA_r, where

$$\Delta A_r = (A_{r+1} - A_r)\,\tau_r .$$

Consider a rectangular pulse where the transition is displaced from its nominal location by ΔT_{nr} such that {pulse area of A_{r+1}} − {pulse area of A_r} = $-\Delta A_r$ that is,

$$A_{r+1}\left(\frac{1}{f_s} - \Delta T_{nr}\right) - A_r\left(\frac{1}{f_s} + \Delta T_{nr}\right) = -\Delta A_r$$

whereby

$$\Delta T_{nr} = +\frac{\tau_r}{2}. \tag{6}$$

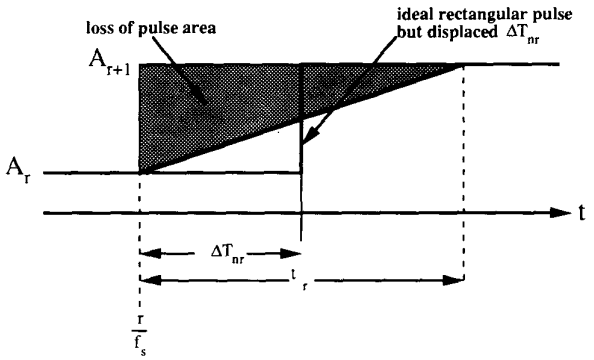


Fig. 13. Two adjacent samples linked by dominant slew-rate distortion.

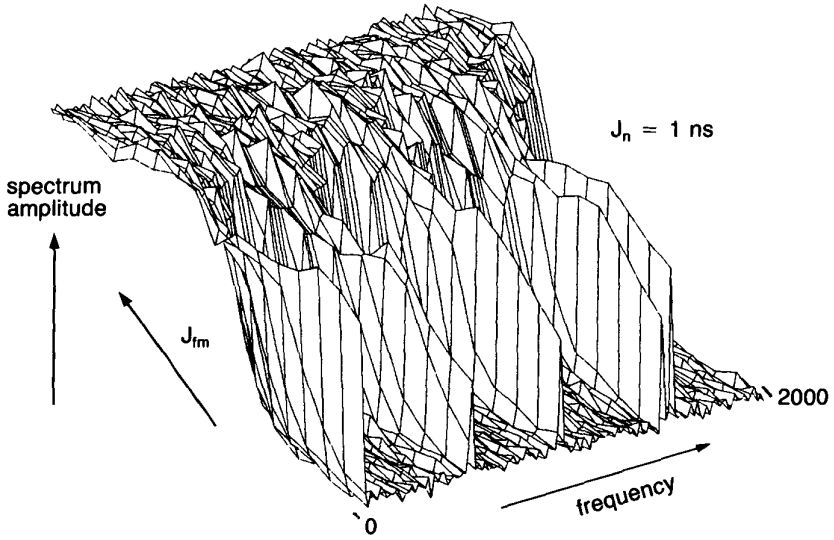


Fig. 11. Family of jitter spectra for varying FM depths; with random jitter.

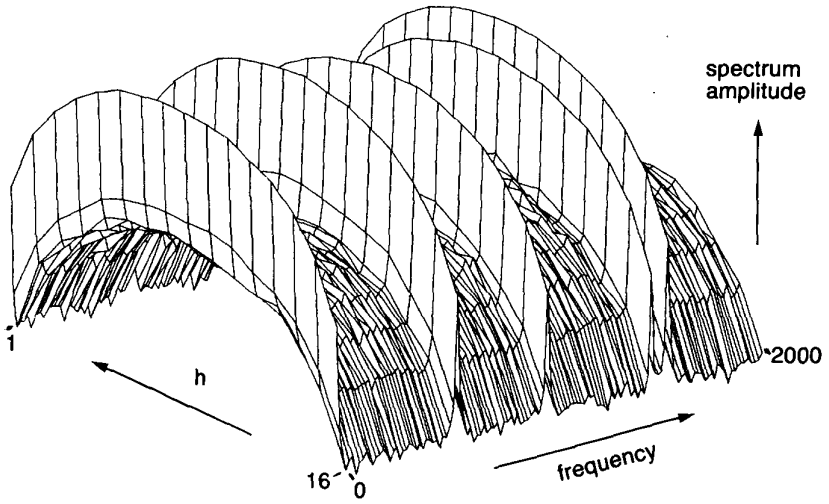


Fig. 12. Family of jitter spectra with varying levels of input signal.

But for a constant slew rate,

$$S = \frac{A_{r+1} - A_r}{\tau_r}$$

and thus

$$\Delta T_{nr} = \frac{A_{r+1} - A_r}{2S}. \quad (7)$$

Although this example is idealized, it demonstrates how an equivalent correlated jitter component ΔT_{nr} can be assigned to the rth sample, and thus the results in Section 1.2 can be used. The analysis ignores spectral changes relating to pulse shape, as these events are of short duration compared with the Nyquist sampling period. The use of an equivalent jitter time defined in association with slew-rate distortion and other related nonlinearities in the current-to-voltage (transresistance) stage of a DAC enables a unification of this class of problem, where Eq. (7) together with Section 1.2 permit specifying the performance.

Equivalent jitter resulting from slew-rate distortion is potentially more serious than random jitter because of the natural correlation between slew limiting and the data samples. Although with appropriate design tighter limits can be achieved, pulse jitter greater than $\approx$ 10 ns has been reported to be of audible significance, and there is anecdotal evidence to support a much tighter specification. However, using the 10-ns criteria and assuming by way of example $\{A_{r+1} - A_r = 500 \text{ mV}\}$, a transresistance amplifier should exhibit a slew rate greater than 25 V/μs.

Even if slew-rate limiting does not occur, an operational amplifier may be close to its open-loop limits during periods of rapid signal transition, and this may contribute momentary "packets" of distortion. There is little doubt that transresistance stages used in DAC systems can contribute distortion that is not fully characterized by observing distortion products when processing band-limited audio signals. Hence this amplifier stage requires band limitation and slew-rate prevention to yield low correlated equivalent jitter, where it is suggested that wide-bandwidth, open-loop transresistance converters are the preferred choice with possible prefiltering to reduce the bandwidth of the DAC output current.

Also, since most DAC transresistance stages operate with 100% pulses, the effect of jitter (random or slew-rate equivalent) is increased. To demonstrate this conjecture, the equations presented in Section 1.2 are used together with Eq. (7) as well as the following data to generate the error spectrum shown in Fig. 14:

Number of samples $N = 4096$
Sampling rate (2 times oversampling) $f_s = 88.2$ kHz
Transresistance amplifier slew rate $S = 50$ V/μs
Sample sequence generator
$A_r = \{\sin(2\pi r f_1/f_s) + \sin(2\pi r f_2/f_s)\}$
Random jitter $J_n = 1$ ns
Correlated jitter $d = 1 \times 10^{-18}$ s

where $f_0 = f_s/N$ Hz, $f_1 = 992 f_0$ Hz, and $f_2 = 512 f_0$ Hz, that is, $f_1 = 21.36$ kHz and $f_2 = 11.025$ kHz. No other correlated jitter source is included.

Finally using similar data, a three-dimensional plot is shown in Fig. 15, where f_1 is scanned linearly in 32 steps from 689 Hz to 22.05 kHz, $f_2 = 11.025$ kHz, and $d = J_n = 1 \times 10^{-18}$ s. The surface shows tracks of intermodulation distortion products, where the calibration of the vertical scale can be estimated from Fig. 14 that correspond to trace 31 where $f_2 = 21.359$ kHz.

3 DAC TOPOLOGY WITH LOW JITTER AND SLEW-RATE SENSITIVITY

3.1 System Topology and Function

To reduce DAC sensitivity to slew rate and jitter, the rapid signal transition at each sample boundary must be

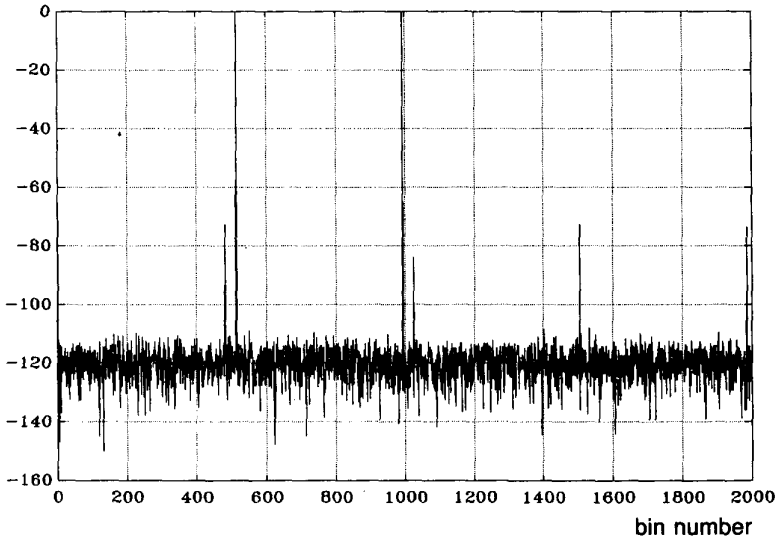


Fig. 14. Distortion spectrum resulting from slew rate used to calibrate Fig. 15.

minimized and a degree of intersample isolation introduced to control the distortion described in Section 1.2. Effectively the operating bandwidth of the DAC should be lowered so that a more "analoglike" conversion is achieved at the gateway of data conversion. The wide bandwidth generally encountered is an artifact of topology that is typified by the current-switching DAC in association with a wide-band transresistance converter.

The proposed topology consists of two time-interleaved DACs operating with mild oversampling, but where the DACs are configured as multiplying converters (MDACs). The reference inputs of the two DACs are raised cosine waveforms with low inherent jitter. The basic system is shown in Fig. 16 and uses a 4× oversampling filter to enable initial interpolation of input data. Fig. 17 shows a series of illustrative waveforms to demonstrate operation.

The output of the 4× oversampling filter is multiplexed alternately between two MDACs ($MDAC_1$ and $MDAC_2$) using sampled latches, where conversion occurs on the alternating data sequences D_1 and D_2. The DACs therefore run at $2f_{ns}$ Hz, and output pulses overlap by $1/f_{ns}$, f_{ns} being the Nyquist sampling frequency.

Although the data applied to each DAC are held constant for two consecutive samples, examination of the respective raised cosine reference waveforms R_1 and R_2 shows each reference voltage to be zero on a conversion edge. Consequently assuming that there is no pulse feedthrough in the MDAC, any jitter on the data edge is attenuated. In a practical system, circuitry would arrange for data to be transformed only when the reference is zero. However, because of the near zero slope of the reference voltage waveform in close proximity to its zero value, the timing of data transition is noncritical. Once the data are latched into an MDAC, the reference voltage (controlling the gain of the MDAC) rises from zero, thus causing the output current I_1 or I_2 to change in direct proportion to, but weighted by, the present data value. When the cosine waveform reaches its peak value, the reference voltage applied to the other MDAC is now at zero, at which instant its data are updated in a similar manner. The process then proceeds at a uniform rate, with data being updated on the corresponding zero of each raised cosine reference signal.

The net result of this process can be summarized as follows.

1) Data conversion only occurs when the reference to an MDAC is zero. Thus the contribution of jitter is minimized. Effectively, the jitter dependence is translated from the digital data to the two raised cosine reference signals.

2) Because the current output of each DAC tracks a raised cosine, the rate of change is reduced compared with rectangular switching. Thus slew-rate-induced distortion and other minor nonlinearities within the transresistance converter are virtually eliminated.

3) Reduction of high-frequency spuriae and the use of 4× oversampling relax the design of the analog reconstruction filter, and the output signal from the converter is more "analoglike."

4) Because a raised cosine consists only of a dc term and a single spectral line, noise filtering to reduce jitter is simplified.

5) Any imbalance in gain between $MDAC_1$ and $MDAC_2$ is of little consequence and only causes a mild increase in the spectral replication at $2f_{ns}$ Hz, which because of 4× oversampling, is located well above the audio band.

6) Reduced bandwidths of signals within the converter mean that circuit layout and parasitic and mutual coupling of circuit elements are less problematic.

Cautionary Note. To achieve the performance specified in the preceding, the cosine weights for all samples should be identical. Consequently the frequency response of the MDAC from reference input to current output must not be code dependent. This is not a fundamental problem, but it does require appropriate attention in the design of the MDAC.

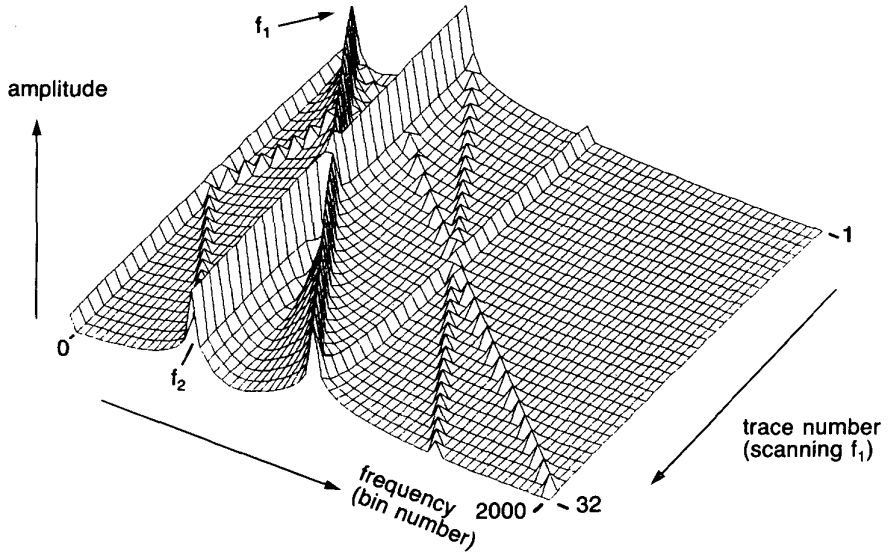


Fig. 15. Scanned distortion spectrum resulting from slew rate.

3.2 Estimate of Jitter Suppression

Fig. 18(a) shows sample reconstruction for a single DAC, where the data conversion timing is optimum. However, if a timing offset T_0 and a superimposed jitter component ΔT_r are introduced at the gateway of data conversion, then the waveform shown in Fig. 18(b) result. In this example only even samples in the oversampled data sequence are shown.

The result of this timing error is a loss in pulse area, which can be estimated as follows. The pulse-area error is given by

$$\Delta A_r = \int_{t=0}^{T_0+\Delta T_r} (A_{r+2} - A_r)\,[1 - \cos(4\pi f_{ns} t)]\,dt\,.$$

After integration,

$$\Delta A_r = (A_{r+2} - A_r)\left\{T_0 + \Delta T_r - \frac{\sin[4\pi f_{ns}(T_0 + \Delta T_r)]}{4\pi f_{ns}}\right\} \quad (8a)$$

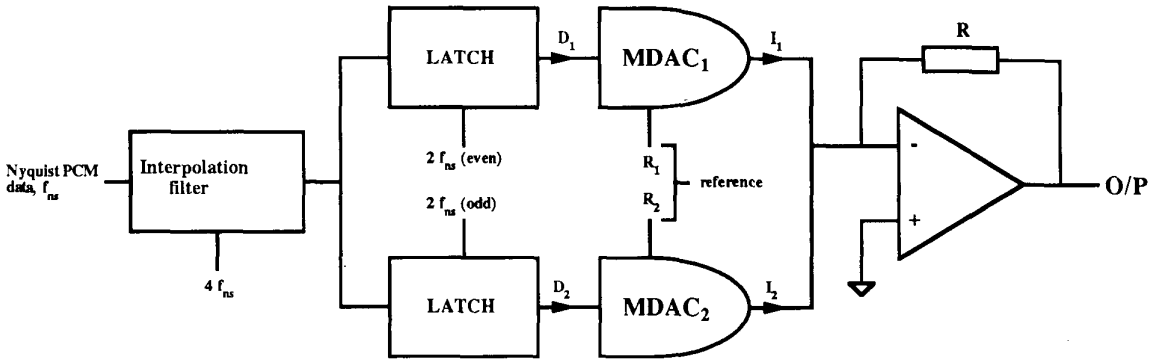


Fig. 16. Basic two-interleaved MDAC topology.

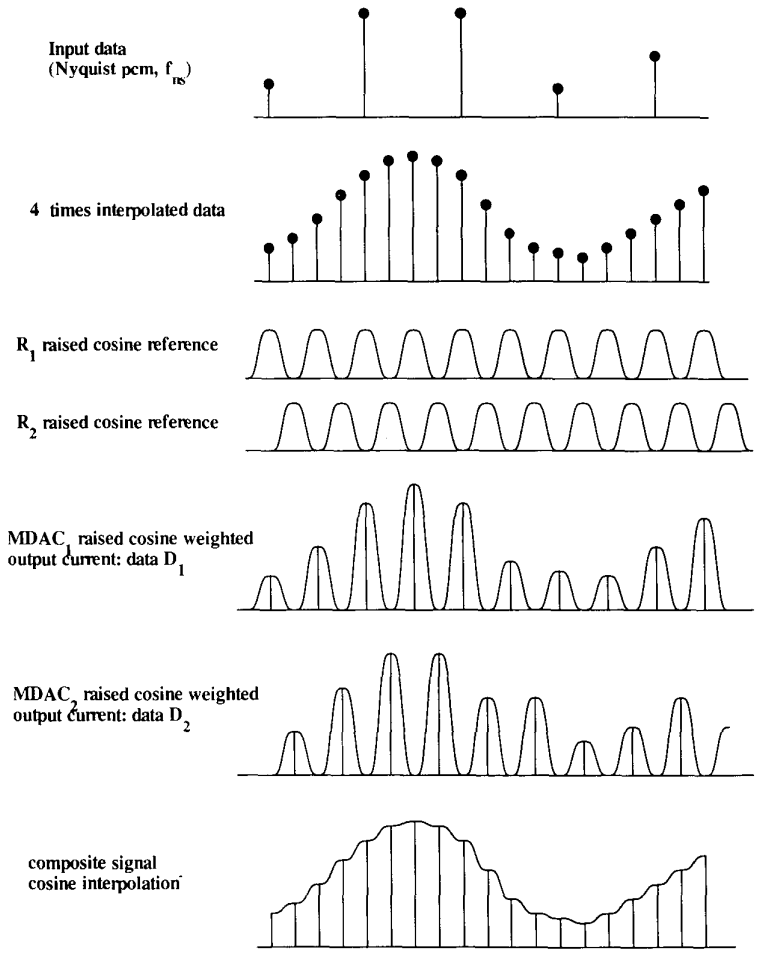


Fig. 17. Illustrative waveforms showing raised cosine interpolation in time-interleaved DAC topology.

which, for $4\pi f_{ns}(T_0 + \Delta T_r) << 1$, simplifies to

$$\Delta A_r = (A_{r+2} - A_r)\left[\frac{8}{3}(\pi f_{ns})^2 (T_0 + \Delta T_r)^3\right]. \quad (8b)$$

Defining an equivalent time offset and jitter error T_{er} for the rectangular pulse shown in Fig. 18,

$$(A_{r+2} - A_r)\, T_{er} = \Delta A_r$$

yielding

$$T_{er} = \frac{8}{3}(\pi f_{ns})^2 (T_0 + \Delta T_r)^3 \quad (9)$$

that is,

$$\frac{T_{er}}{T_0 + \Delta T_r} = \frac{8}{3}[\pi f_{ns}(T_0 + \Delta T_r)]^2. \quad (10)$$

Eq. (9) estimates the equivalent timing error using the raised cosine sample format compared with the case using rectangular samples, whereas Eq. (10) reconfigures this result to show the corresponding reduction in dependence of the timing errors. For example, let

$$(T_0 + \Delta T_r) = 200 \text{ ns}, \qquad f_{ns} = 44.1 \text{ kHz}$$

whereby the effective reduction in timing error $\approx 2 \times 10^{-3}$. This demonstrates that a DAC topology with a response or settling time of only 200 ns is adequate.

The analysis demonstrates a remarkable reduction in sensitivity to jitter within the digital data stream. This improvement is partially dependent on low jitter in the raised cosine waveform and a reference voltage that accurately attains a zero value at the minima of the raised cosine function. However, the form of the raised cosine waveform with only two spectral lines, dc and $2f_{ns}$ Hz, means that band-pass filtering can achieve a low inherent jitter performance that is considerably more effective than smoothing a high-frequency rectangular clock, as the equivalent noise bandwidth can be made lower because of the lower number of contributing harmonics. Also the band-limited raised cosine reference waveforms can be applied directly to the MDACs without additional noise-inducing counters and logic circuits. These factors, together with an almost total independence of digital data jitter, are the principal attributes of this new DAC topology.

3.3 Spectral Response of Raised Cosine Modulated Samples and Requirements on Analog Reconstruction Filters

The requirements for analog signal recovery subsequent to digital-to-analog conversion can be determined by analyzing the combination of 4 × oversampling and the overlapping raised cosine weighting that is associated with each sample. The impulse response $h_c(t)$ of the time-limited raised cosine generator can be expressed as

$$h_c(t) = 0.5\{1 + \cos(4\pi f r_{ns} t\} \operatorname{rect}_{1/2f_{ns}}(t). \quad (11)$$

The Fourier transform $F_c(f)$ then follows as

$$F_c(f) = 0.5 \int_{-1/4f_{ns}}^{1/4f_{ns}} [1 + \cos(4\pi f_{ns} t)]\, e^{-j2\pi ft}\, dt$$

that is,

$$F_c(f) = \frac{1}{8f_{ns}}\left\{\operatorname{sinc}\left[\pi\left(\frac{f}{2f_{ns}} - 1\right)\right] + 2\operatorname{sinc}\left(\frac{\pi f}{2f_{ns}}\right) + \operatorname{sinc}\left[\pi\left(\frac{f}{2f_{ns}} + 1\right)\right]\right\} \quad (12)$$

where $\operatorname{sinc}(x) = \sin(x)/x$.

The time and the corresponding Fourier transform of the time-limited raised cosine pulse are shown in Fig. 19. The Fourier transform shows that there is a significant response to $3f_{ns}$, but for $3f_{ns}$ and above there is attenuation. However, because a 4 × oversampling filter is prescribed, the first spectral replication in the final reconstructed signal is centered on $4f_{ns}$ Hz and extends $\pm f_{ns}/2$ Hz. Consequently the inherent attenuation offered by the raised cosine waveform at $4f_{ns}$ (noting that at exactly $4f_{ns}$ Hz the Fourier transform is zero) significantly suppresses the first spectral replication and thus relaxes the design of the analog recovery filter. Indeed, the spectrum in Fig. 19 suggests that the analog filter can be designed to have a band-reject response centered on $4f_{ns}$ Hz or, alternatively, a twin resonant circuit with rejection bands centered on $3.5f_{ns}$ and $4.5f_{ns}$, respectively, followed by a mild low-pass filter response to reduce out-of-band noise and spuriae.

Finally, because of the form of the raised cosine transform $F_c(f)$ described by Eq. (12) and to enable a flat audio passband, mild linear-phase equalization should be included in the oversampling filter to match the inverse response over the frequency band of $0-0.5f_{ns}$ Hz. It may also be expedient to include a minor correction for the analog reconstruction filter transfer function, although in practice this will be small.

3.4 MDAC Nonlinearity in Reference Signal Path

Nonlinearity within an MDAC can be modeled by assuming a perfect DAC combined with a dynamic nonlinear network in the reference input, as shown in Fig. 20. Because the modified MDAC has now been desensitized to distortion components generated at the data transition, the residual errors are solely dependent on the accuracy of pulse amplitude reconstruction and the nonlinearity within the circuitry associated with the reference (gain-defining) input. If the MDAC is assumed ideal, then nonideality can be modeled by a nonlinear network in the reference channel where pulse amplitude errors can be accounted for by allowing the data input to modulate this network. We thus identify two possibly interrelated error mechanisms, which can cause distortion in the reconstructed output. However, this model

is also useful as it allows us to define succinctly the conditions that prevent distortion from entering the critical audio band. If the only nonlinearity is in the reference input channel, and the data input in no way alters this nonlinearity, then the result is only an addition of harmonic distortion to the raised cosine waveform. Thus at the output of the MDAC we would expect to observe a modest level of spectral replication of the input about these harmonics, which can readily be removed by the output analog reconstruction filter and thus is of little consequence. It is only when the data sequence modifies the nonlinearity that level reconstruction errors occur, which will result in output distortion. However, this is fundamental to all multilevel DACs, and this system is no exception to this class of distortion.

3.5 Estimate of Maximum Rate of Change of Signal at Output of Transresistance Converter

Consider two adjacent samples A_r and A_{r+1} separated by the oversampled time interval $1/4f_{ns}$. The DAC attempts to edit these samples by a half-cosine wave interpolation, as shown in Fig. 17. The amplitude of this half-cosine segment is therefore $\{0.5(A_{r+1} - A_r)\}$ such that over the sample interval $1/4f_{ns}$ the reconstructed signal is

$$V_0 = A_r + 0.5(A_{r+1} - A_r)\,[1 - \cos(4\pi f_{ns}t)]\,.$$

The maximum slope of the output signal is therefore

$$\left.\frac{dV_0}{dt}\right|_{max} = 2\pi f_{ns}(A_{r+1} - A_r)\,. \qquad (13)$$

Assume a maximum amplitude-coded sine wave of frequency $f_{ns}/2$ Hz that has the analog form

$$V_0 = A_m \sin(\pi f_{ns}t)$$

and consider adjacent samples of V_0 such that the amplitude separation of A_{r+1} and A_r is maximized when using $4\times$ oversampling (that is, $4f_{ns}$ Hz). Thus

$$A_r = -A_m \sin\left(\frac{\pi}{4}\right)$$

$$A_{r+1} = A_m \sin\left(\frac{\pi}{4}\right).$$

Hence from Eq. (13) the minimum slew rate S_{min} of the transresistance converter is

$$S_{min} = 4\pi f_{ns}A_m \sin\left(\frac{\pi}{4}\right). \qquad (14)$$

By way of an example, let $A_m = 2\sqrt{2}$ V and $f_{ns} = 44.1$ kHz, whereby $S_{min} = 1.12$ V/μs.

This basic analysis shows that for a standard 2 V_{rms} output signal the maximum rate of change of the output signal for the transresistance converter is constrained. Consequently a performance commensurate with low in-band distortion is simpler to achieve and should be compared with the example given in Section 2.

4 EXPERIMENTAL RAISED COSINE DAC

An experimental raised cosine DAC is shown in Fig. 21, which uses commercially available parts. The design employs a Micro Power Systems MP7616 16-bit CMOS four-quadrant multiplying DAC having the basic architecture shown in Fig. 22. (Although this is a 16-bit device, it is only of marginal performance for high-quality applications offering a current settling time of 2

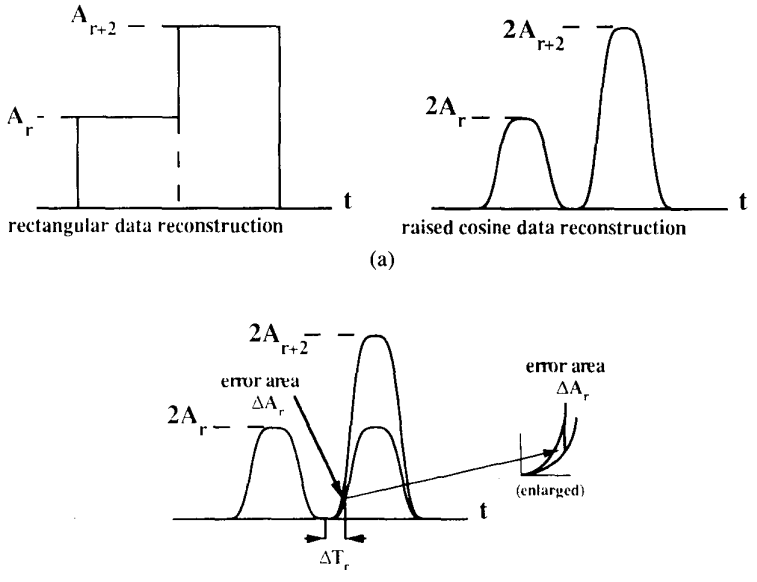


Fig. 18. (a) Rectangular and raised cosine sample reconstruction (samples weighted to have same area). (b) Effect of timing offset and jitter on reconstructed raised cosine samples.

μs to 0.01% of FSR. However it was the only device available at the time of experimentation.) To lower tolerance on resistor matching, the MP7616 features 15 equivalued current sources with inputs decoded from the four MSBs. The remaining 12 bit are then converted using a binary weighted tree. A key feature of this DAC in the present application is the bipolar reference input, which is driven by one of the two raised cosine waveforms.

SPDIF serial digital data are decoded by a Yamaha receiver (YM3623), and the sampling rate is increased eight times using a Burr Brown DF1700. The oversampled data are next converted to a parallel format and, via two alternately clocked 16-bit latches, input to the two 16-bit MDACs. Complementary parallel data allow the use of in-phase raised-cosine waveforms as defined in Fig. 21, where complementary dc offsets result in a zero-mean output current when the two MDAC outputs are summed, thus simplifying the design of the transresistance converter as no dc correction is required.

Raised cosine waveform phase alignment is maintained by including the cosine waveform generator within a PLL, as shown in Fig. 23. The voltage-controlled oscillator (VCO) of the PLL drives two analog gates to produce a symmetrical square wave, which is subsequently band limited by a second-order bandpass filter. Two in-phase signals are formed on secondary windings coupled to the tuned circuit, which are dc offset to form the two raised cosine waveforms. The PLL and the bandpass filter are also instrumental in reducing jitter.

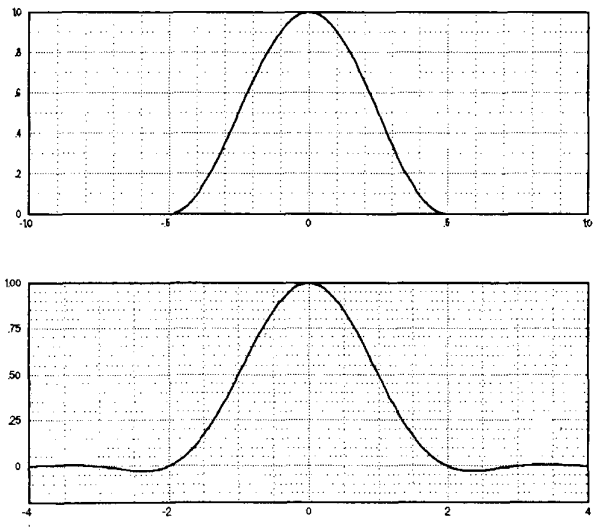


Fig. 19. Raised cosine waveform in both time and frequency domains.

This system validated the operation of the raised cosine DAC, and measurement confirmed cosinusoidal interpolation between adjacent samples, thus relaxing the slew-rate requirements for the current-to-voltage converter. In this sense the DAC can be seen to filter the MDAC output current waveform, but without compromising noise performance, which occurs when a filter is placed between the DAC and the transresistance converter. Also such filters do not reduce the jitter present in the source data. However, in this experimental model the resolution of the DAC and its settling time were limiting factors.

5 CONCLUSION

This paper has described a technique of time interleaving two MDAC converters using complementary raised cosine generators applied to the reference input. The effect of this process is to replace the normal rectangular pulses in a 4 × oversampled converter with raised cosine weighted samples that are time limited to span two consecutive samples in the oversampled data stream. The effect of this process is a reduced sensitivity to data timing jitter and transition distortion as well as a reduction in the slew-rate requirement of the transresistance amplifier. It was also shown that the requirements of the analog recovery filter at the DAC are relaxed.

The elimination of edge-transition distortion and the transformation of jitter dependence from the DAC data clocks to the raised cosine generators are seen as pivotal in the design, as is the reduced bandwidth of the signal presented to the transresistance stage. A signal (the reference generator output) that consists of only two spectral lines (dc and $2f_{ns}$ Hz) is simpler to filter to suppress noise and spuriae, the source of jitter, than in the case of a broad-band square wave. Even if a square wave timing signal is comb filtered to contain only the fundamental and its harmonics, there is still a finite noise bandwidth associated with each harmonic which is greater than that of the raised cosine.

To support the proposal for a DAC with low jitter sensitivity, an analysis was presented that describes the mechanism by which jitter can introduce distortion into the audio band. It was shown that jitter that mapped to pulse-area distortion as well as producing timing errors resulted in greater low-frequency distortion. For jitter that only mistimed a sample event, the distortion spec-

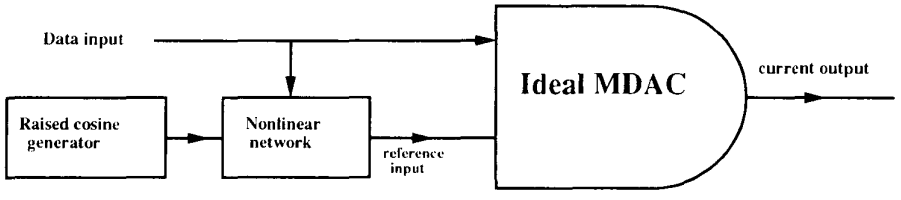


Fig. 20. Basic model of MDAC nonlinearity.

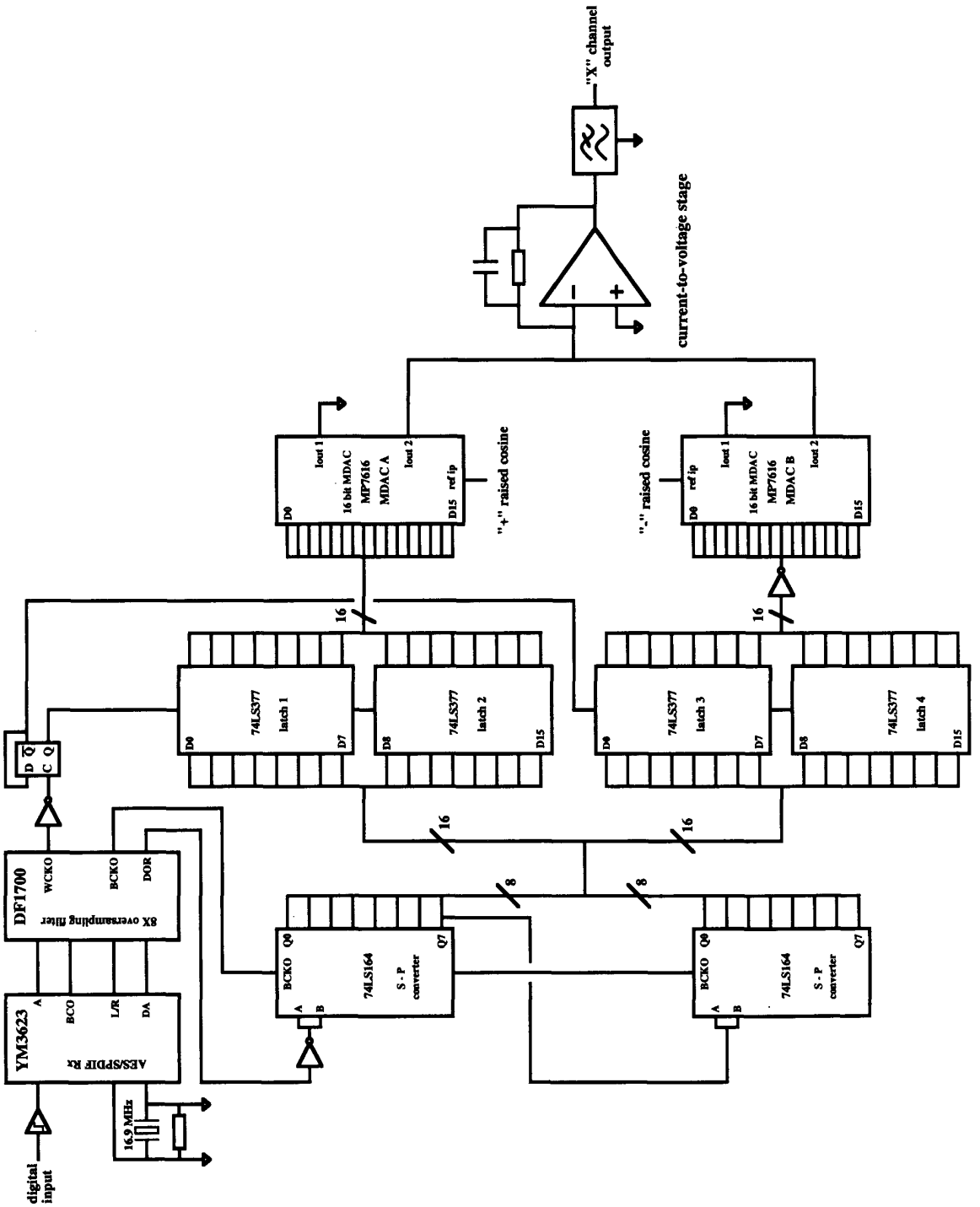


Fig. 21. Raised cosine experimental DAC.

trum was proportional to the frequency, whereas the inclusion of area modulation extended this response to dc. Analysis and computer simulation revealed that the distortion was more problematic where a correlation exists between timing error and program material. Indeed, in many practical digital systems the use of PLLs with inadequate timing recovery can yield correlation, even though this process can be highly nonlinear. For example, where a PLL responds to a change in the data sequence, even though this is in coded form, correlation can exist and a highly complicated jitter spectrum result, which even though of very low level, does not necessarily fall under the auditory mask causing perturbations of the detection thresholds within a number of critical bands [13]. It can be argued that an inherent design limitation of the SPDIF serial code is the nonscrambling of serial data by coding to break the correlation between audio data and bit-pattern-induced jitter. If scrambling were used, any resulting jitter that is related to the serial bit pattern would be decorrelated, and therefore would produce only a noiselike residue of benign character. This is a conjecture developed in a supporting paper [8].

Example results for both correlated and random jitter were included as well as the effect of adding low-frequency modulation. Fig. 8 demonstrated that the correlated and random jitter components did not intermodulate and can be considered essentially additive for a constant-amplitude input sequence. However, the inclusion of low-frequency (sinusoidal) modulation of the jitter sequence mimicking a low-frequency error in a PLL, for example, produced significant levels of intermodulation. Figs. 10 and 11 demonstrated this interaction both with and without a random jitter sequence. However, because all the jitter-induced distortion spectra are dependent on the amplitude of the signal sequence, the spectral levels should all be read with respect to the signal level, whereby if the signal is reduced, the distortion changes in direct proportion. To demonstrate this inherent characteristic, Fig. 12 showed the distortion spectrum as a function of signal level, where the form of the distortion remains the same, but in direct proportion to the signals. This is true whether or not there is correlation between signal and jitter. To illustrate this feature, Fig. 12 included both random and correlated components.

A powerful extension of the jitter analysis was to consider edge transition distortion resulting from slew-rate limiting in the transresistance converter, or indeed inherent within the DAC, and to translate this to an equivalent edge jitter when using 100% rectangular samples. The transformation revealed that edge jitter, slew rate, and related transition distortion fall into a common regime and that a similar analysis procedure can be used. However, with DAC transition distortion the correlation with the signal will almost certainly be higher, implying a greater subjective significance. Although slew rate is a dominant distortion, it should be recalled that the operational amplifier, at the edge transition, is operating near open loop, so although the transition may appear well

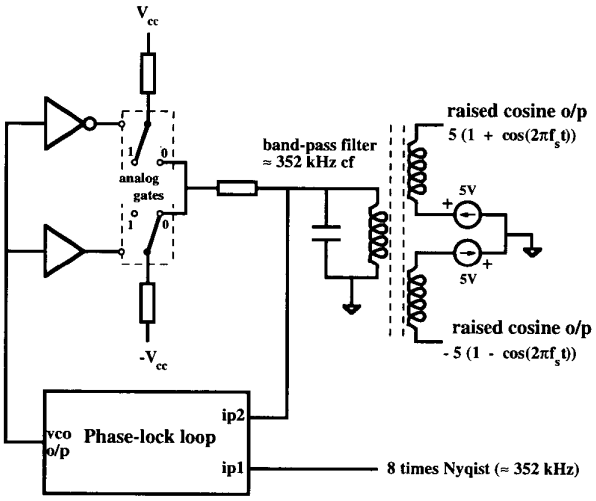


Fig. 23. Raised cosine generation using bandpass filter and phase-locked loop.

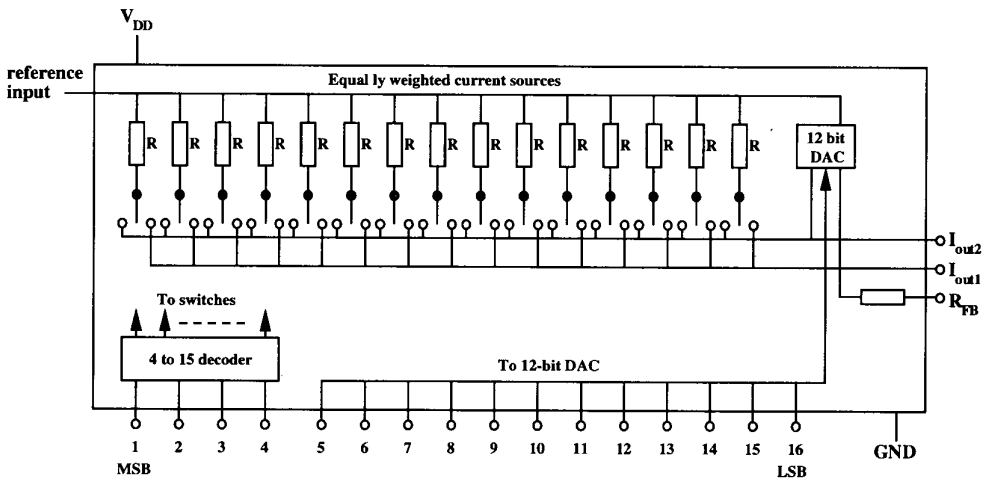


Fig. 22. MP7616 four-quadrant multiplying DAC.

behaved, there may nevertheless be some distortion induction. In this paper the results presented are relatively extreme and do not take account of any softening of the DAC transitions, either internally or by using a transresistance stage with a capacitive feedback element or current input filter. However, the simulations do indicate complicated distortion spectra that are increased in magnitude due to the fundamental correlation between signal and jitter equivalence, as illustrated in Fig. 15, where it can also be observed that the distortion terms are not well matched to psychoacoustic masking thresholds, thus possibly gaining in subjective significance.

It should be noted that when correlation between signal and jitter was considered and the distortion calculated by Eq. (5), the differential of the signal sequence was multiplied by the jitter ΔT_r, which was made directly proportional to the signal. However, for slew-rate-induced distortion the jitter equivalence described by Eq. (7) is proportional also to the differential of the input sequence. Therefore the resulting distortion is proportional to the square of the differential of the input, a subtle but possibly significant difference that implies a greater intermodulation distortion dependence on high-frequency signal components.

The time-interleaved dual DAC topology is potentially less sensitive to many of these problems. Provided the MDACs can achieve accurate level reconstruction and their output/reference input frequency response is not code dependent, then even if jitter exists and data transition distortion would normally occur with rectangular samples, these errors are of little consequence, as was demonstrated by the equivalent timing displacement estimated in Section 3.2. Thus the relatively low bandwidth excitation of the transresistance converter together with the relaxation of the analog filter topology in association with a standard 4× or 8× oversampling filter in the digital domain should result in near theoretic performance.

Section 4 presented an experimental system to confirm operation, although performance restrictions of the available four-quadrant MDAC should be noted. It is possible that an MDAC may exhibit code-dependent distortion, although the relaxation of slew-rate-dependent distortion is potentially of greater benefit. Also, although analog filters can be used to band-limit the output waveform of a DAC prior to the transresistance stage, hence reduce slew-rate-induced distortion, this does not address jitter or distortion present within the DAC settling period after conversion. Such circuits usually imply a high-frequency noise penalty due to a shunt impedance at the transresistance input.

There are DACs having 20-bit amplitude resolution that were designed for high-quality digital audio and military systems. If these can be modified to include access to the reference input and thus allow operation as a precision MDAC, there now exists a means to virtually eliminate the vestiges of a number of inherent imperfections, which, although of low level, can still pervade DAC systems and represent an ultimate performance bound.

6 ACKNOWLEDGMENT

The author would like to thank Andrew McCarthy and Phillipe Dolman for their work on constructing a prototype DAC as part of their B.Eng. program in the Department of Electronic Systems Engineering at the University of Essex.

7 REFERENCES

[1] M. O. J. Hawksford, "An Introduction to Digital Audio," in *Proc. 10th Int. AES Conf.* (London, 1991 Sept.), pp. T3–T42.

[2] J. R. Stuart, "Estimating the Significance of Errors in Audio Systems," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 1011 (1991 Dec.), preprint 3208.

[3] S. Harris, "The Effects of Sampling Clock Jitter on Nyquist Sampling Analog-to-Digital Converters, and on Oversampling Delta–Sigma ADCs", *J. Audio Eng. Soc.*, vol. 38, pp. 537–542 (1990 July/Aug.).

[4] P. van Willenswaard, "Industry Update," *Stereophile*, vol. 13, pp. 78–83 (1990 Nov.).

[5] J. A. Atkinson, "Jitter, Bits and Sound Quality," *Stereophile*, vol. 13, pp. 179–181 (1990 Dec.).

[6] R. Harley, "Industry Update," *Stereophile*, vol. 14, pp. 38–45 (1991 Sept.); vol. 16, p. 65 (1993 Feb.); vol 16, pp. 47–91 (1993 Sept.).

[7] E. Meitner and R. Gendron, "Time Distortions within Digital Audio Equipment Due to Integrated Circuit Logic Induced Modulation Products," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 992.

[8] C. Dunn and M. O. J. Hawksford, "Is the AES/EBU/SPDIF Digital Audio Interface Flawed?," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1040 (1992 Dec.)., preprint 3360.

[9] J. Dunn, "Jitter: Specification and Assessment in Digital Audio Equipment," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1040 (1992 Dec.), preprint 3361.

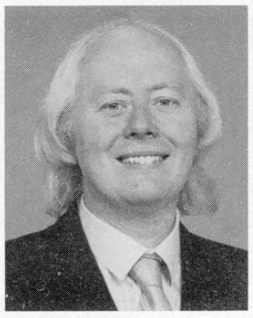
[10] J. Dunn, "Considerations for Interfacing Digital Audio Equipment to the Standard AES-3, AES-5, AES-11," in *Proc. 10th Int. AES Conf.* (London, 1991 Sept.), pp. 115–126.

[11] R. D. Fourre, "Jitter, Jitter, Jitter . . . ," Application Note AP-03, Ultra Analog Inc., Fremont, CA (1992 Sept.).

[12] M. O. J. Hawksford, Letter in response to R. Adams, "Comments on 'Chaos, Oversampling, and Noise-Shaping in Digital-to-Analog Conversion,'" *J. Audio Eng. Soc. (Letters to the Editor)*, vol. 38, pp. 767–768 (1990 Oct.).

[13] J. R. Stuart, "Predicting the Audibility, Delectability, and Loudness of Errors in Audio Systems," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, pp. 1010–1011 (1991 Dec.), preprint 3209.

THE AUTHOR



Malcolm Omar Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at the University of Essex, where his interests encompass audio engineering, electronic system design, and signal processing. Professor Hawksford studied electrical engineering at the University of Aston in Birmingham where he gained a First Class Honours B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship and investigated the application of delta modulation for color television and the development of a time-compression/time-multiplex system for combining luminance and chrominance signals. During his time at Essex University, he has undertaken research on amplifier studies, digital signal processing, and loudspeaker systems. Since 1982 research into digital crossover systems and loudspeaker equalization has been pursued which has led to an advanced digital and active loudspeaker system being produced by the University Company, Wivenhoe Enterprises, under the name Essex Audio. Research has also encompassed oversampling and noise shaping as a means of analog-to-digital and digital-to-analog conversion that includes digital linearization of PWM encoders.

Professor Hawksford has published in the *Journal of the Audio Engineering Society* on topics that include error correction in amplifiers, oversampling techniques, and MLS techniques. His supplementary activities include writing for *Hi-Fi News and Record Review* and designing high-end analog and digital audio equipment. He is a chartered engineer and is a Fellow of the AES, the Institution of Electrical Engineers, and the Institute of Acoustics. He is also technical adviser to *HFN and Record Review* and a technical consultant to *LFD audio, UK*.

Transparent Differential Coding for High-Resolution Digital Audio*

M. J. HAWKSFORD, *AES Fellow*

Centre for Audio Research and Engineering, University of Essex, UK CO4 3SQ

A coding method using cascaded stages of exact differentiation with equalization is presented as an alternative to sigma–delta modulation (SDM) Unlike SDM, the model is inherently linear and can operate losslessly together with an exceptionally wide audio bandwidth. Bit rates are competitive and signal processing can be performed between successive stages of coding. Applications include future ultrahigh-capacity DVD and bridge the debate between DSD and LPCM.

0 INTRODUCTION

A method of using cascaded stages of differential coding together with noise shaping and equalization is described as an alternative format to linear pulse code modulation (LPCM) and to sigma–delta modulation (SDM). LPCM is the basis code that underpins most digital audio systems and is now incorporated into the new established DVD-audio standard as proposed by the DVD consortium WG-4. The advantages of LPCM are well established and can be summarized by observing that the only fundamental distortions are brickwall band limitation resulting from uniform sampling together with additive noise introduced because of quantization and dither. In a system that is correctly implemented and functioning, neither of these processes results in correlated distortion.

It has also become topical to consider SDM as a means of transporting or storing digital audio, partly because of the widespread use of 1-bit converters for both analog-to-digital conversion (ADC) and digital-to-analog conversion (DAC) [1]. The argument in favor says that with simple linkage of ADC and DAC, the absence of decimation and oversampling filters eliminates processing-related errors, offers a wider bandwidth, and therefore allows greater system transparency. The counter argument recognizes that this is only valid for 1-bit converters and is therefore limited in appeal and application. ADC and DAC technology now embraces oversampling and noise shaping in conjunction with low-resolution uniform quantizers with randomization methods to decorrelate residual hardware-induced errors [2], offering performance advantages over the 1-bit converter.

Multibit converters are theoretically linear devices as long as certain conditions are met. Provided a nonsaturating, uniform quantizer is incorporated and correctly formed dither precedes the quantizer, linearity is theoretically achievable, commensurate with proper supporting processing architectures. In contrast, there is no equivalent case for SDM where, because of the saturating nature of the 1-bit quantizer, there is no known theorem that guarantees linear operation. However, in presenting this argument it is recognized that with dither and advanced loop design, low distortion is possible [3]. Some observations about the range of linear operation of SDM are made in Section 1.

The quest for more signal bandwidth, combined with a welcome reduction in processing requirements, is at the heart of the DVD-audio specification, where against all technical and political odds, a maximum sampling rate of 192 kHz has been adopted. Even this high sampling rate is overshadowed by SDM, whose bandwidth extends to one-half the serial bit rate, all be it with the presence of gross quantization noise.

Once the sampling rate of a system is sufficiently high, the option of using noise shaping to reduce the sample resolution is attractive [4]–[7], where, provided the uniform requantizer is not overloaded, it is possible to retain linear operation. The work presented here takes a hybrid approach. It encompasses the advantages of using high sampling rates with noise shaping but applies differential coding to achieve greater efficiency. It will be shown that a performance exceeding SDM can be obtained at commensurate bit rates, but with the linearity

* Presented at the 107th Convention of the Audio Engineering Society, New York, 1999 September 24–27; revised 2001 April 27

advantage of LPCM and with better overload performance.

1 COMPARISONS OF DM AND SDM AND OBSERVATIONS ON LINEARITY OF 1-BIT SYSTEMS

The concept of the 1-bit coder was termed delta modulation (DM) [8] and precedes the wide adoption of LPCM. In its simplest form it consists of a comparator with a negative feedback loop incorporating an integrator into the feedback path, as shown in Fig. 1.

The classical transformation of this topology from DM to SDM was first presented by Inose and Yasuda in 1962/1963 [9], [10] and is shown in evolutionary form in Fig. 2. The reason why this transformation is reproduced is that the DM form has a closer relationship to LPCM. For example, it is possible to mimic first-order DM operation using an open-loop model with slew rate constraints applied to enable the slope overload condition to be incorporated. Slope overload is entered where the output code is a sequence of all 1's or all 0's and the error signal in the first-order loop exceeds 1 quantum. Since this model operates using a uniform quantizer, then, provided the slope overload threshold is not exceeded, dither can be added and linear operation inferred as per LPCM. Earlier work has also shown that in the non-slope-overload condition, DM is equivalent to time-quantized phase modulation [11], [12].

Consequently coding linearity in terms of a 1-bit coder is defined where the reconstructed signal can at some stage be configured as LPCM and where during quantization appropriate dither is applied and the signal is constrained so that no clipping or slope-limiting distortion

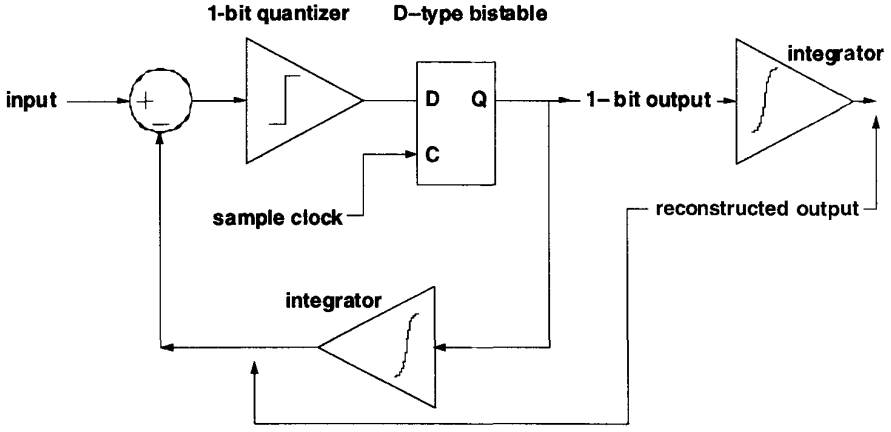


Fig. 1. First-order (single-integrator) delta modulator (DM).

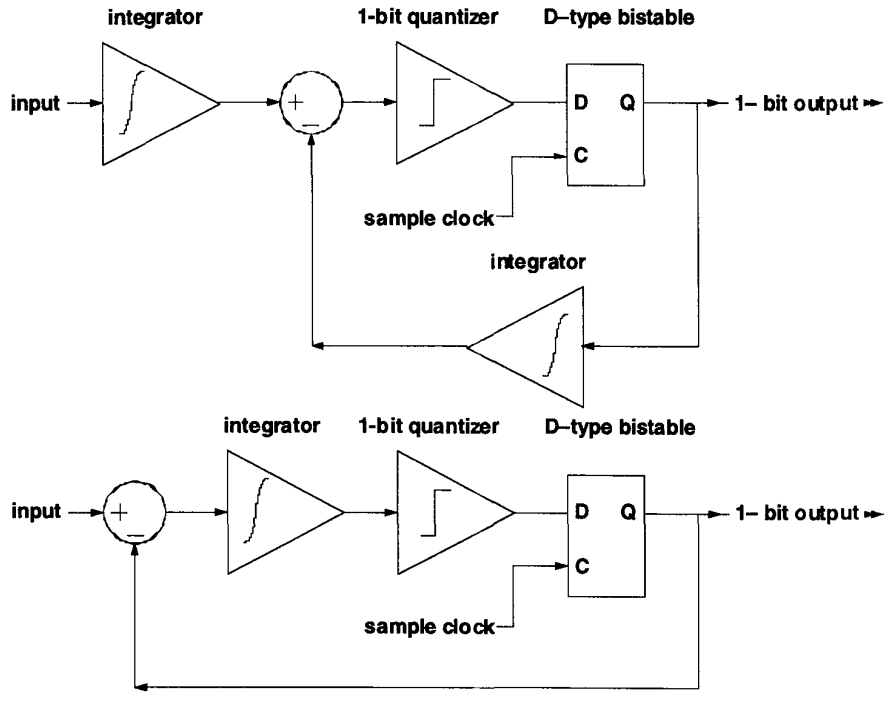


Fig 2 Reconfiguration of DM to SDM.

of any form occurs.

To illustrate this definition, Fig. 3 shows an open-loop DM that includes slope-limitation circuitry [13]. In essence, it is two interleaved flash converters consisting of a bank of comparators and D-type bistables interleaved on alternate clock cycles to mimic the quantization of first-order DM. However, the D-type bistables are also connected vertically to act as an up–down thermometer-style shift register such that the upward and downward progression of 1 pulses is constrained by the clock rate. This implements exactly the slope-overload condition of a first-order DM. The multibit, multilevel output code of the vertical register is then logically differentiated to form the binary output sequence. The two interleaved quantizers are shown explicitly in Fig. 4(a), although in this configuration the slope-overload circuitry has been omitted. However, by introducing a half-quantum offset on alternate samples [12] the same operation, hence idle channel performance, can be achieved using a single quantizer, as shown in Fig. 4(b). In Fig. 4(a) both a positive and a negative ramp is superimposed onto the input of each respective quantizer, which then simulates the idle channel behavior of first-order DM. However, providing the input signal with dither is constrained such that the differentiated quantizer output sequence does not exceed the limits $+1$ or -1, the coder is linear within the coding envelope of DM. If on the other hand the differentiated output sequence exceeds this limit, then DM would demand slope limitation, and this would imply an additional error, which would not be bounded by the error of a linear quantizer. It is this mechanism that is at the core of identifying the nonlinear behavior of DM and by inference that of SDM.

SDM and DM can normally accommodate a second integrator to improve the coding performance, although more than two integrators require careful loop design to ensure stability, which is a direct consequence of slope overload constraining the output. Interestingly, when the models of either Fig. 3 or Fig. 4 include integration in a feedback loop [13], this apparently first-order loop is actually that of second-order DM or SDM. Fig. 5 shows the additional integrator in the forward path, whereas Fig. 6 replaces the open-loop DM structure with an equivalent first-order feedback loop. Finally, in Fig. 7 this is reconfigured to form classic second-order SDM, where the thread of equivalence should be observed, although this latter form includes slope-overload limiting. Consequently, the linear regime of a 1-bit coder can be determined by considering the equivalent model that embeds one or more uniform quantizers and limits the input excitation such that the modulus of the differential of the output does not exceed unity. This applies for both the single- and the double-integration model.

If the slope-overload condition is removed, allowing the bound on the signal differential to be relaxed, then provided that dither is applied correctly at the input to each quantizer, linearity can be inferred. Eliminating the slope-overload condition also allows higher order noise shaping to be applied. It is here that a combination

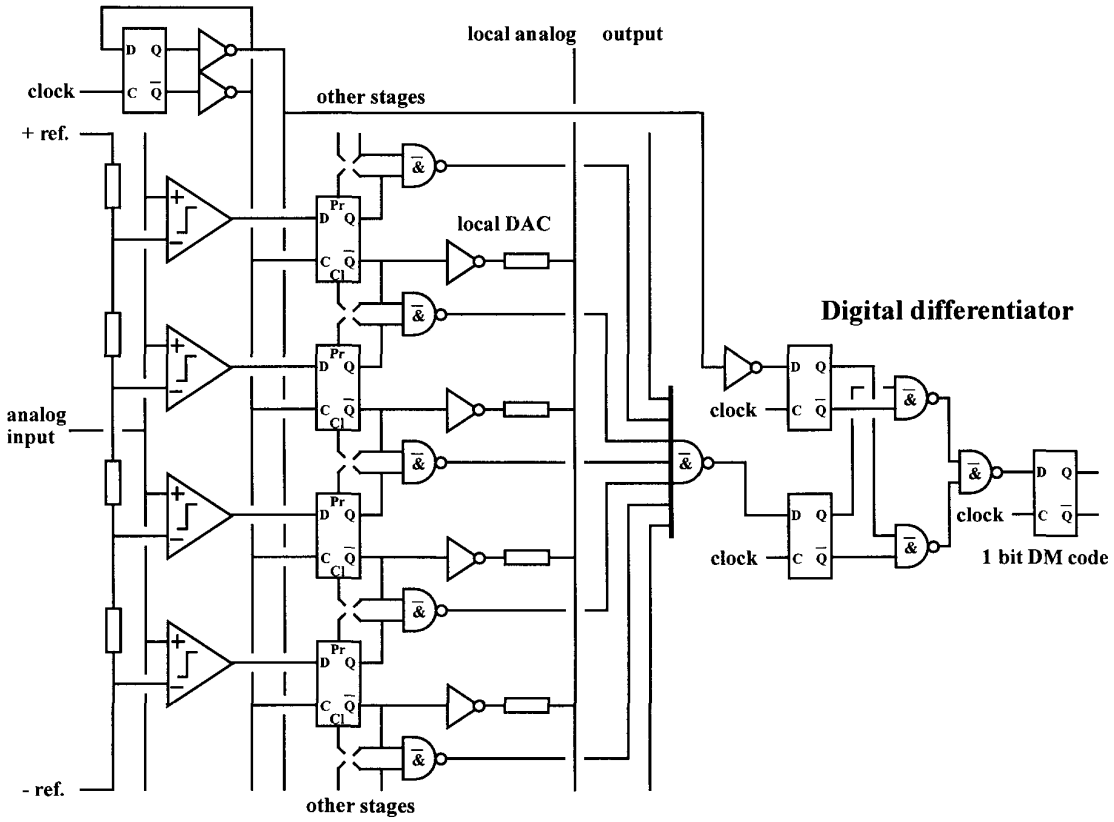


Fig. 3 Open-loop, first-order DM with slope overload circuitry.

of multilevel code and noise shaping offers a fundamental advantage over systems using only a 1-bit code as it enables linear encoding.

2 OVERSAMPLING, NOISE SHAPING, AND EQUALIZATION

The application of noise shaping has been researched in depth for a range of applications that include ADC, DAC, and PWM together with signal requantization as part of a more general signal processing architecture. Provided a signal processor includes uniform quantization with optimal dither, then an exchange between amplitude resolution and sample rate can be made [4]–[7]. Also, by including complementary pre- and deemphasis equalization the relationship between noise floor and amplitude clipping can be modified.

A front-end encoder with a complementary decoder is shown in Fig. 8. It uses an equalizer cascaded with a kth-order noise shaper. We assume here that the source information is encoded with LPCM and that the sampling rate is $8f_s$. The aim is to use a sufficiently high sampling rate such that most of the bandwidth advantage claimed of SDM is achieved, but with the additional advantage of linearity implied by using multilevel uniform quantization with dither. In its basic form the recovered output is derived using a complementary deemphasis filter in cascade with the noise shaper output.

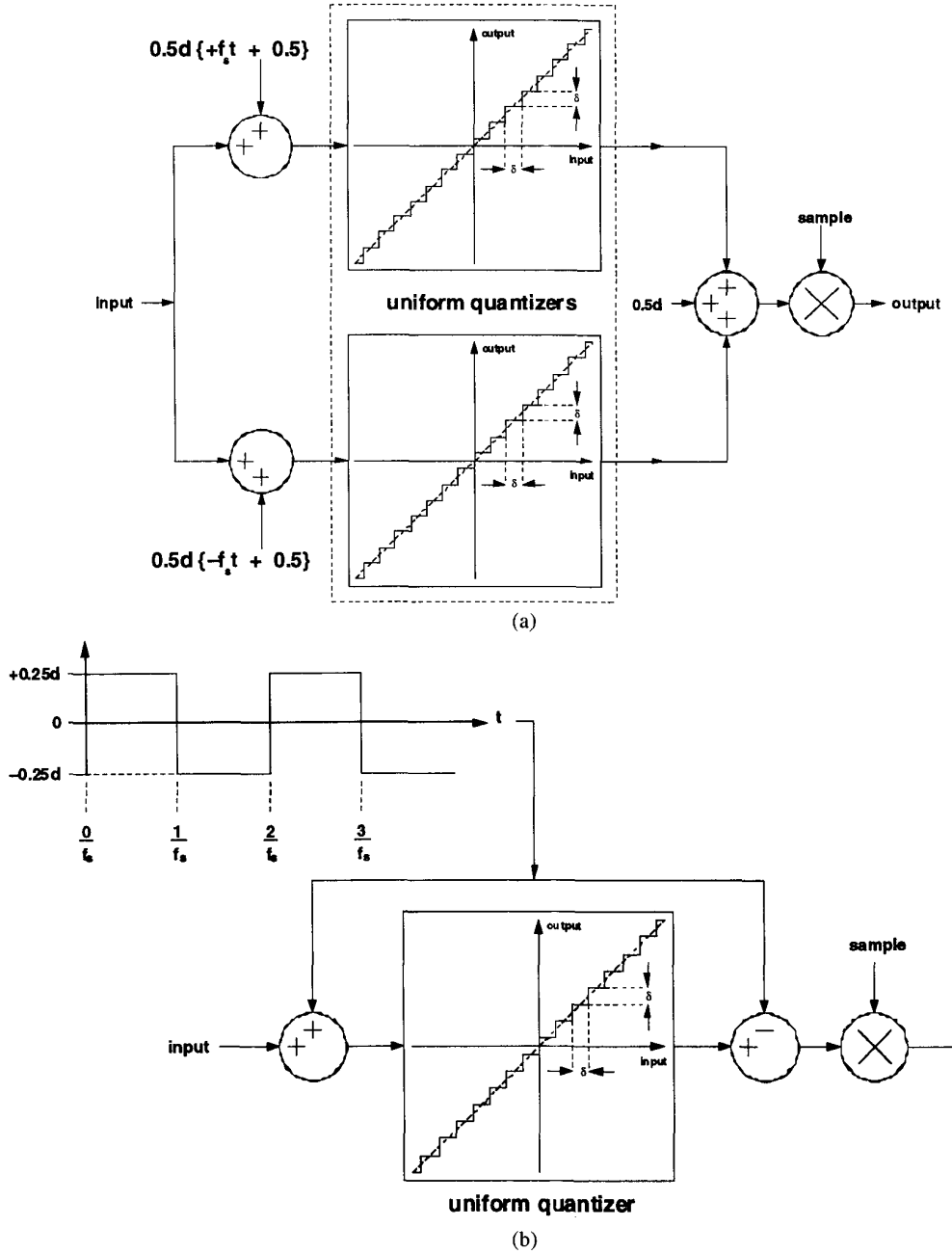


Fig. 4. Open-loop, first-order DM without slope overload circuitry

The aim of this processor is to mimic as closely as possible the performance of a 24-bit LPCM system, except for a relaxation in the high-frequency overload margin to match the characteristic of real-world audio signals. As such, the noise spectrum over the 0- to 24-kHz band should be comparable with (or better than) the noise floor of 24-bit LPCM, although this can be relaxed in the ultrasonic region. A direct approach is to cascade an equalizer and a noise shaper, as shown in Fig. 8. The output data of the equalizer need only be truncated to a word length compatible with the noise shaper input word length (such as 32 bit); so with proper design minimal compromise is implied. However, using the structure of Fig. 9(a), the pre- and deemphasis networks are both driven by the same quantized signal, whereby quantization can be incorporated into the encoder loop together with its own dither signal. By synchronizing encoder–decoder dither (or using no dither), the same signal can be recovered at the output of the side chain. Complementary equalization is achieved using the classic feed-

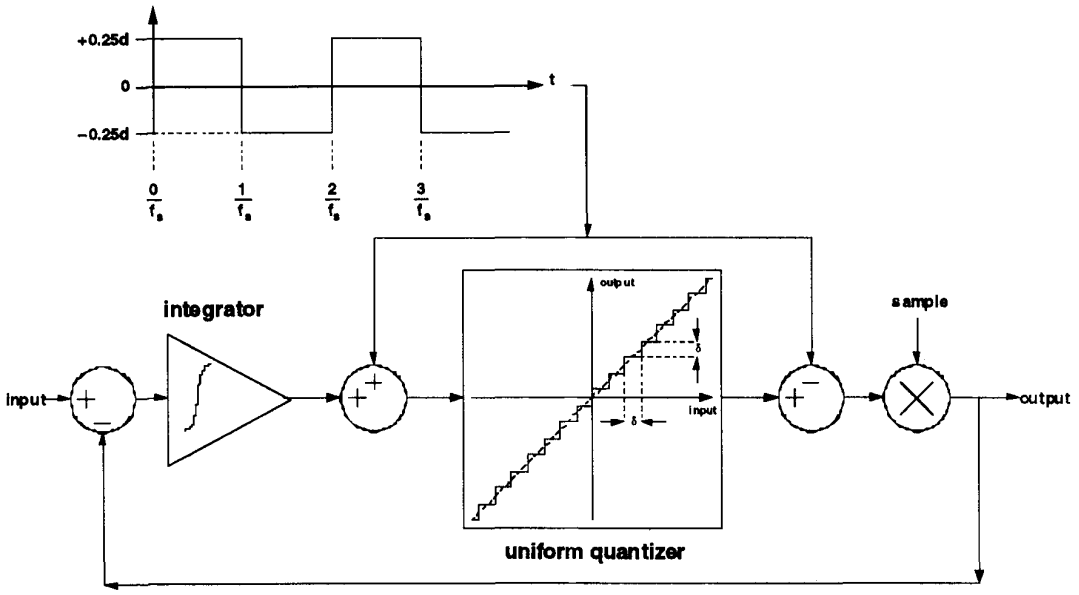


Fig 5. Open-loop DM enclosed with single-integrator feedback

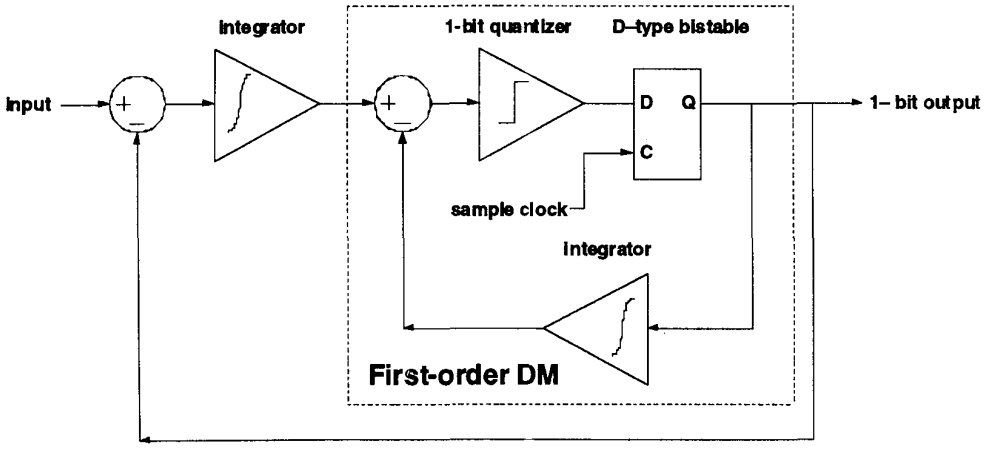


Fig 6 Equivalent second-order DM (integrator in forward path)

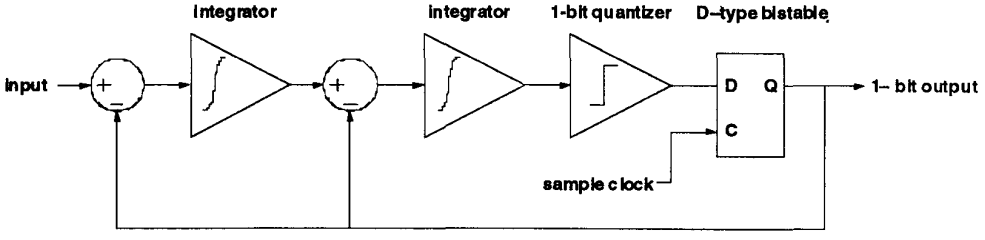


Fig 7 Classic second-order DM

back–feedforward topology, which is conceptually similar to that first used in the Dolby-A noise-reduction system, although this was an analog realization. To investigate the effect of the two quantization processes within this system, forward and reverse noise sources q_f and q_r are shown in the additive noise model of Fig. 9(b). An additive noise model is supported because both quantizers are uniform (although different) and optimal dither is assumed. However, a difficulty encountered with this technique is possible additional constraints on stability, especially when high-order noise shapers are used. As such the cascaded equalizer may prove more tractable and has been employed in the simulations presented in Section 5.

A characteristic of the noise shaper topology shown in Figs. 8 and 9 is that the signal transfer function is unity. This is achieved by including a feedforward path directly to the input of the quantizer and also delaying the main input by one sample period in order to compensate for the unit sample delay required in the feedback path. This process is demonstrated in the following analysis together with complementary equalization.

The intermediate sequence $V_{int}(z)$ can be expressed in terms of the input sequence $V_{in}(z)$ and noise sources q_f and q_r as

$$V_{int}(z) = \frac{1}{1 + z^{-1}E(z)}\left[V_{in}(z) + \frac{q_f}{1 + z^{-1}H(z)} - q_r z^{-1}\right] \quad (1)$$

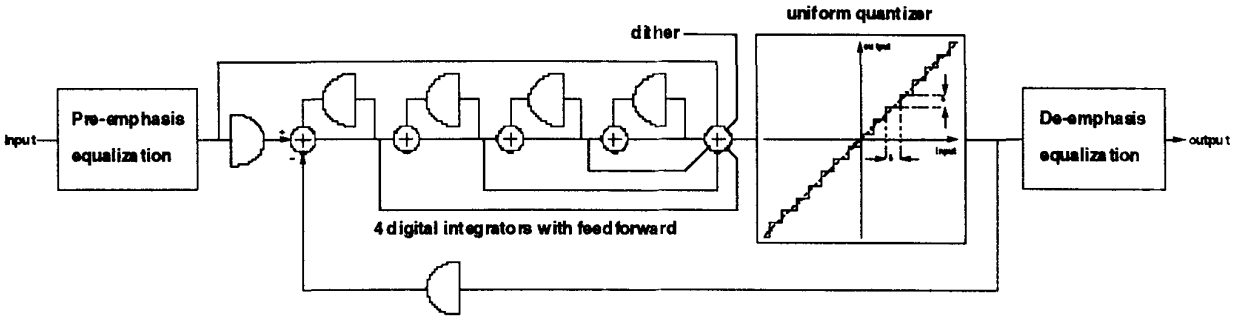
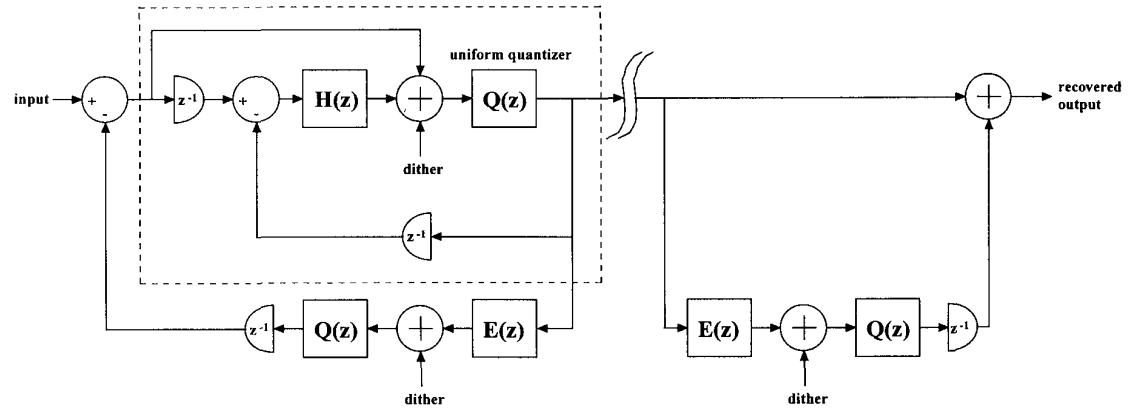
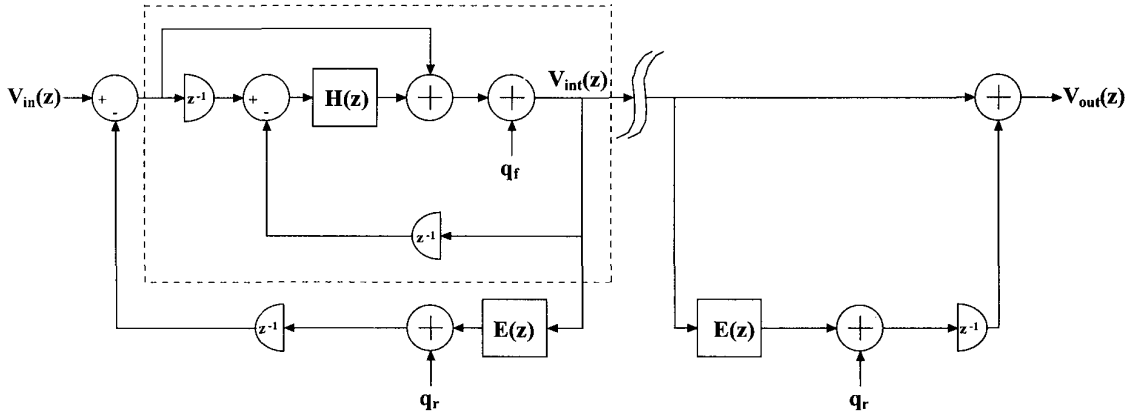


Fig 8 Cascade of preemphasis, noise shaping, and deemphasis



(a)



(b)

Fig. 9. Conceptual system combining equalization and noise shaping together with noise model

while the recovered output $V_{out}(z)$ is

$$V_{out}(z) = V_{int}(z)[1 + z^{-1}E(z)] + z^{-1}q_r . \quad (2)$$

Hence substituting for $V_{int}(z)$, the overall input–output function is

$$V_{out}(z) = V_{in}(z) + \frac{q_r}{1 + z^{-1}H(z)} \quad (3)$$

Provided the same error signal q_r appears in both the encoder and the decoder, then the final output is independent of the truncation noise in the side chain. This suggests that dither is not required in the truncation of the side chain output, removing the need for dither synchronization.

Eq. (3) confirms that the overall signal transfer function is unity, whereas the noise shaping transfer function is $[1 + z^{-1}H(z)]^{-1}$. However, inspection of the intermediate signal shows that $E(z)$ performs preemphasis given by the function $[1 + z^{-1}E(z)]^{-1}$, which has a direct effect on overload performance. Overload can be specified by placing a frequency-domain bound on the intermediate signal $V_{int}(f)$ weighted by a function $W(f)$, such that

$$V_{int}(f)\,|W(f)| \leq \lambda \quad (4)$$

where λ is a constant. However, to determine $|W(f)|$, the transfer function of the intermediate coding stage must be determined, which depends directly on the number of cascaded differentiators [see Eq. (7)]. Fig. 10 shows a possible method of implementation for the filters $H(z)$ and $E(z)$.

3 LOSSLESS DIFFERENTIAL CODING

Section 2 described a method that combined equalization and noise shaping. In this section greater efficiencies are explore by employing lossless multistage differential coding, where our investigation incorporates up to three stages together with overload correction. Differential encoding requires subsequent integration to recover the signal, so it is critical that there be no errors between encode and decode stages, including saturation or requantization. In practice a small amount of side channel information is required to reset the integrators in order to account for errors. Also, effective "ac coupling" should be employed in the decoder to eliminate long-term signal drift and to protect against startup transients. However, it will be shown in Section 4 that certain distortions are permissible, provided the area under the encoded sequence does not change over a given period of time and that errors in the higher order integral waveforms are accommodated.

Fig. 11 shows a two-stage lossless processor that uses complementary differentiation and integration, although additional stages of differentiation and integration can be added. The respective encoding and decoding z trans-

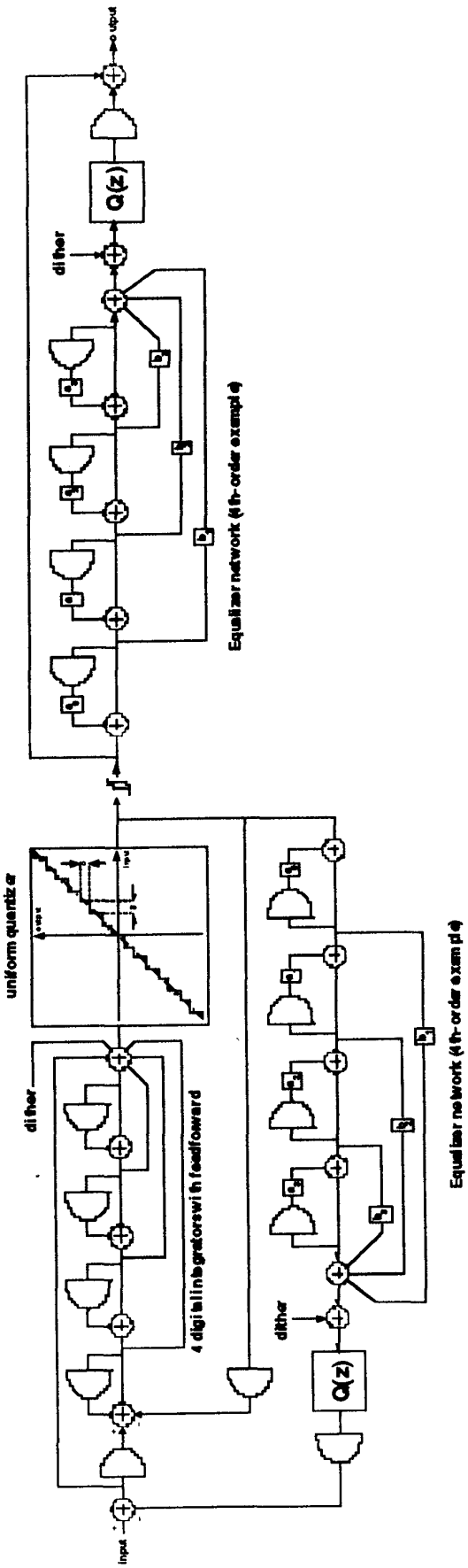


Fig. 10. Composite pre- and deemphasis plus noise shaping.

forms are

$$T_{en}(z) = (1 - z^{-1})^2 \quad (5)$$

$$T_{dec}(z) = (1 - z^{-1})^{-2} . \quad (6)$$

From Eq. (5) the encoding magnitude frequency response $|EN(f)|$ is given as

$$|EN(f)| = 2(1 - \cos(2\pi fT)) . \quad (7)$$

As the coder differentiates the output of the noise shaper, there is a coding advantage in band-limiting the output of the noise shaper to minimize the rate of change. A simple process can average over two adjacent samples, and this option was included in the model. However, there is an interesting problem. As the signal is band-limited, it produces longer words at the output, so the difference between adjacent samples in terms of quanta may increase. A balance therefore has to be found.

4 SOFT AMPLITUDE LIMITiNG USING TIME-DISPERSIVE CORRECTION

For the output bit rate of an N-stage differential encoder to be bounded, the output word length must be limited, implying amplitude clipping and error in the recovered signal. However, as decoding can include up to $N = 3$ cascaded integrators (see Section 3), any uncontrolled modification of the encoded signal can cause the recovered signal to diverge. Hence when clipping distortion occurs, pulse area must be conserved so that the integrated signal remains stable and converges to the required signal level. However, a simple control of the signal average, although necessary, is insufficient as it is shown that errors in the integral waveforms that result from pulse dispersion must also be taken into account.

Introducing appropriately metered time dispersion into signal elements that experience overload can conserve pulse area. To demonstrate the correction process employed in the encoder, Fig. 12 illustrates the principle [14]. In the first waveform a pulse is shown to exceed the overload threshold where the error component is shaded. The first example uses a single backward pulse correction procedure where the excess pulse amplitude of sample n is transmuted and added to the sample $n + 1$. Consequently when pulses n and $n + 1$ are considered together, their total area is conserved. However, although the area under this curve is correct and results in the first integral converging to the correct value, there is a finite loss of area under the first integral. Consequently the second integral does not converge to the correct value (should two or more stages of differentiation–integration be used).

The second example shows a similar procedure, but here half the excess pulse amplitude is added to sample $n + 1$ whereas half is added to sample $n - 1$. This process yields a time-symmetric dispersion of the error, with the error remaining symmetrically centered on sample n, where the effect is similar to symmetrical slew-rate distortion. However, because pulses are amplitude quantized and must remain so when the error dispersion is added, dividing the overload error into two equal parts requires the error to be an even number of quanta. Consequently if the overload error is an odd number of quanta, then the error is increased artificially by one quantum prior to division. The waveforms shown in Fig. 12 reveal that there is no longer an error in the area under the first integral waveform, just a small redistribution of the waveform in time. However, extending to the second integral, as illustrated in Fig. 13, shows that although the correct amplitude is reconstructed after the third pulse, an error in the area under the curve remains, implying a convergence error in the third integral waveform. This requires additional processing to secure the accuracy of the third integral, which is relevant if three stages of differentiation–integration are used.

To correct for convergence error in the third integral, a symmetrical five-pulse substitution sequence is used, as shown in Fig. 14. In all the integral calculations that follow the waveforms are integrated from the sequence beginning at sample $n - 2$ to the sequence ending at sample $n + 2$. For acceptable error correction then samples $n + 2$ and above in the third integral waveform must be identical to the third integral of the nonclipped waveform. The integrals of the five-pulse sequence at sample $n + 2$ are evaluated and compared to the integrals of the nonclipped sample as follows.

For the first integral evaluated at sample $n + 2$.

$$L = a_0 + 2(a_1 + a_2) .$$

For the second integral evaluated at sample $n + 2$,

$$L = a_0 + 2(a_1 + a_2) . \quad (8)$$

For the third integral evaluated at sample $n + 2$,

$$6L = 6a_0 + 13a_1 + 16a_2 . \quad (9)$$

Defining the error in sample n as e and equating e to the sum of the substitution pulses in samples $n - 1$, $n - 2$, $n + 1$, and $n + 2$,

$$e = 2(a_1 + a_2) . \quad (10)$$

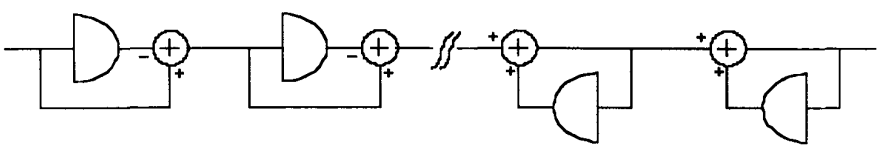


Fig. 11 Two-stage complementary differentiation and integration

Solving Eqs. (8)–(10) yields the amplitudes of the five substitution pulses,

$$a_0 = L - e \tag{11}$$

$$a_1 = \frac{2}{3}e \tag{12}$$

$$a_2 = -\frac{1}{6}e\,. \tag{13}$$

Eq. (13) reveals a weighting of 1/6 in the error e. As a consequence, errors must be quantized to multiples of six quanta to avoid further quantization distortion if an

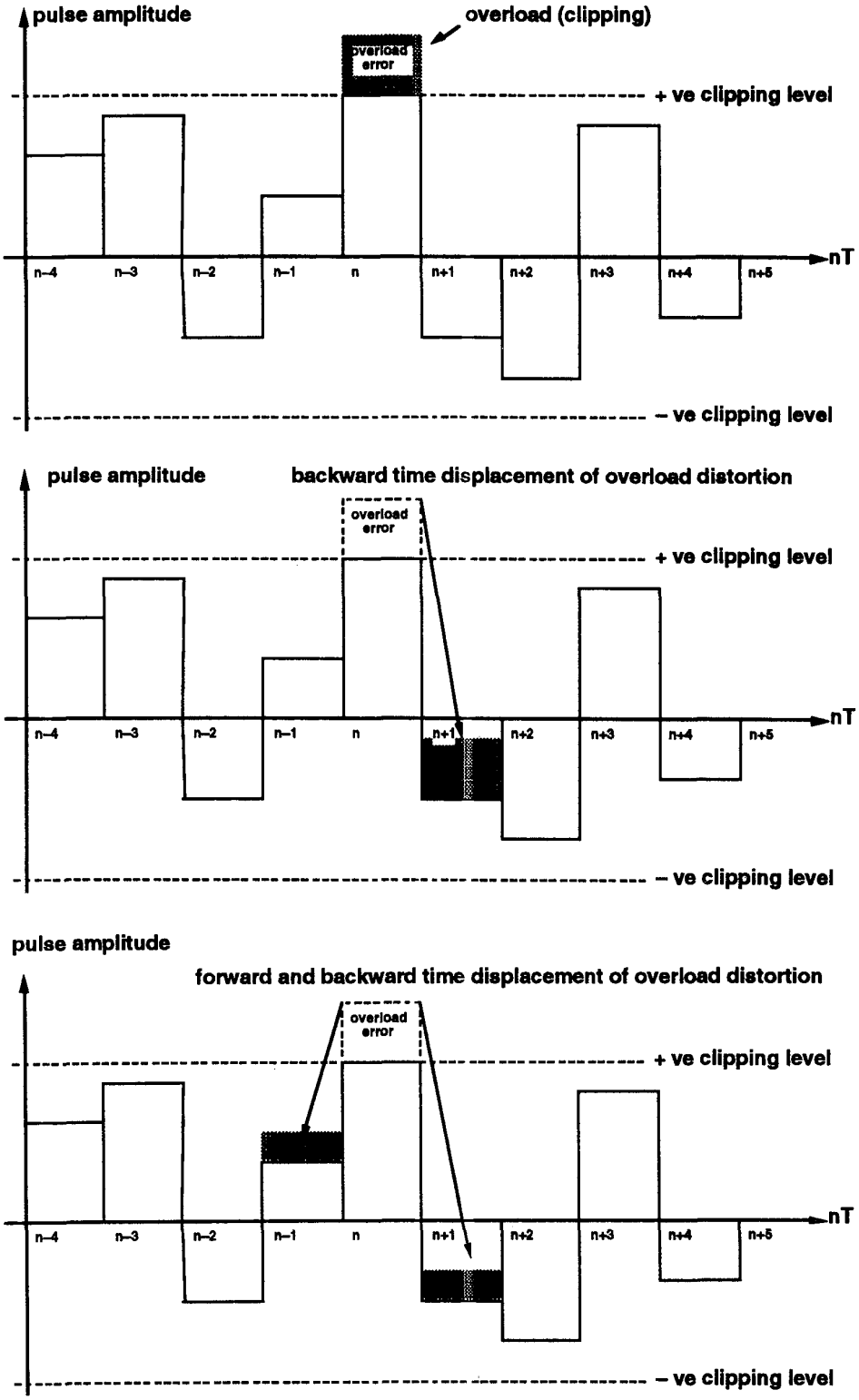


Fig 12 Time dispersive correction of an overload error

increase in output resolution is to be avoided.

The correction procedure operates as follows. First the undistorted differentials are computed (either one stage, two stages, or three stages, depending upon choice) from the output of the noise shaper. The resolution of the output code then establishes the upper and lower clipping levels. Where an overload error occurs, the absolute value of a sample is reduced by the nearest multiple of six quanta. The four remaining substitution pulses are then calculated and summed with the adjacent samples in the output sequence. The resulting waveform is scanned again for overload errors. If any remain, the

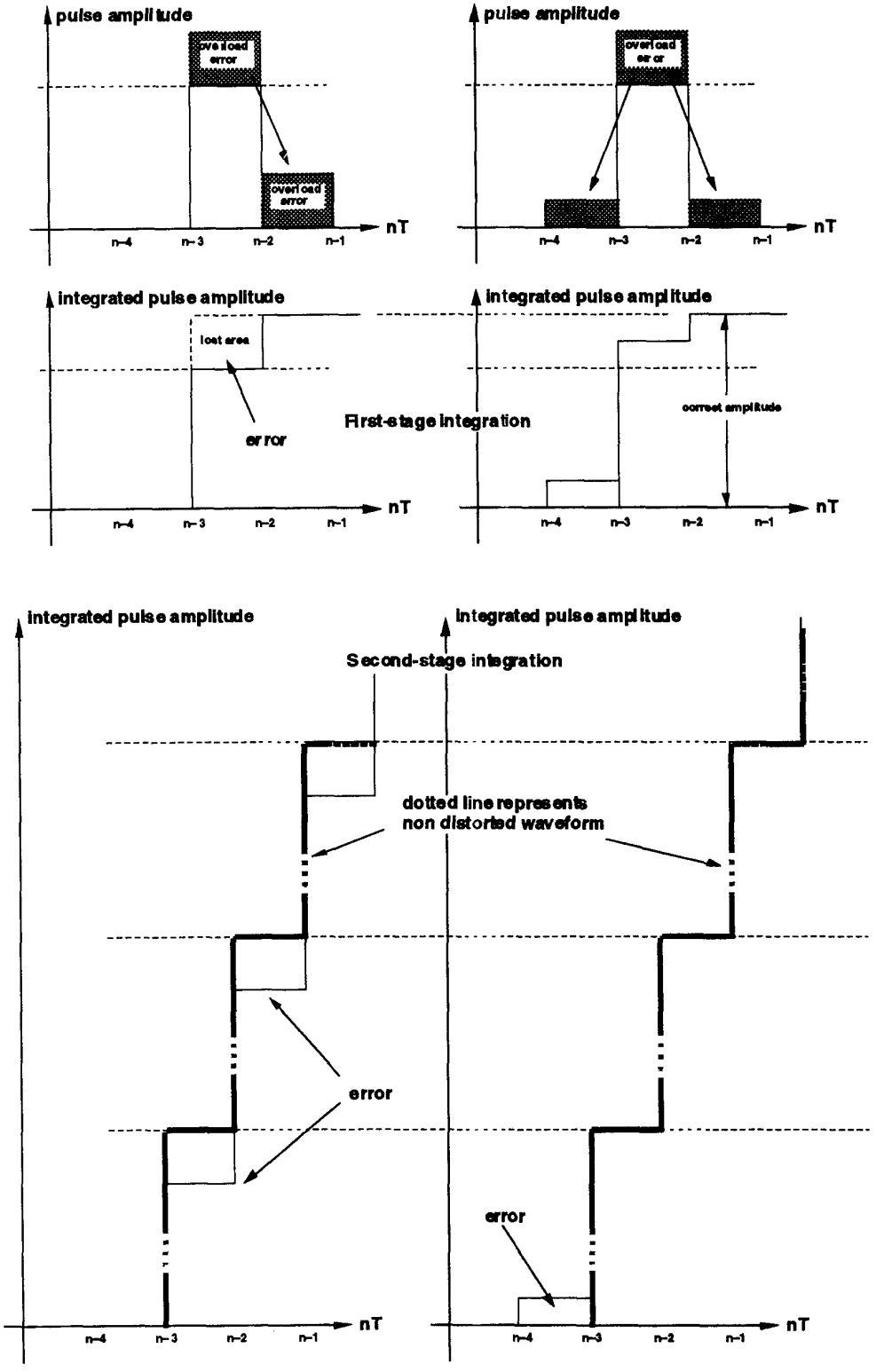


Fig 13. Two-stage integration performance.

procedure is repeated as many times as required until the resulting output falls within the overload limits. As noise shaping is used prior to differentiation, there can be considerable interpulse differences such that if one pulse is, say, a high positive value, then the adjacent pulses are usually negative and can accommodate the error without overload. This factor helps accommodate the dispersive error sequences.

The procedure can be extended to the fourth integral using a symmetrical seven-pulse substitution. Using a approach similar to that used before (but performed from $n - 3$ to $n + 3$), it follows that

$$e = 2(a_1 + a_2 + a_3) .$$

For the first integral evaluated at sample $n + 3$,

$$L = a_0 + 2(a_1 + a_2 + a_3) .$$

For the second integral evaluated at sample $n + 3$,

$$L = a_0 + 2(a_1 + a_2 + a_3) .$$

For the third integral evaluated at sample $n + 3$,

$$10L = 10a_0 + 21a_1 + 24a_2 + 29a_3$$

and for the fourth integral evaluated at sample $n + 3$,

$$4L = 4a_0 + 9a_1 + 12a_2 + 17a_3 .$$

Then,

$$a_0 = L - e \tag{14}$$

$$a_1 = -0.5e \tag{15}$$

$$a_2 = 1.7e \tag{16}$$

$$a_3 = -0.7e . \tag{17}$$

The factor 1.7 implies that a minimum value for e is 10 quanta, making the pulse weightings in the substitution sequence $a_0 = L - 10$, $a_1 = 5$, $a_2 = 17$, and $a_3 = 7$. However, the coefficient a_2 being greater than a_0 implies gain, so there is an increased probability that this substitution could actually push adjacent samples into clipping. As a result, this higher order correction was not pursued in the present study.

5 SIMULATION AND RESULTS

The coder and decoder, including equalization, noise shaping, differentiation–integration, and clipping correction, were simulated in MATLAB. (See Appendix for listing.[1]) However, it is not intended to give an exhaustive search of the coding options but to illustrate

[1] For MATLAB Simulation Program in Appendix, open this paper title in *Supplementary Material to Papers* on AES website at http //www aes org/journal/suppmat/

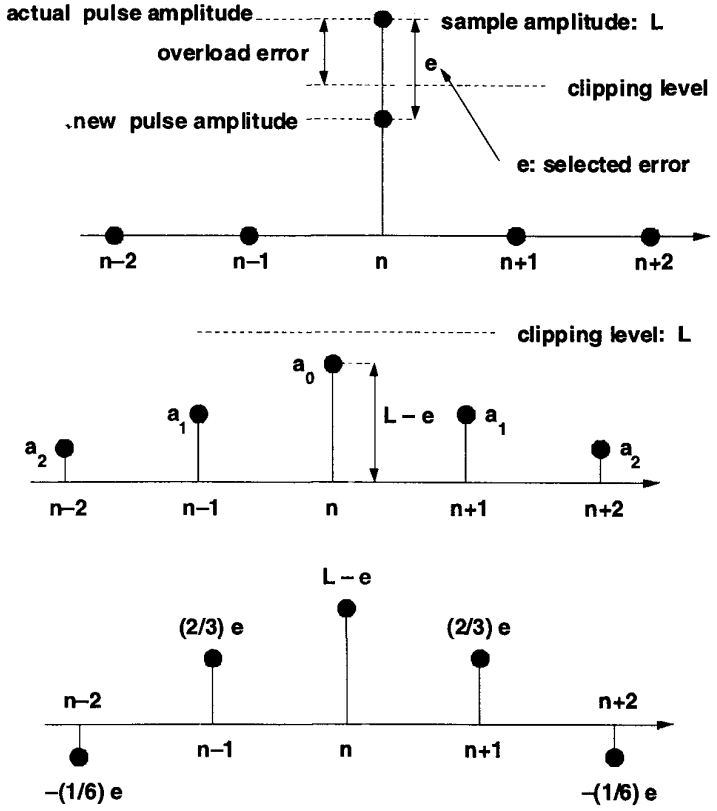
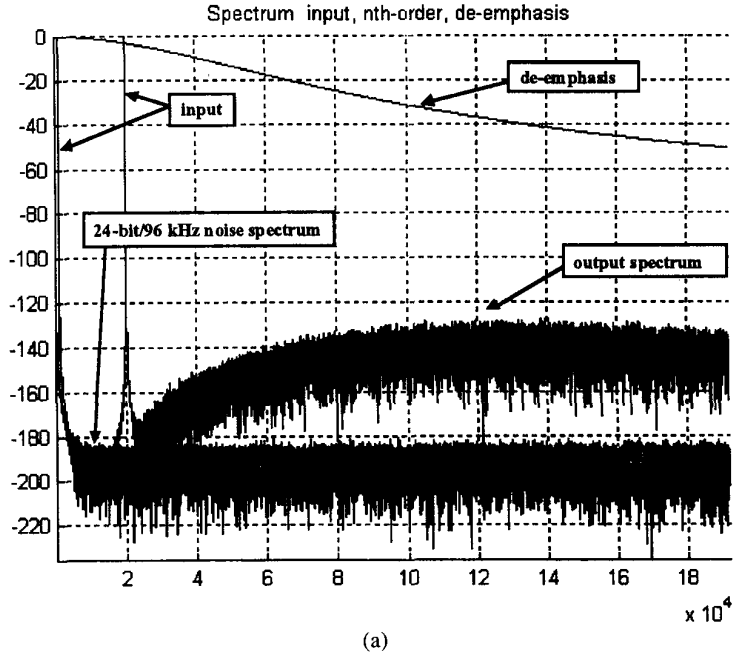


Fig. 14 Five-pulse substitution for third-integral error correction.

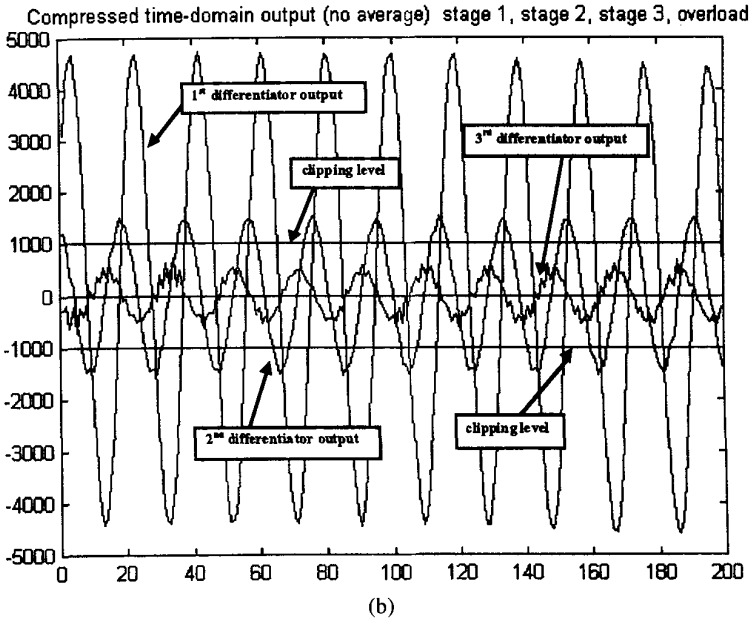
the performance achievable and in particular to illustrate the performance of the overload option using both the three- and the five-pulse substitution procedure. The spectral plots show the line spectrum of the input together with the noise floor of the coding–decoding process, including emphasis and deemphasis. To interpret the noise floor, a 24-bit/96-kHz LPCM dithered reference signal is generated and scaled so that both the output of the recovered signal and that of the reference signal have the same maximum amplitude. The first set of results is shown in Fig. 15 and corresponds to the following data:

```
order = 6        %noise shaper order
A1 = 10000       %input amplitude
A2 = 10000       %input amplitude
fin1 = 1000      %input frequency of A1
fin2 = 20000     %input frequency of A2
fs = 96000       %digital audio sampling frequency (96 kHz)
m = 4            %oversampling ratio (relative to fs)
v = 16           %vector length, bits
qout = 11        %output word length of differentiator
qin = 24         %reference signal resolution
```

In this example the input consists of two equal amplitude sine waves (that is, A1 = A2 = 10 k) of respective frequencies 1 kHz and 20 kHz and the system sampling rate is set at 4 times 96 kHz. The parameters are selected



(a)

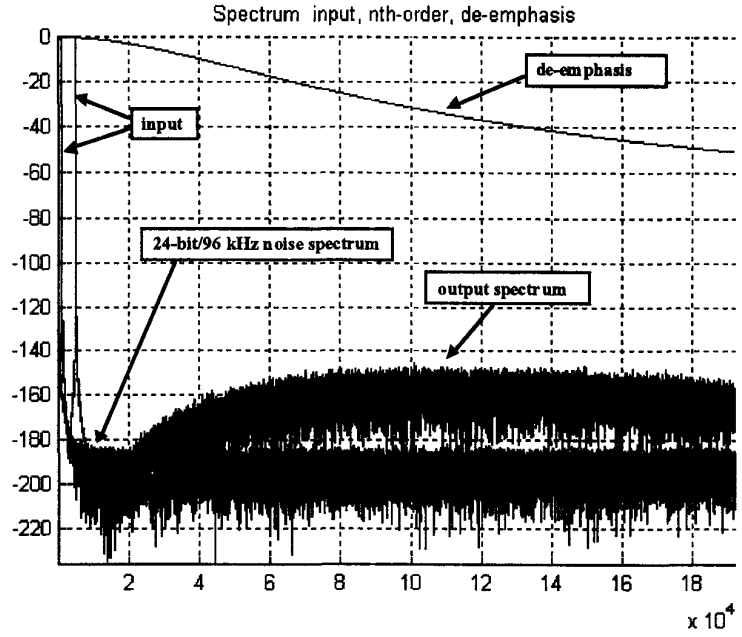


(b)

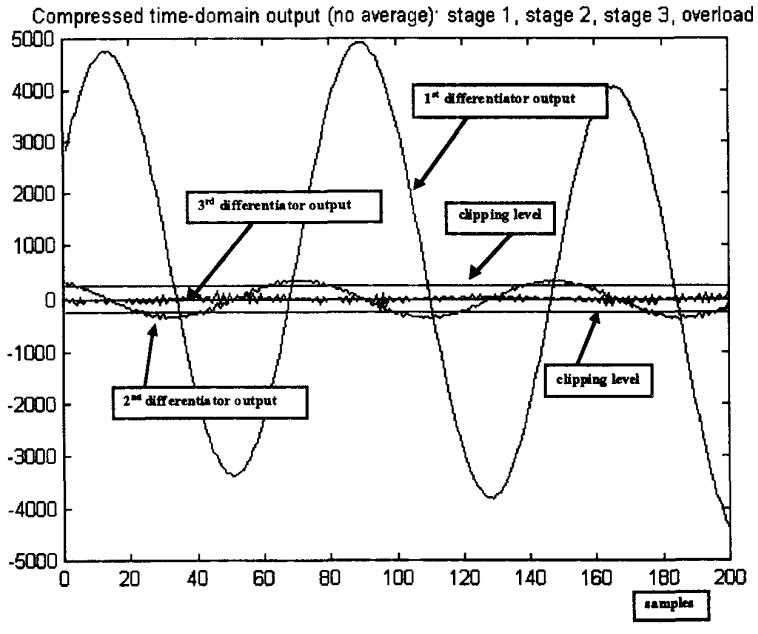
Fig. 15. (a) Recovered spectrum of coder–decoder (bandwidth 0 to $f_s/2$ Hz). (b) Output of first, second, and third differentiators.

to achieve a performance close to 24-bit LPCM at 20 kHz, where the channel bit rate is 4.2240 Mbit/s. Note how in Fig. 15(b) the output of the third differentiator is much lower than that of the second differentiator and remains within the clipping bounds determined by the output word length (q_{out}). However, because of sixth-order noise shaping and differential coding a dynamic range well in excess of 24 bit is achievable at lower frequencies. In Fig. 16 similar results are computed but with input signal frequencies of 5 kHz and 1 kHz. Here the input levels have undergone a five-fold increase in level (that is, A1 = A2 = 50 k), the noise shaper is reduced to fifth order and the output word length is reduced from 11 bit to 9 bit. Fig. 16(b) reveals a lower output level from the third differentiator where this signal is dominated now mainly by noise shaper activity rather than the signal. These factors imply that a lower bit rate can be used before clipping distortion occurs, where in the second simulation the bit rate is now reduced to 3.4560 Mbit/s.

To demonstrate the effect of clipping and the two proposed forms of error correction that can be applied after the third differentiator, the simulations are repeated with the same 5-kHz and 1-kHz signals but with the input signal raised in level. It was observed that a 6-dB increase gave occasional clipping so the signal was



(a)

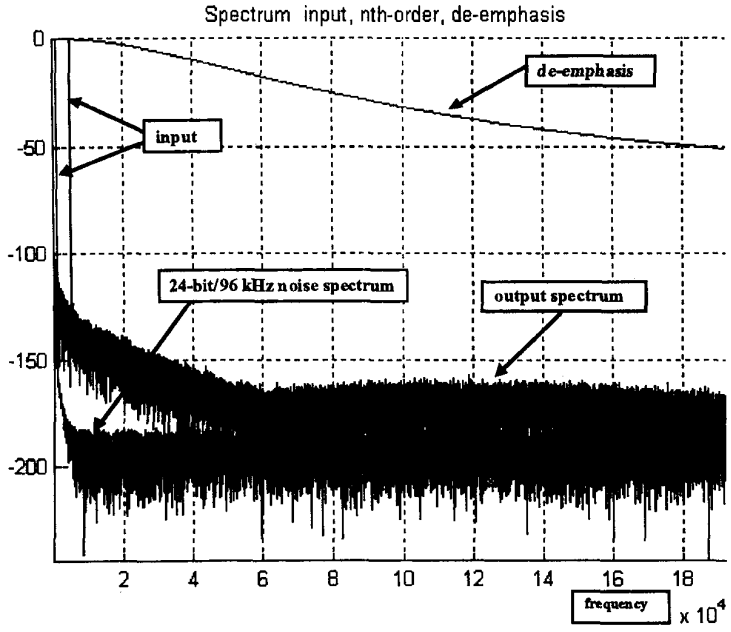


(b)

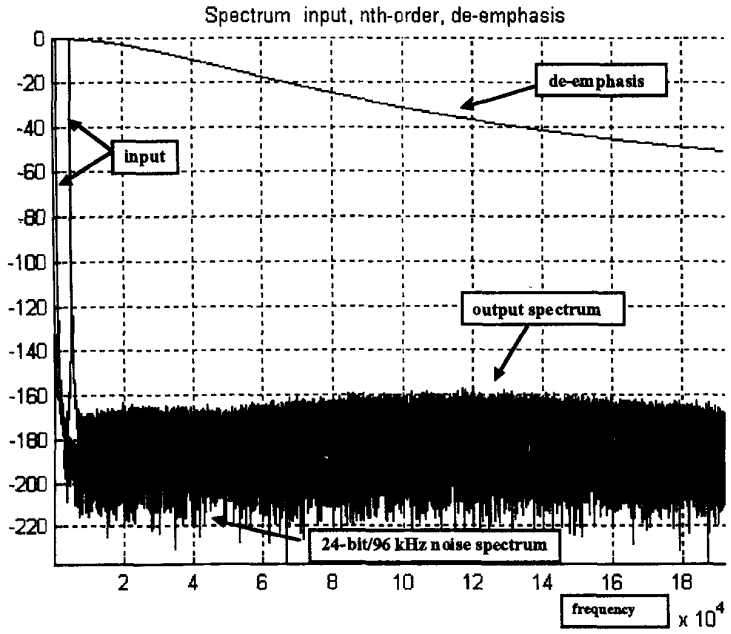
Fig 16. (a) Recovered spectrum of coder–decoder (bandwidth 0 to $f_s/2$ Hz). (b) Output of first, second, and third differentiators

raised a further 6 dB (that is, A1 = A2 = 200 k). The spectral results for error correction with three-pulse substitution (clip = 1) are shown in Fig. 17(a) while those with 5-pulse substitution (clip = 2) are shown in Fig. 17(b). It is evident that the five-pulse substitution gives a superior error spectrum and remains noiselike, even though the output waveform of the third differentiator is clipped. In both simulations the program indicated only a single-error correction iteration (program outputs sx) so error dispersion is limited here to only one or two samples respectively on either side of any samples exceeding overload. In Fig. 18 the simulation is repeated using five-pulse substitution but with the input increased a further 6 dB (that is, A1 = A2 = 400 k). At this level the program required two iterations to converge and was close to its ultimate overload limit. Using five-pulse substitution and assuming the maximum number of iterations permitted is $\widehat{sx}$, then the minimum sample look ahead required for overload correction is $2\widehat{sx}$ where $\widehat{sx} > 5$ is defined as gross overload.

An important attribute of the coding technique is that the error spectrum is not correlated with the signal and that even when clipping does occur, the error is mainly modulation noise. As such, the simulations reveal a use-



(a)



(b)

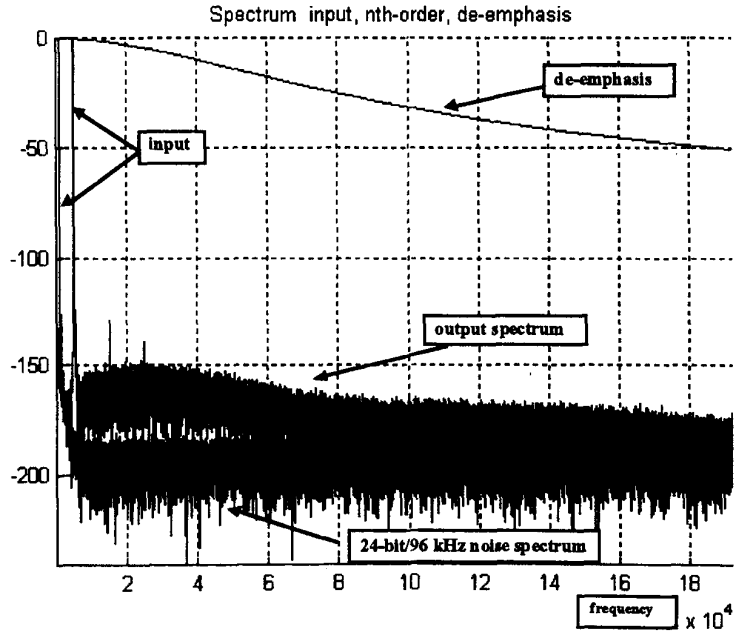
Fig. 17. (a) Recovered spectrum, three-pulse substitution in error correction (+12 dB) (b) Recovered spectrum, five-pulse substitution in error correction (+12 dB).

ful insensitivity to clipping distortion.

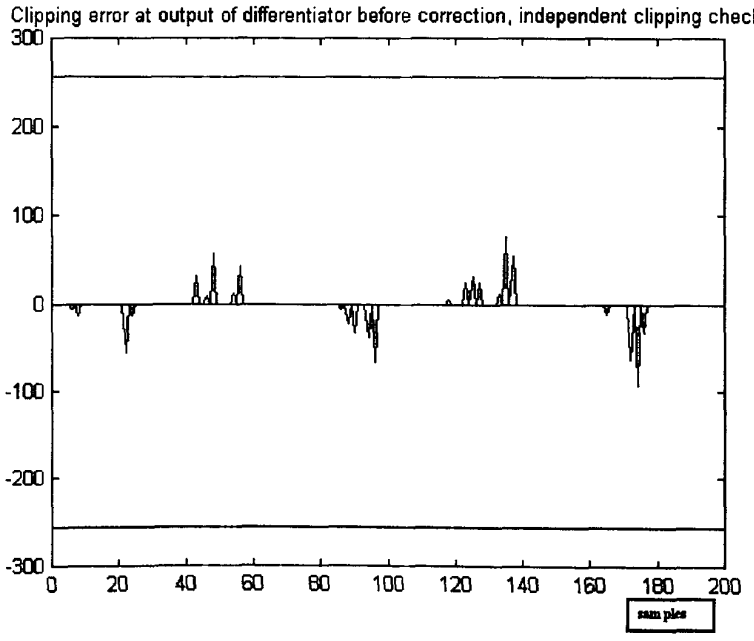
Finally, a more modest design is simulated where comparisons are made with a 20-bit LPCM system. Here the channel bit rate is 1.7280 Mbit/s and is achieved by reducing the sampling rate to 2 times 96 kHz, the noise shaper order to 4, but retaining an output word length of 9 bit. Good coding is obtained at 5 kHz as shown in Fig. 19 together with the advantage of a graceful clipping performance. However, the high-frequency dynamic range is restricted significantly and requires the input to be lowered by 36 dB at 20 kHz compared with 5 kHz to prevent the onset of clipping.

6 CONCLUSION

This engineering report reviewed the fundamental cause of nonlinear distortion in bit-stream coders and then proceeded to describe a method of signal coding that combines noise shaping with differential coding to reduce word lengths. A particular feature of the study was the controlled time dispersion to improve overload performance. Results were presented to confirm the validity of the process and to demonstrate tolerance to overload. A key feature is that for midrange audio signals the error spectrum remains substantially noiselike



(a)



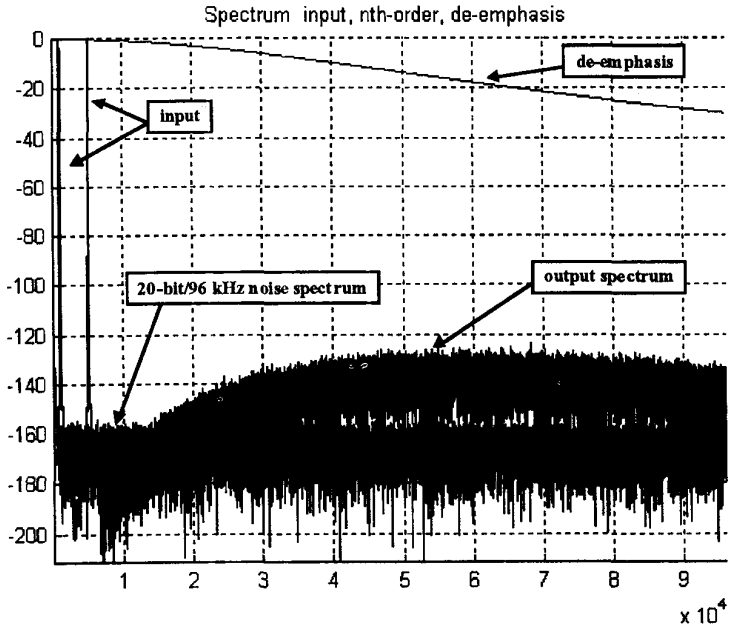
(b)

Fig 18 (a) +18 dB signal overload. (b) Clipping error at output of differentiator before correction (+18 dB).

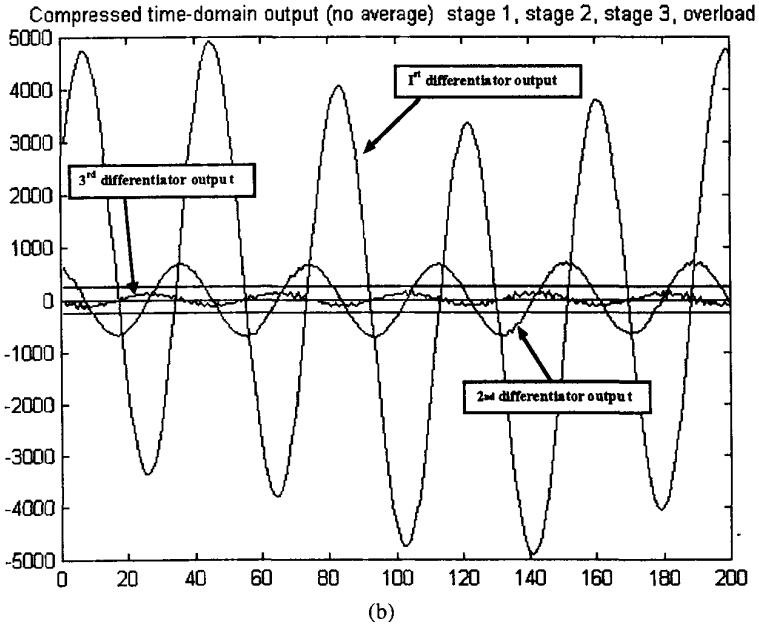
and does not exhibit the usual characteristics of clipping distortion. This distortion performance results from the use of three-stage integration following clipping, which gives a spectral weighting advantage and together with the five-pulse dispersive correction maintains good waveform fidelity.

There is a wide range of system parameters that can be selected, and there is interplay between noise shaper order, oversampling ratio, and output word length (prior to differentiation) selection. However, as a guiding principle, when the output resolution (q_{out}) is set, the order of the noise shaper can be adjusted such that the random activity at the output of the third integrator spans approximately one half the signal range. Also, the adjacent sample-averaging feature proved only advantageous where the output signal of the differentiator stage was dominated by the noise shaper activity. Once the signal level increased above the noise shaper "noise," it had a more coherent form, so averaging increased the signal level.

It is anticipated that improvements can be made in the noise shaping transfer function by introducing transmission zeros to lower the quantization distortion in the 20-kHz region. Also, a better design of pre- and deemphasis characteristics could then be achieved. It is also evident that for systems using lower bit rates the main compromise is in high-frequency overload. However, because the probability of occurrence of high-frequency, high-amplitude signals is lower, variable-bit-rate and possibly



(a)



(b)

Fig. 19. (a) Lower specified coder aimed at 20-bit LPCM (1.7280 Mbit/s). (b) Output of first, second, and third differentiators.

Huffman coding could prove attractive by assigning wider word lengths under these circumstances.

Finally, the results confirmed the low-distortion characteristics inherent in a coder employing LPCM as opposed to a 1-bit code. However, the gains in the low-frequency dynamic range achievable with complementary differential–integral processing are significant, and this is considered an important feature for channels that aspire to high data rates.

7 REFERENCES

[1] P. C. Eastty, C. Sleight, and P. D. Thorpe, "Research on Cascadable Filtering, Equalisation, Gain Control, and Mixing of 1-Bit Signals for Professional Audio Applications," presented at the 102nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 410 (1997 May), preprint 4444.

[2] R. J. Van de Plassche, "Dynamic Element Matching for High Accuracy Monolithic D/A Converters" *IEEE J. Solid State Circ.*, SC-11, pp. 795–800 (1976 Dec.).

[3] C. Dunn and M. Sandler, "A Comparison of Dithered and Chaotic Sigma–Delta Modulators," *J. Audio Eng. Soc.*, vol. 44, pp. 227–244 (1996 Apr.).

[4] H. A. Spang and P. M. Schultheiss, "Reduction of Quantizing Noise by Use of Feedback," *IRE Trans. Commun. Sys.*, pp. 373–380 (1962 Dec.).

[5] S. K. Tewksbury and R. W. Hallock, "Oversampled Linear Predictive and Noise-Shaping Coders of Order $N > 1$," *IEEE Trans. Circ. Sys.*, vol. CAS-25, pp. 437–447 (1978 July).

[6] M. O. J. Hawksford, "Chaos, Oversampling, and Noise Shaping in Digital-to-Analog Conversion," *J. Audio Eng. Soc.*, vol. 37, pp. 980–1001 (1989 Dec.).

[7] J. R. Stuart and R. J. Wilson, "Dynamic Range Enhancement Using Noise-Shaped Dither at 44.1, 48, and 96 kHz," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 646 (1996 July/Aug.), preprint 4236.

[8] R. Steele, *Deltamodulation Systems* (Pentech Press, 1975).

[9] H. Inose, Y. Yasuda, and J. Murakami, "A Telemetering System by Code Modulation—Delta Sigma Modulation," *IRE Trans. Space Electron. Telem.*, vol. SET-8, p. 204 (1962 Sept.).

[10] H. Inose and Y. Yasuda, "A Unity Bit Coding Method by Negative Feedback," *Proc. IEEE*, vol. 51, p. 1524 (1963 Nov.).

[11] J. E. Flood and M. J. Hawksford, "Exact Model for Deltamodulation Processes," *Proc. IEE (London)*, vol. 118, pp. 1155–1161 (1971).

[12] M. J. Hawksford, "Unified Theory of Digital Modulation," *Proc. IEE (London)*, vol. 121, pp. 109–115 (1974 Feb.).

[13] M. J. Hawksford, "Deltamodulation Coder Using a Parallel Realisation," in *Proc. 37th IERE Conf.* (1977 Sept.), pp. 547–557.

[14] M. O. J., Hawksford, "A Comparison of Two-Stage 4th-Order and Single-Stage 2nd Order Delta–Sigma Modulation in Digital-to-Analogue Conversion," in *Proc. IEE Conf. on Analogue to Digital and Digital to Analogue Conversion* (Swansea, UK, 1991 Sept.), publ. 343, pp. 148–152.

8 BIBLIOGRAPHY

D. J. Goodman, "The Application of Delta Modulation to Analogue-to-PCM Encoding," *Bell Sys. Tech. J.*, vol. 48, pp. 321–343 (1969 Feb.).

J. E. Flood and M. J. Hawksford, "Adaptive Delta–Sigma Modulation Using Pulse Grouping Techniques," in *Proc. 23rd IERE Conf. on Digital Processing of Signals in Communications* (Loughborough University, 1972 Apr.), pp. 445–462.

J. D. Everard, "Improvements to Delta–Sigma Modulators when Used for PCM Encoding," *Electro. Lett.*, vol. 12/15 (1976 July).

E. M. Deloraine, Van Mierlos, and Derjavitch, "Méthodes et système de transmission par impulsions," French patent 932.140 (1947/1948).

F. de Jager, "Deltamodulation, a Method of PCM Tranmission Using 1-Unit Code," *Philips Res. Rep.*, vol. 7, pp. 442–466 (1952).

J. F. Schouten, F. de Jager, and J. A. Greefkes, "Deltamodulation, a New Modulation System for Telecommunications," in Dutch, *Philips Tech. Tidjdschr.*, vol. 13, p. 249 (1951 Sept.), *Philips Tech. Rev.*, vol. 13, p. 237 (1952 Mar.).

H Van de Weg, "Quantising Noise of a Single-Integration Deltamodulation System with an N-Digit Code," *Philips Res. Rep.*, vol. 8, p. 367 (1953).

THE AUTHOR

hawksford

Malcolm Hawksford is Director of the Centre for Audio Research and Engineering, and a Professor in the Department of Electronic Systems Engineering at Essex University, where his interests encompass audio engineering, electronic circuit design, and signal processing.

Professor Hawksford gained a B.Sc. with First Class Honours in 1968 from Aston University, Birmingham, U.K. and was also awarded a Ph.D. in 1972. His Ph.D program was sponsored by a BBC Research Scholarship and investigated delta modulation and delta—sigma modulation (now commonly known as "bitstream" coding) for color television and produced a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system. His research encompasses analog and digital systems with a strong emphasis on audio systems, including loudspeaker technology. Since 1982, his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University A first in 1986 was a prototype system that was sponsored by a research contract from Canon and was demonstrated at the Canon Research Centre in Tokyo Much of this work has appeared in the *Journal of the Audio Engineering Society* together with a substantial number of contributions at AES conventions. His research has also encompassed oversampling and noise shaping techniques applied to analog-to-digital and digital-to-analog conversion, the linearization of PWM encoders, and 3-dimensional spatial audio and telepresence, including multichannel sound reproduction.

Professor Hawksford is a recipient of the publications award of the Audio Engineering Society for his paper entitled "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design." This award was made for the best contribution by an author of any age to volumes 45 and 46 of the AES *Journal*. Professor Hawksford is a chartered engineer, a Fellow of the AES, the IEE, and the IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founding member of the Acoustic Renaissance for Audio (ARA) technical committee. He is also a technical consultant for NXT, U K

PROCEEDINGS

THE INSTITUTION OF ELECTRICAL ENGINEERS

Volume 118 Electronics

Exact model for delta-modulation processes

Prof. J. E. Flood, D.Sc., C.Eng., F.I.E.E., and M. J. Hawksford, B.Sc.

Indexing terms: Delta modulation, Modelling

Abstract

A mathematical model is described which generates a pulse waveform identical to that of a single-integrator delta modulator, provided that the input signals to the latter do not cause slope overloading. The model uses analogue techniques of angle modulation and sampling to generate time- and amplitude-quantised signals, thus readily lending itself to exact analysis. The delta-modulation process is treated in a general manner that is equally applicable to delta-entry and sigma-entry systems. By this means, a central delta modem is defined which includes both pulse modulation and local decoding (prior to final filtering) of the pulse waveform. The model formulation is such that it is equivalent to the delta modem. The equivalence of the delta modem and model is proved analytically. It has also been verified by simulation on a digital computer and demonstrated experimentally. The model can be extended to simulate a double-integration network, provided that the necessary prediction is included. The model can also be extended to represent pulse-code modulation, because a linearly quantised p.c.m. signal can be obtained by suitable sampling of the output of a delta modulator.

List of symbols

$A(\omega)$ = transfer function of second integrator in double-integration process
C = number of digits in binary code of p.c.m. system
$D(t)$ = instantaneous signal level, at time t, to delta modulator
D_{max} = maximum amplitude of $D(t)$, when $D(t)$ is sinusoidal
d = sampling delay time in p.c.m. model
$E(t)$ = error signal in quantisation process at time t
f = signal frequency
f_c = maximum signal frequency
G = constant multiplier
H = sampling rate of p.c.m. model
K = constant, in phase-modulation process
M = number of complete rotations of the phase of the phase-modulated carrier from zero time to Nth sample
N = Nth sample in delta-modulation and model process
N_{N1} = number of negative pulses in delta-modulation process from zero time to Nth sample
N_{N2} = number of negative pulses in model process from zero time to Nth sample
N_{P1} = number of positive pulses in delta-modulation process from zero time to Nth sample
N_{P2} = number of positive pulses in model process from zero time to Nth sample
N_S = Nyquist sampling rate
P = delta-modulator and model-clock rate
P_c = p.c.m.-system pulse rate
$P_1(t)$ = pulse pattern of delta-modulation process
$P_2(t)$ = pulse pattern of model process
R = positive integer
r = small positive number
R_S = signal range in delta-modulation and p.c.m. processes
$S_1(t)$ = accumulated output of local integrator in delta modulation
$S_1'(t)$ = remote accumulated output, including transmission errors in delta-modulation process
$S_2(t)$ = accumulated output of local integrator in model process
$S_2'(t)$ = remote accumulated output, including transmission errors, in model process
$s(t)$ = normalised signal input to delta–sigma modulator and frequency-controlled model
t = time
T_1 = time constant of first integrator in double-integration process
x = instantaneous amplitude of phase-modulated sinusoid
X = amplitude of phase-modulated sinusoid
δ = general delta pulse
ϕ = excess phase having range 0–2π
$\Phi(t)$ = phase function in phase-modulation process
ω = angular frequency

Paper 6516 E, first received 19th January and in revised form 2nd July 1971
Prof. Flood is Head of, and Mr. Hawksford is with, the Department of Electrical Engineering, University of Aston in Birmingham, Sumpner Building, 19, Coleshill Street, Birmingham B4 7PB, England

1 Introduction

In the study of delta modulation (ΔM) and of p.c.m., the analysis of the noise structure generated by these modulation processes becomes tedious to describe mathematically owing to quantisation errors. It is useful to consider the possibility of alternative equivalent systems which lend themselves more readily to analysis.

Usually, textbook treatments[1] compare analogue-modulation methods to digital-modulation methods. However, they stress the differences between all analogue methods on the one hand and all digital methods on the other. Thus, similarities and equivalences between certain analogue and digital methods are overlooked.

Consider single-integration delta modulation[2] with a perfect integration process, i.e. infinite memory. If the input signal

remains constant, the output-pulse pattern is a sequence of equal amplitude and area pulses, but of alternating sign. Increasing the input signal causes an increase in the rate of 1 pulses, and a reduction in the rate of 0 pulses. Decreasing the input signal causes the opposite effect. It should be noted that the total rate of 1 and 0 pulses together is a constant, being the clock rate of the modulator. The difference between the rate of 1 pulses and the rate of 0 pulses is proportional to the slope of the input signal. Thus, the modulation process can be seen to be similar to a form of discrete-pulse-phase modulation, since, in pulse-phase modulation (p.p.m.), the rate of pulses is also proportional to the slope of the modulating signal.

Similarly, the relationship between pulse-frequency modulation and delta–sigma[3] modulation ($\Delta\Sigma$M) can be argued, since the only difference between the phase-controlled and the frequency-controlled system is the position of the integration process. In pulse-frequency modulation (p.f.m.), a positive modulation signal results in an output pulse rate greater than the unmodulated pulse rate, and a negative modulating signal results in an output pulse rate less than the unmodulated pulse rate. In $\Delta\Sigma$M, the rate of transmission of 1 pulses is similarly greater than the unmodulated rate when the modulating signal is positive, and less than the unmodulated rate when the modulating signal is negative.

In p.p.m. and p.f.m., the output pulses are not constrained to fixed time slots, whereas in ΔM and $\Delta\Sigma$M they are constrained by the clock pulse generator to discrete equispaced intervals. The similarities between the analogue and the digital methods would become even greater if the p.p.m. and p.f.m. signals were quantised by controlling the timing of their output pulses to coincide with the 'clock' pulses.

The object of this paper is to show that models derived from p.p.m. and p.f.m. can generate exactly the pulse trains produced by ΔM and $\Delta\Sigma$M. These models should be of use in analysing the structure of quantisation noise generated by digital modulation processes.

2 Model for single-integration delta-modulation process

Fig. 1A shows a block schematic diagram of both a delta modulator and a delta–sigma modulator. With the switches Sw_A and Sw_B set in position 1, a delta-entry modulator is represented, and with the switches in position 2, a delta–sigma-entry modulator is represented. It can readily be appreciated that the overall operation of the two systems is identical except for the transposition of the linear integration process between the input and output stages. A scaling factor P is introduced at the sigma entry, and this, together with an integrator having unit time constant, generates a signal of slope P when the normalised sigma-entry input signal is a maximum. This corresponds to the delta modulator having an input signal of maximum slope. The value of this slope is also P, because, when the delta modulator generates a continuous train of pulses at the clock rate P, integration of this pulse train results in a signal increasing with slope P. Whichever system is employed, the operation of the central delta modem remains unaltered.

The proposed model for representing the above system is shown in Fig. 1B.

(*a*) The input signal $D(t)$ to the delta modem is used to phase-modulate a sinusoidal carrier of frequency $P/2$ to a maximum frequency deviation of $\pm P/2$ (block A, Fig. 1B).

(*b*) The phase-modulated carrier is converted to a naturally sampled p.p.m. signal (block B). The zero crossings of the phase-modulated carrier at which the phase-modulated carrier has a positive slope define the position of the leading edge of the standard pulse (block C). The standard pulses, thus generated, form the p.p.m. signal. The length of these pulses is exactly $1/P$, and the height is of magnitude 2.

A d.c. level of magnitude 1 is subtracted from the above pulse waveform as shown in Fig. 2B.

The waveform generated thus swings from -1 to $+1$.

(*c*) The p.p.m. signal is now quantised by 'time slotting'. The time axis is divided into equally spaced time slots of duration $1/P$, where P is the equivalent delta-modulator clock rate. If the leading edge of a standard pulse occurs within a time slot, a 1 pulse is generated at the end of the time slot. If no leading edge occurs, a 0 pulse is generated. Thus, a series of discrete pulses is generated. This time slotting, illustrated in Fig. 2 by traces *b*, *c* and *d*, is made equivalent to a sampling process.

The length of the standard pulse is made identical to the duration of a time slot, hence, the length is $1/P$. If leading edge of a standard pulse falls within a time slot, the standard pulse is in a '1' state at the end of that time slot. The end of the time slot defines the sampling point in an equivalent delta-modulation process: thus the p.p.m. signal is sampled by a delta sampling pulse at this instant. If a standard pulse is present, then a $+\delta$ output occurs. If no pulse is present, i.e a leading edge has not occurred within the time slot, then the sample output is $-\delta$.

Thus, an output pulse train $P_2(t)$ is generated (Fig. 1B) which corresponds to the pulse train $P_1(t)$ in Fig. 1A, where the positions of the delta samples coincide with the 'clock' positions of the delta modulator.

3 Analytical demonstration of equivalence

3.1 Assumptions made

The following assumptions are made:

(*a*) Integration processes are perfect, so that there is no leakage. This is valid, since digital integration can be performed to a high degree of accuracy.

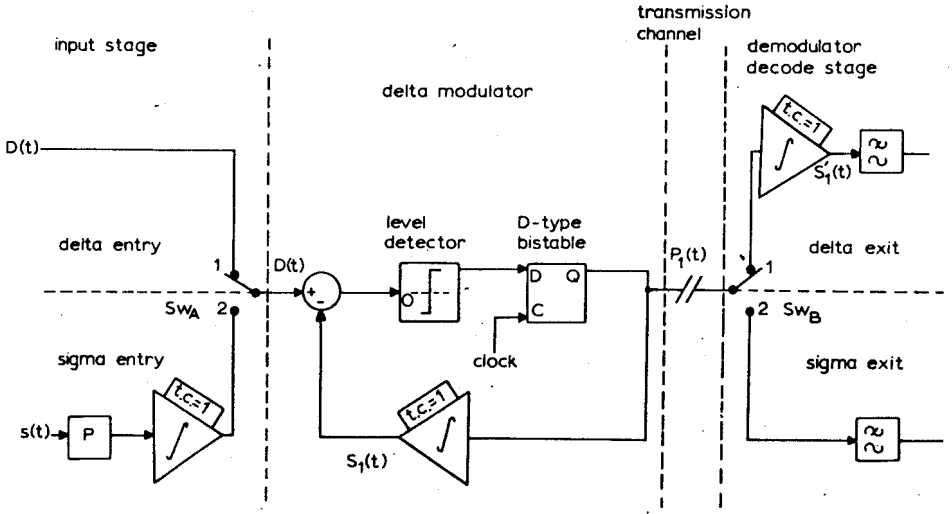


Fig. 1A

Generalised structure for ΔM and $\Delta\Sigma M$ digital process

(*b*) In the delta modulator of Fig. 1A, the comparator performance is such that

error $\geqslant 0$, comparator output high

error < 0, comparator output low.

(*e*) At no time does the input signal exceed the overload conditions.

(*f*) If there are no error pulses in the transmission channel, similar waveforms are present at the sending and receiving terminals of Figs. 1A and 1B; i.e.

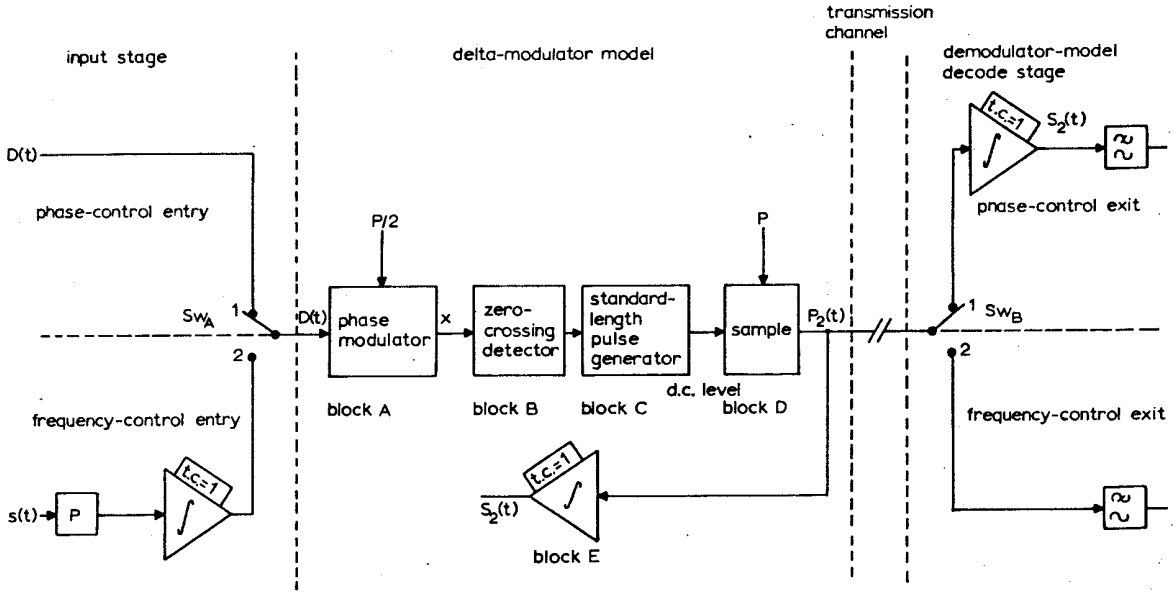


Fig. 1B
Generalised model for ΔM and ΔΣM

(*c*) In the sampling process of the model of Fig. 1B, if the positive-slope zero crossing coincides with the delta sampling function, then a 1 pulse appears at the output of the model. This is compatible with (*b*).

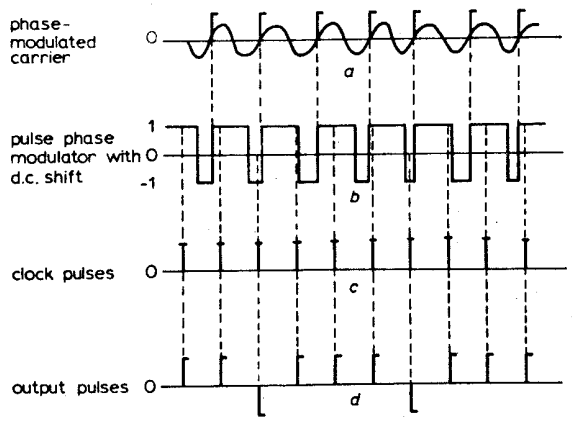


Fig. 2
Waveforms in model process
a P.P.M. signal formed from a phase-modulated carrier
b P.P.M. with standard pulse of length $1/P$ enabling time slotting of the leading edge to the nearest clock-pulse position, where p.r.f. is P
c Clock pulses
d Time-slotted version of the p.p.m. which corresponds to the output pulse pattern $P_2(t)$

(*d*) The initial conditions of the pulse-summing integrators in both the model and the delta modulator are set at $+0{\cdot}5$ during the first time slot, so that

$$S_1(t) = S_2(t) = 0{\cdot}5$$

for $0 \leqslant t < 1/p$

The initial conditions of the remaining integrators are set at zero for $t = 0$.

It is also assumed that all input modulating signals to both model and delta modulator are initially zero at $t = 0$. Thus, at commencement of processing, the systems will not be overloaded.

$$S_1(t) = S_1'(t)$$

$$S_2(t) = S_2'(t)$$

Since, in general, this does not occur, the analysis is only concerned with the delta modem. A locally decoded output prior to filtering is observed, thus allowing transmission errors to be ignored.

(*g*) The time slot associated with the Nth sample is defined over the interval

$$\frac{(N-1)}{P} < t \leqslant \frac{N}{P}$$

3.2 Delta-modem feedback network

This system is described with reference to Fig. 1A. If the pulse-rate is P, (p.p.s.), the sampling points are spaced at intervals of $(1/P)$ s.

Consider the conditions at the Nth sample. The input signal is given by

$$D(t) = D\left(\frac{N}{P}\right) \quad \ldots\ldots\ldots\ldots \quad (1)$$

Since the modulator at no time goes into slope overload, the output $S_1(t)$ must lie within $\pm$ one quantisation step of $D(t)$; i.e.

$$S_1(t) = D(t) - E(t) \quad \ldots\ldots\ldots\ldots \quad (2)$$

where $-1 < E(t) < 1$.

This assumes that the quantisation step is normalised to 1. That is, since $P_1(t)$ is composed of δ functions, and since all integrators have unit time constants, the output $S_1(t)$ changes by steps of 1.

Hence, at time t, where

$$t = \lim_{r \to 0} \left| \frac{N}{P} + r \right. \quad \ldots\ldots\ldots\ldots \quad (3)$$

$$S_1\left(\frac{N}{P}\right) = D\left(\frac{N}{P}\right) - E\left(\frac{N}{P}\right) \quad \ldots\ldots\ldots \quad (4)$$

The initial condition on the integrator is set during the first time slot ($N = 1$), so that

$$S_1\left(\frac{N}{P}\right) = 0{\cdot}5$$

Thus, $S_1(t)$ can oscillate symmetrically about zero if $D(t) = 0$. The pulse at $t = 0$ is not included in the count.

If, at time t, there have been N_{P1} positive pulses and N_{N1} negative pulses,

$$S_1\left(\frac{N}{P}\right) = N_{P1} - N_{N1} + 0{\cdot}5 \quad . \quad . \quad . \quad . \quad . \quad (5)$$

and $$N = N_{P1} + N_{N1} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

At the instant just after sampling, the error is within ± 1. The pulse pattern $P_1(t)$ is uniquely defined, since each pulse is generated by reference to the error at that sampling point.

3.3 Delta-modem model

This system is shown in Fig. 1B. Block A, the first stage of the model, is a phase modulator, modulating a carrier function of frequency $P/2$, where P is the delta modulator p.r.f.

A general expression for this carrier is

$$x = X\cos\{\pi Pt + \Phi(t)\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where $\Phi(t)$ is a linear function of the input signal $D(t)$. Put

$$\Phi(t) = KD(t) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

The signal $D(t)$ must be constrained so that the frequency deviation does not exceed $\pm P/2\ 2(\pi)$. This restriction is in accordance with the slope overload criterion, and, to keep the phase rotation of eqn. 7 positive or zero,

$$\left|\frac{d}{dt}\Phi(t)\right|_{max} = \frac{P}{2}(2\pi) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

i.e. $$\left|K\frac{d}{dt}D(t)\right|_{max} = \frac{P}{2}(2\pi) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

The maximum slope of $D(t)$ is given when the delta modulator of Fig. 1A is producing either all 1 or all 0pulses. Thus, assuming unit step height,

$$\left|\frac{d}{dt}D(t)\right|_{max} = P \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

Substituting eqn. 11 into eqn. 9 gives

$$\left|K\frac{d}{dt}D(t)\right|_{max} = K\left|\frac{d}{dt}D(t)\right|_{max}$$
$$= KP$$
$$= \frac{P}{2}(2\pi)$$

Hence, $K = \pi$.

Substituting K into eqn. 8, and, subsequently, $\Phi(t)$ into eqn. 7, gives

$$x = X\cos[\pi\{Pt + D(t)\}] \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

The phase-modulated carrier is now converted to a naturally sampled p.p.m. signal by observing the positive-going zero crossings of x which occur whenever the phase of x passes through $(-\pi/2 + 2M\pi)$; M is a positive integer, being the Mth positive zero crossing from $t = 0$.

At time t, defined by eqn. 3, let the number of complete positive rotations of the phase of x be M, and let ϕ be the excess phase. Then,

$$\phi + (2M\pi - \pi/2) = \pi(Pt + D(t)) \quad . \quad . \quad . \quad (13)$$

where $0 \leqslant \phi < 2\pi$.

At time t, N samples have occurred. Thus, substituting from eqn. 3 into eqn. 13,

$$\left(\frac{\phi}{\pi}\right) + 2M - 0{\cdot}5 = P\left(\frac{N}{P}\right) = D\left(\frac{N}{P}\right)$$

therefore,

$$(2M - 0{\cdot}5) + \left(\frac{\phi}{\pi}\right) = N + D\left(\frac{N}{P}\right) \quad . \quad . \quad . \quad (14)$$

Again, N_{P2} positive pulses have occurred, and N_{N2} negative pulses have occurred.

Also, set an initial condition of $+0{\cdot}5$ during the $N = 1$ time slot to bring the model initially into alignment with the delta modulator. Therefore,

$$S_2\left(\frac{N}{P}\right) = N_{P2} - N_{N2} + 0{\cdot}5 \quad . \quad . \quad . \quad . \quad . \quad (15)$$

But, $$N = N_{P2} + N_{N2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

Thus, eqns. 15 and 16 generate

$$S_2\left(\frac{N}{P}\right) = N_{P2} - (N - N_{P2}) + 0{\cdot}5$$

Therefore,

$$S_2\left(\frac{N}{P}\right) = 2N_{P2} - N + 0{\cdot}5 \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

In eqn. 14, M represents the number of positive rotations of the phase of x, excluding the rotation in the $N = 0$ time slot. Since the signal does not exceed slope overload, all the rotations of the phase are detected, and pass through the sampling process. Since each complete rotation is represented by a positive pulse,

$$M = N_{P2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (18)$$

Substituting from eqn. 18 into eqn. 14, and rearranging,

$$2N_{P2} - N + 0{\cdot}5 = D\left(\frac{N}{P}\right) - \left(\frac{\phi}{\pi}\right) + 1$$

Thus, substituting into eqn. 17 gives

$$S_2\left(\frac{N}{P}\right) = D\left(\frac{N}{P}\right) + \left(-\frac{\phi}{\pi} + 1\right) \quad . \quad . \quad . \quad . \quad (19)$$

The error term is $\left(-\frac{\phi}{\pi} + 1\right)$, since $0 \leqslant \phi < 2\pi$, and, hence, $0 \leqslant \frac{\phi}{\pi} < 2$ and $-1 \leqslant \left(1 - \frac{\phi}{\pi}\right) < 1$.

Thus, eqn. 19 states that the accumulated output at the Nth sample is equal to the modulating signal to an accuracy of ± 1. This is identical to the delta modulator.

Since eqn. 19 holds for all integer values of N, a unique pulse pattern $P_2(t)$ is generated.

Hence, with the appropriate choice of initial conditions for the model and delta-modulator integrator and of the initial phase of the phase-modulated carrier,

$$S_1\left(\frac{N}{P}\right) = S_2\left(\frac{N}{P}\right)$$

and, consequently,

$$P_1\left(\frac{N}{P}\right) = P_2\left(\frac{N}{P}\right)$$

for $N = 1, 2, 3$ etc. . . .

Since the modulating signal $D(t)$ and the accumulated signal in both processes are identical at each sampling point, the error signal is also identical. Consequently, both systems generate the same noise structure.

Thus, time-quantised pulse-phase modulation is in every way identical to delta modulation with a single integrator, providing correct initial conditions are observed and slope overloading does not occur.

If switches Sw_A and Sw_B in Figs. 1A and 1B are in position 2, an integrator is introduced at the input of both the delta

modulator and the model. A multiplier P is also introduced so that the magnitude of the input signal can be independent of the system parameters, i.e.

$$|s(t)|_{max} = 1$$

Thus, the required slope of $D(t)$ is controlled by the multiplier P.

In the model of Fig. 1B the input signal to the phase modulator is

$$D(t) = P\int_0^t s(t)dt \quad \ldots\ldots\ldots \quad (20)$$

The input signal thus controls the frequency of the carrier x instead of its phase. In the system of Fig. 1A, insertion of the integrator converts the system from a delta modulator to a delta–sigma modulator. Thus, delta–sigma modulation is exactly equivalent to a time-quantised pulse-frequency modulation process.

The model is thus applicable either to delta modulation or delta–sigma modulation. The only limitation is that slope overloading for ΔM and amplitude overloading for $\Delta\Sigma M$ are not represented. These effects may be included by adding suitable limiters at the inputs to the model.

The model should be useful for analysing quantising noise in ΔM systems. The quantising noise generated by the sampling process in the model causes considerable overlap of the sidebands about the sampling harmonics. This explains why the idle-channel noise spectra observed by Iwersen[4] and Laane[5] appear to be phase modulated. It should also be noted that noise is added to the baseband signal prior to quantising due to sideband distortion in the angle-modulation system.

4 Method of simulating double inegration with prediction, using model

Fig. 3A shows a double-integration delta modulator with a prediction network. Prediction is established by

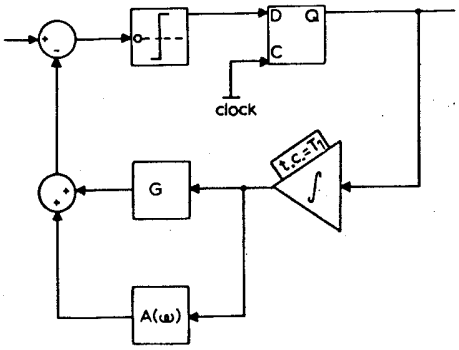


Fig. 3A
Double-integration delta modulation with prediction

summing a fraction G of the first integral to the output of the second integrator whose transfer function is $A(\omega)$. On rearranging the loop, the equivalent network of Fig. 3B is

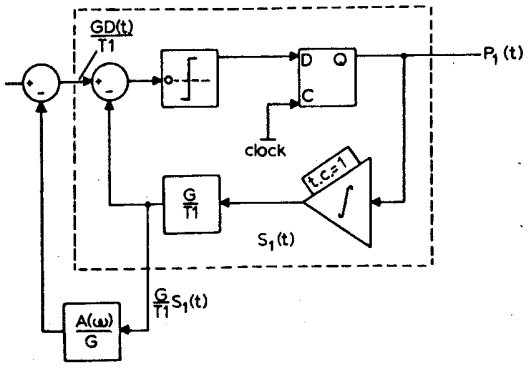


Fig. 3B
Equivalent networks to Fig. 3A

obtained. Here the first integrator time constant has been normalised, and this is compensated for by a multiplying factor $1/T_I$.

In Fig. 3B, a single-integrator delta modulator is shown. The signals $D(t)$, $P_1(t)$, $S_1(t)$ refer to the equivalent signals as shown in Fig. 1A. It is, therefore, possible to replace the feedback network, shown within the dotted lines, by the model equivalent, as illustrated in Fig. 4A. The input weighting factor T_I/G, scales the signal so that equivalence is obtained. Since $A(\omega)$ is a linear network, the network can be simplified to that shown in Fig. 4B.

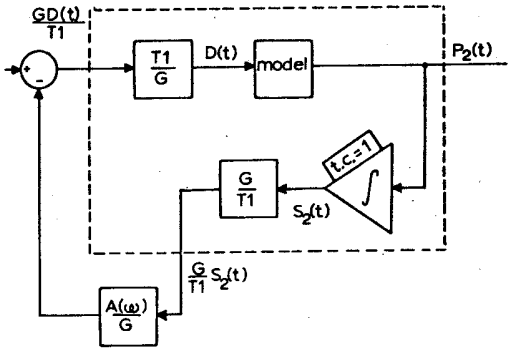


Fig. 4A
Equivalent double-integration system using model

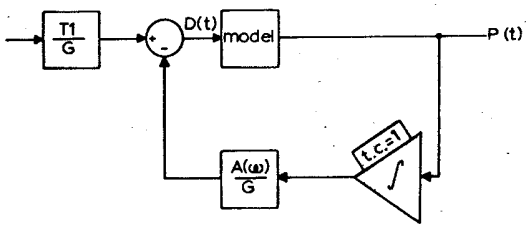


Fig. 4B
Simplified representations of system model

The application of the model to the double-integration process reveals that, whereas the single-integration system can be realised on open loop, the double integration requires feedback control. This is one of the factors contributing to the improved noise performance available with this system. Here, the pulse step is adapted by a linear network, the degree of adaption being under feedback control. It is also of interest to note that, for an equivalent system to exist, prediction is essential. In practice, this is always used, as the double-integration system is unstable without prediction.

The model should prove useful in allowing a study to be made of the nature of $A(\omega)$ necessary for the stability criteria to be met. The feedback loop also explains why the noise spectrum with double integration is more continuous than that of the single-integration modulator when encoding simple periodic functions. The model generates a wideband line spectrum for a sinusoidal input. The feedback path integrates this spectrum and feeds its many components back to the modulator input; thus, the modulator has a wideband input signal which results in a nearly continuous output spectrum. The buildup of the spectrum could alternatively be realised by an iterative open loop.

5 Extension of model to p.c.m. with uniform quantisation

It has been shown that, at each sampling point of the delta modulator output (i.e. $S_1(t)$ at intervals $1/P$), the decoded signal is quantised to within ± 1 step height of the modulating signal $D(t)$, providing that the slope of $D(t)$ does not cause an overload condition.

Hence, for delta modulation,

$$|\text{error}|_{\Delta M} < 1 \text{ (step height)}$$

In a delta modulator, the quantised output from the integrator is limited to a certain amplitude range, which is a function of signal frequency and the modulator 'clock' rate. (Assume a mean signal level of zero.) It is a characteristic of delta modulation that the amplitude range increases with decreasing signal frequency, having a theoretical infinite range at d.c. and a minimum range at the highest frequency component f_c.

In comparison, however, p.c.m. has a 'flat' signal-amplitude/frequency characteristic, handling the same amplitude range at d.c. as at the highest frequency component f_c. Thus, in order that a delta modulator may encode the same signal as a p.c.m. system, it is necessary that the amplitude range at the highest frequency f_c be greater than, or equal to, the signal range in the p.c.m. system.

In a p.c.m. system, the encoding accuracy is to within one-half of a quantisation step.

Hence, for p.c.m.,

$$|\text{error}|_{p.c.m.} \leqslant 0{\cdot}5 \text{ (quantisation step)}$$

To make the ΔM and p.c.m. processes compatible, the quantisation step of the p.c.m. is made twice that of the delta-modulator step height. Hence, in p.c.m., the quantisation step is designated a value, 2.

If at the zero sample of a delta modulator the integrator output is at an even level, it is observed that at even samples of the delta modulator the integrator output is at an even level, and on odd samples it is on odd levels. To convert the ΔM integrated output to the quantised p.a.m. signal of p.c.m., the ΔM output is sampled at the Nyquist (or greater) rate by delta pulses which are coincident with the ΔM samples. Also, by sampling at even ΔM samples, only even levels are generated for the p.a.m. signal. These coincide with the p.c.m. quantisation levels. Thus, there is a requirement that the ΔM 'clock' rate be a positive, even, integer multiple of the Nyquist sampling rate (or higher rate if used).

In practice, this method of generating p.c.m. can be used as a p.c.m. encoder.[6,7] It is arranged in the ΔM that integration is performed by an up/down counter and digital/analogue converter. At each Nyquist sample, the number stored in the counter, excepting the smallest digit, forms the p.c.m. pulse pattern, which is then transmitted over the next sample interval.

Consider the required parameters for a ΔM to perform the encoding of a signal which is compatible with a p.c.m. system, where

$$\text{quantisation step of } \Delta M = 1$$
$$\text{quantisation step of p.c.m.} = 2$$

Let

$$D(t) = D_{max} \cos(2\pi ft) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

For a unit quantisation step, the maximum value of D (to avoid slope overload) for a given pulse rate P and signal frequency is

$$D_{max} = \frac{P}{2\pi f} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

If the maximum frequency to be encoded is f_c, the total signal range of $D(t)$ is $2D_{max}$. Therefore the total signal range is

$$R_S = \frac{P}{\pi fc} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

In a p.c.m. system, the maximum-signal-amplitude/frequency response is flat.

Thus, if R_S is the signal range of the p.c.m. system, and f_c is the highest frequency component, for the delta modulator to be able to define completely the signal range without overload,

$$\frac{P}{\pi fc} \geqslant R_S$$

i.e.

$$P \geqslant \pi f_c R_S \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (24)$$

The Nyquist sampling rate N_S for the p.c.m. system is given by

$$N_S = 2f_c \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

In practice, a higher sampling rate H is used, where

$$H \geqslant N_S \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (26)$$

A further condition is imposed as previously discussed, i.e.

$$\frac{P}{H} = 2R \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (27)$$

where R is a positive integer.

If, in the p.c.m. system, a C digit code is used,

$$R_S = 2 \times 2^c \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (28)$$

Thus, the p.c.m. pulse rate P_c is given by

$$P_c = HC \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (29)$$

Since in the ΔM quantisation procedure the error is only within ± 1 at a time t defined by eqn. 3, the sampling at the rate H must be such that it occurs with a fixed delay d after the time t, as shown in Fig. 5B. Since the ΔM clock interval is $1/P$, it is clear that

$$d < 1/P$$

which corresponds to the condition in eqn. 27.

Thus, the addition of the sampling process (shown in Figs. 5A and B) on the integrator output of a ΔM followed by

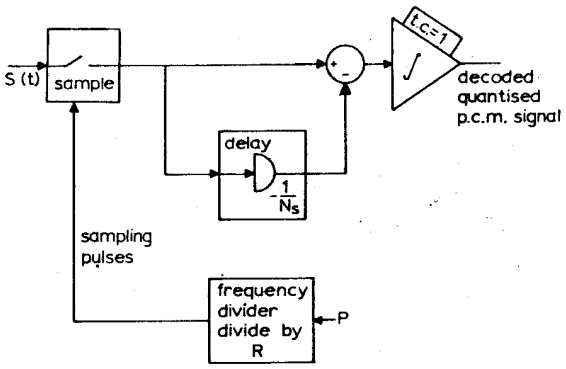


Fig. 5A
Extension of model to p.c.m.

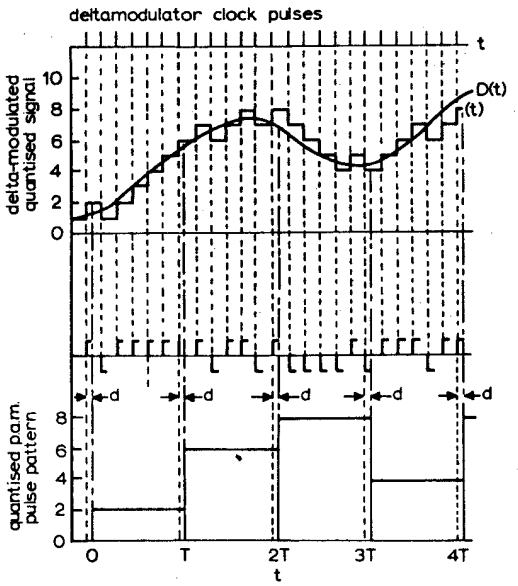


Fig. 5B
Waveforms in buildup of p.c.m. by a delta-modulator encoder

a sample and hold circuit allows the spectral structure of a p.c.m. system to be analysed. Since the model described has been shown to be equivalent to ΔM, this procedure holds equally for both the ΔM and $\Delta\Sigma M$ method.

6 Computer simulation and experimental work

The theory was checked by simulating a delta modulator and the model on a digital computer. The output signals were compared when each system received the same input signal.

The delta modulator was simulated by noting the error at each pulse and generating an output pulse accordingly. The integrated output was simulated by adding these output pulses, giving 0 pulses a weight of -1.

The model was simulated by generating the values of a phase-modulated carrier and noting the zero crossings having positive slope. The instantaneous value of the carrier was evaluated at ten equally spaced intervals during every time slot. On comparing the values in the sequence, a crossing could be detected, and was followed by the generation of a 1 pulse at the end of the time slot. The integrator output was again simulated by addition of the output pulses.

Operation of these systems was simulated for an input signal consisting of the sum of eight sinusoids of equal amplitudes but different frequencies. At each sample of the input signal, the output signals of the delta modulator and the model and the error signals were computed. This was carried out for a sequence of 999 samples, and at no time was there any difference between the outputs of the two systems.

In addition to the computer simulation, an experimental demonstration was carried out. A $\Delta\Sigma M$ system model, as shown in Fig. 1B, was constructed. A carrier of 200 kHz was frequency-modulated to a maximum deviation of ± 5 kHz. Modulation with a second carrier of 205 kHz was used to shift the central frequency down to 5 kHz, thus producing a 5 kHz carrier modulated to a maximum deviation of ± 5 kHz. This frequency-modulated carrier was squared and applied to a time quantiser using t.t.l. integrated logic circuits.

The time quantiser consisted of two counters, each counting up to four. One counter was driven by the f.m. waveform and the other by a train of 10 kHz clock pulses via an inhibit gate. When the two counters produced coincident outputs, the clock pulses were inhibited. When the f.m. waveform crossed zero with positive slope, the counter driven by it counted 'one'. This opened the inhibit gate, causing the clocked counter also to count 'one' and reach the same count. Since the frequency of the f.m. waveform did not exceed the clock pulse rate, the two counters could track, and output pulses were produced whenever the counters were not coincident, and thus a time-quantised output signal was provided.

The output waveform from this model was compared on an oscilloscope with that from a conventional delta–sigma modulator using a single integrator. When both systems had the same input signal, consisting of a sinewave of frequency within the range 50–500 Hz, they produced similar output pulse trains.

7 Conclusions

A model has been developed which describes the behaviour of single-integration delta modulation and delta–sigma modulation. It has been shown that delta modulation is equivalent to a process of time-quantised pulse-phase modulation and delta–sigma modulation is equivalent to a process of time-quantised pulse-frequency modulation. The model can also be inserted in a suitable network so that a double-integration network is obtained.

The most important result of this analysis is that the single-integration delta modulator does not require feedback. Until now, the feedback feature of delta modulation has been considered a fundamental concept of this pulse-modulation technique. As a result, it has been almost impossible to obtain an exact analysis for this process. The properties of the model, however, invalidate this previous assumption, and equally demonstrate that delta modulation can be readily analysed using the well established mathematics of Fourier analysis and sampling. Only when double integration is used is the feedback path essential. Even here, the networks are linear and consequently the same mathematics apply, although the problem is somewhat more complex.

The model does not only have theoretical application. It can provide a practical method of delta modulation. Severe difficulties are encountered in designing conventional delta modulators using clock rates of the order of 100 MHz, which are required for the transmission of television signals. A system using time-quantised angle modulation should be easier to design for these high digit rates.

The model has also been extended to represent pulse-code modulation with uniform quantisation. The p.c.m. signal is generated by sampling the ΔM quantised signal at the Nyquist rate, using a zero-order hold circuit.

8 Acknowledgments

The financial support of the British Broadcasting Corporation, in the form of a research studentship, together with the continued interest of members of the BBC research staff, is greatly appreciated.

The co-operation of J. Hodgson and D. Arrowsmith in the design and construction of the demonstration model is gratefully acknowledged.

9 References

1 PANTER, P. F.: 'Modulation, noise and spectral analysis' (McGraw-Hill, 1965)
2 DE JAGER, F.: 'Delta modulation, a method of p.c.m. transmission using the 1-unit code', *Philips Res. Rep.*, 1952, **7**, p. 442
3 INOSE, H. L., and YASUHIKO YASUDA, Y.: 'A unity bit coding method by negative feedback', *Proc. Inst. Elec. Electron. Eng.*, 1963, **51**, pp. 1524–1535
4 IWERSEN, J. E.: 'Calculated quantisation noise of single integration deltamodulation coders', *Bell Syst. Tech. J.*, 1969, **48**, pp. 2359–2389
5 LAANE, R. R.: 'Measured quantising noise spectrum for single integration deltamodulation coders', *ibid.*, 1970, **49**, pp. 191–195
6 MCKIBBIN, J., and STRADLING, J. B. M.: 'The use of deltamodulation in pulse transmission systems', *Inst. Eng. Aust. Elec. Eng. Trans.*, 1970, **6**, pp. 35–39
7 GOODMAN, D. J.: 'The application of deltamodulation to analog-to-p.c.m. encoding', *Bell Syst. Tech. J.*, 1969, **48**, pp. 321–343

Unified theory of digital modulation

M. J. Hawksford, B.Sc., Ph.D.

Indexing term: Pulse-code modulation

ABSTRACT

A theory of signal quantisation is developed from first principles that describes quantisation using analogue-modulation techniques. The nonlinearities, which are fundamental to digital source encoding, are expressed by two sampling processes, which directly lead to solutions in terms of amplitude and phase modulation. The analysis derives a related set of mathematical models that deterministically describe the input/output transfer characteristics of uniform and nonuniform p.c.m., uniform delta modulation (d.m.) and uniform d.p.c.m.

1 INTRODUCTION

Although much detailed information is now available on pulse-code modulation (p.c.m.),[1,2] delta modulation (d.m.)[3,4,5,6] and differential-pulse-code modulation (d.p.c.m.), very little of the documentation attempts to produce a general theory that describes the interrelation of particular digital modulators and their relationship to other forms of modulation. This paper is an attempt to establish a more unified approach to digital and analogue techniques.

The analysis derives a closely related set of mathematical models that deterministically describe the input/output transfer characteristics of uniform and nonuniform p.c.m., uniform delta modulation and uniform d.p.c.m. The advantage of these models is their ability to relate digital-modulation processes to analogue-modulation processes, which are generally more tractable to further analysis.

The applications of these mathematical models occur where the exact spectral distribution of the signal error is to be calculated; e.g. when digitally encoding television signals the correct choice of sampling frequency can reduce the subjective impairment of the signal errors by suitable positioning of the signal errors in the frequency domain. The models also produce a framework in which different digital-modulation techniques can be compared, thus aiding the choice of a modulator system for a particular application.

The paper develops the basic signal-distorting processes of digital modulation by using two sampling functions that, after manipulation, produce the basic model described in terms of phase modulation and amplitude modulation. The basic model is then applied to nonuniform p.c.m. and uniform differential systems, the latter being treated in two classes depending on the parameters of the modulator. Finally, it is also shown that delta modulation is a restricted case of the more general d.p.c.m., and that the same model applies to both systems.

2 SIGNAL QUANTISATION

Signal quantisation is the process of converting an analogue signal into a sequence of discrete packages or quanta and it is a fundamental process to any digital source encoder. This Section develops the theory of signal quantisation, and produces a set of equations that represent the transfer characteristics of both uniform and nonuniform quantisation.

There are two basic subdivisions of signal quantisation; time quantisation or time sampling and amplitude quantisation. The theory of time sampling is well documented; however, the basic results are summarised, because they form an interesting comparison with amplitude quantisation, and the results are incorporated in the mathematical models of p.c.m., delta modulation, and d.p.c.m. that are presented in this paper.

The basic process of time sampling is shown in Fig. 1. Sampling is achieved by the multiplication of the analogue signal $\phi(t)$ by a delta-comb function, which samples at predetermined time instants. The result of sampling is the time-sampled signal $S_T(t)$, where

$$S_T(t) = \phi(t) \sum_{R=-\infty}^{\infty} \delta\left(t - \frac{R}{f_S}\right) \quad (1)$$

Paper 7089 E, received 6th August 1973

Dr. Hawksford is with the Department of Electrical Engineering Science, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, England

If the sampling frequency f_S is time invariant, the sampling is uniform. Fig. 1 also shows the reconstructed signal $S_{Th}(t)$ that is derived from the sampled signal $S_T(t)$ by using a zero-order hold function, where

$$S_{Th}(t) = \int_{x=0}^{t} \left\{ S_T(x) - S_T\left(x - \frac{1}{f_S}\right) \right\} dx \quad (2)$$

and x is a time variable.

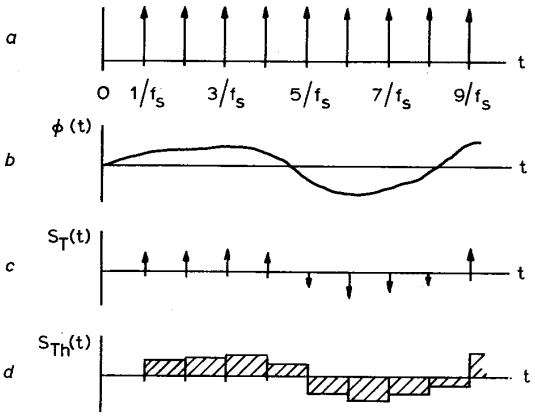


Fig. 1

The process of time quantisation using uniform sampling

a Sampling function
b Analogue signal
c Sampled signal
d Sampled signal with zero-order hold

It now will be shown that a similar sampling process can be defined for amplitude quantisation. Amplitude quantisation is the method of approximating a continuously variable signal to a signal that is only defined at exact amplitudes. This approximation process can again be analysed using sampling theory, but instead of sampling the analogue signal at predetermined time instants, the signal is sampled at predetermined amplitudes. The amplitude-sampled signal $S_a(\phi(t))$ is defined by multiplying the analogue signal $\phi(t)$ by a delta-comb function, where

$$S_a(\phi(t)) = \phi(t) \sum_{N=-\infty}^{\infty} \delta(\phi(t) - 2\pi N) \quad (3)$$

The development of the amplitude-quantised signal q(t) is shown in Fig. 2, where the sample spacing has been normalised to 2π and the delta-comb function has a weighting coefficient of 1. The quantised signal q(t) is the result of applying a hold function to $S_a(t)$, except that the function is defined along the amplitude axis θ, thus

$$q(t) = \int_{\theta=0}^{\phi(t)} \{S_a(\theta) - S_a(\theta - 2\pi)\} d\theta \quad (4)$$

Amplitude quantisation is therefore similar to time quantisation, since both require a sampling process, and both processes are defined as uniform if the sampling functions

have constant intersample spacing. However, if the amplitude-quantised signal q(t) is expressed as a function of time, the period of the hold function is dependent on the slope of $\phi(t)$. For example, the greater the rate of change of $\phi(t)$, the shorter the hold periods: thus, the signal-level transitions of the quantised signal are phase modulated by the function $\phi(t)$. The result of phase modulation of the sampling pulses, where the sampling frequency is zero when the signal $\phi(t)$ is constant, is inevitably signal impairment, and this distortion is termed quantisation distortion. This latter observation shows a disimilarity between amplitude and time quantisation: in the latter case, if the sampling frequency is chosen to be above the Nyquist rate, no aliasing distortion results in the baseband.

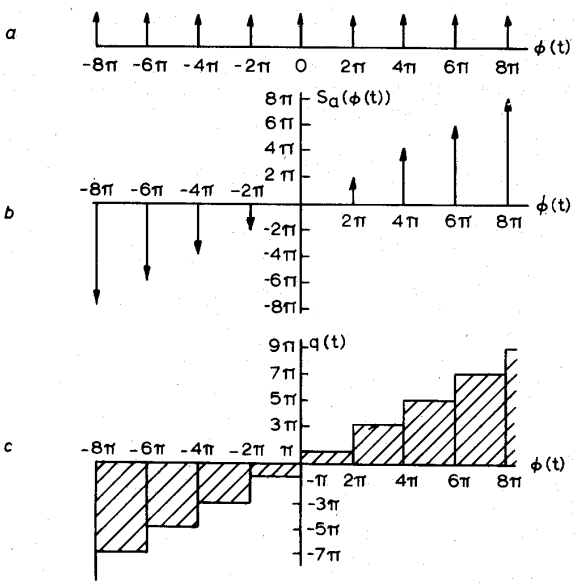


Fig. 2

The process of amplitude quantisation using amplitude sampling

a Sampling function, $\sum_{N=-\infty}^{\infty} \delta(\phi(t) - 2\pi N)$

b Sampled signal, $\phi(t) \sum_{N=-\infty}^{\infty} \delta(\phi(t) - 2\pi N)$

c Sample and held signal, $\int_{\theta=0}^{\phi(t)} \{S_a(\theta) - S_a(\theta - 2\pi)\} d\theta$

Consider now the development of the equations that describe the amplitude-quantisation distortion $\epsilon(t)$. The signal $\epsilon(t)$ is the instantaneous difference between the input signal $\phi(t)$ and the quantised signal q(t), i.e.

$$\epsilon(t) = \phi(t) - q(t) \tag{5}$$

The signal q(t) was derived by operating on the signal $S_a(\phi(t))$ by a hold function, which was defined by eqn. 4. However, substituting for $S_a(\phi(t))$ from eqn. 3, then

$$q(t) = \int_{\theta=0}^{\phi(t)} \left\{ \theta \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N) - (\theta - 2\pi) \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi(N-1)) \right\} d\theta$$

which simplifies to

$$q(t) = 2\pi \int_{\theta=0}^{\phi(t)} \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N) d\theta \tag{6}$$

Since the uniformly quantised signal is derived from a periodic-sampling process, by applying Fourier analysis a series solution for q(t) results, i.e.

$$q(t) = \phi(t) + 2 \sum_{N=1}^{\infty} \frac{1}{N} \sin \{N\phi(t)\} \tag{7}$$

The derivation and properties of eqn. 7 are given in Appendix 8.1, where it is also shown that if

$$\phi(t) = 2\pi N + \Delta\phi \tag{8}$$

then

$$2 \sum_{N=1}^{\infty} \frac{1}{N} \sin \{N\phi(t)\} = (\pi - \Delta\phi) \tag{9}$$

for $0 < \Delta\phi < 2\pi$.

It follows directly from eqns. 5, 7 and 8 that

$$\epsilon(t) = -2 \sum_{N=1}^{\infty} \frac{1}{N} \sin \{N\phi(t)\} \tag{10}$$

and

$$\epsilon(t) = -\pi + \Delta\phi \tag{11}$$

where $\epsilon(t)$ clearly has the desired signal range of $-\pi < \epsilon(t) < \pi$.

Eqns. 7 and 10 express the amplitude-quantised signal q(t) and the instantaneous quantisation distortion as a function of the analogue signal $\phi(t)$; together, they define the amplitude-quantisation transfer characteristic shown in Fig. 2. Eqn. 10 shows exactly the relationship of amplitude quantisation to phase modulation, where $\epsilon(t)$ is expressed as a discrete Fourier series of phase-modulated sinusoids.

The analysis has shown how an analogue signal is converted into a discrete signal in which the quanta spacings are uniform both along the time axis and the amplitude axis. Consider now the more general case where the amplitude quantisation is nonuniform and the amplitude sampling levels are defined in an arbitrary manner. Again it is possible to determine a Fourier series to describe the amplitude-quantisation process; however, for the series to be discrete it is necessary for the quantisation process to be periodic when it is expressed as a function of $\phi(t)$. This periodicity infers that the expressions describing the quantisation-transfer function must only be applied over a restricted range of $\phi(t)$, i.e. $|\phi(t)| < \pi$, a condition that can reasonably be imposed in practice, and also that the transfer characteristic must have odd symmetry about $\phi(t) = 0$.

Fig. 3 illustrates the nonuniform-quantisation process, and shows the periodicity of the quantised signal, for a repetition frequency along the $\phi(t)$ axis of $1/2\pi$. Only five signal comparison levels are shown over the signal range $|\phi(t)| < \pi$, but this can be extended to any arbitrary number, the only restriction being that all comparison levels are in the range $-\pi$ to π. The comparison levels shown in Fig. 3 are defined by the coefficient set $\{\ldots, -a_2, -a_1, a_1, a_2, \ldots\}$ and the approximated output levels produced by the quantiser are defined by the set $\{\ldots, -b_3, -b_2, -b_1, -b_0, b_0, b_1, b_2, b_3, \ldots\}$.

Initially, the quantisation-transfer characteristic can be expressed in terms of the unit-step function $H[\phi(t) - x]$ over the input-signal dynamic range $|\phi(t)| < \pi$, i.e.

$$q_n(t) = b_0 H[\phi(t)] + \sum_{r=1}^{n} \{(b_r - b_{r-1}) H[\phi(t) - a_r]\} - b_0 H[-\phi(t)] - \sum_{r=1}^{n} \{(b_r - b_{r-1}) H[-\phi(t) - a_r]\} \tag{12}$$

where $q_n(t)$ is the quantised signal produced by a nonuniform selection of {a} and {b} coefficients.

The analysis of the nonuniform-quantisation characteristic can be simplified by decomposing the characteristic into a set of uniform-quantisation characteristics, which are then recombined by superposition. Fig. 4 shows how the rth term of eqn. 12 can be expressed as the sum of two shifted uniform-transfer characteristics $q_{1r}(t)$ and $q_{2r}(t)$. The quantisation-transfer characteristic $q_{r1}(t)$, which is shifted by a_r, follows from modifying eqn. 7 whereby

$$q_{r1}(t) = \frac{(b_r - b_{r-1})}{2\pi} \left(\{\phi(t) - a_r\} + 2 \sum_{N=1}^{\infty} \frac{1}{N} \sin [N\{\phi(t) - a_r\}] \right) \tag{13}$$

and similarly,

$$q_{r2} = \frac{(b_r - b_{r-1})}{2\pi}\left(\{\phi(t) + a_r\} + 2\sum_{N=1}^{\infty}\frac{1}{N}\sin[N\{\phi(t) + a_r\}]\right) \quad (14)$$

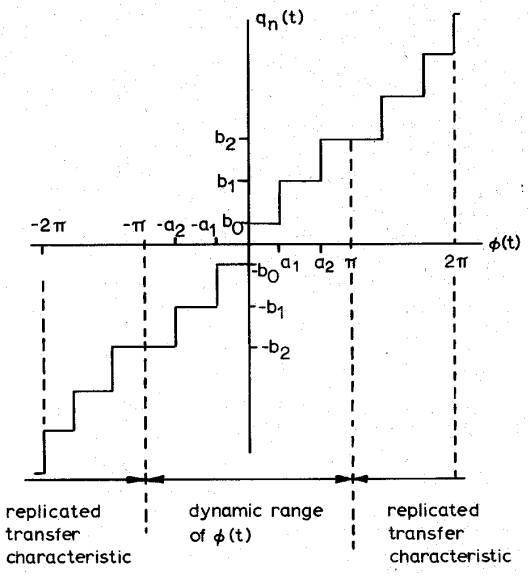


Fig. 3

Nonuniform amplitude-quantisation transfer characteristic

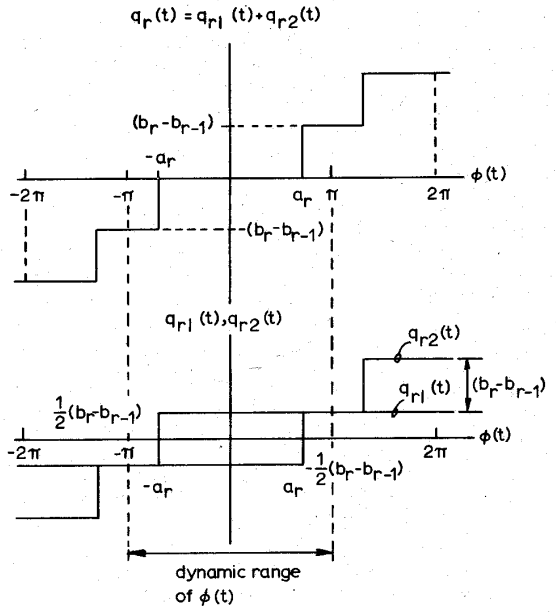


Fig. 4

Decomposition of nonuniform quantisation into two superimposed shifted uniform quantisation characteristics

Hence, $q_r(t)$ is calculated by superposition, where

$$q_r(t) = q_{r1}(t) + q_{r2}(t) \quad (15)$$

and from eqns. 13 and 14 this becomes after simplification

$$q_r(t) = \frac{(b_r - b_{r-1})}{\pi}\left[\phi(t) + 2\sum_{N=1}^{\infty}\frac{1}{N}\sin\{N\phi(t)\}\cos(Na_r)\right] \quad (16)$$

From Fig. 4, it follows that for $|\phi(t)| < \pi$

$$q_r(t) = (b_r - b_{r-1})(H[\phi(t) - a_r] - H[-\phi(t) - a_r]) \quad (17)$$

Hence, comparing eqn. 17 with eqn. 12,

$$q_n(t) = q_0(t) + \sum_{r=1}^{n} q_r(t)$$

which, after substitution of $q_r(t)$ from eqn. 16 and noting that

$$q_0(t) = \frac{b_0}{\pi}\left[\phi(t) + 2\sum_{N=1}^{\infty}\frac{1}{N}\sin\{N\phi(t)\}\right]$$

simplifies to

$$q_n(t) = \frac{b_n}{\pi}\left[\phi(t) + 2\sum_{N=1}^{\infty}\frac{1}{N}\sin\{N\phi(t)\}\left\{\frac{b_0}{b_n} + \sum_{r=1}^{n}\frac{(b_r - b_{r-1})}{b_n}\cos(Na_r)\right\}\right] \quad (18)$$

The nonuniform-quantisation distortion $\epsilon_n(t)$ is therefore $\frac{b_n}{\pi}\phi(t) - q_n(t)$, i.e.

$$\epsilon_n(t) = -\frac{2}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\{N\phi(t)\}\left\{b_0 + \sum_{r=1}^{n}(b_r - b_{r-1})\cos(Na_r)\right\} \quad (19)$$

However, when applying eqns. 18 and 19, the following restrictions must be observed, i.e.

$$|\phi(t)| < \pi \quad (20)$$

and

$$0 < \{a\} < \pi \quad (21)$$

Eqns. 18 and 19 describe the nonuniform-amplitude quantisation characteristic and quantisation distortion, respectively. Comparing these results with eqns. 7 and 10, they can be seen to be of similar form except that the harmonics of the series are amplitude modulated by a set of coefficients, which in turn are dependent upon the coefficient sets $\{a\}$ and $\{b\}$, thus the phase-modulation functions remain.

3 ANALOGUE MODELLING OF OPEN-LOOP P.C.M.

P.C.M. is a digital method of signal transmission that transmits the signal quanta derived from simultaneous amplitude and time quantisation by digital encoding. Since, however, the objective is to produce a mathematical model that represents the distortion due to quantisation and not transmission errors, it is not necessary to consider the details of digital coding.

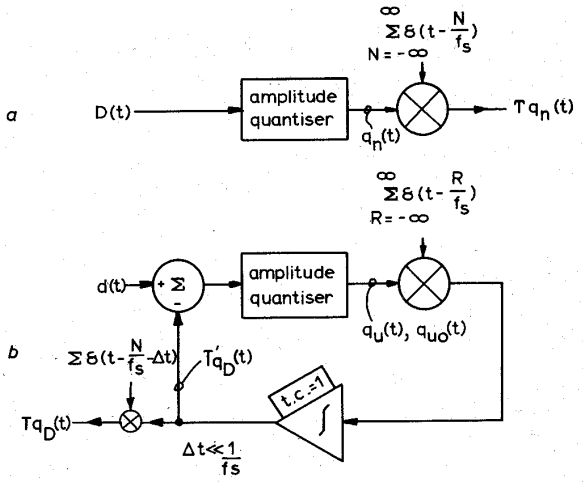


Fig. 5

Basic models for generating the quantisation distortion due to source encoding

a P.C.M. model
b D.P.C.M. model

Fig. 5a illustrates a signal processor that simulates exactly the signal distortion introduced by source encoding when using p.c.m. The model is formed by a cascaded-amplitude quantiser and time quantiser, and is applicable to both non-unform- and uniform-amplitude quantisation: the time quantisation, however, is assumed to be uniform. The output signal $Tq_n(t)$ of the model is presented as a p.a.m. signal, where

$$Tq_n(t) = q_n(t) \sum_{R=-\infty}^{\infty} \delta\left(t - \frac{R}{f_s}\right) \tag{22}$$

Thus, considering the general case of nonuniform-amplitude quantisation by substitution of $q_n(t)$ from eqn. 18:

$$Tq_n(t) = \frac{b_n}{\pi}\left[\phi(t) + 2\sum_{N=1}^{\infty}\frac{1}{N}\sin\{N\phi(t)\}\left\{\frac{b_0}{b_n} + \sum_{r=1}^{n}\frac{(b_r - b_{r-1})}{b_n}\cos(Na_r)\right\}\right] \times \sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) \tag{23}$$

If the input signal $\phi(t)$ is expressed in terms of D(t), where

$$D(t) = \frac{b_n}{\pi}\phi(t) \tag{24}$$

from eqn. 23,

$$Tq_n(t) = D(t)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) + \left[\frac{2}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left\{N\pi\frac{D(t)}{b_n}\right\}\left\{b_0 + \sum_{r=1}^{n}(b_r - b_{r-1})\cos(Na_r)\right\}\right]\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) \tag{25}$$

where

$$|D(t)| < b_n \tag{26}$$

Similarly, by scaling eqn. 7, the uniform amplitude-quantised signal $q_u(t)$ is derived, where

$$q_u(t) = D(t) + \frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left\{2\pi N\frac{D(t)}{\Delta}\right\} \tag{27}$$

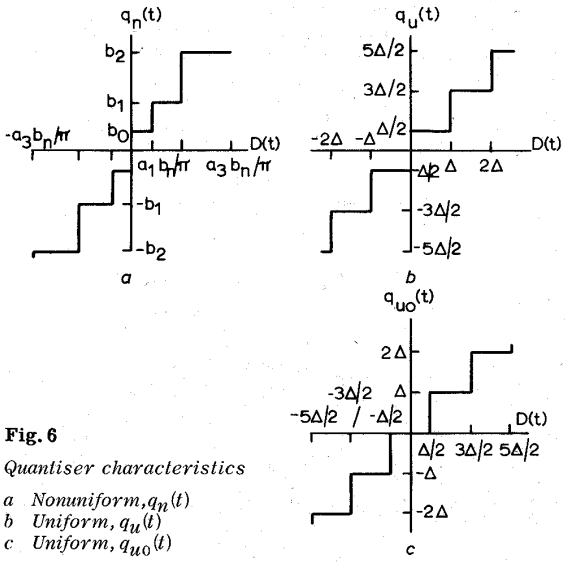


Fig. 6

Quantiser characteristics

a Nonuniform, $q_n(t)$
b Uniform, $q_u(t)$
c Uniform, $q_{uo}(t)$

and Δ is the signal difference between quantised signal levels. Thus, the time- and amplitude-quantised signal $Tq_u(t)$ follows directly as:

$$Tq_u(t) = D(t)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) + \left[\frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left\{2\pi N\frac{D(t)}{\Delta}\right\}\right]\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) \tag{28}$$

Eqns. 25 and 28 represent exactly the overall quantisation process of p.c.m., where the nonuniform- and uniform-quantisation characteristics are shown in Fig. 6a and 6b and the sampling frequency is f_s. The equations are expressed such that the second term in each equation represents the quantisation distortion. Thus, a deterministic model of open-loop p.c.m. is formed, which allows the exact analysis of the source-encoding functions to be performed.

4 ANALOGUE MODELLING OF UNIFORM-DIFFERENTIAL ENCODING SYSTEMS

Differential encoders are characterised by the digital signal that conveys coded information, this information representing the quantised signal differences between adjacent samples: thus it follows that the reconstructed output signal is formed by summing all the past sample differences. Since the adjacent sample-difference signals are quantised, in general the differential encoder requires a feedback loop to prevent the accumulation of signal errors. However, for the special case of uniform-amplitude quantisation an open-loop model can be derived, which simulates exactly the signal quantisation of the encoder.

Differential encoders are usually subdivided into two groups: delta modulation and d.p.c.m. However, the analysis presented in this paper treats the two systems identically, and it demonstrates that the model derived correctly describes both systems. The analysis also shows the relationship of uniform-differential encoding to uniform p.c.m., thus producing a unified theory.

There are two forms of uniform-amplitude quantisation that can be used to quantise the intersample differences: these are illustrated in Fig. 6b and Fig. 6c. The quantisation transfer characteristic is placed in the forward path of the differential encoder as shown in Fig. 5b.

The differential encoder that uses the transfer characteristic of Fig. 6c produces the simpler model. To yield minimum quantisation-error power, the quantised signal $Tq_D(t)$ is controlled so that it only assumes integer multiples of the quantum spacing Δ. The characteristic of Fig. 6c also determines that the quantised intersample differences are integer multiples of Δ. It can therefore be concluded that the open-loop model of Fig. 5a simulates exactly the signal quantisation of the differential encoder, when using the transfer characteristic of Fig. 6c.

The transfer characteristic of Fig. 6c can be analytically derived by shifting the uniform characteristic of Fig. 6b, which is described by eqn. 27, along a 45° axis. The amplitude-quantised signal $q_{uo}(t)$ then becomes

$$q_{uo}(t) = d(t) + \frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{d(t)}{\Delta} + 0{\cdot}5\right\}\right]$$

where d(t) is the analogue input to the differential encoder. Thus by sampling

$$Tq_D(t) = d(t)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) + \left(\frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{d(t)}{\Delta} + 0{\cdot}5\right\}\right]\right)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_s}\right) \tag{29}$$

Thus, the model can be seen to be equivalent to the open-loop p.c.m. system described by eqn. 28, except for the shifted quantisation transfer characteristic.

Consider now the feedback encoder that uses the transfer characteristic illustrated in Fig. 6b. It follows directly from this characteristic that the quantised-output signal $q_u(t)$ is of the form $(M\Delta + \Delta/2)$, where M is an integer. Since it is the quantiser output that determines the inter-sample changes of the integrated-output signal $Tq_D(t)$, the following observations can be made:

(a) After an odd number of samples, the signal $Tq_D(t)$ can have changed only by an odd integer multiple of half-quantum spaces $\Delta/2$.

(b) Similarly, after an even number of samples, the signal $Tq_D(t)$ can have changed only by an even integer multiple of half-quantum spaces.

These conditions force the integrated output of the differential encoder to have an oscillatory component: e.g. in the idling state, the integrated output signal $Tq_D(t)$ will produce a square wave of one half the sampling frequency with a peak-to-peak amplitude of $\Delta/2$. The mathematical model describing differential encoding therefore has to be modified to accommodate this oscillatory mode.

The open-loop mathematical models produced so far have been based on phase modulation. This phase modulation process has the property that, when the input signal is constant, the output signal has zero frequency. Consequently, the output frequency of the phase-modulation process swings either positive or negative, the frequency of swing being directly proportional to the slope of the input signal.

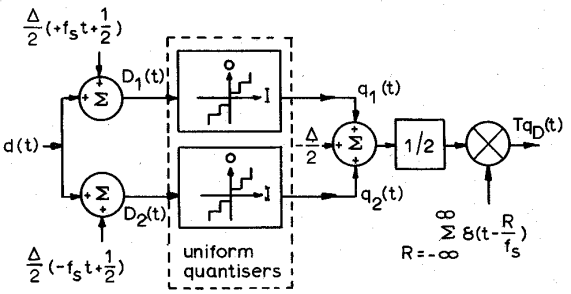


Fig. 7

Model equivalent of d.p.c.m. encoder when using the transfer characteristic of Fig. 6b

The mathematical model, which simulates the differential encoder using the transfer characteristic of Fig. 6b, uses a modified phase-modulation process, the centre frequency of which is shifted from zero to one half the sampling frequency, when considering the fundamental sinusoid in the series. The system shown in Fig. 7 requires the use of two phase modulators with effective frequency shifts of $f_s/2$ and $-f_s/2$, respectively. As before, phase modulation is the result of amplitude quantisation, the frequency shift being produced by the addition of a ramp signal to the input of each quantiser. The input signals to each quantiser are $D_1(t)$ and $D_2(t)$, respectively, where

$$D_1(t) = \Delta\left\{\frac{f_s}{2}t + \frac{1}{4} + \frac{d(t)}{\Delta}\right\} \tag{30}$$

$$D_2(t) = \Delta\left\{\frac{-f_s}{2}t + \frac{1}{4} + \frac{d(t)}{\Delta}\right\} \tag{31}$$

Since the idling state of the differential encoder is oscillatory, the integrated-output signal $Tq_D(t)$ is offset by $-\Delta/4$. This results in a signal that oscillates symmetrically about the quantiser-comparison levels. This offset is compensated for by the addition of a constant signal $\Delta/4$ to the input of each quantiser, as shown by eqns. 30 and 31.

The validity of the model illustrated in Fig. 7 is demonstrated in Appendix 8.2, where the relationship to pulse-length modulation is given as a corollary. The analysis derives the final quantised signal, which is given by

$$Tq_D(t) = d(t)\sum_{R=-\infty}^{\infty}\delta\left(t-\frac{R}{f_s}\right) + \left(\frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{d(t)}{\Delta}+\frac{1}{4}\right\}\right]\cos(N\pi f_s t)\right)\sum_{R=-\infty}^{\infty}\delta\left(t-\frac{R}{f_s}\right) \tag{32}$$

Eqn. 32 shows that the phase-modulated sinusoids of the amplitude-quantisation function are each modulated onto a cosinusoidal carrier, the frequency of which is proportional to the harmonic number of the sinusoid.

Eqn. 32 represents exactly the quantisation distortion introduced by differential encoding when using the transfer characteristic of Fig. 6b. The model is applicable to both delta modulation and d.p.c.m. However, for the model to represent delta modulation the slope of the input signal d(t) must be limited. A d.m. encoder is only capable of transmitting one out of two incremental changes between samples, i.e. $\Delta/2$ or $-\Delta/2$. Thus, for delta modulation, the input signal must never have a slope that exceeds $\Delta f_s/2$, i.e.

$$\left|\frac{d}{dt}\{d(t)\}\right| \leq \frac{\Delta}{2}f_s \tag{33}$$

Applying this condition to $D_1(t)$ and $D_2(t)$, as defined by eqns. 30 and 31, then

$$0 \leq \frac{d}{dt}D_1(t) \leq \Delta f_s \tag{34}$$

$$0 \geq \frac{d}{dt}D_2(t) \geq -\Delta f_s \tag{35}$$

The conditions stated by the inequalities of expr. 34 and 35 infer that for delta modulation the quantised signals $q_1(t)$ and $q_2(t)$ (Fig. 7) always produce a rising-staircase waveform and a falling-staircase waveform for a step change of $\Delta/2$ and $-\Delta/2$, respectively. However, for d.p.c.m. the slope of the staircase waveforms can change sign, and the step change is of a general form $(M\Delta + \Delta/2)$.

The analysis, therefore, leads to the mathematical model for delta modulation represented in terms of phase modulation, which was the subject of a previous paper.[7] The uniform d.p.c.m. model includes this previous model with the extension that the phase modulator is allowed to produce a negative output frequency.

5 CONCLUSIONS

The analysis has derived a closely related set of mathematical models, which exactly simulate the quantisation distortion generated by uniform and nonuniform p.c.m., uniform delta modulation and uniform d.p.c.m.

Equations were derived to describe the amplitude quantisation as a harmonically related series of phase-modulated sinusoids, and the time quantisation as a harmonically related series of amplitude-modulated sinusoids. These equations then formed the basis of a deterministic model of uniform and nonuniform p.c.m., the latter requiring the series of phase-modulated sinusoids to be weighted by a set of constant coefficients.

Uniform d.p.c.m., which included delta modulation, was subdivided into two groups depending on whether the quantised output was nonoscillatory or oscillatory. When the output was nonoscillatory, the uniform p.c.m. model could be applied directly. However, the oscillatory d.p.c.m. required each harmonic of the phase-modulated sinusoids to be amplitude modulated by a sinusoid, the frequency of which was proportional to the harmonic number of the phase-modulated sinusoid. This latter model described both d.p.c.m. and delta modulation, except that delta modulation required the input signal to be slope limited.

The analysis therefore described a unified theory that care-

fully linked together various digital-modulation techniques with analogue-modulation techniques, and thus produced a framework that illustrated how specific digital modulators distorted the input signal during source encoding.

6 ACKNOWLEDGMENTS

The constructive discussion and continued interest of Prof. K. W. Cattermole in the work presented in this paper is gratefully acknowledged.

7 REFERENCES

1 CATTERMOLE, K. W.: 'Principles of pulse code modulation' (Iliffe, 1969)
2 PANTER, P. F.: 'Modulation, noise and spectral analysis' (McGraw-Hill, 1965)
3 De JAGER, F.: 'Delta modulation, a method of p.c.m. transmission using the 1-unit code', Philips Res. Rep., 1952, **7**, p. 442
4 IWERSEN, J. E.: 'Calculated quantisation noise of single integration deltamodulation coders', Bell Syst. Tech. J., 1969, **48**, pp. 2395-2389
5 LAANE, R. R.: 'Measured quantisation noise spectrum for single integration deltamodulation coders', ibid., 1970, **49**, pp. 191-195
6 GOODMAN, D. J.: 'The application of deltamodulation to analogue-to-p.c.m. encoding', ibid., 1969, **48**, pp. 321-343
7 FLOOD, J. E. and HAWKSFORD, M. J.: 'Exact model for delta-modulation processes', Proc. IEE, 1971, **118**, (9), pp. 1155-1161

8 APPENDIXES

8.1 Derivation and properties of the quantisation error function $\epsilon(t)$

This Appendix develops the analytical expression for the uniformly quantised signal q(t) when expressed as a function of $\phi(t)$.

The uniform-quantisation transfer function is illustrated in Fig. 2c. Since the transfer characteristic is the result of a uniform-sampling process, it is possible to derive a discrete series to express q(t) as a function of $\phi(t)$. The derivation is as follows:

It was shown in Section 2 (eqn. 6) that the quantisation signal was related to $\phi(t)$ by the integral

$$q(t) = 2\pi \int_{\theta=0}^{\phi(t)} \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N)\,d\theta$$

Let

$$Z(\theta) = \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N) \tag{36}$$

where $Z(\theta)$ is a periodic function of θ and forms a delta-pulse sequence with a fundamental frequency of $1/2\pi$. The function $Z(\theta)$ can be expressed as a discrete Fourier series

$$Z(\theta) = \sum_{N=-\infty}^{\infty} a_N \exp\{j(N\theta)\} \tag{37}$$

The coefficients a_N are calculated from the discrete Fourier integral

$$a_N = \frac{1}{\psi} \int_{-\psi/2}^{\psi/2} Z(\theta) \exp\left\{-j\left(2\pi\frac{\theta}{\psi}\right)N\right\} d\theta \tag{38}$$

where ψ is the fundamental period 2π. Thus substituting for $Z(\theta)$ from eqn. 36 and putting $\psi = 2\pi$, then

$$a_N = \frac{1}{2\pi} \int_{\theta=-\pi}^{\pi} \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N) \exp(-jN\theta)\,d\theta$$

$$= \frac{1}{2\pi}$$

Thus $Z(\theta)$ can be expressed as a Fourier series, where

$$Z(\theta) = \frac{1}{2\pi} \sum_{N=-\infty}^{\infty} \exp\{j(N\theta)\}$$

Hence, from eqns. 6 and 36,

$$q(t) = \int_{\theta=0}^{\phi(t)} \sum_{N=-\infty}^{\infty} \exp\{j(N\theta)\}\,d\theta \tag{39}$$

Eqn. 39 is identical to eqn. 6 and describes the quantisation transfer characteristic. However, by rearrangement,

$$q(t) = \int_{\theta=0}^{\phi(t)} \left\{1 + 2 \sum_{N=1}^{\infty} \cos(N\theta)\right\} d\theta$$

and finally integrating

$$q(t) = \left\{\theta + 2 \sum_{N=1}^{\infty} \frac{1}{N} \sin(N\theta)\right\}_{\theta=0}^{\phi(t)}$$

leads to eqn 7; thus

$$q(t) = \phi(t) + 2 \sum_{N=1}^{\infty} \frac{1}{N} \sin\{N\phi(t)\}$$

Now, substituting for q(t) from eqn. 6, it follows that

$$-2 \sum_{N=1}^{\infty} \frac{1}{N} \sin\{N\phi(t)\} = \phi(t) - 2\pi \int_{0}^{\phi(t)} \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N)\,d\theta \tag{40}$$

If the delta pulses of eqn. 6 are symmetrically defined, i.e.

$$2\pi \int_{2\pi M}^{(2\pi M+\Delta\phi)} \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N)\,d\theta \equiv \pi$$

and

$$2\pi \int_{2\pi M}^{(2\pi M-\Delta\phi)} \sum_{N=-\infty}^{\infty} \delta(\theta - 2\pi N)\,d\theta \equiv -\pi,$$

where M is an integer. Putting $\phi(t) = 2\pi M + \Delta\phi$ (eqn. 8) where $0 < \Delta\phi < 2\pi$, eqn. 9 follows:

$$2 \sum_{N=1}^{\infty} \frac{1}{N} \sin\{N\phi(t)\} = (\pi - \Delta\phi)$$

The function expressed in eqn. 9 is shown in Fig. 8, where it is seen to represent $-\epsilon(t)$ expressed as a function of $\phi(t)$.

The signal $\epsilon(t)$ is periodic in $\phi(t)$ with a repetition period of 2π. An alternative derivation would have been to commence with $\epsilon(t)$ expressed in terms of $\phi(t)$, and to derive the discrete Fourier series of this signal, the result of which is expressed by eqn. 10. This latter analysis forms a cross check on the proof presented in this Appendix.

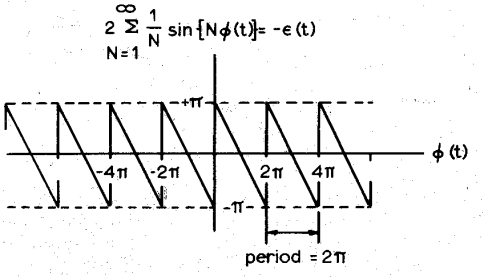


Fig. 8
$-\epsilon(t)$ expressed as a function of $\phi(t)$

8.2 Verification of the mathematical model describing differential encoding

This Appendix verifies the exactness of the mathematical model (Fig. 7) that simulates the differential-encoding system using the transfer characteristic of Fig. 6b.

The quantisation transfer characteristic, which is used in both signal paths of the model of Fig. 7, is deterministically described by eqn. 27. The input signals to the quantisers are $D_1(t)$ and $D_2(t)$, which are defined by eqns. 30 and 31. Thus, applying eqn. 27, the amplitude-quantised signals $q_1(t)$ and $q_2(t)$ are given by

$$q_1(t) = \Delta\left\{\frac{f_S}{2}t + \frac{1}{4} + \frac{d(t)}{\Delta}\right\} + \frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{f_S}{2}t + \frac{1}{4} + \frac{d(t)}{\Delta}\right\}\right] \quad (41)$$

and

$$q_2(t) = \Delta\left\{-\frac{f_S}{2}t + \frac{1}{4} + \frac{d(t)}{\Delta}\right\} + \frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{-\frac{f_S}{2}t + \frac{1}{4} + \frac{d(t)}{\Delta}\right\}\right] \quad (42)$$

The final quantised output signal $Tq_D(t)$ is defined in Fig. 7 as

$$Tq_D(t) = 0{\cdot}5\left[q_1(t) + q_2(t) - \frac{\Delta}{2}\right]\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) \quad (43)$$

Thus, substituting for $q_1(t)$ and $q_2(t)$ from eqns. 41 and 42, and after simplification,

$$Tq_D(t) = d(t)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) + \left(\frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{d(t)}{\Delta} + \frac{1}{4}\right\}\right]\right.$$
$$\left.\cos(N\pi f_S t)\right)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) \quad (44)$$

which also has the equivalent form

$$Tq_D(t) = d(t)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) + \left(\frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{d(t)}{\Delta} - \frac{1}{4}\right\}\right]\right.$$
$$\left.\cos\{N\pi(f_S t+1)\}\right)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) \quad (45)$$

The expressions for $Tq_D(t)$ as defined by eqns. 44 and 45 are identical, but their presentation allows the quantisation error on even and odd samples to be more readily observed. Let the sampling instants be defined as

$$t = \frac{R}{f_S} \quad (46)$$

where R is an integer. Hence, substituting for t from eqn. 46 in eqns. 44 and 45, for even samples, i.e. R even,

$$Tq_D\left(\frac{R}{f_S}\right) = d\left(\frac{R}{f_S}\right) + \frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{1}{\Delta}d\left(\frac{R}{f_S}\right) + \frac{1}{4}\right\}\right] \quad (47)$$

and for odd samples, i.e. R odd,

$$Tq_D\left(\frac{R}{f_S}\right) = d\left(\frac{R}{f_S}\right) + \frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{1}{\Delta}d\left(\frac{R}{f_S}\right) - \frac{1}{4}\right\}\right] \quad (48)$$

The phase difference between the sinusoids of eqns. 47 and 48, i.e. between adjacent samples, results in a displacement of the quantisation transfer characteristic along a 45° axis; the shift is illustrated in Fig. 9.

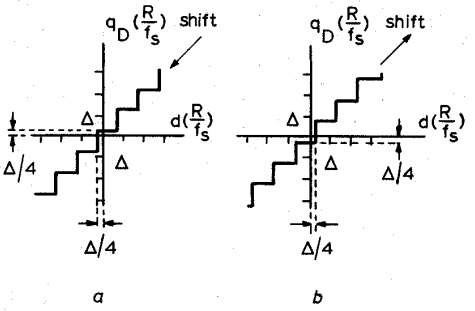


Fig. 9

Shifted quantisation transfer characteristic represented by eqns. 47 and 48

a R even
b R odd

The shifting of the transfer characteristic by $\pm\Delta/4$ produces the oscillatory output of the d.p.c.m. encoder as well as maintaining the encoding accuracy within the range $\pm\Delta/2$. The mathematical model of Fig. 7 therefore simulates exactly the performance of the differential encoder that uses the transfer characteristic of Fig. 6b.

8.3 Corollary

It is possible, by observing eqns. 44 and 45, to produce an equivalent model that shows the relationship between differential encoding and pulse-length modulating (p.l.m.). The quantised signal $Tq_D(t)$, as defined by eqns. 44 and 45, can be expressed in terms of a function $\sigma(t)$, where

$$Tq_D(t) = d(t)\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) + \left(\frac{\Delta}{\pi}\sum_{N=1}^{\infty}\frac{1}{N}\sin\left[2\pi N\left\{\frac{d(t)}{\Delta} + \sigma(t)\right\}\right]\right)$$
$$\sum_{R=-\infty}^{\infty}\delta\left(t - \frac{R}{f_S}\right) \quad (49)$$

where in general $\sigma(t)$ can be any function provided that for even R

$$\sigma\left(\frac{R}{f_S}\right) = \frac{1}{4}$$

and for odd R

$$\sigma\left(\frac{R}{f_S}\right) = -\frac{1}{4}$$

If $\sigma(t)$ is a triangular wave, the p.l.m. equivalent system is produced as shown in Fig. 10. It should, however, be noted that the p.l.m. in this system does not exhibit overload, but generates an output signal that changes in discrete steps.

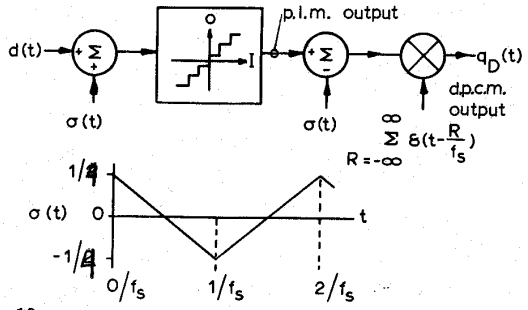


Fig. 10

Equivalent model showing relationship to p.l.m.

PAPERS

Time-Quantized Frequency Modulation, Time-Domain Dither, Dispersive Codes, and Parametrically Controlled Noise Shaping in SDM*

M. O. J. HAWKSFORD, *AES Fellow*

Centre for Audio Research and Engineering, University of Essex, UK CO4 3SQ

Time-domain quantization (TDQ) and noise shaping applied to linear frequency modulation (LFM) offers an alternative although unconventional means of generating uniformly sampled 1-bit code with characteristics similar to that generated by feedback sigma–delta modulation (SDM). Fundamental insight into the SDM process emerges by exploiting the relationship between uniform quantization and phase modulation. Linearity and output noise spectra are benchmarked against linear pulse-code modulation (LPCM) and error spectra derived by comparing 1-bit SDM against an identical feedback loop but without quantization. Sony FF class SDM is discussed and an example is shown to achieve almost noiseless performance from 0 Hz to 30 kHz by incorporating parametrically determined noise shaping stabilized by step-back in time closed-loop control.

0 INTRODUCTION

The principal objective of this paper is to strengthen the linkage between sigma–delta modulation (SDM) and a coding technique based on linear frequency modulation (LFM) and time-domain quantization (TDQ), here designated LFM–SDM, which can produce uniformly sampled binary pulse sequences with properties similar to SDM. The author became aware of this relationship during the period of 1968 to 1971, when the equivalence of single-integrator delta modulation with time-domain quantized phase modulation and of first-order SDM with time-domain quantized frequency modulation was established [1], [2]. Although at that time SDM was directed mainly toward encoding video signals [3], with only passing reference to audio, applications to high-performance audio conversion, especially in the context of Direct Stream Digital (DSD), would have to wait a further 25 years for Super Audio CD (SACD) [4] to emerge.

The motivation for the present study stems from a renewed debate on SDM, which is attempting to understand in much greater detail its fundamental operation. In particular there is the critical question of linearity and how the performance of binary SDM compares with multilevel SDM with uniform quantization. It is shown that by exploiting the new class of feedback encoder [5], using parametrically controlled noise shaping with step-back in time stabilization [6], extraordinarily low levels of in-band quantization noise can be achieved, countering much of the negative criticisms on linearity proffered, for example, by Lipshitz and Vanderkooy [7]–[9]. Also the Trellis algorithm [10] has gained prominence where refinements reported by Janssen and Reefman [11] achieve efficient and stable coding solutions that can circumvent the need for dither. Fundamentally there is no reason why the Trellis algorithm with appropriate cost function descriptors should not yield results that are superior to those of the parametric coder as Trellis represents a comprehensive search of coding solutions. However, here the SDM examples presented use only parametric SDM.

Although it is well established that a uniform amplitude quantizer can be rendered linear with the inclusion of amplitude domain dither with a triangular probability density function (TPDF) [12], there is a major difficulty in applying this theory to feedback SDM [13] because the inclusion of TPDF dither together with any finite level of input signal dictates more than two quantization levels. Nevertheless, high-order SDM encoders using suboptimum dither levels according to linear pulse-code modulation (LPCM) theory demonstrably achieve low correlated distortion due in part to extreme chaotic-like loop activity [14]. Consequently part of the rationale for exploring the LFM–SDM model is to gain an alternative perspective to dither, especially as it is shown that the process can be applied in the time domain rather than the amplitude domain. Reefman and Janssen [15] have since published evidence in support of this approach.

*Manuscript received 2003 January 13; revised 2004 April 13.

In all forms of SDM the audio signal is embedded directly within the uniformly sampled 1-bit output pattern[1] such that after jitter minimization and output pulse standardization [16], only a low-pass filter is required for signal recovery. However, other types of converter technology also exhibit this property, including pulse-amplitude modulation within an LPCM kernel augmented by noise shaping, oversampling, and decorrelation [17], where linearity can be guaranteed within the arithmetic resolution of this class of system when incorporating uniform quantization and TPDF dither.

The paper commences by formulating mathematical descriptions of time-domain sampling and uniform amplitude quantization based on harmonic series of amplitude- and phase-modulated carriers. These ideas are then linked to the process of LFM and adapted to form virtually distortion-free pulse-density modulation (PDM). This theoretical framework, when extended by the introduction of TDQ, establishes a gateway for unifying analog and digital modulation methods from which the complete LFM–SDM model can be derived. Later sections develop further the concept of TDQ and time-domain dither, where Matlab[2] simulations reveal that binary code similar to that formed by first- and second-order SDM can be generated. At this juncture it should be emphasized that the purpose of studying the LFM–SDM model is not specifically to forge optimized SDM code but to explore the process of time-domain dither as a means of decorrelating quantization distortion as well as building an alternative and more holistic philosophical approach to SDM. Finally the paper applies the parametric coder with step-back in time stabilization to generate an extreme example of noise shaping and thus demonstrate that SDM coding artifacts within the audio band can be engineered to virtually zero.

0.1 Abbreviations

DSD	Direct Stream Digital
LFM	Linear frequency modulation
LFM–SDM	Sigma–delta modulation model derived using LFM and time-domain quantization
LPCM	Linear pulse-code modulation
NSTF	Noise-shaping transfer function
NTSI	Natural time sampling instants
PDM	Pulse-density modulation
PSZC	Positive-slope zero crossing
SACD	Super Audio Compact Disc
SDM	Sigma–delta modulation
TDQ	Time-domain quantization
TPDF	Triangular probability density function

1 SDM MODELING USING LINEAR FREQUENCY MODULATION

If the behavior of SDM is observed with respect to the frequency of occurrence of 1's and 0's in the output bit stream, it is evident that their average rate of occurrence is proportional to the amplitude of the input signal level. Also, because the combined rate of occurrence of both 1 and 0 pulses is fixed, then if the number of 1 pulses increases, the number of 0 pulses must fall. Thus in practice only the sequence of either 1 pulses or 0 pulses need be determined. This suggests intuitively that SDM output pulses are related to individual cycles in single-carrier frequency modulation and in particular to LFM, where by definition the instantaneous carrier frequency changes as a linear function of the modulating signal. However, fundamental differences between LFM and SDM exist, especially as the output pulses of the latter are constrained to discrete time instants. Hence if a high degree of equivalence is to be established, quantization must be included in the process.

The discussion commences by formalizing the fundamental processes of time sampling and amplitude quantization in terms of analog modulation. Following conventional sampling theory [18], amplitude modulation describes time sampling (a quantization process directed along the time axis) whereas phase modulation, less well known, describes amplitude quantization (a process directed along the signal amplitude axis). Thus together they establish unification between analog and digital modulation processes. The core mathematical descriptors of TDQ and amplitude quantization can be described succinctly by two harmonic series.

1.1 Time Sampling → Amplitude Modulation

Consider a discrete and uniformly spaced (directed along the time axis t) series of Dirac sampling pulses $\mathrm{sam}(t)$ of sampling frequency f_s Hz, which by Fourier series analysis can be expressed as

$$\mathrm{sam}(t) = 1 + 2\sum_{r=1}^{\infty} \cos(2\pi r f_s t).$$

If $\phi(t)$ is the input signal then the resulting sampled signal $s(t)$ is given as $s(t) = \phi(t)\mathrm{sam}(t)$, whereby

$$s(t) = \phi(t) + 2\sum_{r=1}^{\infty} \phi(t)\cos(2\pi r f_s t). \quad (1)$$

This equation reveals the well-known result of time-domain sampling being described by a harmonic series of amplitude-modulated carriers, each having identical and symmetrical sidebands, where, provided the terms $\phi(t)$ and $2\sum_{r=1}^{\infty} \phi(t)\cos(2\pi r f_s t)$ are orthogonal, linear filtering can retrieve $\phi(t)$ without distortion.

1.2 Amplitude Quantization → Phase Modulation

A similar analysis can be applied to the process of amplitude quantization. Consider a uniform quantizer with quantum δ and input signal $\psi(t)$. The quantization error $\varepsilon(t)$ is equal to the difference between the quantized output $q[\psi(t)]$ and the input $\psi(t)$, where

$$\varepsilon(t) = q[\psi(t)] - \psi(t).$$

[1] Although appropriate pulse-forming techniques are required to lower the sensitivity to clock jitter and pulse amplitude noise.

[2] Matlab is a trade name of MathWorks Inc.

$\varepsilon(t)$ forms a periodic sawtooth waveform when expressed as a function of $\psi(t)$, which by Fourier series analysis results in a series expansion where,

$$q[\psi(t)] = \psi(t) + \frac{\delta}{\pi}\sum_{r=1}^{\infty}\left\{\frac{1}{r}\sin\left[2\pi r\frac{\psi(t)}{\delta}\right]\right\} \quad (2a)$$

that is, the quantization error $\varepsilon(t)$ is given by

$$\varepsilon(t) = \frac{\delta}{\pi}\sum_{r=1}^{\infty}\left\{\frac{1}{r}\sin\left[2\pi r\frac{\psi(t)}{\delta}\right]\right\} \quad (2b)$$

To demonstrate the veracity of this analysis, Fig. 1 illustrates a synthesis of the uniform quantization characteristic based on Equation (2a) for fundamental only and for summations taken over 3, 10, and 1000 harmonics in the series. It is evident that as the number of harmonics used in the summation increases, the accuracy to which the ideal quantization characteristic is matched improves progressively. Eq. (2b) reveals that the quantization error $\varepsilon(t)$ resulting from uniform amplitude quantization forms a harmonic series of phase-modulated sinusoids, as the argument of each sine term is a linear function of input $\psi(t)$. Consequently amplitude modulation maps to time sampling as phase modulation maps to amplitude quantization, where together they encapsulate the process of LPCM.

If the input signal $\psi(t)$ in Eq. (2b) is assumed to have zero mean, then the carrier frequency of each sinusoidal term in the series has a center frequency of 0 Hz, whereas with SDM the idle channel bit pattern has an average frequency of $f_{sdm}/2$ Hz, with f_{sdm} being the SDM sampling frequency in hertz. Imagine therefore a linear ramp function with constant slope λ when expressed as a function of time superimposed on the input signal $\psi(t)$. The argument $\theta(t)$ of the fundamental sinusoidal term ($r = 1$) in the series in Eq. (2b) is then

$$\theta(t) = 2\pi\frac{[\psi(t) + \lambda t]}{\delta} = 2\pi\frac{\psi(t)}{\delta} + \pi f_{sdm}t. \quad (3)$$

Eq. (3) shows that for the idle channel case, $\psi(t) = 0$, the fundamental term in Eq. (2b) has a constant frequency $f_{sdm}/2$, requiring $\lambda = 0.5\delta f_{sdm}$. This condition forces $\varepsilon(t)$ to have a repetition frequency set precisely to one-half the sampling rate of SDM, which matches the basic . . .010101. . . idle pattern.

The development of the LFM–SDM model exploits at its core the frequency modulation behavior of $\varepsilon(t)$ when

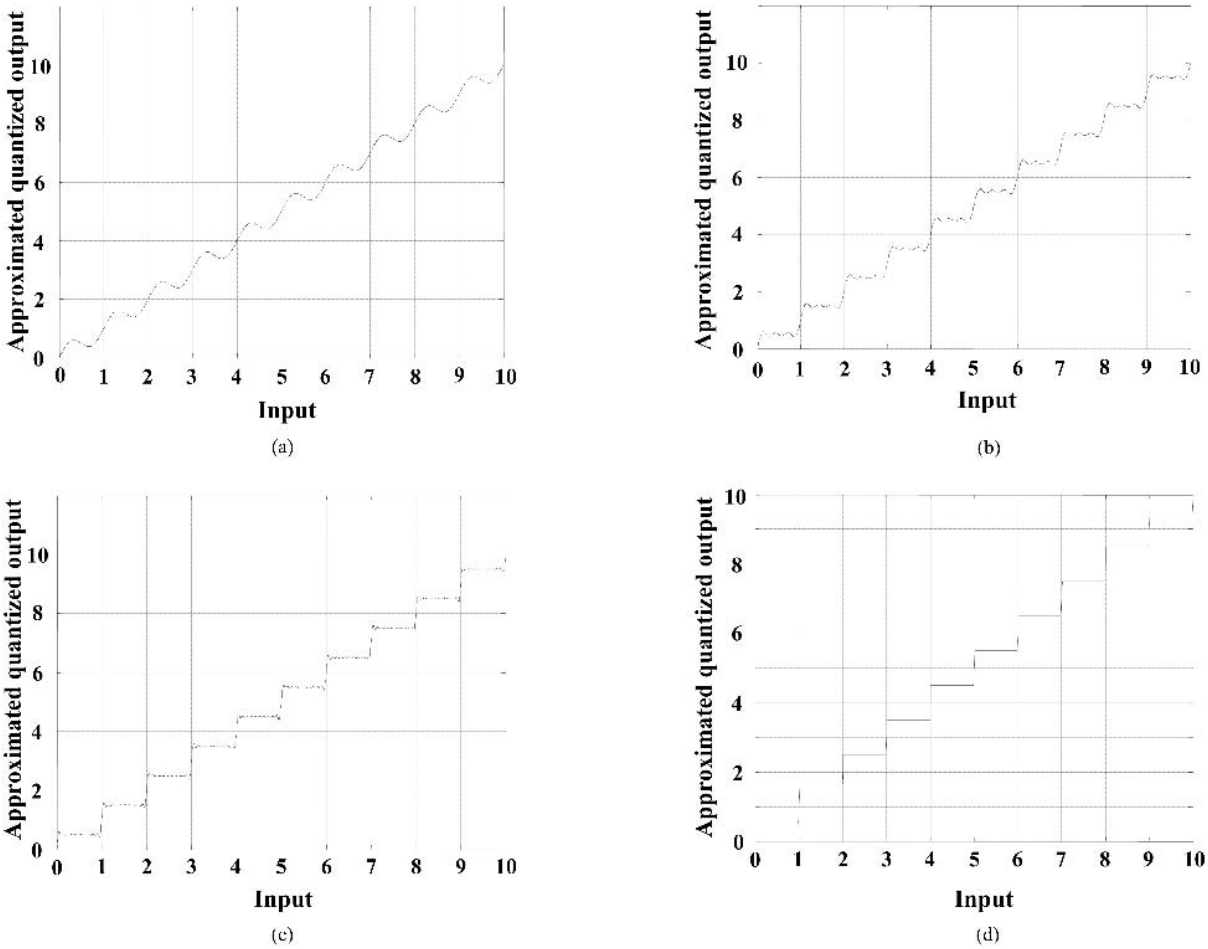


Fig. 1. Synthesized uniform quantization characteristic (output versus input). (a) Fundamental only. (b) Summation: 3 harmonics. (c) Summation: 10 harmonics. (d) Summation: 1000 harmonics.

the input signal $\psi(t)$ is applied together with the ramp signal to shift the idle channel frequency from dc to $f_{sdm}/2$ Hz. The higher terms ($r > 1$) in the series expansion in Eq. (2b) also become separated out in frequency. Because $\varepsilon(t)$ is a phase-modulated sawtooth wave, each step transition in this waveform can be used to define a reference point in time. If only the fundamental component ($r = 1$) of the series in Eq. (2b) is considered, then these reference points correspond to the positive-slope zero crossings (PSZCs) of the fundamental sine-wave term with center frequency $f_{sdm}/2$ Hz. Because in terms of uniform amplitude quantization each PSZC represents natural time sampling instants (NTSIs),[3] where the quantization error is zero, the NTSIs are endowed with properties that are central to achieving virtually zero distortion with a PDM waveform. Effectively the model requires that at each NTSI a reference pulse be generated with time-invariant characteristics having constant amplitude and width. The resulting pulse sequence then forms PDM. It follows that when the PDM sequence is low-pass filtered, the modulating input signal can be recovered. However, at this stage when making comparisons with SDM it is important to note that there is yet no quantization of the signal in time. The NTSI pulse locations are free to associate with their optimum time coordinates. Effectively what has been generated is SDM code, where the pulses, although constrained in number per second, are unconstrained in time. It is shown in Section 2 that NTSIs require quantization in time in order to conform to the SDM code. However, if the question is mooted, "What are the optimum locations in time for SDM pulses to be located?" then the LFM model in terms of NTSI defines formally those instants. This observation is a fundamental characteristic of the SDM process, and Matlab simulations presented later in this section support the analysis.

The discussion has shown so far that an amplitude quantizer when modeled by Fourier in terms of amplitude can be described by phase modulation. To simplify the formulation of the model, the modulation process is next abstracted purely in terms of equivalent LFM, and the core elements of this scheme are shown in Fig. 2. LFM has the characteristic that the instantaneous output frequency is proportional to the input signal amplitude where, mirroring earlier discussion, the LFM center frequency (that is, the LFM output frequency corresponding to zero modulating input amplitude) is set equal to $f_{sdm}/2$ Hz. This condition is necessary to ensure that the number of NTSIs formed by LFM matches on average those produced in SDM, where the NTSIs are associated with PSZCs of the LFM output signal and are defined without recourse to feedback. To complete the mapping from LFM to SDM code Fig. 2 illustrates basic time quantization. Where for each NTSI falling within an SDM sample window of $1/f_{sdm}$ a 1 pulse is placed at the SDM sampling instant. Otherwise if no NTSI occurs within the sample window, then a 0 pulse is allocated. The process of TDQ and the application of time-domain noise shaping are developed in Section 2 and types of dither statistics and their effect on linearity are discussed in Section 3.

1.3 Analysis and Performance of LFM in Context of LFM-SDM Model

This subsection presents a formal definition and study of LFM and shows by spectral analysis that PDM based on the computation of NTSIs, the precursor to a full LFM–SDM model, has extremely low intrinsic nonlinearity. Consider a phase-modulated sine wave $s_{lfm}(t)$ of amplitude

[3]The author conceptualizes these instants as "magic moments in time."

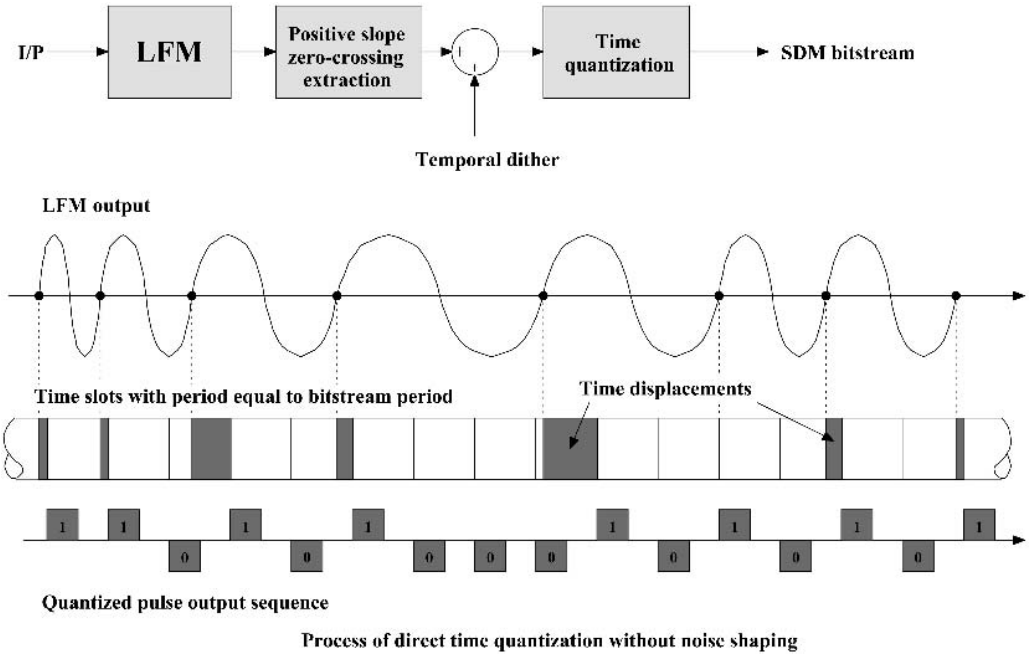


Fig. 2. Time-domain quantized linear frequency modulation model of SDM.

A with instantaneous phase $\theta(t)$ expressed as a function of time,

$$s_{\text{lfm}}(t) = A \sin[\theta(t)]. \quad (4)$$

The instantaneous angular frequency $2\pi f(t)$, where the frequency $f(t)$ is itself a function of time, is defined as the rate of change of phase $\theta(t)$ with respect to time, with $\theta(t)$ being the argument of the frequency-modulated sine wave, that is,

$$f(t) = \frac{1}{2\pi}\frac{\mathrm{d}\theta(t)}{\mathrm{d}t}.$$

Integration gives

$$\theta(t) = 2\pi \int_{t=0}^{t} f(t)\,\mathrm{d}t.$$

Substituting for $\theta(t)$ in Eq. (4),

$$s_{\text{lfm}}(t) = A \sin\left[2\pi \int_{t=0}^{t} f(t)\,\mathrm{d}t \right]. \quad (5)$$

For LFM the instantaneous frequency $f(t)$ is directly proportional to the input signal $y(t)$. Hence if $Y_{\max}$ represents the maximum amplitude of $y(t)$ and $f_{\text{sdm}}/2$ Hz the half SDM sampling frequency when $y(t) = 0$, then $f(t)$ is defined as

$$f(t) = \frac{f_{\text{sdm}}}{2}\left[1 + \frac{y(t)}{Y_{\max}} \right]. \quad (6)$$

Hence by substituting for $f(t)$ from Eq. (6), $s_{\text{lfm}}(t)$ follows from Eq. (5),

$$s_{\text{lfm}}(t) = A \sin\left\{ \pi \int_{t=0}^{t} f_{\text{sdm}}\left[1 + \frac{y(t)}{Y_{\max}} \right] \mathrm{d}t \right\}$$

giving

$$s_{\text{lfm}}(t) = A \sin\left\{ \pi f_{\text{sdm}} t + \frac{\pi f_{\text{sdm}}}{Y_{\max}} \int_{t=0}^{t} y(t)\,\mathrm{d}t \right\}. \quad (7)$$

Eq. (7) formally describes the process of LFM, where the argument $\theta(t)$ of the sine wave is a function of the integral of the input signal $y(t)$, allowing the instantaneous frequency to be linearly proportional to the input signal amplitude. Inspection shows that if $y(t) = Y_{\max}$, then the instantaneous output frequency is f_{sdm} Hz, whereas if $y(t) = -Y_{\max}$, the frequency becomes zero, and if $y(t) = 0$, the frequency is $f_{\text{sdm}}/2$ Hz. $Y_{\max}$ therefore is the maximum input signal amplitude to maintain the output frequency in the range of 0 to f_{sdm} Hz. This also coincides with the ultimate overload limits of SDM. However, it is desirable to prevent the instantaneous carrier frequency from falling within the audio frequency band and creating excessive distortion. Thus in practice the peak amplitude of $y(t)$ should be less than the maximum. If the minimum and maximum carrier frequencies are specified, then the limits on $y(t)$ can be calculated using Eq. (6).

In order to derive a PDM signal using the LFM model it is necessary to calculate the time coordinates t_r of NTSI. Since NTSIs occur at PSZCs of the frequency-modulated sine wave, solutions for t_r occur where $\theta(t_r) = 2\pi r$ for integer r. Hence from Eq. (7) the time coordinates t_r of NTSI can be calculated from

$$t_r + \frac{1}{Y_{\max}} \int_{t=0}^{t_r} y(t)\,\mathrm{d}t = \frac{2r}{f_{\text{sdm}}}. \quad (8)$$

In the domain of discrete signal processing it is a difficult task to seek exact solutions to Eq. (6) for a general input signal $y(t)$. However, the problem can be visualized with reference to Fig. 3, where the instantaneous phase $\theta(t)$ is plotted as a function of time. In this graph each NTSI is associated uniquely with each integer r provided the slope of the graph is greater than zero, a condition guaranteed if $-Y_{\max} < y(t) < Y_{\max}$, which limits the LFM frequency range to 0–f_{sdm} Hz.

In this study a four-stage approach was adopted to solve Eq. (8) in terms of t_r:

1) The LFM signal is computed at discrete and uniform time instants at a rate of σf_{sdm} Hz, where the oversampling factor is $\sigma > 1$ in order to give a finer time resolution (for example, $\sigma = 8$).

2) The computed samples are then scanned to identify negative-to-positive polarity transitions in the LFM output to identify the presence of a PSZC.

3) Linear interpolation is performed on either side of any detected PSZC to achieve a closer approximation to each t_r value.

4) Finally an iterative error-driven procedure is used to converge toward the optimum solution for each t_r.

A Matlab subroutine (see Appendix 1) was written based on these four stages to calculate each value of t_r. The signal processing steps in the program are summarized here:

- Convert the oversampled sinusoidal LFM signal to a square wave using a squaring (sign) function.
- Compute a difference signal between adjacent samples that is nonzero only when a zero-crossing transition occurs between samples.
- Interrogating the sign of the nonzero intersample difference signal, only the PSZCs are selected.
- Apply the Matlab *sort* function to the difference sequence to eliminate all zero differences and determine a vector $zr(1{:}L)$ that contains only progressively sequenced sample numbers, describing samples just prior to each PSZC.
- Knowing the time coordinate $zr(r)$ of the nearest sample that just precedes the rth PSZC in the sequence and $zr(r) + 1$, the coordinate of the sample that just follows the PSZC, two samples[4] of the LFM output signal $\widetilde{s_{\text{lfm}}}[zr(r)]$ and $\widetilde{s_{\text{lfm}}}[zr(r) + 1]$ taken on each side of the PSZC are calculated and a more accurate time location t_{ra} is estimated by linear interpolation,

[4]The notation $\widetilde{s_{\text{lfm}}}[j]$ refers here to a sample data domain, where j is an integer sample number.

$$t_{ra} = zr(r) - \frac{\widetilde{s_{\mathrm{lfm}}[zr(r)]}}{\widetilde{s_{\mathrm{lfm}}[zr(r)+1]} + \widetilde{s_{\mathrm{lfm}}[zr(r)]}}. \quad (9)$$

- The approximate NTSI value t_{ra} is then substituted back into the LFM signal expression in Eq. (7) to form an error signal zerror (see Appendix 1) defined as zerror $= s_{\mathrm{lfm}}(t_{ra}) - s_{\mathrm{lfm}}(t_r) = s_{\mathrm{lfm}}(t_{ra})$ since $s_{\mathrm{lfm}}(t_r) \equiv 0$, that is, if $t_{ra} = t_r$, then zerror $= 0$. A new estimate for t_{ra} is then made using the recursion expression

$$t_{ra} \Rightarrow t_{ra} + 1.5 \text{ zerror}. \quad (10)$$

This iterative procedure is repeated (typically 100 times) until zerror has converged down to an acceptable level. The coefficient of 1.5 in Eq. (10) was selected by experiment based on convergence and final accuracy.

To estimate the attainable accuracy of this procedure, the Fourier transform of a sequence of N unit impulses located at each NTSI was calculated over a period of the input signal, where the input signal frequency and the LFM center frequency were selected for a harmonic ratio. However, because NTSIs are asynchronous there are no uniform sampled data for the fast Fourier transform, so a direct spectral calculation based on the delay elements was performed following earlier reported practice [19], [20], where the PDM spectrum out(f) is given by the summation,

$$\mathrm{out}(f) = \sum_{r=1}^{N} \mathrm{e}^{-\mathrm{j}2\pi f t_r}. \quad (11)$$

To illustrate the signal distortion resulting from the LFM process, four example spectra are displayed in Fig. 4, each derived by using an input of two equal-amplitude sinusoids of frequencies 19 and 20 kHz, where the normalized amplitudes for the four simulations are 0.30, 0.25, 0.1, and 0.01, respectively (1 representing 100% modulation depth). In all examples the LFM center frequency is set to 1.4112 MHz, exactly one-half the SACD data rate. Typical final error values in computing NTSI locations are estimated to lie in the range of 10^{-7} to 10^{-12}, the higher error levels occurring only for extreme modulation levels. The result shown in Fig. 4(a) is the only spectrum to reveal audio-band distortion where the two sinusoidal inputs combine to give a modulation depth of about 0.6. By reducing the peak amplitude to 0.5 significantly lower levels of LFM sideband spectral spillage are revealed within the audio band. Also, in Fig. 4(d), where the LFM sidebands are narrow, spectral replication is evident at harmonics of the sampling frequency of 2.8224 MHz, which is consistent with the SACD sampling rate.

These results demonstrate a spectral sideband spread resulting from LFM where, except for extreme modulation, there is negligible spillage within the signal band. As such the computational precision and significance of NTSIs as true natural sampling data points is validated. It is considered a critical factor, especially as NTSIs relate to instants of zero distortion when using the equivalent uniform quantizer model and linear ramp input to form LFM. Consequently it can be concluded that virtually all distortion results from the process of TDQ, and there is negligible signal degradation when converting from input signal to NTSI sequence.

2 TDQ IN LFM–SDM MODELING

Section 1 considered the formation of PDM based upon techniques of LFM. In order to create discrete SDM code,

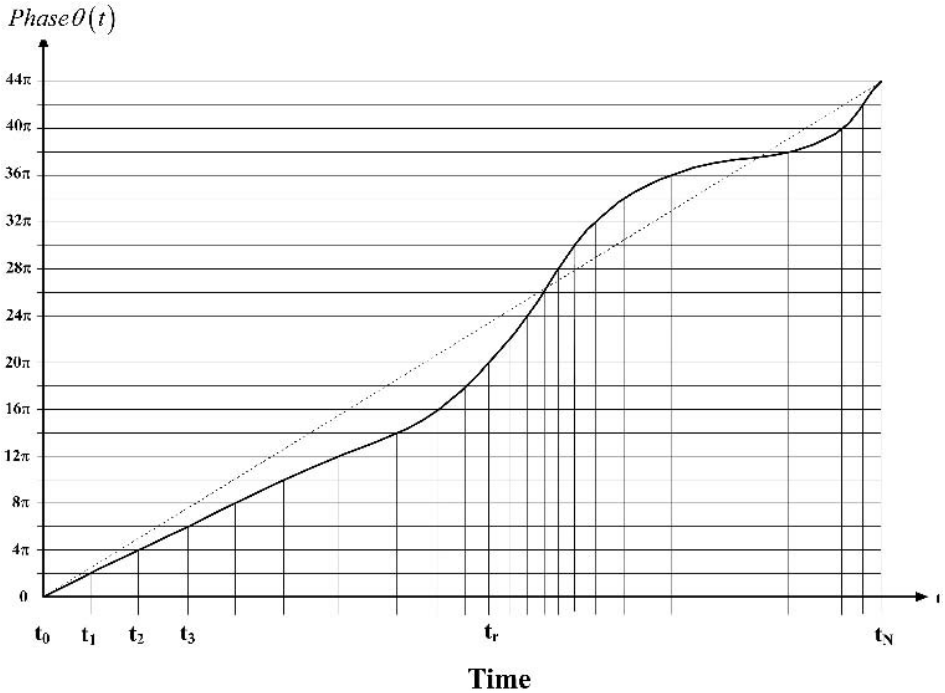


Fig. 3. Seeking natural sampling solutions for NTSI $\{t_r\}$.

the NTSIs derived from the LFM model must be redistributed in time to form a uniformly sampled data sequence. This process is designated TDQ and represents the process by which digitization is achieved. As with any quantization process, rounding errors are produced that form quantization distortion. Consequently residual correlation of distortion and signal is a critical issue, and it is in this domain that the effectiveness of dither strategies should be judged. An obvious difference between this class of quantization and amplitude quantization is that the process is directed in time rather than amplitude. It inevitably takes on characteristics akin to timing jitter [21] with an error spectrum that rises with frequency. In LPCM perturbing a single sample amplitude is a memoryless op-

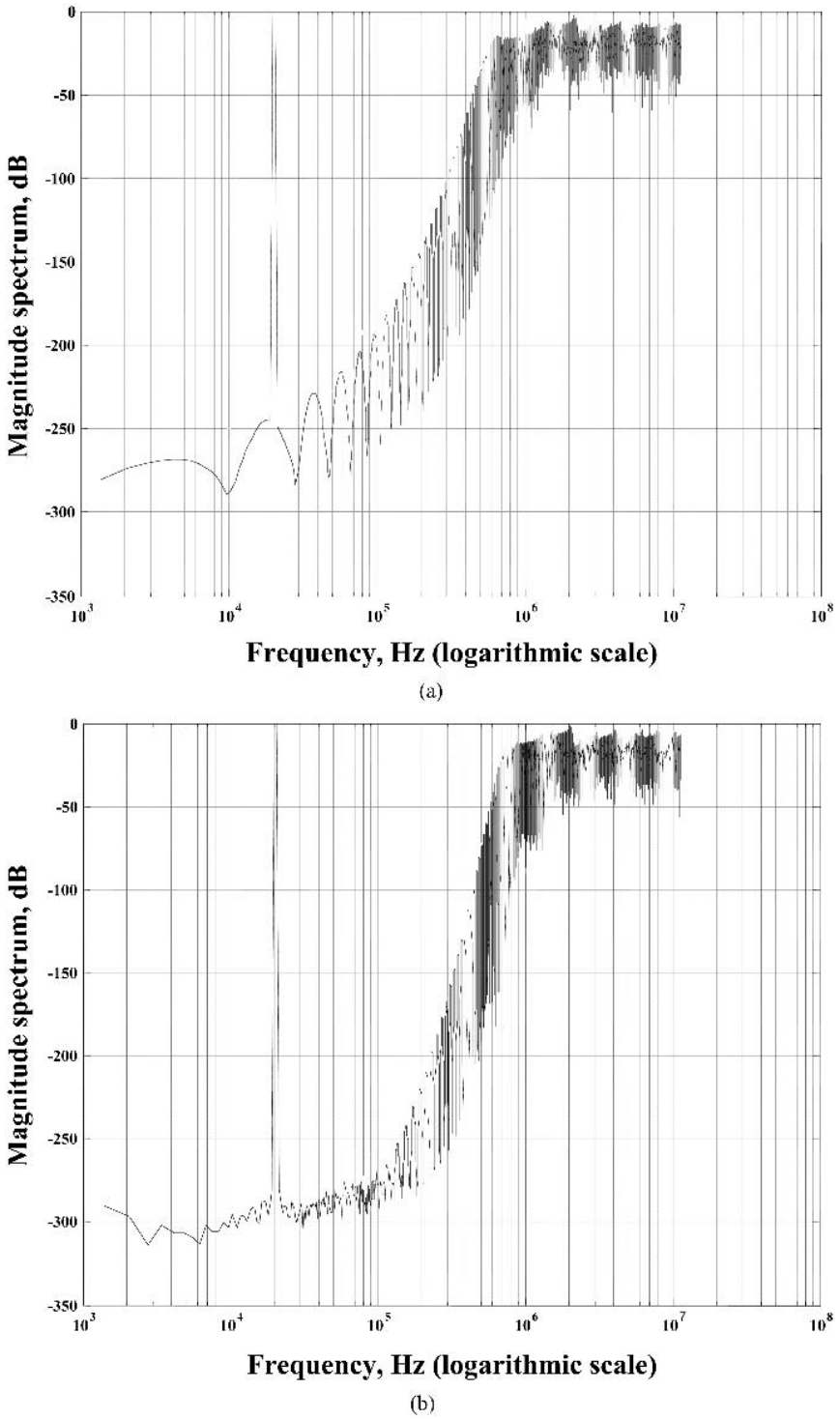


Fig. 4. Spectrum of NTSI derived pulse sequence. (a) $A = 0.30$. (b) $A = 0.25$. (c) $A = 0.10$. (d) $A = 0.01$.

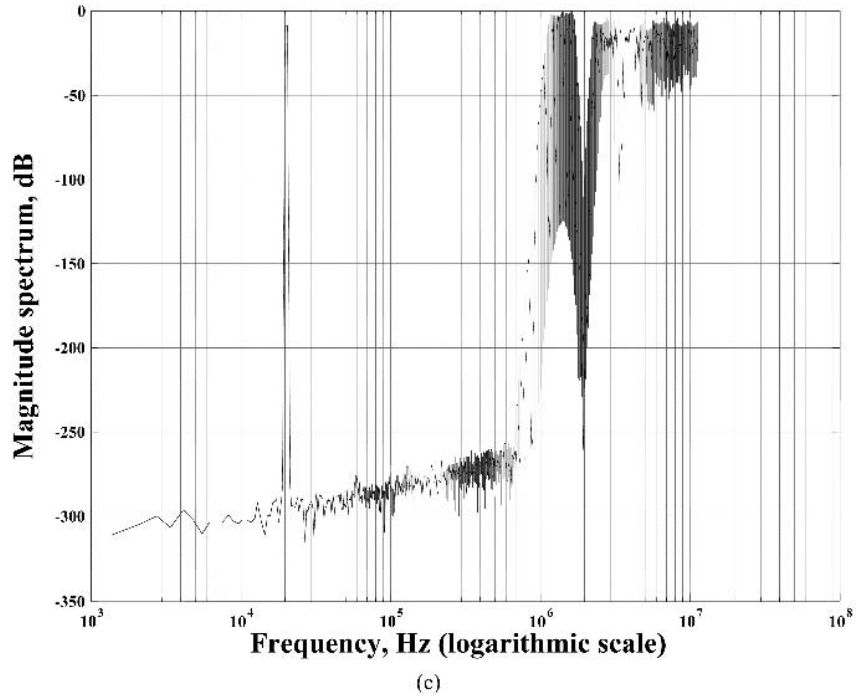
eration, whereas a time displacement of a sample produces both frequency- and phase-dependent distortion. It is therefore likely that sample perturbations in time lead to a more effective jitter strategy. In any event there are differences to be observed.

This section discusses techniques of TDQ used to relocate NTSIs to form a uniformly sampled SDM pulse sequence. Once the NTSI time coordinates t_r are determined, for example, by following the procedures outlined in Section 1, quantization in time can be performed using a combination of time-domain dither and temporal noise shaping. Such a process takes a sequence $\{t_r\}$ as input and determines a quantized output sequence $\{tq_r\}$ such that if

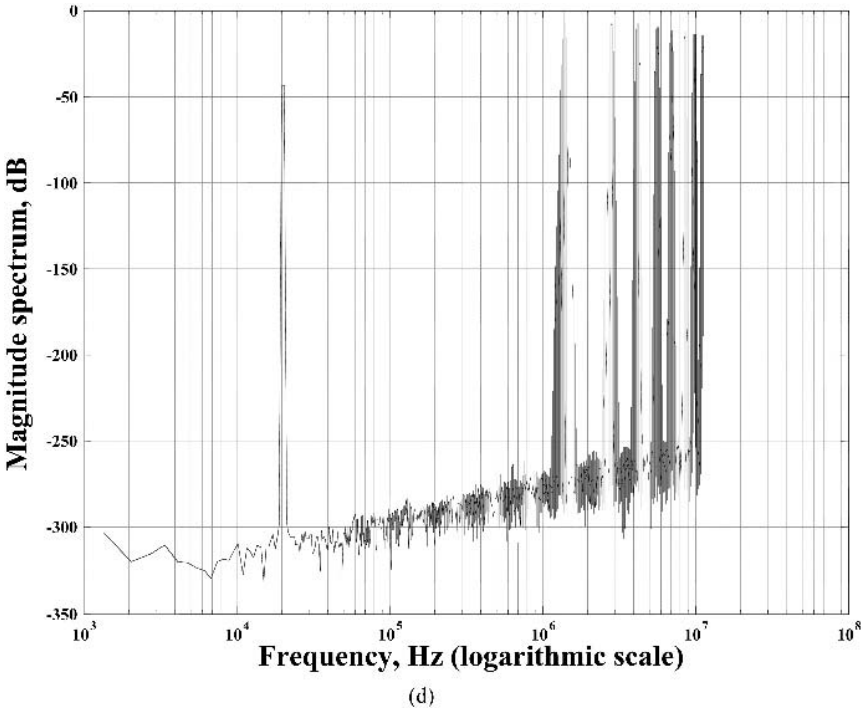
$$\text{NTSI} \Rightarrow \{t_r\}_{r=0}^{N} \tag{12a}$$

then

$$\{tq_r\}_{r=0}^{N} \Rightarrow \text{TQ}[\{tq_r\}_{r=0}^{N}] = \text{TQ}[\text{NTSI}] \tag{12b}$$



(c)



(d)

Fig. 4. *Continued*

where TQ[. . .] represents the generalized operation of TDQ, and $\{tq_r\}$ are the resulting time coordinates of the uniformly sampled binary SDM data. Because the quantization process is directed in time rather than amplitude, after quantization $\{tq_r\}$ may not be sequenced in natural sample order, such as [. . . 10 13 17 **15** 19 . . .]. There is a requirement that the quantized sequence $\{tq_r\}$ produce progressively sequenced binary pulses. Methods of dealing with noncausal sequences and sequences containing coincident pulses, as part of the TDQ process, are discussed in Sections 2.1 to 2.4.

A basic TDQ process using only time-domain dither and a uniform quantizer is illustrated in Fig. 5. Fig. 6 is an extension of Fig. 5 and includes temporal noise shaping configured to embrace the quantizer. Both approaches mirror the techniques used in conventional LPCM signal processing and therefore need little further explanation other than to point out that the input and output sequences are the time coordinate vectors $\{t_r\}$ and $\{tq_r\}$, respectively. However, it is significant that from the perspective of linearity [7]–[9], [22] the time quantizer input range has no imposed limits due to the natural progression of time, unlike feedback SDM where sample amplitude quantization is restricted to two levels. Nevertheless, the greater the order of the temporal noise shaper, the greater becomes the maximum time shifting of pulses away from their optimum NTSIs, so in practice it is expedient to include a degree of time displacement limitation, as discussed in Section 2.4. All the uniform quantization processes discussed here include TPDF dither to decorrelate quantization jitter noise from the signal, making the process linear but noisy [12]. The effectiveness of this expediency is discussed in Section 3.

The following sections discuss factors and strategies that relate to the TDQ process with respect to generating SDM code, where special emphasis is placed on correcting improperly sequenced temporal data streams as defined by the vector $\{tq_r\}$. The techniques presented include both

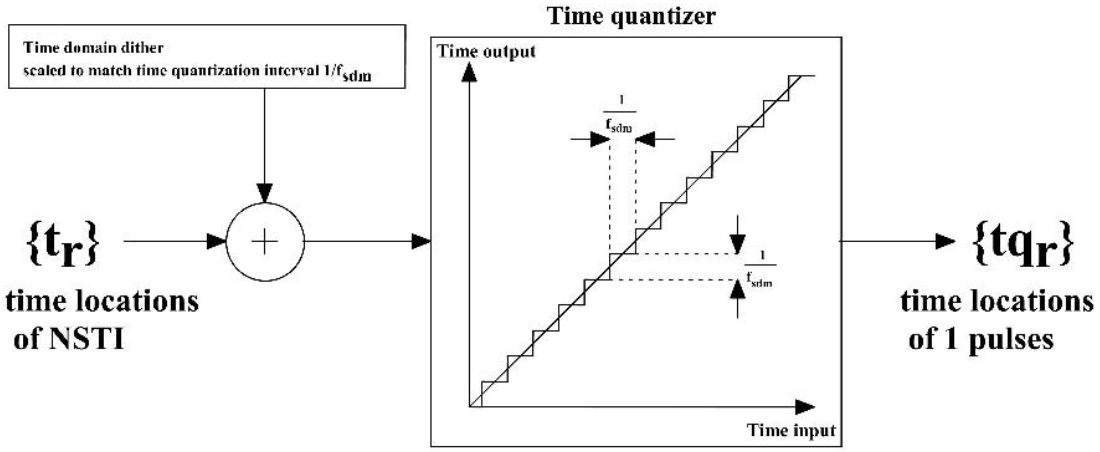


Fig. 5. TDQ including additive time-domain dither.

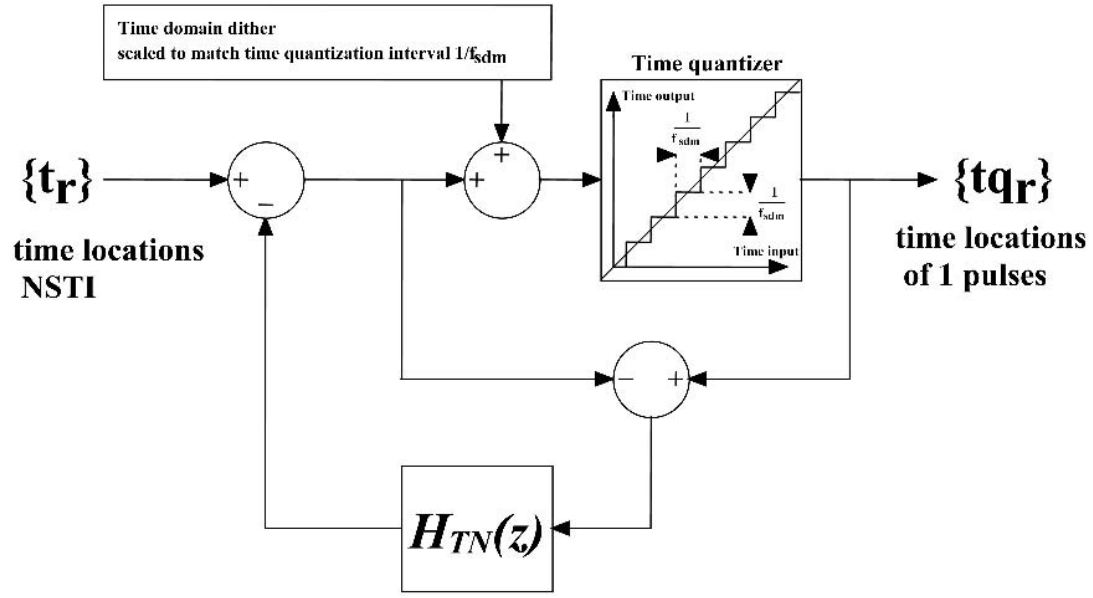


Fig. 6. TDQ with temporal noise shaping and time-domain dither.

open-loop and closed-loop solutions. Section 2.5 presents example spectra based on the correction procedures classified as types 1 to 4.

2.1 Open-Loop Temporal Correction: Type 1

The occurrence of coincident and incorrectly sequenced pulses can be corrected by using a time-sorting procedure [23], [24] that is applied open loop after the temporal noise shaper. This process guarantees a positive arithmetic time progression of pulses, seeks out both dual and multiple coincident pulses, and translates any coincident samples into a near-symmetric bidirectional pulse distribution, all achieved with no loss of pulse area. This algorithm has restrictions in that it does not observe the spectral error introduced in the time domain. Thus it does not yield a useful noise-shaping advantage. However, it does offer code conversion that is appropriate for simple time-domain quantizers with low-order noise shaping and as such reveals some unique features.

The type 1 algorithm is defined as follows. Consider a vector $[X_r]_{r=1}^{N}$ of length N. Here the elements represent time-domain quantized sample locations of 1 pulses, where both sequence order and coincident pulse errors have occurred. The corrected and time-sorted vector $[Y_r]_{r=1}^{N}$ devoid of noncoincident pulses can be computed using the Matlab *sort* function, which rearranges a modified version of the input vector so that the element values are ranked in order of their magnitude (observing that elements take only positive values in this case as time flows from zero),

$$[Y_r]_{r=1}^{N} = \text{sort}([X_r]_{r=1}^{N} - [1{:}N]) + [1{:}N] \tag{13}$$

where $[1{:}N]$ represents the vector $[1 \quad 2 \quad 3 \quad \ldots \quad r \quad \ldots \quad N]$. Because of the addition and subtraction of the vector $[1{:}N]$ the net area under the input and output vectors is invariant, which is important as TDQ and any subsequent postprocessing should not add to or subtract from the average number of 1 pulses.

To demonstrate the validity of this algorithm, three examples are presented, where the vector length is selected arbitrarily as $N = 10$.

Example 2.1.1: One Cluster of Two Coincident Pulses

Define the output vector $[X]$ of the temporal noise shaper as

$$[X_r]_{r=1}^{10} = [2 \quad 5 \quad 8 \quad 11 \quad 9 \quad 9 \quad 13 \quad 15 \quad 16 \quad 19].$$

Subtracting vector [1:10],

$$[Z_r]_{r=1}^{10} = [1 \quad 3 \quad 5 \quad 7 \quad 4 \quad 3 \quad 6 \quad 7 \quad 7 \quad 9].$$

Reorder to form a positive progression in the time sequence,

$$\text{sort}[Z_r]_{r=1}^{10} = [1 \quad 3 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 7 \quad 7 \quad 9].$$

Finally, add back the vector [1:10] to produce the corrected vector $[Y]$,

$$[Y_r]_{r=1}^{10} = [2 \quad 5 \quad 6 \quad 8 \quad 10 \quad 12 \quad 14 \quad 15 \quad 16 \quad 19].$$

The sum of the sample coordinates for both vectors $[X]$ and $[Y]$ is 107, and the respective SDM sequences sdm_x and sdm_y derived from $[X]$ and $[Y]$ are

$$\text{sdm}_x = [0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 2 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1]$$

$$\text{sdm}_y = [0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1].$$

The SDM error vector then follows as

$$[\text{sdm}(r)_x - \text{sdm}(r)_y]_{r=1}^{10} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 2 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0].$$

Example 2.1.2: One Cluster of Three Coincident Pulses

Consider a second example, where

$$[X_r]_{r=1}^{10} = [2 \quad 5 \quad 8 \quad 9 \quad 13 \quad 13 \quad 13 \quad 16 \quad 18 \quad 19]$$

which results in

$$\text{sdm}_x = [0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 3 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1]$$

$$\text{sdm}_y = [0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1]$$

producing an error vector

$$[\text{sdm}(r)_x - \text{sdm}(r)_y]_{r=1}^{19} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0 \quad 2 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0].$$

Example 2.1.3: Three Clusters of Coincidence

Finally consider a third example where

$$[X_r]_{r=1}^{10} = [2 \quad 2 \quad 8 \quad 9 \quad 13 \quad 13 \quad 13 \quad 16 \quad 19 \quad 19]$$

which results in

$$\text{sdm}_x = [0 \quad 2 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 3 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 2 \quad 0]$$

$$\text{sdm}_y = [1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1]$$

producing an error vector

$$[\text{sdm}(r)_x - \text{sdm}(r)_y]_{r=1}^{20} = [-1 \quad 2 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0 \quad 2 \quad 0 \quad -1 \quad 0 \quad 0 \quad -1 \quad 2 \quad -1].$$

Observe in each example how the time-sorted pulse locations have unique values arranged in arithmetic time progression. Also, the areas under the input and output sequences are invariant, which is demonstrated by the sum of each SDM error vector being zero.

2.2 Open-Loop Temporal Correction: Type 2

The technique presented in Section 2.1 used pulse redistribution to correct pulse coincidence in the noise-shaper output, otherwise multilevel pulses would occur. An earlier study [13] proposed a technique of pulse redistribution to account for quantizer overload by substituting

any overload sample with a short symmetric pulse sequence. A symmetrical mapping function has the intrinsic property of producing only spectral amplitude error whereas an asymmetric function introduces both amplitude and phase error. In the present application correction based on symmetrical pulse mapping requires an iterative procedure to convert multiple coincident pulses into a binary sequence as the map itself may give rise to multilevel samples. The only symmetrical mapping function found to give robust convergence when used in an iterative procedure is defined as follows.

Consider a uniformly quantized sequence $\{\mathrm{sdm}(x)\}$ derived from multilevel SDM code that incorporates a mid-riser quantizer with quantum interval Δ. $\{\mathrm{sdm}(x)\}$ is to be down-converted to a binary sequence with respective amplitudes -0.5Δ and 0.5Δ. For sample $\mathrm{sdm}(x)$ and adjacent samples $\mathrm{sdm}(x-1)$ and $\mathrm{sdm}(x+1)$ the substitution map is then applied,

$$\mathrm{sdm}(x) \Rightarrow \mathrm{sdm}(x) - 2\Delta\ \mathrm{sign}[\mathrm{sdm}(x)]$$

$$\mathrm{sdm}(x-1) \Rightarrow \mathrm{sdm}(x-1) + \Delta\ \mathrm{sign}[\mathrm{sdm}(x)]$$

$$\mathrm{sdm}(x+1) \Rightarrow \mathrm{sdm}(x+1) + \Delta\ \mathrm{sign}[\mathrm{sdm}(x)]$$

where $\mathrm{sign}(k) = 1$ for $k > 0$, $\mathrm{sign}(k) = -1$ for $k < 0$, and $\mathrm{sign}(0) = 0$. Mapping is applied sequentially to the multilevel sequence, where the current sample must include any substitutions imposed by the previous calculation. However, a single pass does not guarantee compliance with a binary sequence. Therefore the process must be repeated until convergence is achieved. A caveat to convergence is that the information signal contained within the multilevel sequence must not overload the down-converted binary sequence and for a finite vector, the net areas before and after correction must be compatible.

To illustrate the mapping procedure, consider two examples. Here the input sequence is constrained to take only positive integer values where $\Delta = 1$.

Example 2.2.1 (Same as Example 2.1.3)

Let the temporal noise-shaper output $[X]$ be

$$[X_r]_{r=1}^{10} = [2\ \ 2\ \ 8\ \ 9\ \ 13\ \ 13\ \ 13\ \ 16\ \ 19\ \ 19]$$

from which uncorrected SDM code is derived as

$$\mathrm{SDM}_0 = [0\ \ 2\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 0\ \ 0\ \ 0\ \ 3\ \ 0\ \ 0\ \ 1\ \ 0\ \ 0\ \ 2\ \ 0].$$

Applying symmetric pulse mapping, after one pass of the mapping function we have

$$\mathrm{SDM}_1 = [1\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 0\ \ 0\ \ 1\ \ 1\ \ 1\ \ 0\ \ 1\ \ 0\ \ 1\ \ 0\ \ 1].$$

which yields an input–output error vector,

$$\mathrm{SDM}_0 - \mathrm{SDM}_1 = [-1\ \ 2\ \ -1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ -1\ \ 2\ \ -1\ \ 0\ \ 0\ \ 0\ \ -1\ \ 2\ \ -1].$$

Comparing this error with the error derived in Example 2.1.3 shows that although similar, the symmetric pulse mapping has achieved one error cluster where the time dispersion is halved, a result that should contribute to lower distortion in the recovered signal.

Example 2.2.2

This second example illustrates SDM code with more densely packed clusters of multilevel pulse, which requires several iterations to form a binary sequence. Let

$$[X_r]_{r=1}^{10} = [2\ \ 2\ \ 4\ \ 5\ \ 5\ \ 6\ \ 8\ \ 8\ \ 8\ \ 11\ \ 13\ \ 13].$$

Translating to multilevel SDM code gives the initial uncorrected sequence as

$$\mathrm{SDM}_0 = [0\ \ 2\ \ 0\ \ 1\ \ 2\ \ 1\ \ 0\ \ 3\ \ 0\ \ 0\ \ 1\ \ 0\ \ 2\ \ 0].$$

In this more extreme example, where the sum of the vector elements is 12 out of a vector of length 14, the symmetric pulse mapping required seven iterations to converge to a binary sequence. The intermediate results are given in Table 1, where each row shows the sequence progression commencing with the initial pulse sequence. The iteration number is given in the left-hand column.

The process is shown to converge after the seventh iteration to form the binary sequence

$$\mathrm{SDM}_7 = [1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 0\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 0\ \ 1].$$

where the overall error vector is

$$\mathrm{SDM}_0 - \mathrm{SDM}_7 = [-1\ \ 1\ \ -1\ \ 0\ \ 1\ \ 0\ \ 0\ \ 2\ \ -1\ \ -1\ \ 0\ \ -1\ \ 2\ \ -1].$$

This method of converting multilevel pulses to binary is applicable not only to the present system employing TDQ,

Table 1. Symmetric pulse mapping showing 7 iterations required for convergence.

0	0	2	0	1	2	1	0	3	0	0	1	0	2	0
1	1	0	1	2	0	2	1	1	1	0	1	1	0	1
2	1	0	2	0	2	0	2	1	1	0	1	1	0	1
3	1	1	0	2	0	2	0	2	1	0	1	1	0	1
4	1	1	1	0	2	0	2	0	2	0	1	1	0	1
5	1	1	1	1	0	2	0	2	0	1	1	1	0	1
6	1	1	1	1	1	0	2	0	1	1	1	1	0	1
7	1	1	1	1	1	1	0	1	1	1	1	1	0	1

but also to multilevel SDM [17], where it forms a multilevel to two-level transformer provided the input signal falls within the coding range of the down-converted binary code.

2.3 Closed-Loop Temporal Correction: Type 3

Rather than applying correction for pulse coincidence as a separate process applied to the output of the temporal noise shaper, correction can be applied directly within the feedback loop. The process is summarized as follows, where $tq(n)$ represents the quantizer output after the nth sample.

In this feedback-controlled system the present value of $tq(n)$ is compared against a time window of past samples and tested for any identical values. If coincidence is detected, then a quantum δ is added to $tq(n)$,

$$tq(n) \Rightarrow tq(n) + \delta. \tag{14}$$

The test is reapplied and if coincidence is detected once again, then a further quantum is added to $tq(n)$, and so on; otherwise the current value is assigned to the output. Using this technique, pulse coincidence is eliminated, and because the process is contained within a feedback loop, any modification to $tq(n)$ is considered an inherent characteristic of the quantizer and partially corrected by feedback. However, there is no constraint placed upon the degree of pulse displacement that can occur, so this may contribute additional noise.

2.4 Closed-Loop Temporal Correction with Constrained Time Dispersion: Type 4

Because sample time displacement in SDM code contributes noise, it is expedient to constrain the peak temporal pulse displacement introduced in the noise shaper. This can be achieved by both constraining the noise-shaper transfer function and limiting the maximum pulse displacement range. To constrain sample time displacement in the temporal noise shaper, a limiter is introduced just before the coincidence detector–corrector described in Section 2.3.

The time quantizer input–output error ddt is calculated at each sampling instance, where if $tr(n)$ and $tq(n)$ are the respective input and output of the temporal noise shaper just after the nth sample is processed, then in Matlab notation,

```
ddt = round((tr(n) - tq(n))/of);
```

where of is a scale factor. If the limiter is set to lim, then a program fragment defines the process as

```
if ddt > lim
tq(n) = tq(n) + (ddt - lim)*of;
elseif ddt < -lim
tq(n) = tq(n) - (ddt + lim)*of;
end
```

2.5 Simulation Examples of Temporal Noise Shaping with Correction, Types 1 to 4

To demonstrate the relative performance of the type 1 to 4 correction techniques, Matlab simulations were performed using the LFM–SDM model with a second-order noise-shaping time quantizer. All simulations used two equiamplitude sine-wave inputs of normalized amplitudes 0.1 with respective frequencies of 19 and 20 kHz. Just one typical spectral result is shown in Fig. 7, since without signal averaging no significant differences were observed between procedures.

To show an extreme example of the capability of the type 2 multilevel-to-binary conversion algorithm described in Section 2.2, a mapping was performed on the output derived from a fourth-order multilevel SDM incor-

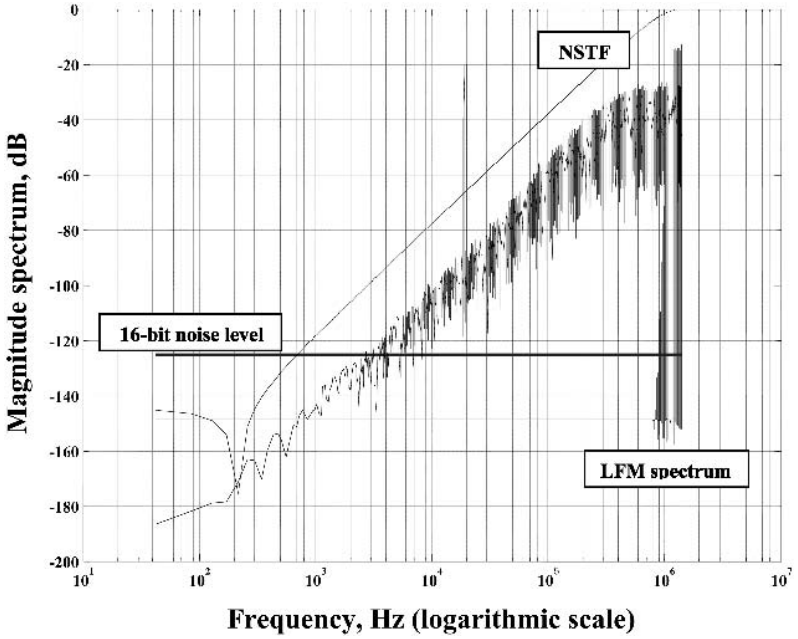


Fig. 7. LFM–SDM output spectrum, typical of type 1 to 4 procedures.

porating a midriser quantizer with quantum $\Delta = 1$. A Matlab simulation was written for the noise shaper that included multilevel quantization and four cascaded integrators with local feedforward paths to achieve stability. A Hankel matrix facilitated compact code design (a segment of the routine is presented in Appendix 2). This is considered an extreme test because most samples require mapping and their values are relatively large. The time-domain results shown in Fig. 8 confirm a successful mapping from an amplitude range of approximately −17.5 to 17.5 down to binary levels of −0.5 and 0.5. The corresponding output spectra before and after mapping are shown in Fig. 9, where although there is substantial degradation, the noise performance remains marginally superior to the SDM spectrum presented in Fig. 7. Also there is minimal evidence of intermodulation distortion products. The remarkable aspect of this simulation is the attainment of convergence to the desired two-level SDM code. However, for mapping to be successful, the input signal applied to the multilevel noise shaper must be limited in peak amplitude

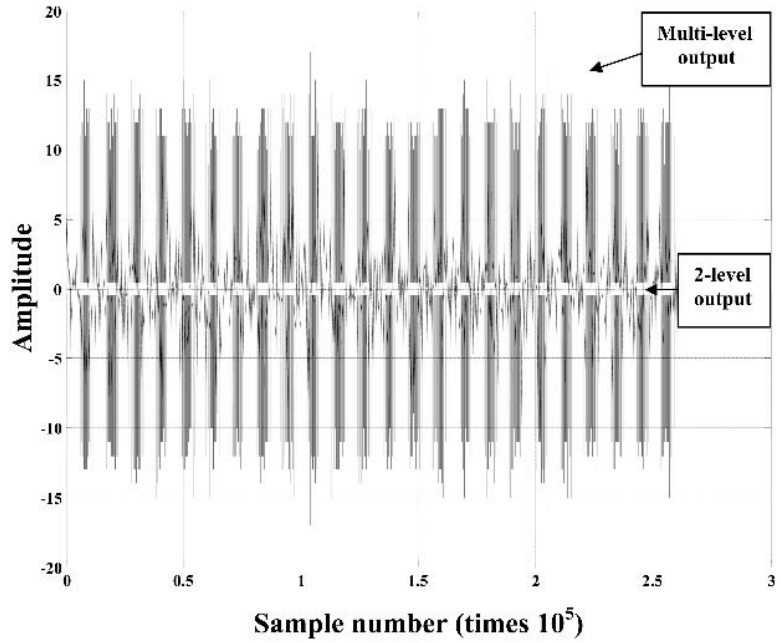


Fig. 8. Time-domain SDM output before and after multilevel to two-level transformation.

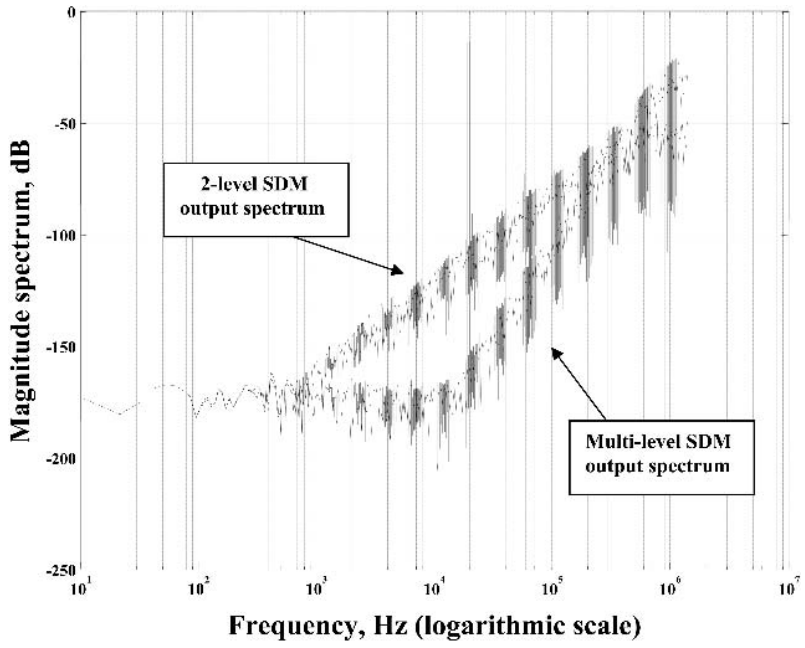


Fig. 9. SDM output spectra before and after multilevel to two-level transformation, 4 integrator noise shaper with input dithered to 24-bit LPCM.

so as to fall within the amplitude range of the two-level code. Also, experiments with other higher order time dispersive codes have shown poor convergence such that the simple map described in Section 2-2 is a special case that has been found to exhibit excellent convergence properties.

3 LINEARITY OF THE LFM–SDM MODEL

Lipshitz and Vanderkooy [8], [9] have demonstrated that with low-order SDM there can be noise modulation, and correlated distortion components manifest as periodic components often buried within the noise floor. It is concluded that this is a failing of the loop quantizer, which is unable to accommodate conventional TPDF dither [12] without incurring overload [13]. Further, it is known that the SDM loop when extended to include a multilevel, uniform quantizer does not exhibit nonlinear distortion within a bound determined by TPDF dither bit depth and finite word length of the SDM coder. The classic comparison procedure is to simulate SDM and study its performance with varying levels of dither addition prior to binary quantization. Tests often use carefully selected sinusoids and/or combinations of low-level dc offset, but it is also instructive to use low-level dc sweeps and to construct three-dimensional noise plots to identify tracks of periodic idle tones that sweep across the frequency space (see Dunn and Sandler [25]).

However, generating SDM code by using the LFM–SDM model gives an alternative perspective on the function of dither. If the coding method is viewed from the generation of NTSI-derived pulse sequences then, as shown in Section 1, it can be concluded that below overload, LFM sideband spillage into the audio band is negligible, as demonstrated in Fig. 3(c) and (d). If TDQ is performed with the addition of temporal dither, then pulse redistribution is randomized and does not lead to correlated distortion artifacts. It could be argued that a small degree of noise modulation may occur as the LFM spectrum expands and contracts with the input signal amplitude and frequency, but such effects are extremely small. This section considers the effect of dither statistics on linearity in the most basic TDQ process and then extends the results to first-order time-domain noise shaping.

3.1 Dither Performance in Zero- and First-Order TDQ

Dither is an intrinsic element in quantization and achieves quantization distortion decorrelation if applied correctly. Much of the pioneering work stems from early digital video processing. This section evaluates three types of dither applied to TDQ with both zero- and first-order noise shaping. Three dither sequences are examined:

- RPDF—Offset rectangular probability distribution spanning one SDM sample period
- TPDF—Offset triangular probability distribution spanning two SDM sample periods
- ATPDF—Asymmetric TPDF spanning two SDM sample periods.

Because dither is directed along the time axis there is no problem with quantizer overload, even though it is desirable to keep the dither amplitude low and because, as discussed in Section 2, it is necessary to prevent coincident sample values as these map to multiamplitude pulses, which are not permitted in binary SDM. Fig. 10 gives histograms of the three dither sequences, where each dis-

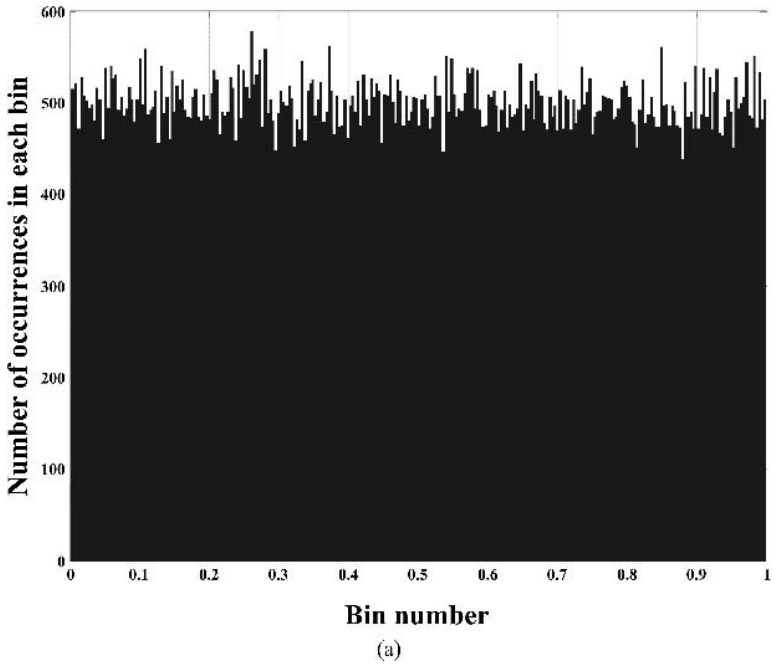
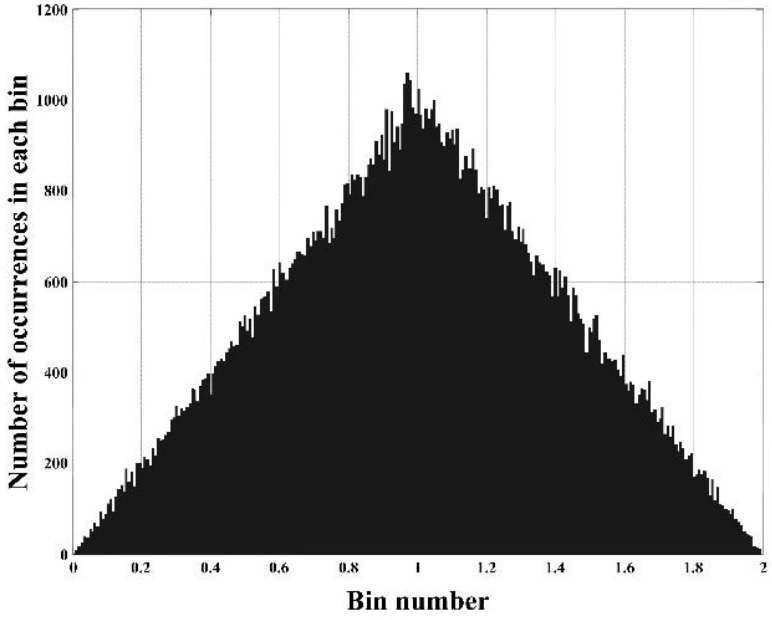


Fig. 10. Histograms of dither sequence. (a) Offset RPDF dither. (b) Offset TPDF dither. (c) Asymmetric TPDF dither.

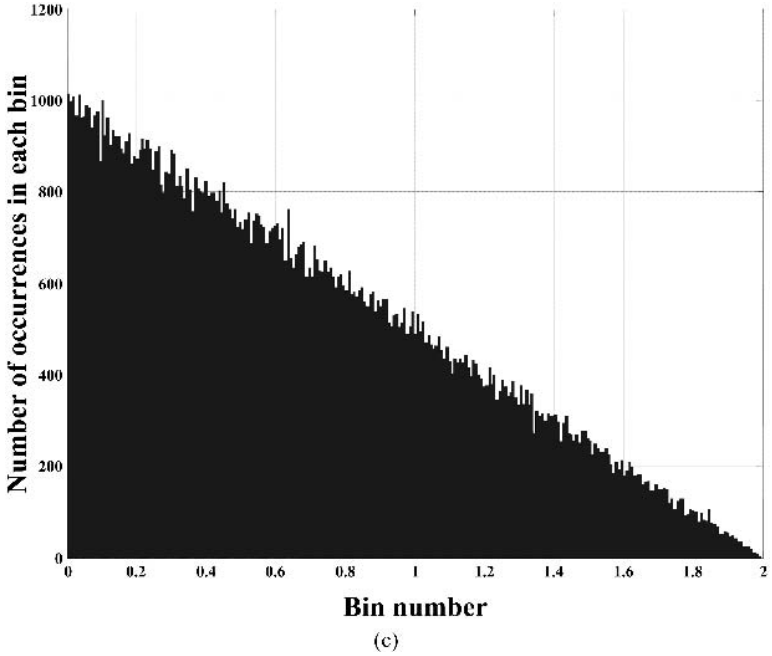
tribution is offset from zero. This expedient biases the quantization error $t_r - tq_r$ such that $tq_r > t_r$.

3.2 Simulation Results

Output spectra were calculated for LFM–SDM for each dither sequence, both with and without noise shaping. All included type 1 code correction described in Section 2.1. A vector of length 2^{18} was used and outputs were averaged 4096 times to expose distortion artifacts. Output spectra and the corresponding $t_r - tq_r$ error histograms for both zero- and first-order TDQ are presented in Figs. 11 and 12. In addition, Fig. 13 shows the output spectrum for second-order temporal noise shaping together with type 1 correction, again using 4096 averages but a vector of length 2^{15}. The LFM spectrum is also shown derived from $\{t_r\}$ to confirm its low audio band distortion. The last spectrum shows that no significant distortion components have emerged above the quantization noise, which has been lowered theoretically by 36 dB compared to the nonaveraged spectrum. This is important not because it yields particularly low levels of noise within the audio band, but because it demonstrates that dither can be applied effectively in the time domain, unlike dither applied in the amplitude domain of a feedback coder, which by necessity has to take a suboptimal level.



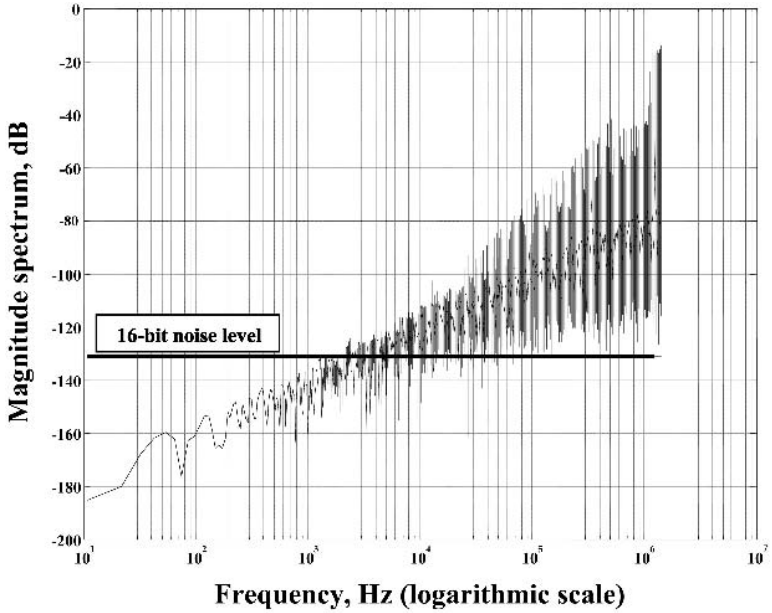
(b)

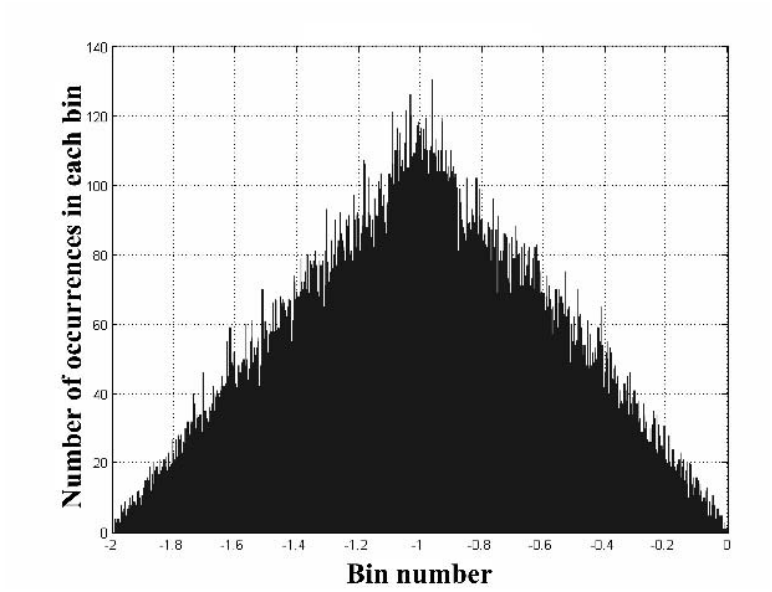


(c)

Fig. 10. *Continued*

The results reveal that dither is effective, but that the type of dither has the greatest effect on distortion for the zero-order examples, where TPDF gives the best result. Except for asymmetric TPDF dither, residual distortion is related to type 1 correction, which because of the dc dither offset, works mainly to suppress coincident pulses since tq_r values occur after t_r values. The asymmetric TPDF dither of Fig. 10(c) gives the worst result, as might be anticipated, where Fig. 11(c) reveals a skewed $t_r - tq_r$ error histogram, implying nonlinearity. However, when a first-order noise shaper is introduced, the statistics of the dither have less significance, whereas if Fig. 12(c) is inspected, reduced distortion together with a virtually symmetric $t_r - tq_r$ error histogram is revealed. In fact all the output spectra using higher order noise shaping reveal much less distortion. Finally a test was performed to explore RPDF dither reduced to a level of 0.35, which is often found effective in high-order SDM applications. Histograms of $t_r - tq_r$ error are shown in Fig. 14 for both zero- and first-order time-domain noise shaping. While the zero-order case exhibits a poorly defined distribution, the first-order example has mapped almost to TPDF spanning two quanta and compares favorably with Fig. 11(a). These results support the observation that dither statistics are less critical in high-order SDM and that a TPDF dither spanning two quanta, although inappropriate when applied in the ampli-





(a)

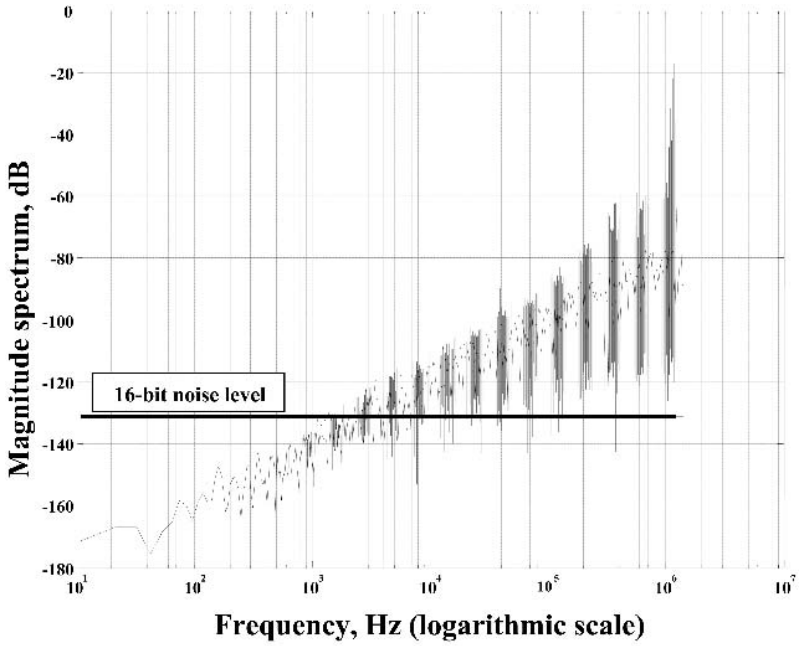
Fig. 11. Zero order. Output spectra (4096 averages) and ($t_r - tq$) histograms. (a) Offset RPDF dither. (b) Offset TPDF dither. (c) Asymmetric TPDF dither.

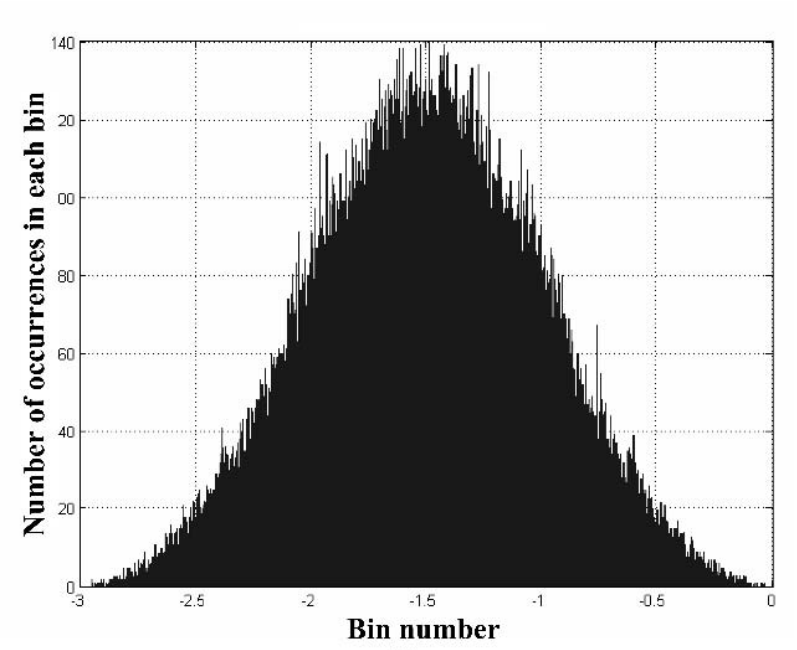
tude domain because of overload and stability problems, still remains effective at a reduced level. The observation that time-domain noise shaping renders an almost symmetric $t_r - tq_r$ error histogram from a highly asymmetric dither is critical to the debate, even though the averaged noise spectrum returned was lowest with symmetric TPDF dither.

However, although LFM–SDM derived SDM offers excellent linearity, it does not in its present form achieve the low in-band noise performance demanded of high-resolution audio. Therefore to explore the limits of noise shaping and linearity as well as the benefit of reduced dependence on dither type with increased loop order, Section 4 investigates high-order SDM, whereas Section 5 presents a novel approach to loop transfer function implementation and stability together with simulations to demonstrate its noise performance and coding linearity.

4 HIGH-ORDER SONY FF STYLE SDM

Takahashi and Nishio [26] have published a fifth-order feedforward (Sony FF) SDM encoder capable of high per-





(b)

Fig. 11. *Continued*

formance in the context of SACD. This encoder is discussed in Section 4.1. Section 4.2 gives consideration to the number of principal integrators[5] in the noise-shaping transfer function, and Section 4.3 presents a ninth-order variant with lower in-band noise.

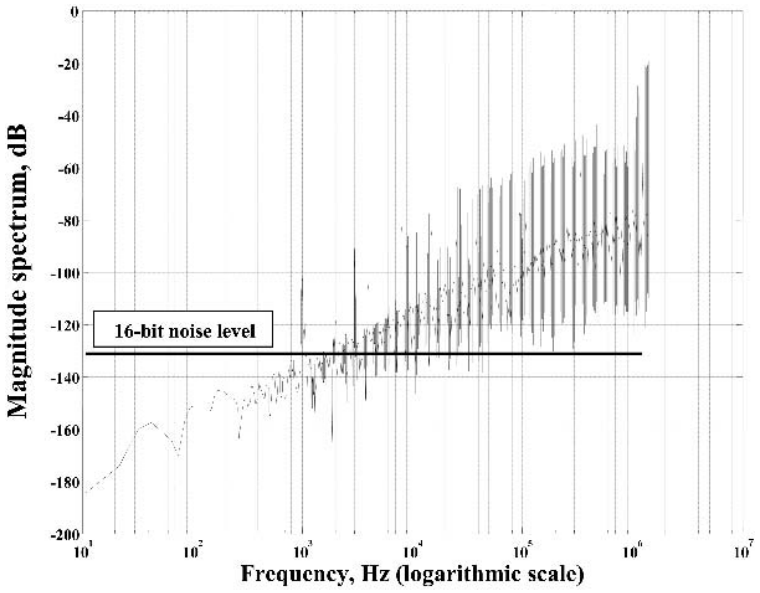
4.1 Fifth-Order Sony FF SDM

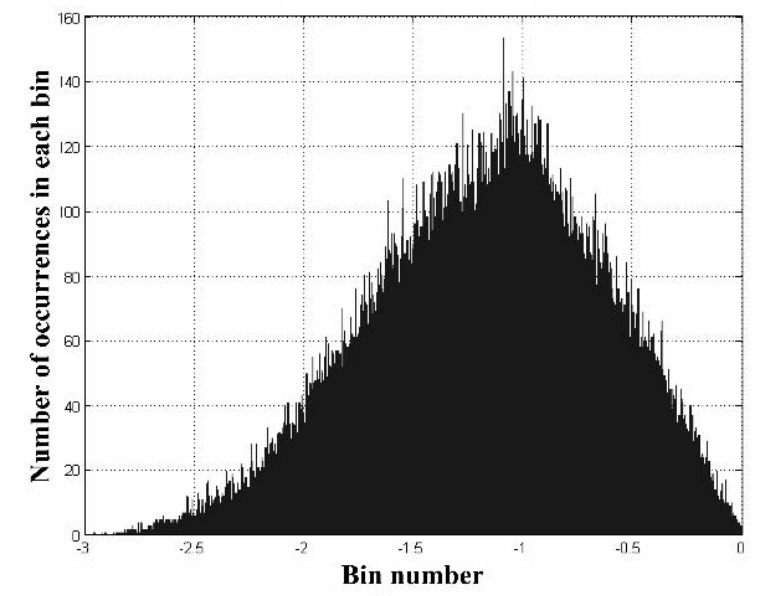
The fifth-order Sony FF SDM topology [26] is shown in Fig. 15, where a low level of audio band noise is obtained by implementing two cascaded biquadratic filter sections. These filters form two sets of real pole pairs within the forward-path transfer function located at 10 and 20 kHz, respectively, which map into corresponding zeros in the noise-shaping transfer function (NSTF). For each pole pair two integrators in the forward path are required. Hence a cascade of five integrators can accommodate two sets of pole pairs, nine integrators accommodate four sets of pole pairs, and so on.

The following analysis determines the NSTF for the fifth-order FF SDM encoder shown in Fig. 15. If $A_5(z)$ is the z-domain transfer function of the forward path, then for this fifth-order loop,

$$\mathrm{NSTF}_5(z) = \frac{1}{1 + z^{-1}A_5(z)}. \tag{15}$$

[5]The term "principal integrator" is used to here to describe the main integrators in the forward path of a SDM feedback coder.





(c)

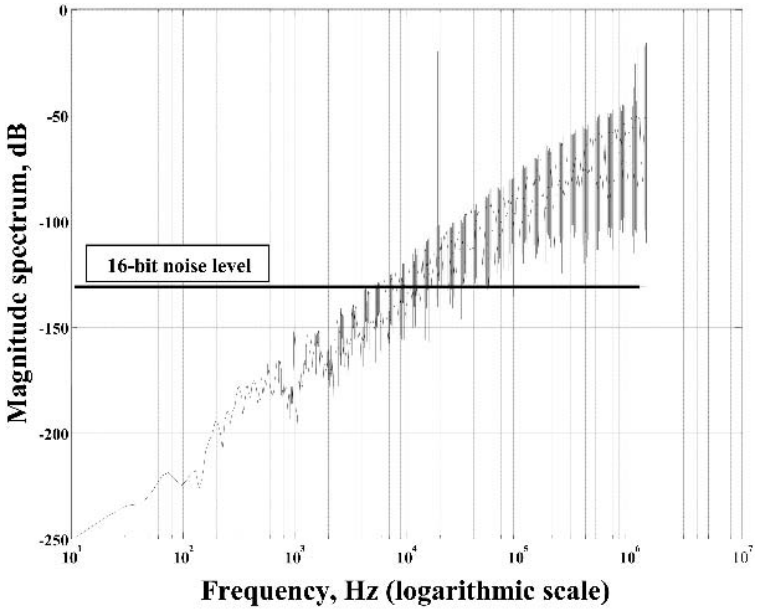
Fig. 11. *Continued*

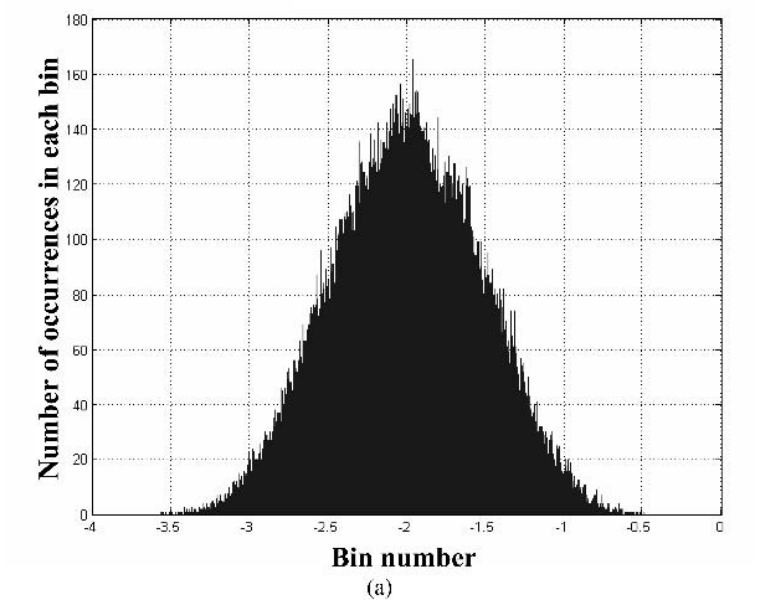
Determining $A_5(z)$ by inspection of Fig. 15, given as

$$A_5(z) = b_1 b_2 b_3 b_4 b_5 \left[\frac{1 + \left(\frac{1 - z^{-1}}{b_1 b_3}\right)\left(1 + b_3 - z^{-1} + \left(\frac{1 + b_5 - z^{-1}}{b_2 b_5}\right)[(1 - z^{-1})^2 - b_3 c_2 z^{-1}]\right)}{(1 - z^{-1})[(1 - z^{-1})^2 - b_3 c_2 z^{-1}][(1 - z^{-1})^2 - b_5 c_4 z^{-1}]} \right]. \quad (16)$$

The denominator of Eq. (16) forms the NSTF numerator and establishes zeros in $\mathrm{NSTF}_5(z)$. The $\{a\}$ and $\{b\}$ coefficients for fifth-order Sony FF SDM [26] are

$b_1 = 1$, $b_2 = 0.5$, $b_3 = 0.25$,
$b_4 = 0.125$, $b_5 = 0.0625$
$c_2 = -0.001953125$, $c_4 = -0.03125$.



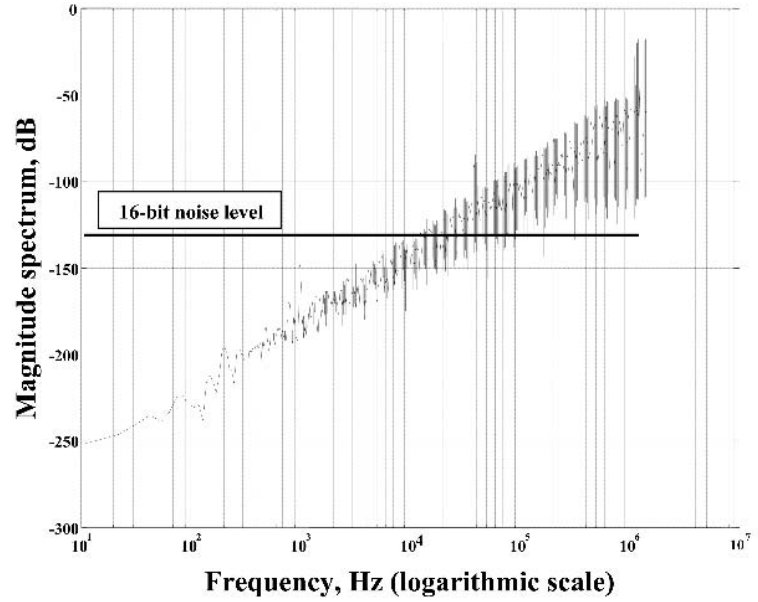


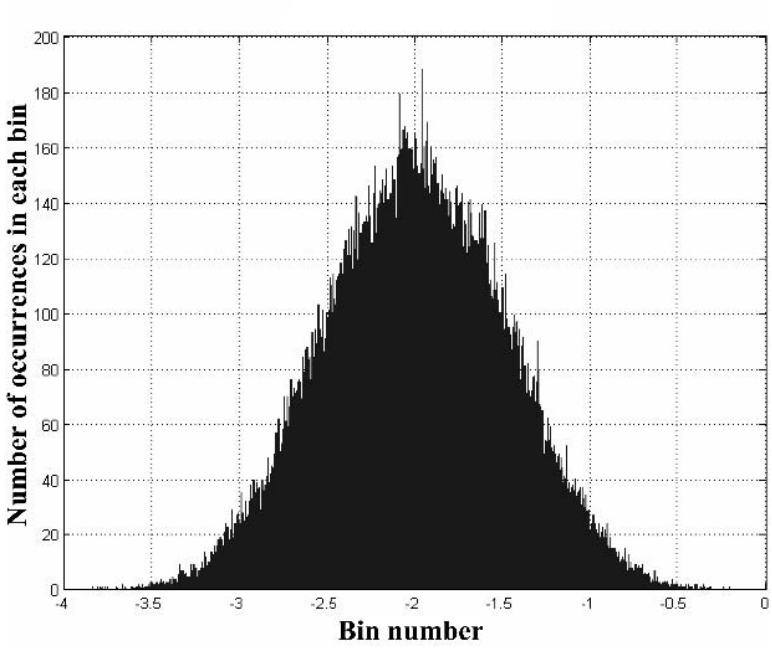
(a)

Fig. 12. First order. Output spectra (4096 averages) and ($t_r - tq$) histograms. (a) Offset RPDF dither. (b) Offset TPDF dither. (c) Asymmetric TPDF dither.

and locate zeros at 10 and 20 kHz, respectively, for a 64*44.1-kHz sampling rate. To confirm this result, a Matlab simulation computed a binary SDM output sequence of 2^{18} samples for a periodic input sequence. Applying the Fourier transform and using a Blackman-squared window[6] to reduce truncation errors, the output spectrum is derived as shown in Fig. 16. To benchmark this spectrum an LPCM spectral noise floor is calculated assuming a 24-bit uniform quantizer with optimum dither and a sampling rate of 44.1 kHz. In this example an input is used consisting of two sine waves, each at −20-dB amplitude and frequencies of 9 and 10 kHz, respectively (where 0 dB corresponds to a sine wave with a peak input amplitude of 1) where the SDM code has a range −1 to +1. This signal is quantized to 24 bit and upsampled to the SDM sampling frequency using Fourier transform techniques. The SDM output spectrum confirms NSTF zeros at 10 kHz and 20 kHz and reveals a noiselike structure with minimal correlated distortion. The noise spectrum corresponds to that of the

[6]The Fourier transform assumes a circular function that is repetitive. However, in SDM the state of the encoder at the end of the sequence is generally distant from its commencement, consequently spectral errors occur that are most noticeable at low frequency and overestimate spectral levels, thus giving a pessimistic estimate of performance. The use of a Blackman-squared window can suppress to a large extent these anomalies, although as a consequence there is now a limit in low-frequency spectral resolution.





(b)

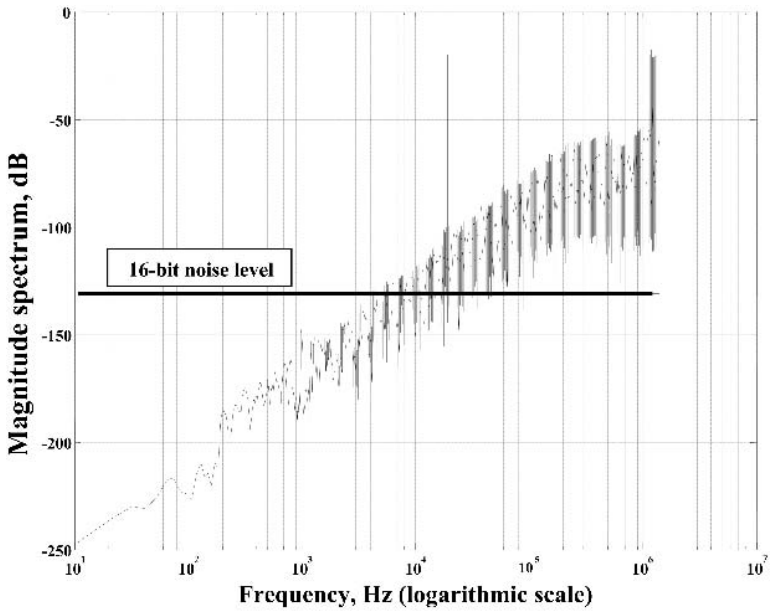
Fig. 12. *Continued*

24-bit LPCM input signal below about 800 Hz, but then rises with frequency although staying just within about a 20-bit LPCM accuracy below 20 kHz.

To examine coding linearity, the Sony FF SDM algorithm is benchmarked against the topology shown in Fig. 17, which uses an identical loop filter to guarantee identical signal transfer functions, but with the quantizer substituted by a unity-gain stage. Both binary and "no quantization" SDM simulations were run simultaneously using identical input sequences, with outputs subtracted to form the error spectrum shown in Fig. 18. Consequently this difference process forms a sensitive indicator for distortion arising purely from restricting a multilevel quantizer to binary, where the computed difference spectrum is shown to be noiselike and shaped by the NSTF with no significant distortion evident within the bound set by the noise in the curve.

4.2 Selecting the Number of Integrators in SDM Feedback Encoders

The range of available NSTFs using only principal integrators is depicted in Fig. 19, where only the feedforward paths have been retained for the purpose of closed-loop stability when using the binary-weighted $\{b\}$ coefficients [26] given in Section 4.1. These curves indicate that for increasing loop order and using just the principal integrators, there is a frequency-dependent bound beyond which no improvement is possible. Hence there is



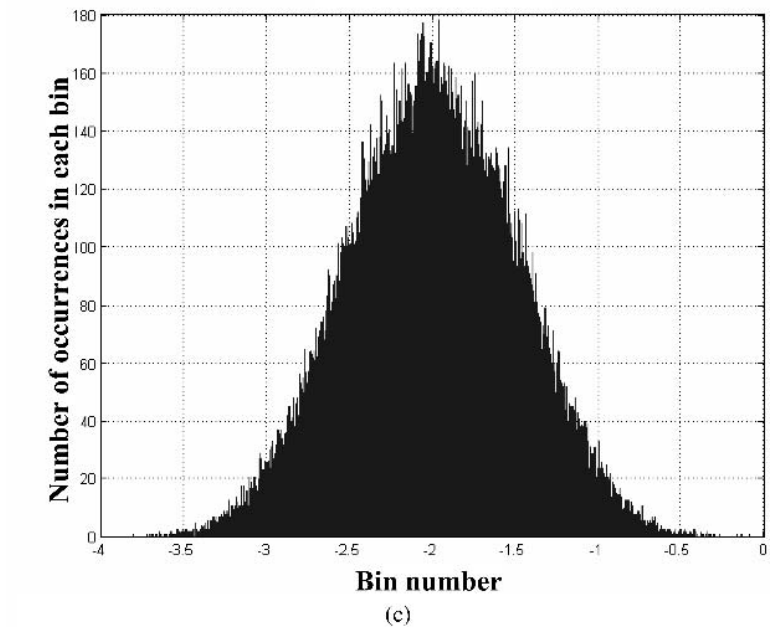


Fig. 12. *Continued*

minimal advantage in using more than five integrators for a system with a 2.8224-MHz sampling rate. To address this limitation it is necessary to introduce frequency-selective boosting of the loop gain, as discussed in Section 4.1, focused on the upper end of the audio spectrum and typically in the band of 10–20 kHz. In this scheme each notch in the NSTF requires a pair of integrators, so if more than two notches are to be employed, then the number of integrators must be increased even though on their own they offer little performance advantage.

4.3 Ninth-Order Modified Sony FF SDM

A ninth-order SDM is presented in this section as an extension of the topology shown in Fig. 15 and uses nine principal integrators with four internal feedback loops, where the $\{b\}$ and $\{c\}$ coefficients are specified as follows:

$$b(1) = 1, \quad b(2) = 0.5, \quad b(3) = 0.25,$$
$$b(4) = 0.0125, \quad b(5) = 0.0625,$$
$$b(6) = 0.03125, \quad b(7) = 0.015625,$$
$$b(8) = 0.0078125, \quad b(9) = 0.00390625$$

$$c(2) = -0.03125/50, \quad c(4) = -0.03125/2,$$
$$c(6) = -0.125, \quad c(8) = -0.125*1.1.$$

Using the same input signal excitation as in Section 4.1, the output spectrum for the ninth-order SDM is shown in Fig. 20, where the LPCM reference is again set at 24 bit to highlight improved performance and to show that 24-bit accuracy is achieved almost to 20 kHz. However, scrutinizing the forward-path transfer function $A_5(j2\pi f)$ presented in Eq. (16), for fifth-order Sony FF SDM it is evident that below the frequency band where the internal feedback loops introduce notches, the NSTF tends to a slope less than would be anticipated using just the principal integrators without local feedback loops. To validate this observation consider a low-frequency approximation by substituting $(1 - z^{-1}) \Rightarrow j2\pi f/f_{sdm}$ in Equation (16) where f_{sdm} is the SACD sampling rate, then as $f \to 0$ it follows that

$$A_5(f)|_{f\to 0} \Rightarrow \frac{b_1 b_2 b_4 f_{sdm}}{j2\pi f c_2 c_4}$$

revealing that the slope of the NSTF tends to first order (6 dB per octave) at low frequency. A similar analysis can be applied to the ninth-order example, where again a first-order slope is encountered at the low-frequency limit. This is considered a significant limitation of this class of forward-path transfer function as the low-frequency coding accuracy in the critical midrange of human hearing is compromised whereas it can be improved substantially together with enhanced linearity, as shown in Section 5.

5 SDM WITH PARAMETRIC NSTF EQUALIZATION

This section describes parametric SDM, which allows the NSTF to be handcrafted using methods similar to parametric equalization [27] and can achieve substantially lower in-audio band noise compared to both fifth-order and modified ninth-order SDM. Parametric SDM addresses both the limited performance enhancement available by using more than five principal integrators (see Section 4.2) and the low-frequency NSTF compromise observed in the Sony FF topology (see Section 4.3).

The basic parametric SDM is shown in Fig. 21. Adding the complete z-domain topology, Fig. 22 illustrates five cascaded biquadratic-type band-pass/low-pass filter sections located ahead of the five principal integrators. A

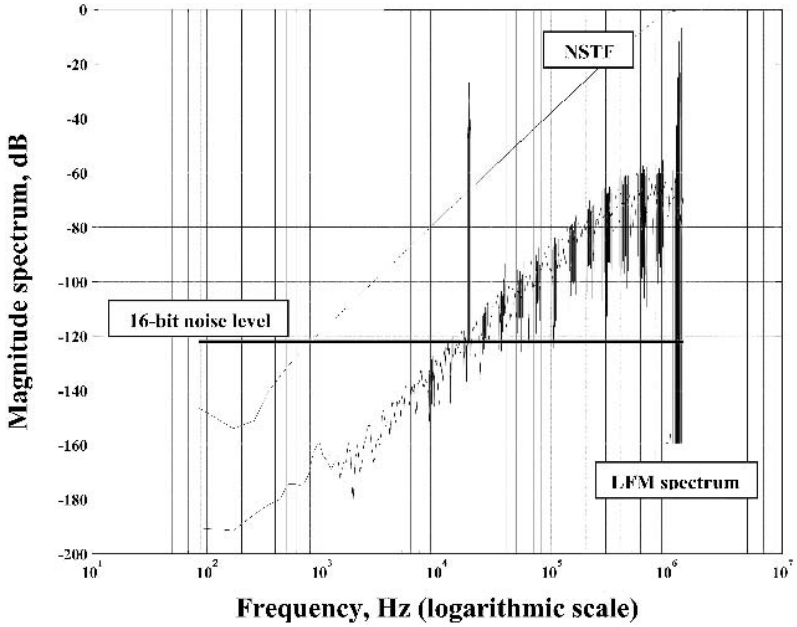


Fig. 13. LFM–SDM output spectrum, second-order noise shaping with RPDF dither, and type 1 correction (4096 averages).

feature of this topology over the Sony FF structure is that if parametric filters designated $P_1(z)$ to $P_5(z)$ have negligible insertion gain, then the forward path transfer function defaults to that of the five principal integrators, with no compromise in the low-frequency NSTF. Each parametric equalizer consists of a second-order biquadratic filter with individual center frequency, Q factor, and low-pass and band-pass insertion gains. The transfer function

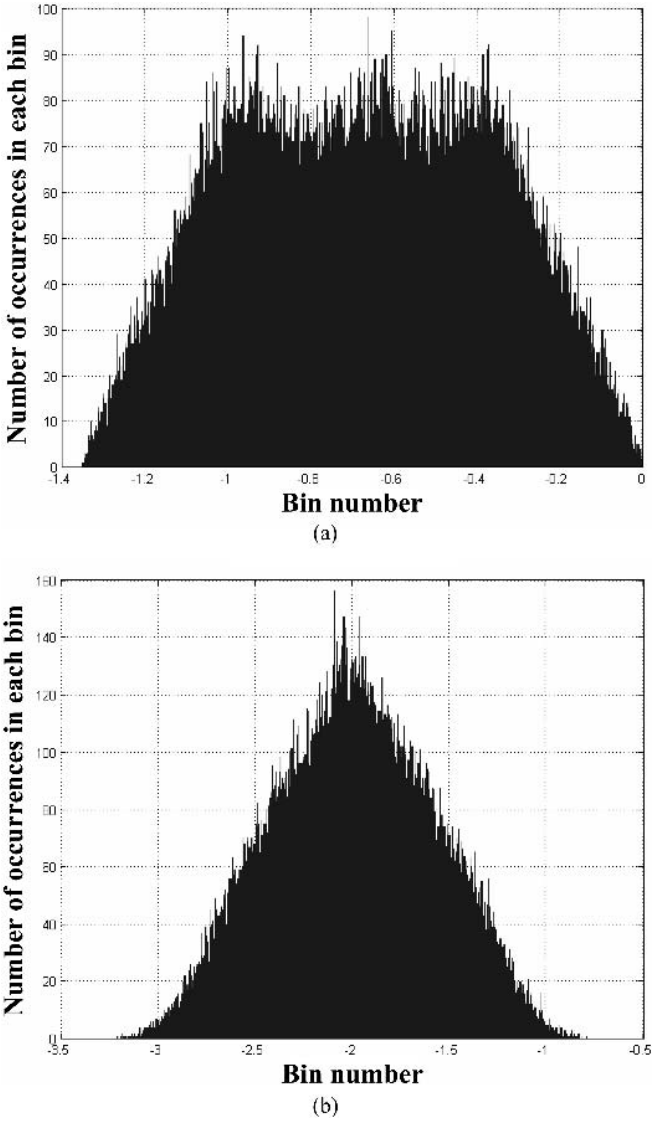


Fig. 14. TDQ ($t_r - tq$) histograms with RPDF dither spanning 0.35 quanta. (a) Zero order. (b) First order.

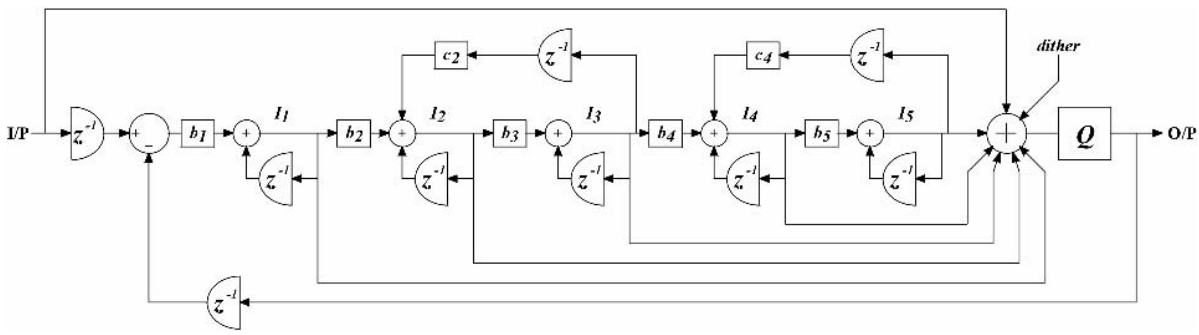


Fig. 15. z-domain description of fifth-order Sony FF SDM [26].

of the r^{th}-stage $P_r(z)$ is defined in the z domain with reference to Fig. 22 as

$$P_r(z) = \frac{g_1(r)k(r)\dfrac{(1-z^{-1})}{p(r)} + g_2(r)}{\dfrac{(1-z^{-1})^2}{p(r)^2} + k(r)z^{-1}\dfrac{(1-z^{-1})}{p(r)} + z^{-1}} \quad (17)$$

where each pair of integrator scale factors $p(r)$ sets the center frequency, $k(r)$ sets the Q factor, $g_1(r)$ the insertion gain of the band-pass filter, and $g_2(r)$ the insertion gain of the low-pass filter. The complete transfer function for N cascaded parametric filters $PT_N(z)$ is

$$PT_N(z) = \prod_{r=1}^{N}[1 + P_r(z)]. \quad (18)$$

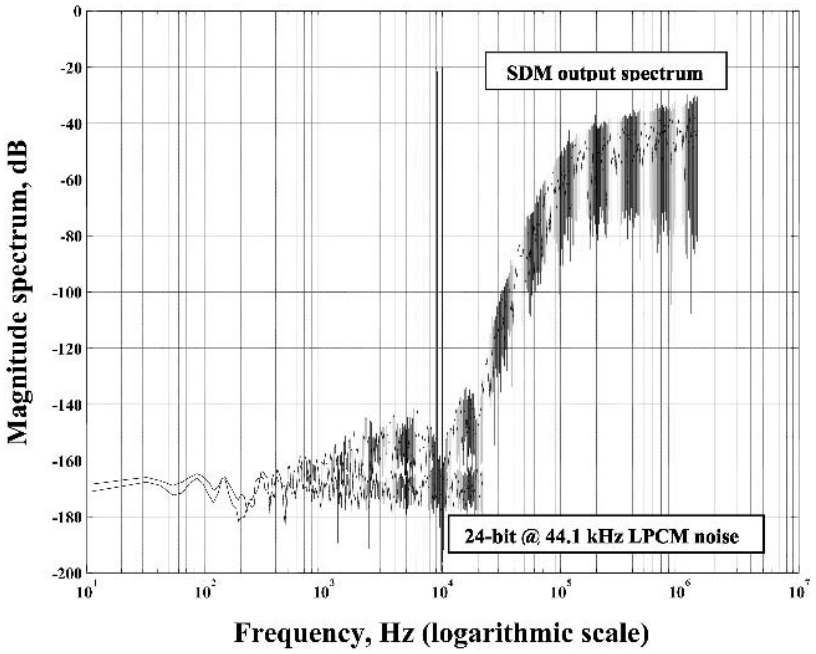


Fig. 16. Fifth-order Sony FF SDM versus 24-bit dithered LPCM reference.

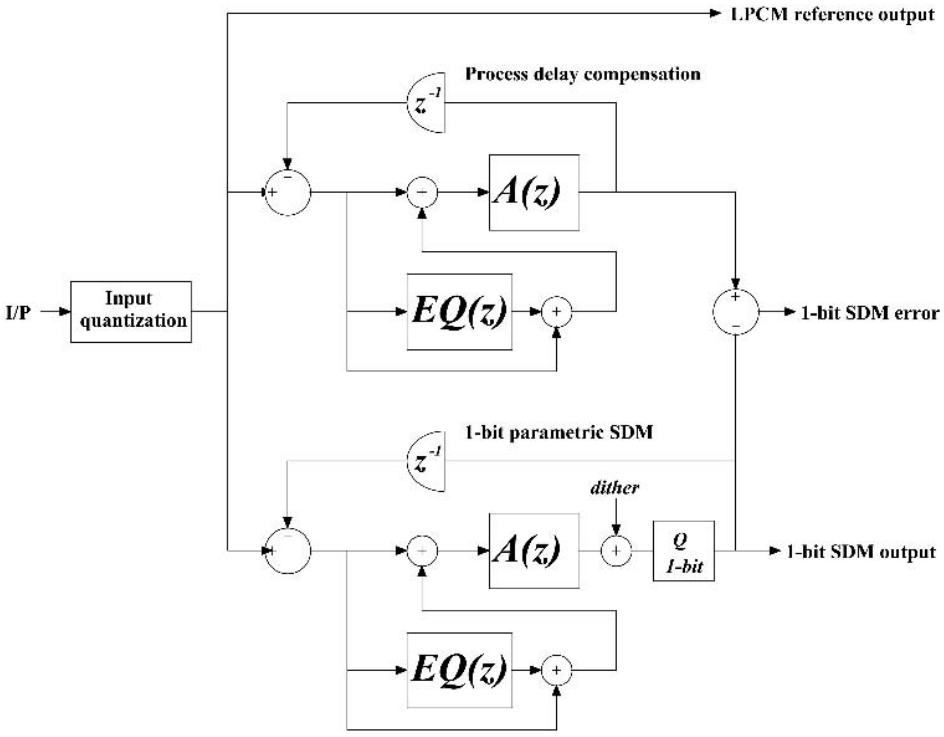


Fig. 17. Input–output SDM error comparison scheme.

Defining $A_P(z)$ as the transfer function of the principal integrators, the NSTF follows as

$$\text{NSTF} = \frac{1}{1 + z^{-1}A_P(z)PT_N(z)}. \quad (19)$$

Experimenting with a variety of NSTFs, where the degree of noise shaping is relatively mild results in stable SDM behavior with a low probability of instability. However, as noise shaping is made more aggressive and the signal level is increased, the probability of instability increases. To improve stability a "step-back in time" technique [6] was implemented. Here a set of criteria are established to interrogate the input to the two-level quantizer from which the onset of instability can be detected, allowing the coder to step back in time with the aid of a state memory to an earlier sample selected depending on pulse group activity. The SDM output at the selected sample is then inverted and a new dither sequence calculated, after which coding progresses again in the forward direction. This process can be repeated more than once if required but is designed to ensure that the statistics of the quantizer input are well behaved such that groups of multiple 1's and 0's in the SDM output are minimized, thus maximizing the 0 to 1 and 1 to 0 transitions; also the peak amplitude is limited. Consequently the probability for the quantizer input to attain a high value is reduced as it must oscillate close to the zero-level threshold. In practice the degree of step-

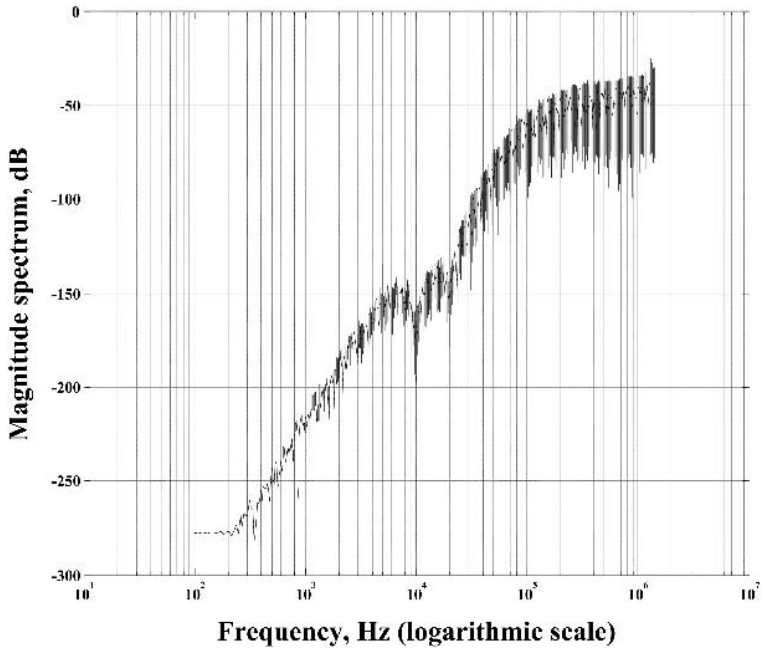


Fig. 18. Error spectrum between two-level SDM and equivalent loop without quantization.

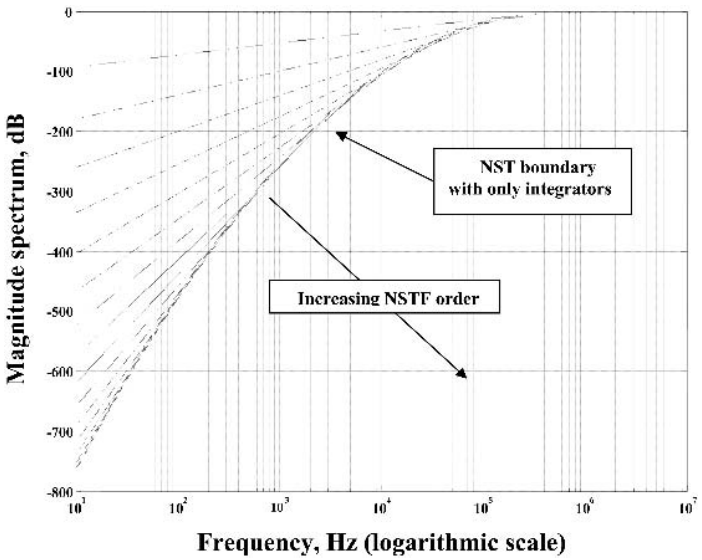


Fig. 19. Family of NSTFs for integrator-feedforward loop (no internal local feedback loops).

back activity proves to be strongly linked with the probability of instability and therefore provides a useful indicator on SDM coding performance.

An example NSTF was crafted for a parametric SDM designed to achieve extremely low in-band noise and distortion performance in the presence of a periodic input signal. This SDM incorporated four principal integrators and 13 parametric filters with center frequencies ranging from 20 to 39 kHz. The input consisted of three sine waves, each of amplitude 0.1 with frequencies of 4, 9, and 20 kHz, respectively, and 2^{18} samples were computed. The SDM output spectrum is shown in Fig. 23, where it can be

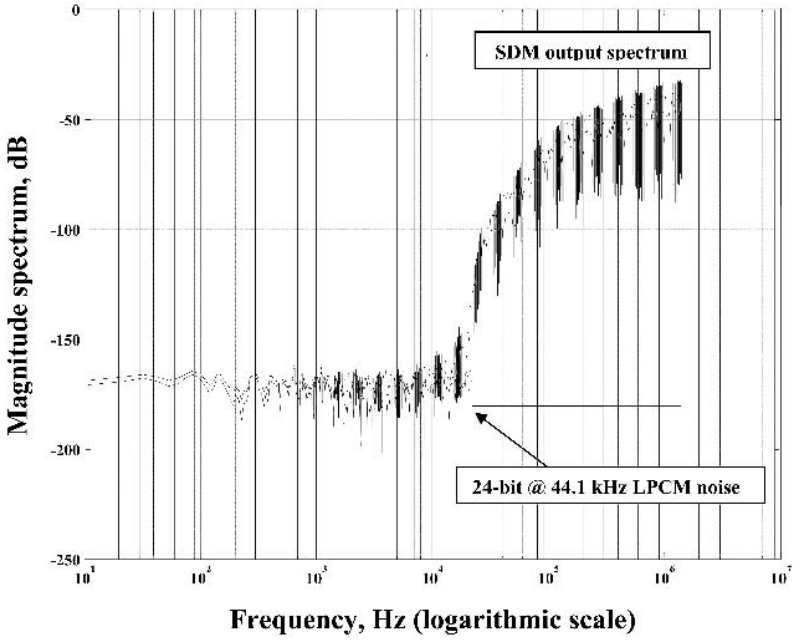


Fig. 20. Ninth-order Sony FF SDM against 24-bit dithered LPCM reference.

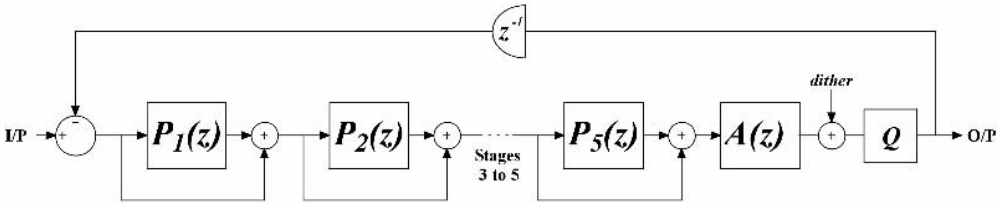


Fig. 21. Outline structure of five-band parametric SDM with filters $P_1(z)$ to $P_5(z)$.

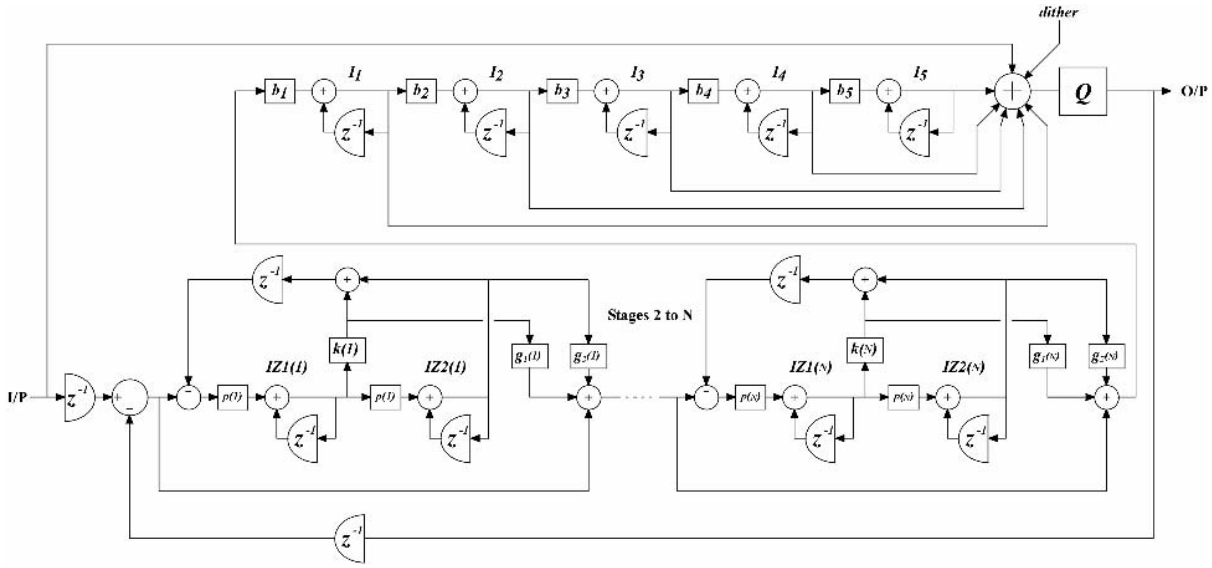


Fig. 22. z-domain N-band parametric SDM with five principal integrators.

concluded that appropriately designed SDM can yield almost zero coding artifacts in the audio band, especially as the low-noise band extends to just above 30 kHz, a result considered exemplary in the context of the SACD sampling rate. To further demonstrate the linearity of this coder, Fig. 24 shows the input–output error spectrum computed using a procedure similar to that described in Section 4.1. The result confirms a coding accuracy to better than 32-bit LPCM with an accuracy that, unlike the Sony FF topology, increases at a rate determined by the four principal integrators. Finally Fig. 25 shows the histogram of the input signal to the two-level quantizer, and the corresponding time-domain plot is shown in Fig. 26. A bounded and smooth distribution is revealed, although evidence of bifurcation about zero in the histogram is just evident, indicating that the selected NSTF is close to instability. This is also confirmed by the high degree of step-back activity observed during the simulation and recorded by the asterisk plots in Fig. 26. Until recently the author did not believe that an SDM coder could achieve such low levels of distortion; however, this result provides substantial evidence to counter earlier criticism.

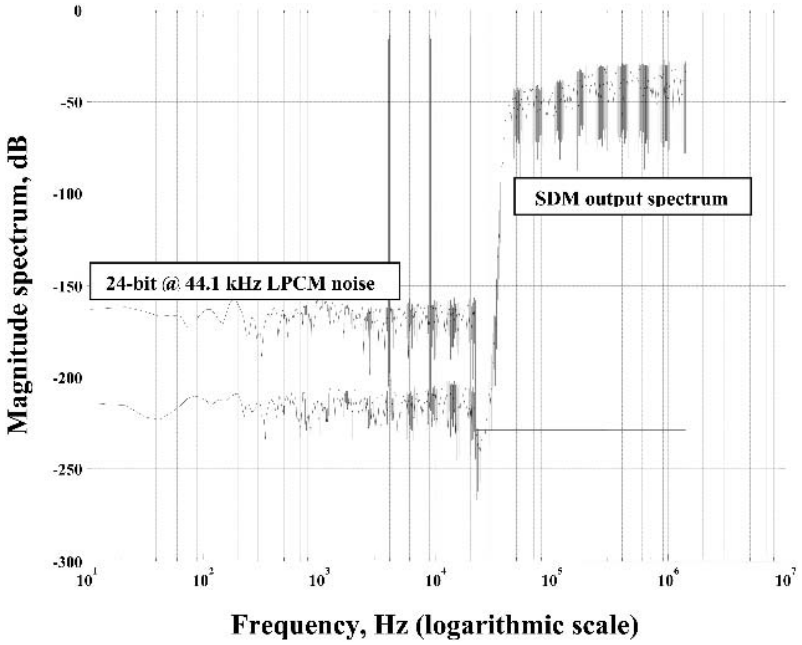


Fig. 23. Output spectrum of parametric SDM. Input quantized to 32 bit (24-bit at 44.1-kHz LPCM reference spectrum shown).

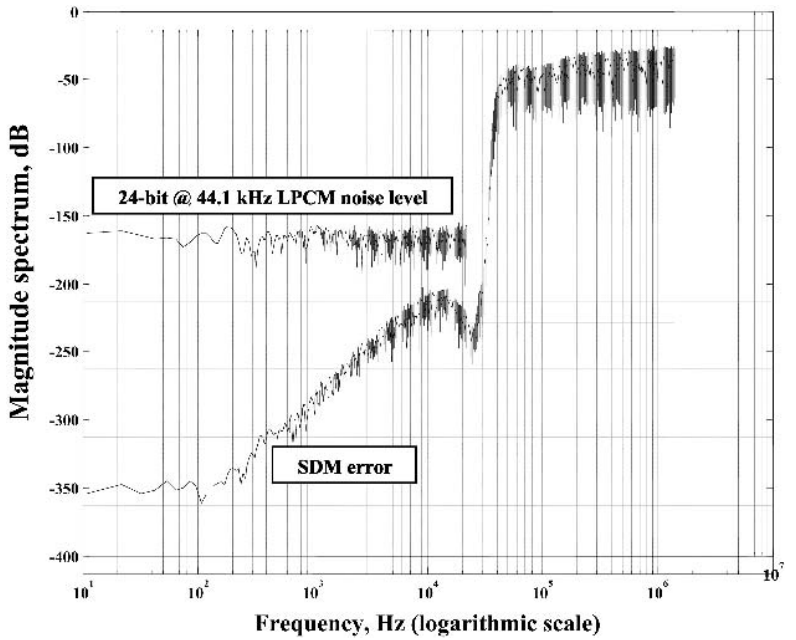


Fig. 24. Input–output error spectrum of parametric SDM. Input quantized to 32 bit (24-bit at 44.1-kHz LPCM reference spectrum shown).

6 CONCLUSION

A number of techniques have been investigated both for the generation of SDM code and to consider aspects of linearity. Further ideas were presented on the application of the LFM–SDM model, and the nature of the LFM spectrum was shown in simulated examples spanning a wide range of input levels. Of significance here is the extremely low level of LFM spectral spillage into the audio band together with the pivotal role of NTSIs. It was argued that if these time instants are used as starting points for forming time-domain quantized SDM, where time-domain TPDF dither is employed to randomize the process, then a high degree of linearity can be achieved. However, the limitation of the present scheme is that although noise shaping can be applied to the process, it is not formed in the signal domain, so the quantization noise spectrum is not minimized appropriately by noise shaping to give low noise in the audio band.

Generating noise is explained in terms of the LFM–SDM model by the deliberate introduction of timing jitter to the NTSI pulse locations. To reduce noise, a means of restricting extreme pulse displacement was considered by limiting the noise-shaping quantizer range. Also, a consequence of time-domain noise shaping is that coincident pulses can be formed. Both open-loop and closed-loop methods of preventing this condition were described to-

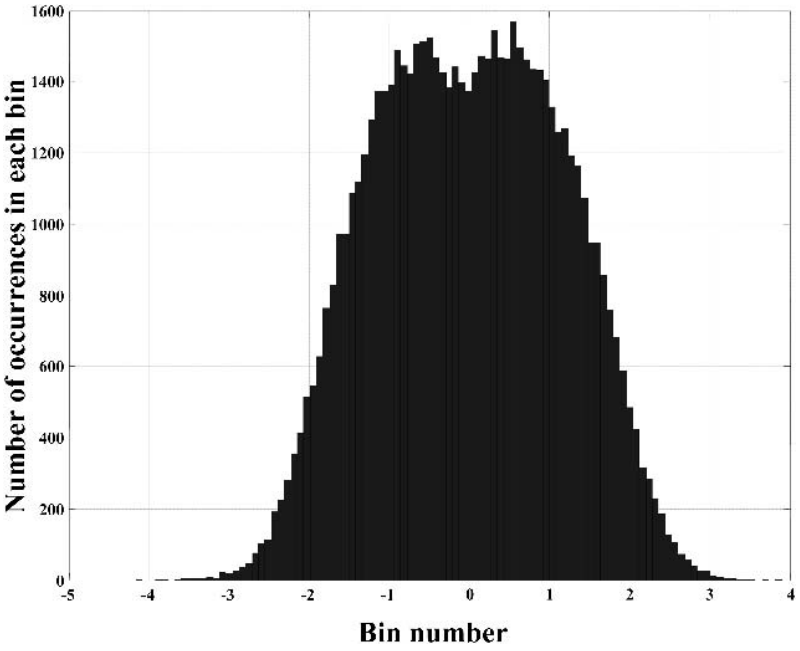


Fig. 25. Histogram of input signal to two-level quantizer in parametric SDM.

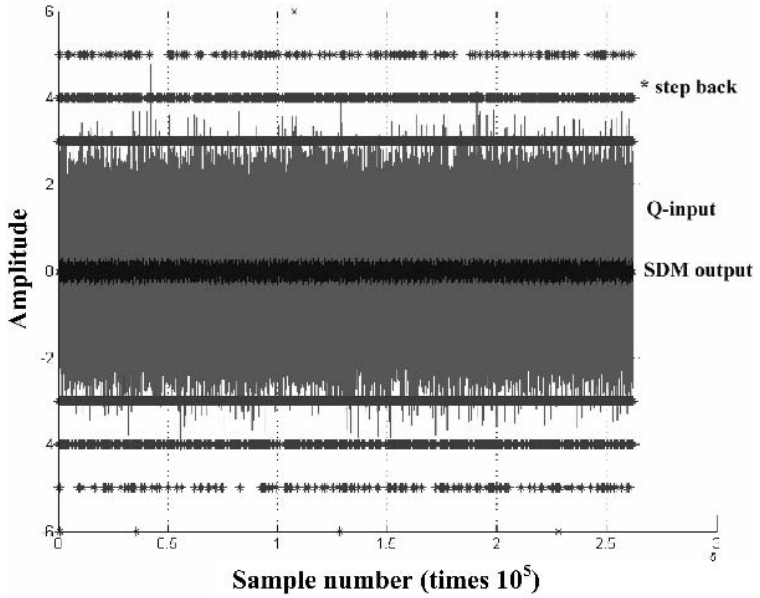


Fig. 26. Two-level quantizer input (grey), step-back instants (*), and SDM output (black) derived from parametric SDM simulation.

gether with a technique of offsetting the dither sequence. The effect of dither was studied, and it was shown that zero-order TDQ was most sensitive to dither statistics, whereas time-domain noise shaping offered greater tolerance to the type of dither.

A novel feature discussed in the paper was the identification of a stable multilevel to two-level mapping algorithm, which when applied within an iterative loop, would converge to the required amplitude resolution, provided the input signal was bounded to fall within the two-level range. Example spectra were presented to show the performance penalty.

An objective of the paper was also to consider the linearity of high-order SDM. A new form of parametrically equalized high-order coder was identified, which showed excellent coding performance within the band of 0–30 kHz. Linearity was investigated by calculating the difference spectrum between two similar coders with identical NSTFs, where the reference used no quantization whereas the other coder employed two-level quantization. Results showed that the error spectrum was shaped by the NSTF, whereas the parametric SDM coder achieved extremely low error levels within the audio band.

Although these results demonstrate that SDM is capable of high performance within the 0–30-kHz band with no significant noise and distortion, the rise in noise spectral density above 30 kHz remains of concern. An exploration using parametric SDM augmented by the step-back in time algorithm suggests that the attained performance is close to the theoretical performance boundary. Even if further reduction in selected regions of the noise spectrum is achieved, the probability of instability increases rapidly, especially at higher signal levels. It is suggested that even the Trellis algorithm may fail under such extreme noise shaping as valid signal trajectories are rendered illusive. In practice it may be prudent to choose a NSTF with a gentler rise in high-frequency noise, though at the expense of audio-band accuracy. Consequently the only means by which the just out-of-audio-band noise can be reduced is either by complementary preemphasis and deemphasis, a process that exacerbates high-frequency overload and coder stability, or preferably by an increase in the SDM sampling rate. It is suggested that an oversampling ratio of 128 rather than the just adequate factor of 64 would meet most criticisms in future high-resolution systems based on SDM.

7 REFERENCES

[1] J. E. Flood and M. J. Hawksford, "Exact Model for Deltamodulation Processes," *Proc. IEE* (*London*), vol. 118, pp. 1155–1161 (1971).

[2] M. J. Hawksford, "Unified Theory of Digital Modulation," *Proc. IEE* (*London*), vol. 121, pp. 109–115 (1974 Feb.).

[3] M. J. Hawksford, "Application of Delta-Modulation to Television Systems," Ph.D. Thesis, University of Aston, Birmingham, UK (1972).

[4] J. Verbakel, L. van de Kerkhof, M. Maeda, and Y. Inazawa, "Super Audio CD Format," presented at the *104th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 46, p. 570 (1998 June), preprint 4705.

[5] M. O. J. Hawksford, "Parametric SDM Encoder for SACD in High-Resolution Digital Audio," in *Proc. Inst. of Acoustics Conf. on Reproduced Sound-18,* vol. 24, pt 8 (2002 Nov.).

[6] M. O. J. Hawksford, "Parametrically Controlled Noise Shaping in Variable State Step-Back Pseudo-Trellis SDM," presented at the *115th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 51, p. 1222 (2003 Dec.), convention paper 5877.

[7] S. P. Lipshitz and J. Vanderkooy, "Why Professional 1-Bit Sigma–Delta Conversion Is a Bad Idea," presented at the *109th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 48, p. 1099 (2000 Nov.), preprint 5188.

[8] S. P. Lipshitz and J. Vanderkooy, "Towards a Better Understanding of 1-Bit Sigma–Delta Modulators," presented at the *110th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 49, pp. 544, 545 (2001 June), convention paper 5398.

[9] S. P. Lipshitz and J. Vanderkooy, "Towards a Better Understanding of 1-Bit Sigma–Delta Modulators—Part 2," presented at the *111th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*). vol. 49, p. 1231 (2001 Dec.), convention paper 5477.

[10] H. Kato, "Trellis Noise-Shaping Converters and 1-Bit Digital Audio," presented at the *112th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 516 (2002 June), convention paper 5615.

[11] E. Janssen and D. Reefman, "Advances in Trellis-Based SDM Structures," presented at the *115th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 51, p. 1257 (2003 Dec.), convention paper 5993.

[12] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and Dither: A Theoretical Survey," *J. Audio Eng. Soc.,* vol. 40, pp. 355–375 (1992 May).

[13] M. O. J. Hawksford, "Transparent Differential Coding for High-Resolution Digital Audio," *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 49, pp. 480–497 (2001 June).

[14] L. Risbo, "Σ–Δ Modulators—Stability Analysis and Optimization," Ph.D. Thesis, Technical University of Denmark (1994 June).

[15] D. Reefman and E. Janssen, "DC Analysis of High Order Sigma–Delta Modulators," presented at the *113th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 969 (2002 Nov.), convention paper 5693.

[16] D. J. Naus, E. C. Dijkmans, E. F. Stikvoort, A. J. McKnight, D. J. Holland, and W. Bradinal, "A CMOS Stereo 16-Bit Converter for Digital Audio," *IEEE J., Solid-State Circuits,* vol. SC-22 (1987 June).

[17] M. O. J. Hawksford, "Chaos, Oversampling, and Noise Shaping in Digital-to-Analog Conversion," *J. Audio Eng. Soc.,* vol. 37, pp. 980–1001 (1989 Dec.).

[18] A. I. Zayed, *Advances in Shannon's Sampling Theory* (CRC Press, Boca Raton, FL, 1993).

[19] M. O. J. Hawksford, "Dynamic Model-Based Linearization of Quantized Pulse-width Modulation for Applications in Digital-to-Analog Conversion and Digital Power Amplifier Systems," *J. Audio Eng. Soc.,* vol. 40, pp. 235–252 (1992 April).

[20] M. O. J. Hawksford, "Linearization of Multilevel, Multiwidth Digital PWM with Applications in Digital-to-Analog Conversion," *J. Audio Eng. Soc.,* vol. 43, pp. 787–798 (1995 Oct.).

[21] C. Dunn and M. O. J. Hawksford, "Is the AES/EBU/SPDIF Digital Audio Interface Flawed?," presented at the *93rd Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 40, p. 1040 (1992 Dec.), preprint 3360.

[22] J. A. S. Angus, "Effective Dither in High-Order Sigma–Delta Modulators," presented at the *111th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 48, p. 1231 (2000 Dec.), convention paper 5478.

[23] M. O. J. Hawksford, "SDM Versus LPCM: The Debate Continues," presented at the *110th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 49, p. 545 (2001 June), convention paper 5397.

[24] M. O. J. Hawksford, "Time-Quantized Frequency Modulation with Time Dispersive Codes for the Generation of Sigma–Delta Modulation," presented at the *112th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 516 (2002 June), convention paper 5618.

[25] C. Dunn and M. Sandler, "A Comparison of Dithered and Chaotic Sigma–Delta Modulators," *J. Audio Eng. Soc.,* vol. 44, pp. 227–244 (1996 Apr.).

[26] H. Takahashi and A. Nishio, "Investigation of Practical 1-Bit Delta–Sigma Conversion for Professional Audio Applications," presented at the *110th Convention of the Audio Engineering Society, J. Audio Eng. Soc.* (*Abstracts*), vol. 49, p. 544 (2001 June), convention paper 5393.

[27] S. J. Orfanidis, "Digital Parametric Equalizer Design with Prescribed Nyquist Frequency Gain," *J. Audio Eng. Soc.,* vol. 45, pp. 444–455 (1997 June).

APPENDIX 1

MATLAB ROUTINE TO SEARCH FOR NTSI USING THE LFM MODEL

The following Matlab subroutine was used in Section 1 to perform the iterative search for each NTSI:

```
% search for PSZC
sd-.5*(1+sign([sign(sm(2:L))-
  sign(sm(1:L-1))0]-.1));

%sort approximate PSZC locations to
  determine their time coordinates
[p q]=sort(sd.*(1:L));
[mx my]=max(q(Lx/2:L));

%zr is a vector that defines the
  sample number of the sample just
  preceding a PSZC
zr=q(my+Lx/2:L).*sign(p(my+Lx/2:L));
clear p q mx my

%stage 1: linear interpolation to
  improve time estimate of PSZC
tr=zr-sm(zr)./(sm(zr+1)-sm(zr));

%stage 2: iterative error correction
  for PSZCs
for x=1:100
zerror=cos(g*(tr*dt-(a1/(h1*w0))*
  cos(h1*w0*tr*dt)-(a2/(h2*w0))*
  cos(h2*w0*tr*dt)));
tr=tr-1.5*zerror;
end
```

APPENDIX 2

MATLAB PROGRAM TO COMPUTE SDM CODE USING G-CASCADED INTEGRATORS

The following routine used in Section 2.5 employs compact code that exploits the Hankel matrix to achieve feedforward stability compensation within the forward path of a *G*-integrator noise shaper. In this routine *ax* is the input vector of length *L, G* is the loop order, and sdm the output vector.

```
% multi-level multi-integrator noise
  shaper: Hankel matrix routine
I = zeros(1,G);
for n = 2:L
er = ax(n) - sdm(n-1);
I(1:G) = I(1:G) + er;
I = sum(hankel(I));
sdm(n) = round(I(1) + dither(n));
end
```

THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and he studied delta modulation and sigma-delta modulation (SDM) for color television applications. During this period he also invented a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, Colchester, UK, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems, with a strong emphasis on audio systems including signal processing and loudspeaker technology. Since 1982 his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. The first one was developed in 1986 for a prototype system to be demonstrated at the Canon Research Centre and was sponsored by a research contract from Canon. Much of this work has appeared in *JAES,* together with a substantial number of contributions at AES conventions. He is a recipient of the AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," for the best contribution by an author of any age for *JAES,* volumes 45 and 46.

Dr. Hawksford's research has encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with special emphasis on SDM and its application to SACD technology. In addition, his research has included the linearization of PWM encoders, diffuse loudspeaker technology, array loudspeaker systems, and three-dimensional spatial audio and telepresence including scalable multichannel sound reproduction.

Dr. Hawsford is a chartered engineer and a fellow of the AES, IEE, and IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA). He is also a technical consultant for NXT, UK and LFD Audio UK and a technical adviser for *Hi-Fi News* and *Record Review.*

Parametrically controlled noise shaping in variable state-step-back pseudo-Trellis SDM

M.O.J. Hawksford

Abstract: Progress is reported in parametrically controlled noise shaping sigma delta modulator (SDM) design. As this SDM structure can provide a higher SNR than normal SDM structures, Philips Research Laboratories questioned whether further improvement could be obtained using techniques inspired by the Trellis SDM. Simulations are used here to illustrate the performance of a parametrically controlled pseudo-Trellis SDM. The technique uses uniquely a variable state step-back approach to mediate loop behaviour that is shown to achieve robust stability in the presence of aggressive noise shaping and high level signals. Comparisons are made with traditional SDM structures and LPCM systems for high-resolution audio applications.

1 Introduction

This paper presents a novel approach to the generation of sigma-delta modulation (SDM) [1–3] code as used at the core of the high-resolution audio format designated super audio CD (SACD) [4]. The discussion is limited to a purely digital implementation configured as a source coder where a high-resolution digital audio signal [5–7] is converted to the 1-bit uniformly sampled data stream mandated by SACD. Specifically the problem of instability [8] is addressed for cases where the closed-loop transfer function is designed for extreme noise shaping together with a high-amplitude input signal that may contain significant ultrasonic components, since it is a feature of SACD that input filtering is not required.

The classic problem of stability is exacerbated in SDM by the presence of binary quantisation within the feedback loop [3] since the normalised output code is only 1 or −1. In relaxed noise shaping applications, and where an extended multilevel quantiser is used [9, 10] then loop stability can be achieved by considering just the standard Nyquist stability criteria and treating the quantiser as a unity gain stage. However, the use of highly constrained two-level quantisation places an additional limitation on the loop behaviour by restricting the number of possible output signal combinations (or trajectories) available within a given time frame. It follows that some output signal bit patterns do not lead to convergence and result in instability, although such patterns are themselves a function of the state of the loop and the input signal. This problem becomes aggravated when the loop is designed to achieve a high degree of noise shaping and also where the input signal demands a high density of either 1 or 0 output pulses. This latter condition limits further the number of available code combinations where ultimately, total instability results when no valid codes remain. Taking this perspective, stability can be related to the identification of valid output code sequences produced within the natural operation of the feedback loop that do not lead to irreversible divergence, observed typically at the input Q_{in} of the quantiser. In conventional SDM [3, 4] where encoding moves forward naively without regard to conditional loop stability, occasionally a binary bit pattern occurs that drives the SDM coder in such a way that feedback correction is impossible. Although, the probability of such instability is difficult to predict, its occurrence can be made extremely low providing the input signal level is not too high and the noise shaping transfer function (NSTF) not too demanding in terms of suppressing in-audio band noise levels. Conventional SDM encoders [6] found at the core of SACD [4] and even enhanced topologies designed to improve linearity [7] normally use continuous and sequential coding supplemented perhaps by a state-reset function activated on detection of instability. However, coding methods that do not seek out stable signal trajectories severely compromise performance especially as best performance in terms of noise shaping and signal handling is achieved often when the probability of instability is relatively high.

Recently, a new approach to SDM termed Trellis [11] has been introduced that embeds a degree of look-ahead computed over an M-sample window. The method calculates a cost function by evaluating the available 2^M code combinations taken over this window and then selects the code that exhibits lowest error. Such an approach yields theoretically the best possible result for a given loop filter and window length, although a full Trellis implementation has a high computational overhead. Since publication of this seminal paper [11] there have been important refinements reported for improving coder efficiency notably in research from Philips Research Laboratories [12, 13] and by Angus [14, 15]. Nevertheless, it appears unproven that in determining whether a particular code segment is optimum that a change in window length or a small change in input signal level may impact upon the computed result.

In this paper a modified Trellis-like algorithm is described that although not as rigorous as a fully populated Trellis implementation, nevertheless is shown capable of extremely accurate coding together with robust stability and tight control of the error statistics as observed at the quantiser input. The coder embeds a state memory together with a variable step-back in time procedure that is activated

IEE Proceedings online no. 20051076

doi: 10.1049/ip-vis:20051076

Paper first received 31st October 2003 and in final revised form 6th August 2004

The author is with the University of Essex, Colchester, Essex, CO4 3SQ, UK

E-mail: mjh@essex.ac.uk

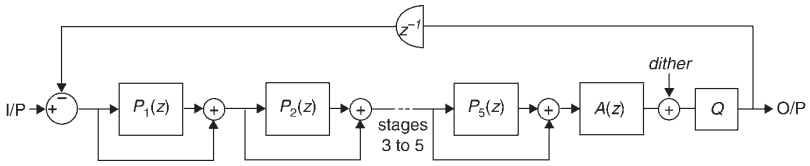


Fig. 1 *Basic SDM topology with series parametric noise shaper stage*

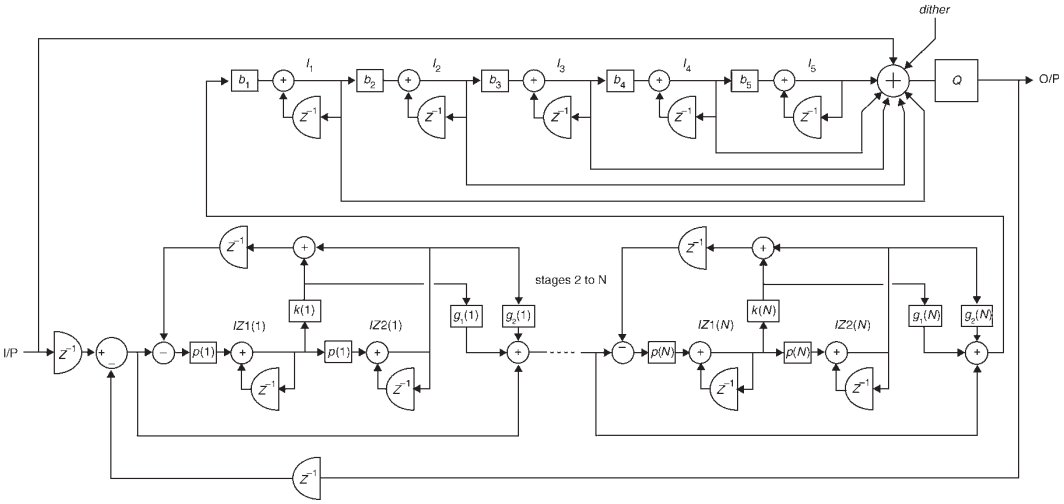


Fig. 2 *Parametric SDM topology showing series parametric stages and principal integrators*

only when certain error criteria are breached. The efficacy of this strategy is evaluated using *Matlab* [Note 1] simulations, which attempt to approach the theoretical performance boundaries in terms of output noise spectrum, low nonlinear distortion performance and high-level input signal handling. In this paper the standard SACD sampling rate f_{sdm} of $64f_s$ is used where $f_s = 44.1$ kHz. However, whereas Trellis [11] employs a fixed analysis time window and seeks to evaluate the best path available, the step-back algorithm uses a wide range, variable step-back procedure. Also, because the step-back procedure is variable, it is not possible to make a direct comparison with formal Trellis other than to judge the performance of the encoder holistically in terms of distortion, noise shaping and signal handling. Consequently, this paper seeks to demonstrate coding performance using the step-back algorithm in the context of a parametrically motivated noise shaper [16], which can then be compared for example, against results from Philips Research Laboratories [12, 13], Angus [14, 15], Sony [6] and Lipshitz and Vanderkooy [17–21]. Results are presented for DC, multi-tone AC and bandlimited random noise input sequences.

2 SDM with parametric noise shaper

Parametric SDM has been reported [16, 22] previously as a method of attaining a fine degree of control over the closed-loop transfer function. The classic noise shaping SDM architecture is shown in Fig. 1 that is extended here to include an additional multistage cascaded equaliser comprising N parametric stages $P_1(z)$ *to* $P_N(z)$ in addition to the principal integrators [22] with feedforward stabilisation. Each stage $P_r(z)$ is a biquadratic filter section where the output is derived from a weighted sum of the two integrator outputs, where this configuration allows both lowpass and bandpass filtered components to be introduced individually into the NSTF. Figure 2 shows the z-domain SDM topology that reveals the structure of the parametric equaliser stages.

Inspection of this SDM topology shows that if parametric filters $P_1(z)$ to $P_N(z)$ have zero insertion gain then because of the local feedforward paths around each filter stage, the SDM forward path transfer function equates exactly to that of the cascaded principal integrators. The series parametric equaliser selected here consists of N second-order filters with resonant frequency, Q-factor (i.e. bandwidth) and insertion gains for both bandpass and lowpass outputs independently pre-selected. The rth biquadratic transfer function $P_r(z)$ is given as

$$P_r(z) = \frac{g_1(r)k(r)\frac{(1-z^{-1})}{p(r)} + g_2(r)}{\frac{(1-z^{-1})^2}{p(r)^2} + k(r)z^{-1}\frac{(1-z^{-1})}{p(r)} + z^{-1}} \quad (1)$$

where integrator scale factor $p(r)$ sets the resonant frequency, $k(r)$ sets the Q-factor and $g_1(r)$ and $g_2(r)$ the respective insertion gains of the bandpass and lowpass parametric responses. The overall transfer function of the series parametric equaliser $PT_s(z)$ then follows as,

$$PT_s(z) = \prod_{r=1}^{N}(1 + P_r(z)) \quad (2)$$

The NSTF of the parametric SDM then has the form,

$$NSTF = \frac{1}{1 + z^{-1}A_p(z)PT_s(z)} \quad (3)$$

where $A_p(z)$ is the transfer function of the principal integrators including their local feedforward paths as shown in Fig. 2. Also, by embedding a one-sample delay in the input path and feeding forward the input signal directly to the quantiser input, then analysis reveals this topology to have a signal transfer function of precisely unity irrespective of the number of parametric stages and principal integrators.

Note 1: *Matlab* is the trade name of MathWorks Inc.

Two major advantages of this topology are that the number of principal integrators is independent of the order of the parametric equaliser and that the parametric equalisers do not compromise the loop gain at low frequency unlike SDM using embedded second-order resonators [6]. It has been shown [16, 22] that increasing the number of principal integrators above five offers little advantage because the noise shaping advantage is only realised at progressively lower frequencies. However, the parametric stages allow the loop gain to be tailored such that they boost the NSTF selectively in the upper range of the audio spectrum. Also, it is shown in Section 4 that aggressive noise shaping transfer functions may be used, where if conventional SDM was employed without step-back correction, the probability of instability is high.

3 SDM output bit patterns and pre-quantiser error statistics

The statistics of both the input signal Q_{in} and output pulse sequence of the two-level quantiser used in SDM reveals useful performance data and especially information on stability. For example, if the coder is approaching instability then Q_{in} exhibits momentary high values together with extended bursts of all-1 or all-0 pulses. To illustrate this observation a standard Sony-FF coder [6] is simulated with a peak input signal of level 0.5 to increase the probability of instability. The classic NSTF of this encoder is illustrated in a later simulation example in Fig. 12 while a short Q_{in} signal segment is shown in Fig. 3. Although the NSTF of the Sony FF SDM is relatively relaxed, the high value of input signal is sufficient to ensure a frequent occurrence of instability. For example, by inspecting Q_{in} in Fig. 3 around sample number 160, an event is indicated where instability almost occurs that is characterised by reduced positive-to-negative oscillation, a local broadening of the waveform and an increase in its amplitude. Just prior to this event a number of more minor anomalies can also be observed. If the corresponding output bit pattern, i.e. sign(Q_{in}), is examined then extended groups of 1 pulse must follow. Consequently, by observing the output bit pattern and in particular determining when an extended burst of all-1 or all-0 pulses occur, the proximity of instability and conditions for its detection (see Section 4) can be determined. Also, additional information can be extracted from Q_{in}, where its maximum amplitude can be moderated by feeding information into the step-back control procedure.

It follows that as the loop filter order is increased and especially with more aggressive noise shaping that in linear terms moves closer to the Nyquist stability bound, then the greater becomes the probability of instability. In an absolute sense SDM always has an indeterminate or fuzzy stability bound that is a consequence of the gross non-linearity introduced by quantisation. Thus setting just a linear bound does not guarantee stability, it is therefore more prudent to refer to the probability of instability, which in turn can be related to the rate of occurrence and growth of higher order pulse groups. Of course it follows that the higher the input signal level the more frequent higher order pulse groups become, where an earlier estimation was made [23] based upon a model of linear frequency modulation (LFM) [24, 25]. Although the basic LFM model excludes the noise-like behaviour of a typical high-order SDM, which adds a random dimension to the occurrence of multiple pulse groups, it does allow threshold input levels to be determined where specific pulse groups become mandatory. These simplified threshold calculations are repeated here:

Consider a process of LFM having centre frequency one half the SACD sampling rate $0.5f_{sdm}$ Hz and instantaneous frequency f Hz where the peak amplitude normalised input signal is $m(t)$, thus

$$f = 0.5f_{sdm}\{1 + m(t)\} \tag{4}$$

A pulse group of length λ samples occurs when λ positive slope zero-crossings (PSZC) fall just within λ consecutive time slots defined by the SACD sampling rate. The input threshold level IT(λ) corresponds to the input signal level $m(t)$ that just produces an output pulse group of λ consecutive 1s or 0s. In Fig. 4 this condition is illustrated for the case where $\lambda = 3$.

Observing the relationship between the instantaneous LFM period and time slots determined by the SACD sampling rate, the condition for the generation of a group of λ pulses occurs when the time between λ consecutive PSZCs, corresponding to $(\lambda - 1)$ cycles of the LFM output, is just less than λ time slots. Now λ time slots occupy a period λ/f_{sdm} as time slots occur with frequency f_{sdm} Hz and $(\lambda - 1)$ cycles of the LFM has a time $(\lambda - 1)/f$, then substituting for frequency from (4), it follows that

$$\frac{\lambda - 1}{0.5P(1 + IT(\lambda))} = \frac{\lambda}{f_{sdm}} \tag{5}$$

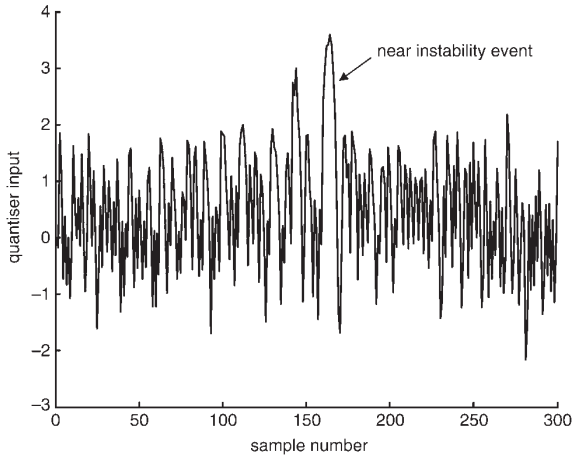


Fig. 3 *Time domain input Q_{in} of SDM quantiser showing an event close to instability*

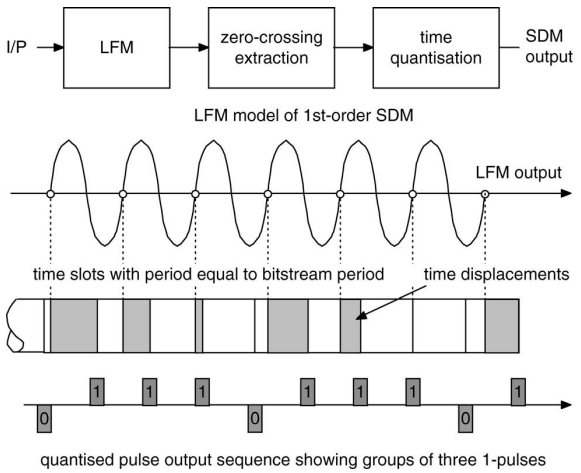


Fig. 4 *LFM model showing onset of $\lambda = 3$ pulse groups*

from which the threshold levels IT(λ) are derived as

$$IT(\lambda) = \frac{\lambda - 2}{\lambda} \tag{6}$$

giving the onset of group 2, 3 and 4 pulse groups as

$$IT(2) = 0 \qquad IT(3) = \frac{1}{3} \qquad IT(4) = \frac{1}{2}$$

These estimates show that for first-order SDM, groups of three pulses do not occur until the input equals and exceeds a normalised input of 1/3 and for groups of four, this is increased to 0.5. Hence, in the step-back algorithm described in Section 4, only groups of three or more pulses are detected as these become mandatory only at high input signal level and where the probability of instability is increasing. Also, it is shown in Section 4 that the step-back algorithm does not necessarily exclude larger pulse groups, only that the system becomes alerted if their presence is detected, additional conditions have to be met to instigate a step-back event.

In this SDM coder, instability was prevented by incorporating additional high-level thresholds in the quantizer, so if demanded multilevel output code is produced. Although infrequent, multilevel signals can be transformed to binary using for example time dispersive code [10, 22], Section 4 develops a more accurate approach of using the step-back algorithm to seek out stable sequences without compromising the noise shaping properties of SDM.

4 Step-back SDM algorithm

If a more aggressive noise shaping characteristic is required than that offered by the Sony FF encoder [6] then instability inevitably occurs with an even greater probability at higher input signal levels; consequently a stable coding method is required that does not employ either output pulse correction or state reset since both these processes degrade noise performance.

The approach followed here is to allow the natural performance of the closed loop to always dictate encoding, where this has the advantage that it is potentially fast if the probability of instability is low. Nevertheless, it is recognised that occasionally the coder will follow a blind alley from which there is either excessive error or, in the case of instability, there is no return as the degrees of freedom within the binary output sequence are too limited. In performing typical simulations of high-order loops this behaviour is generally observed where occasionally the coder would crash and require a forced reset or a repeat of the simulation, whereon using the same input signal a successful computation can be obtained. In the proposed step-back encoder strategy, if a blind alley is detected then the following sequence of events is activated:

- the encoder steps-back in time and takes earlier data from the state memory;
- the corresponding output sample is inverted by flipping 1 to 0 or 0 to 1;
- a new quantiser input dither sequence (see [22] for discussion and references on dither) is then applied to randomise the process;
- sequential coding then progresses in a forward direction.

This process can be repeated as many times as necessary where the depth of step-back is controlled as a function of the severity of the problem. In all cases the quantiser input Q_{in} remains properly controlled by feedback although it is monitored principally through the output pulse sequence of the SDM encoder. A critical aspect of this process is that by recalculating a new dither sequence and controlling the step-back process so this represents a full and proper coordination of all SDM state variables, forward progress is always seen as a natural progression without any coding discontinuity. This continuity condition means there is no deterioration in the coder performance and corresponding degradation in the output spectrum because this is controlled completely by the loop filter. The process recognises there are many valid paths that can be used in forming the output code where it chooses to select a path that avoids poor behaviour as observed in the input signal of the quantiser; see for example Fig. 3. In a full Trellis implementation [11] the approach taken is to calculate integrated errors for all the paths taken over a fixed window and to select the one with the lowest error. The problem here is that by changing the window length by just one sample, may lead to a different result. Also, it is conjectured that provided the behaviour of the quantiser input signal is bounded in peak value and does not exhibit significant gaps of the type shown in Fig. 3, then normal closed-loop behaviour achieves valid results. Consequently, step-back correction is used as a means of maintaining stability and is not a formal part of the NSTF. On the other hand, a Trellis implementation attempts to both forge stable performance and try to lower the output noise. In the step-back algorithm these aspects are segregated with improvements in noise shaping being applied to the closed-loop filter and then the step-back procedure being used to steer a stable path through the encoding procedure when it is required.

Hence, with low-demand loop filters the step-back algorithm is used as an alternative to a forced reset and is active only occasionally. However, with more aggressive noise shaping and where high-level signals are encountered, step-back operates much more frequently and consequently slows down the coding performance. Nevertheless, provided a coding path can be identified, the loop in effect operates continually with minimal noise modulation.

The condition to activate a step-back procedure is detected indirectly by observing the output bit pattern. It is argued that the occurrence of 1 to 0 and 0 to 1 transitions should be frequent and bursts of all-1 and all-0 pulses minimised. If bursts of identical valued pulses are observed in terms of the input signal to the loop quantiser, then often it follows that this signal shows a divergent behaviour that if too persistent can cause instability. While the SDM encoder is operating, groups of pulses are detected and the following conditional functions applied to detect and activate a step-back procedure. Let the variable *back* be set to control the state step-back range that is set typically to 3 (with a maximum value of 10). Thus the value of *back* can range from 3 to 10 and controls each incremental step-back. However, it should be noted that the algorithm allows multiple and sequential step-back procedures so a deep state variable memory is used to store the past encoder states, where the state variable memory depth is typically set to 256. Also, if the input level is high then step-back depth is increased according to (3) to accommodate naturally occurring longer pulse groups.

Stage 1

If the nth SDM output sample is $sdp(n)$ (where $sdp(n) = 1$ or -1) and the binary quantiser input Q_{in} is $ss(n)$ such that

$$sdp(n) = \text{sign}(ss(n))$$

then a variable *errdet* is calculated by summing the SDM output over an interval *back* as

$$errdet = \text{sum}(sdp(n - back + 1 : n))$$

Stage 2
To activate the step-back procedure three conditions must be passed:

if abs($errdet$) > back − 1 *condition 1*
if abs($ss(n)$) > abs($ss(n-1)$) *condition 2*
if sign($ss(n)$) = sign($ss(n-1)$) *condition 3*

The SDM coder then steps back in time by *back* samples by reloading the appropriate variables from the state memory shift register. However, after step-back the new current output code undergoes a forced inversion in an attempt to drive the loop filter in the opposite direction and thus prevents excessive build up of Q_{in}. The loop then progresses normally although a new value of dither is used to introduce a stochastic element into the process.

Studying the three conditions reveals the methodology of this process. Initially, condition 1 identifies an excessive number of like-valued pulses within a block of size *back* samples. Secondly, if the current value of $Q_{in}(n)$ is greater than the previous value $Q_{in}(n-1)$ this indicates divergent behaviour while the final criteria seeks to identify whether the current and past samples have the same sign. Hence, the test fails to activate a step-back procedure if a current sample is greater than the previous sample but has opposite sign, as this would signal either a 1 to 0 or 0 to 1 transition, which is good behaviour in SDM. It is observed that instability is more likely to occur when a decision to maintain a burst of like valued pulses is made where a small change in $Q_{in}(n)$ would have caused a pulse of opposite sign. The effect of applying these conditions and the consequence of the step-back process with corrective sample inversion are shown in two examples in Fig. 5.

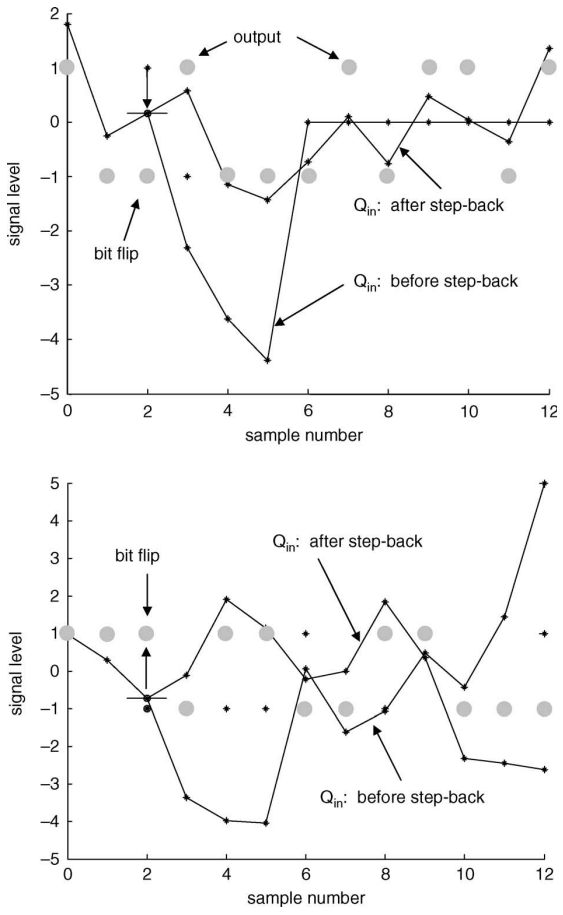


Fig. 5 *Two examples of Q_{in} and SDM output pulse sequence before and after step-back*

In each example (Fig. 5) correction takes place at the indicated sample number 2. Observe how at sample 2 the value of Q_{in} is close to the 0-level quantiser threshold while at samples 3, 4 and 5 relatively high levels are attained producing a burst of like-valued output pulses. The key to correction is that under these conditions, a complementary decision should have been made at sample 2. Consequently, in the step-back in time procedure the state of the system returns to sample 1, a new dither sequence is generated, the binary output at sample 2 inverted compared to the previous pass followed by the resumption of forward encoding. The two examples shown in Fig. 5 illustrate the effect of SDM output pulse inversion together with subsequent successful progression in forward encoding, where in each case a bit flip is shown at sample 2.

As forward encoding progresses the conditions for step-back are always applied. Thus there may be just one step-back event or indeed under extreme encoding there can be many. If multiple step-back events are encountered then provision is made to increment the degree of step-back (set by variable *back*), where experiment has converged on a limited range 11 > *back* > 2. Thus with successive step-back activity the value of *back* is incremented upwards, while when a successful forward encoding progression is achieved, *back* is reset to its minimum value of typically 3.

Stage 3
In addition to the three conditions defined in Stage 2, an extra condition is imposed that takes control of the variable step-back parameter if the magnitude of Q_{in} exceeds a defined threshold (typically set to 5). Under this latter condition state step-back is set to maximum and then the step-back variable reset to its initial condition. This process acts so as to limit the peak magnitude of Q_{in} especially under conditions where instability is probable. The idea here is to step back over an extended temporal region where poor encoding is experienced and then progress once more in a forward direction now using new dither values over a range of samples. As such, new signal trajectories are discovered that are likely to allow feedback control to steer a well-behaved coding path. Note that this process can be repeated as often as necessary and combined with the shorter more normal step-back procedure, so if there is an extreme coding condition, the random nature of the dither and variable step-back options allows new signal paths to be searched. It is important to observe that when successful encoding progresses in a forward direction, it is seen as continuous with no resets or discontinuities in the state variables. As such modulation artefacts are avoided.

5 SDM encoder performance

The step-back SDM encoder was evaluated using a Matlab simulation performed over a discrete block of data to generate the 1-bit output code. The input to the SDM routine was 32-bit LPCM data that was generated initially at 44.1 kHz sampling, then upsampled by a factor of 64 (to match standard SACD practice) and finally re-quantised with appropriate dither to 32 bit resolution. As a benchmark a 24-bit random noise sequence, again sampled at 44.1 kHz was generated and upsampled by 64 times. This noise spectrum could then be displayed alongside the spectrum of the SDM output sequence to enable an immediate comparative interpretation of coding performance with

respect to LPCM. In all spectral calculations a Blackman-squared window was applied in the time domain prior to the application of the Fourier transform to deal with the non-periodic nature of SDM output code even when the input signal is periodic.

Three sets of simulations were performed using an input sequence composed of three equal amplitude input sine waves of amplitudes 0.1 and of frequencies 4 kHz, 9 kHz and 20 kHz and represent a peak modulation index close to 0.3 which is considered high although not extreme. The analysis frame was set at 2^{16} samples where the SDM sampling rate is 2.8224 MHz. Dither was applied at the SDM quantiser input with weighting 0.35 (i.e. rectangular PDF noise with a range −0.35 to 0.35, see [22]); this was used to break down residual limit cycles and also to add an element of randomness to the step-back process. Each simulation used a NSTF ranging in form from relatively mild through to aggressive, where the latter was positioned close to the edge of instability and attempted to achieve the widest possible frequency band with extremely low levels of quantization noise in the frequency band 0 to 20 kHz. The three selected theoretical NSTFs are shown in Fig. 6. The corresponding performance simulations are presented in Figs. 7–9 to where the SDM with a mild NSTF has five principal and five parametric stages, the more aggressive NSTF has four principal and ten parametric stages while the aggressive NSTF has four principal and 13 parametric stages. The loop filter coefficients used in each SDM simulation are given in the Appendix, Section 9, and the computed results consist of three graphs:

- Figs. 7*a*, 8*a* and 9*a* present the SDM output spectra and include a 24-bit@44.1 kHz LPCM noise spectrum to facilitate the interpretation of absolute noise levels.
- Figs. 7*b*, 8*b* and 9*b* show quantiser input Q_{in} input signal where asterisks are used to indicate each incident of step-back activity as coding proceeds. Observe as the NSTF becomes more aggressive there is substantial increase in the step-back activity.

All results reveal an extraordinarily low level of quantisation noise from DC to approximately 30 kHz where there is negligible evidence of intermodulation or harmonic related distortion confirming a high degree of linearity. A study of the statistics of quantiser inputs Q_{in} in each case also revealed well-formed and symmetrical histograms.

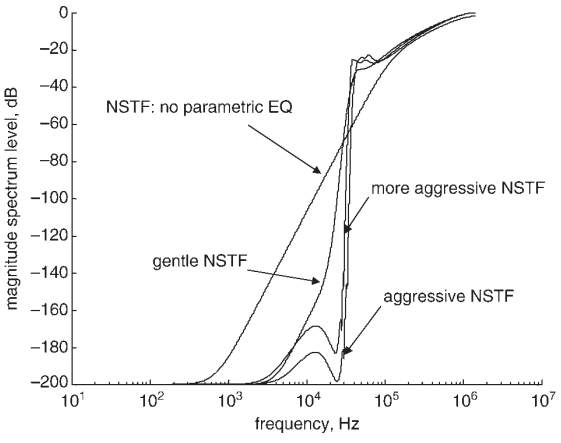


Fig. 6 *Selected noise shaping transfer functions (NSTFs) used in simulation*

a SDM output spectrum and noise reference level
b Corresponding quantiser input Q_{in} and step-back activity

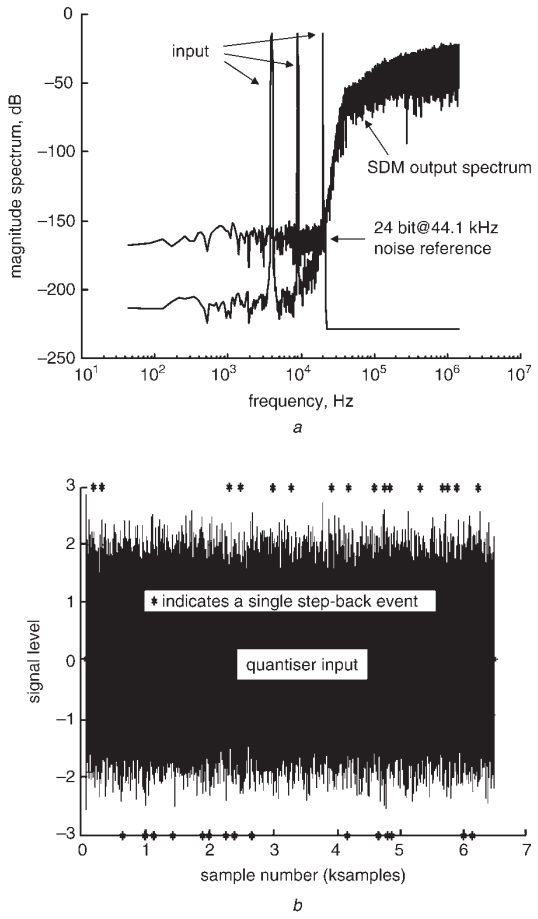


Fig. 7 *Simulation performance data for SDM with 'gentle' noise shaping*

a SDM output spectrum and noise reference level
b Corresponding quantiser input Q_{in} and step-back activity

To explore the linearity of the SDM encoder, a simulation was performed using a wideband noise input excitation with rectangular PDF spanning −0.4 to 0.4, sampled at 44.1 kHz and then upconverted by 64 to the SDM sampling rate. Results calculated over 2^{16} samples are shown in Fig. 10 where the output spectrum reveals a virtually undistorted replica of the brickwall input noise spectrum ranging from DC to 22.05 kHz. Observe there is no intermodulation distortion in the broadband output – input error spectrum and the quantisation noise spectrum above 22.05 kHz falls to the residual SDM noise level determined by the NSTF and follows for example a similar path to that shown in Fig. 9*a*.

To test further the linearity of the system signal averaging was applied to lower uncorrelated noise artifacts in an attempt to expose nonlinear distortion. Quantisation of the input signal was changed from 32 bit to 48 bit to expose the SDM quantisation noise spectrum and a two sine wave excitation employed with signal amplitudes 0.1 and respective frequencies 19 kHz and 20 kHz, vector length was 2^{16} samples. Figure 11 presents the output spectrum after 1028 synchronised averages where uncorrelated dither sequences were applied to the quantiser input during each simulation. 1028 averages theoretically reduce uncorrelated noise components by 30 dB, where the virtual absence of distortion components below 40 kHz confirms that a 0.35 level of rectangular PDF dither is about optimum.

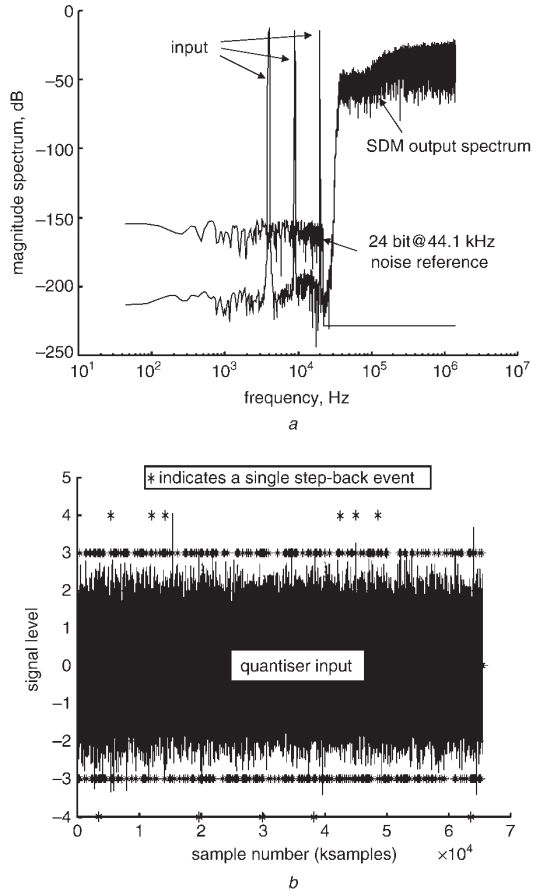


Fig. 8 *Simulation performance data for SDM with 'more aggressive' noise shaping*

a SDM output spectrum and noise reference level
b Corresponding quantiser input Q_{in} and step-back activity

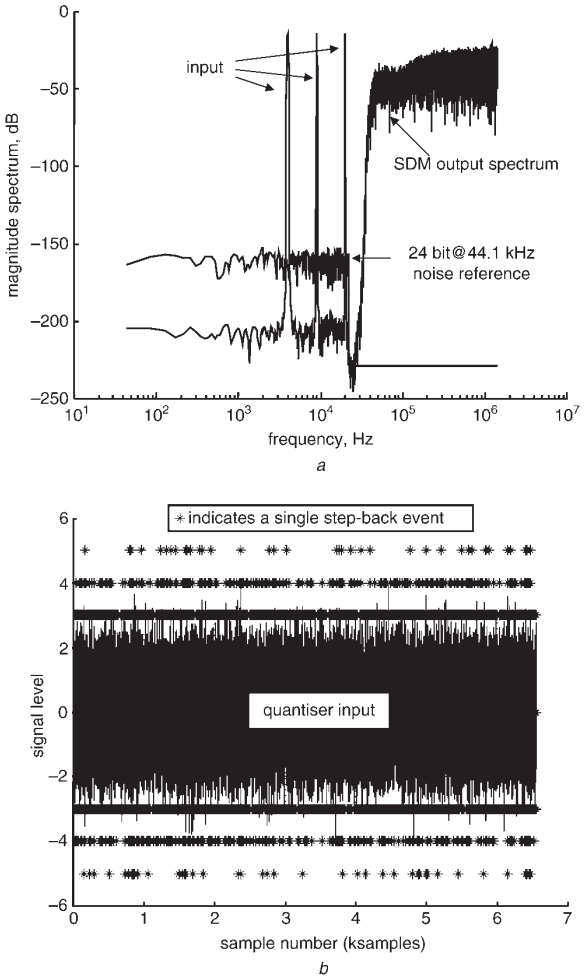


Fig. 9 *Simulation performance data for SDM with 'aggressive' noise shaping*

However, the averaging exposes sidebands distributed around the SDM sampling frequency resulting from frequency modulation [22, 24, 25].

The step-back algorithm can be applied to other SDM structures such as the Sony FF encoder [6]. Figure 12*a*, 12*b* shows the output spectrum and quantiser input signals respectively for this class of SDM. The input signal here consists of two sine waves each of amplitude 0.35 with frequencies 19 kHz and 20 kHz. However, observe that the SDM noise spectrum lies significantly above the 24-bit noise reference level and the noise performance below the two notch frequencies does not improve so rapidly. This effect was reported earlier [16, 22] and is a consequence of the method by which the notch frequencies are realised using local feedback around pairs of integrators. Figure 12*b* reveals substantial step-back activity as a consequence of the high peak input signal set at 0.7, although because of the relaxed NSTF, this activity decays rapidly as the input signal level is reduced. Nevertheless, the encoder shows excellent behaviour in the formation of the quantiser input signal with no excessive peaks or gaps with a noise-like structure. A further SDM example is presented in Fig. 13*a*, 13*b* having a milder high frequency NSTF offering reduced high frequency noise although at the expense of a degraded in-band noise performance. Again observe the 24-bit reference spectrum and compare it against the spectrum in Fig. 12*a*. Although the noise level within the audio band is higher than the aggressive noise shaper examples shown in

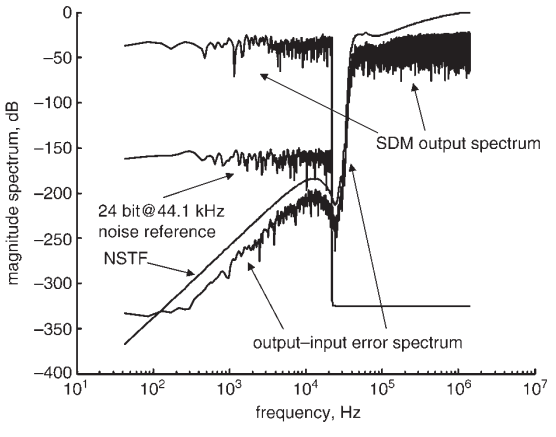


Fig. 10 *'Aggressive' noise shaping, rectangular PDF noise input (−0.4 to 0.4)*

Figs. 10 and 11, there is an improvement in high-frequency noise performance that surpasses the Sony FF encoder. Also, the relative system complexity is more modest where five principal integrators and just two parametric stages are used. However, the penalty of this NSTF tuning is a greater reliance upon step-back activity as shown in Fig. 13*b* where simulation reveals significant step-back activity remains even when the input signal level is reduced. The Sony FF

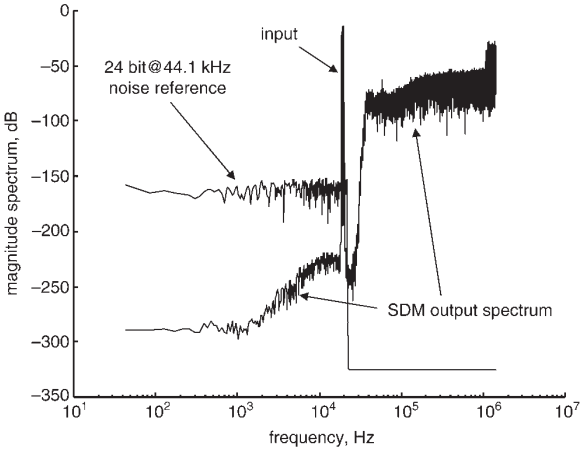


Fig. 11 *Output spectrum after 1028 averages (LPCM input quantised to 48 bit)*

a SDM output spectrum and noise reference with high-level input signal
b Corresponding quantiser input Q_{in} and step-back activity

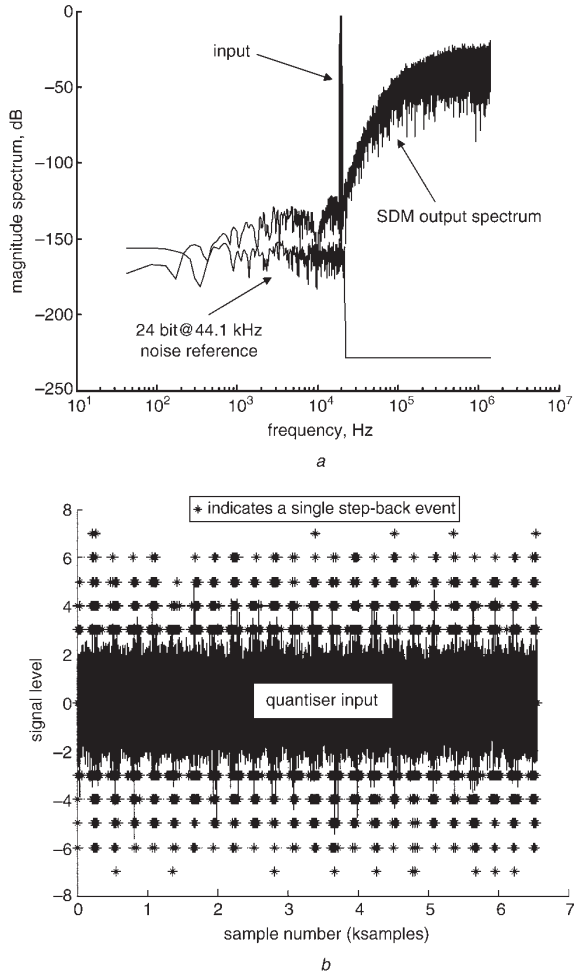


Fig. 12 *Simulation performance data for SONY FF SDM using step-back algorithm*

a SDM output spectrum and noise reference with high-level input signal
b Corresponding quantiser input Q_{in} and step-back activity

encoder is therefore more relaxed and less demanding in this respect and supports faster coding.

Finally, results are presented to explore closed-loop stability as a function of a pure DC input. On initial

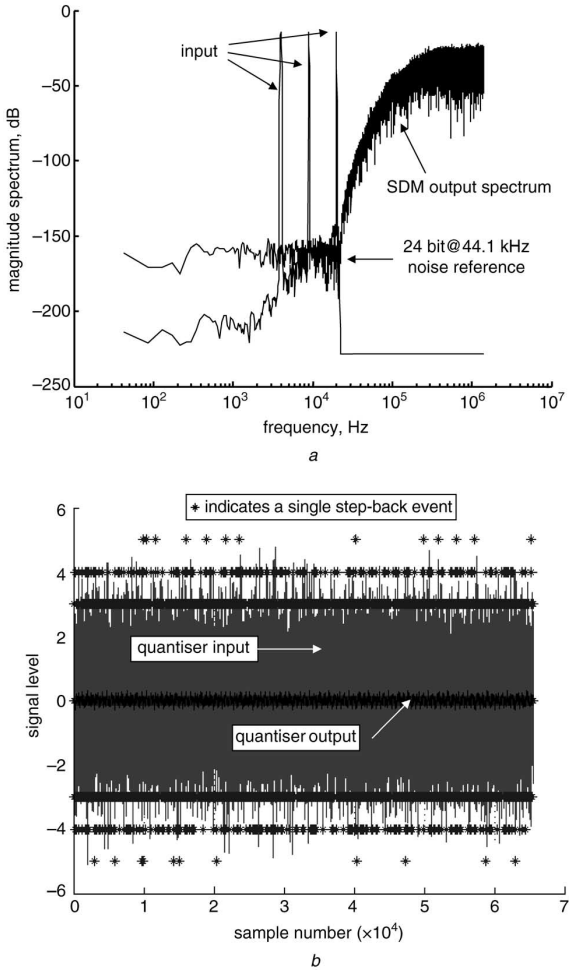


Fig. 13 *Simulation performance data for parametric SDM with 'more relaxed' high-frequency noise shaping derived using five principal integrators and two parametric stages*

consideration it may appear that a more severe test would use a high-frequency high-amplitude input, however, such signals by their nature do not occupy maximum level continuously. On the contrary, a DC input with a level comparable to the peak amplitude of a sinewave demands that on average the output bit pattern must contain a higher ratio of 1 to 0 pulses. Consequently, as DC input level increases the degrees of freedom to form pulse patterns reduce, a factor that eventually both bounds loop stability and impacts upon step-back activity. Three sets of simulations were performed using the three example NSTFs defined in the Appendix, Section 9, where for each SDM the maximum DC input (approximated to 0.05 increments) was determined over 2^{16} samples that just allowed step-back control to maintain stability. Also, under these extreme conditions of near instability the distributions of step-back occurrences as a function of step-back depth were recorded and are presented in Tables 1 to 3 together with distributions corresponding to respective inputs reduced by 6 dB and 12 dB from their maximum. In each case output spectra remained similar to those shown in Figs. 7 to 10 although at the highest DC levels the Q_{in} histograms become skewed as were the distributions on step-back activity acquiring bias depending on DC input polarity, otherwise their form remained similar to the AC simulations. These results confirm step-back control is

Table 1: Step-back occurrence rates: 'Aggressive' NSTF maximum DC level = 0.2

	Step-back depth (samples)					
DC	3	4	5	6	7	8
Occurrences						
0.20	19507	5007	468	5	0	0
0.10	9014	912	71	0	0	0
0.05	6467	455	33	0	0	0

Table 2: Step-back occurrence rates: 'More aggressive' NSTF maximum DC level = 0.4

	Step-back depth (samples)					
DC	3	4	5	6	7	8
Occurrences						
0.40	6	0	22660	1463	21	0
0.20	1895	58	9	0	0	0
0.10	645	9	1	0	0	0

Table 3: Step-back occurrence rates: 'Relaxed' NSTF maximum DC level = 0.55

	Step-back depth (samples)					
DC	3	4	5	6	7	8
Occurrences						
0.55	6	0	11154	1745	176	12
0.27	416	10	0	0	0	0
0.13	29	0	0	0	0	0

effective with dc inputs although some compromise on maximum input level with the most aggressive noise shaper was observed. It is recommended therefore that in audio applications an additional low-frequency servo amplifier is used to steer the SDM into its central coding region with the highpass filter response imposed on the signal transfer function having a 3 dB break frequency chosen in the sub-audio 1 to 10 Hz band.

6 Conclusions

This paper has presented a new approach to SDM employing a state memory to enable a step-back in time function to be implemented. The technique recognises that because of gross nonlinearity within SDM and limitations of binary output code, encoding can enter blind alleys that lead directly to instability. Such states can be identified by inspecting output pulse groups and behaviour of the quantiser input signal. When detected a variable step-back procedure directs the SDM back to an earlier state, inverts the previous output at that instant and then proceeds forward applying a new dither sequence. This process can be repeated many times and is not dependent on fixed block architecture, indeed under difficult conditions deep step-back can occur so a state memory of 256 samples is incorporated. Thus it is possible to undo mistakes in the past that are not recognised as such until their consequences become evident. A key feature is always to let the feedback loop steer forward coding with no discontinuities imposed by forced state reset that cause modulation artefacts. Simulation results demonstrate the coding potential in the combined use of parametric noise shaping and step-back control in terms of audio band noise levels and distortion. Results show virtually noiseless and distortionless performance is possible up to about 30 kHz commensurate with the standard SACD sampling rate. Also, in all simulations the use of a feedforward path from SDM input to the quantiser input as shown in Fig. 2, yields an exact signal transfer function of unity. The method builds upon earlier work on SDM parametric coding which showed how to implement aggressive noise shapers, although instability is problematic at higher signal levels. Earlier work [10, 22] had described strategies to combat instability using multilevel output code combined with time dispersive correction, although non-linear distortion results in output spectrum degradation. The present technique eliminates such corrective processing and enables a smooth forward progression in encoding under proper closed-loop control.

There are inevitable comparisons to be drawn with the Trellis algorithm. Observing a full Trellis implementation, the present approach is computationally more efficient especially in situations where only minor step-back activity is demanded. However, there has been important progress made which leads to a more efficient Trellis implementation [12–15]. The ultimate question is whether Trellis leads to an optimal solution. Initial thoughts suggest that a comprehensive search of all signal combinations and choosing the lowest cost function must lead to the best solution. However, in practice it is not that clear-cut as changing the Trellis analysis block size and making even minor changes to the NSTF and loop dither inevitably converge towards a different bit pattern. Also, there is the possibility of extremely small coding artefacts akin to modulation distortion, related to block size. The step-back procedure has no regular block structure and appears as a continuous and virtually non-granular process with no evidence that step-back correction ever occurred. In the step-back algorithm noise shaping and step-back control are distinctly separate functions. The former shapes the noise spectrum while the latter identifies valid signal trajectories that allow stable encoding.

It is observed that for aggressive NSTF the number of stable signal paths fall and become more difficult to identify as confirmed by increased levels of step-back activity. Therefore, it is proposed that a better concept for stability is to refer to the '*probability of instability*' rather than just a single state or bound where instability can occur. The step-back algorithm takes advantage of this stochastic behaviour and itself becomes a good indicator of loop stability jointly in terms of NSTF and signal level. This increase in step-back activity forms a measure of the proximity to instability and can be observed in Figs. 7*b* to 9*b*.

The results presented here have been used to explore further the boundaries of parametric encoding as this allows an efficient way to specify the NSTF. However, as shown the technique can readily be applied to other NSTF architectures and is, for example, an excellent companion to the standard Sony FF SDM [6] where it prevents failure due to occasional instability. Used here step-back eliminates the need for forced state reset, thus improving encoding and because this algorithm uses relatively relaxed noise shaping there is only a small time penalty as the rate of step-back activity would be low, where a computation performed over 2^{20} samples required only 10 incidences of step-back when the peak-input signal was reduced to 0.02.

7 Acknowledgments

The author wishes to thank the Audio Engineering Society for giving permission to adapt a convention paper for

publication by the IEE. Also I wish to acknowledge the encouragement and support to publish this work that was given by Derk Reefman and Erwin Janssen from Philips Research Laboratories, Eindhoven. Their interest and expertise in this research was a source of great inspiration.

8 References

1 de Jager, F.: 'Delta modulation – a method of PCM transmission using the one unit code', *Philips Res. Rep.*, 1952, **7**, pp. 442–466
2 Steele, R.: 'Delta modulation systems' (Pentech Press, London, 1975)
3 Inose, H., and Yasuda, Y.: 'A unity bit coding method by negative feedback', *Proc. IEEE*, 1963, **51**, pp. 1524–1535
4 Verbakel, J., van de Kerkhof, L., Maeda, M., and Inazawa, Y.: 'Super audio CD format'. 104th AES Convention, Amsterdam, May 1998, paper 4705
5 Reefman, D.: 'Why Direct Stream Digital is the best choice as a digital audio format'. 110th AES Convention, Amsterdam, May 2001, paper 5396
6 Takahashi, H., and Nishio, A.: 'Investigation of practical 1-bit delta-sigma conversion for professional audio applications'. 110th AES Convention, Amsterdam, May 2001, paper 5392
7 Reefman, D., and Janssen, E.: 'Enhanced sigma delta structures for super audio CD applications'. 112th AES Convection, Munich, May 2002, paper 5616
8 Risbo, L.: 'Σ−Δ modulators – stability analysis and optimization'. PhD thesis, Technical University of Denmark, 1994
9 Hawksford, M.O.J.: 'Chaos oversampling and noise shaping in digital-to-analog conversion', *J. Audio Eng. Soc.*, 1989, **37**, (12), pp. 980–1001
10 Hawksford, M.O.J.: 'Transparent differential coding for high-resolution digital audio', *J. Audio Eng. Soc.*, 2001, **49**, (6), pp. 480–497
11 Kato, H.: 'Trellis noise-shaping converters and 1-bit digital audio'. AES112th convention, Munich, 2002 March 10-13, paper 5615
12 Harpe, P., Reefman, D., and Janssen, E.: 'Efficient trellis-type sigma delta modulator'. 114th AES Convention, Amsterdam, The Netherlands, 22–25 March 2003, paper 5845
13 Janssen, E., and Reefman, D.: 'Advances in trellis based SDM structures'. 115th AES Convention, New York, 10–13 October 2003, paper 5993
14 Angus, J.A.S.: 'Tree based lookahead sigma delta modulators'. AES 114th Convention, Amsterdam, The Netherlands, 22–25 March 2003, paper 5825
15 Angus, J.A.S.: 'Efficient algorithms for look-ahead sigma-delta modulators'. 115th AES Convention, Amsterdam, New York, October 2003, paper 5950
16 Hawksford M.O.J.: 'Parametric SDM encoder for SACD in high-resolution digital audio'. Proc. Institute of Acoustics Conf. on Reproduced Sound-18, Vol. 24, Part 8, November 2002
17 Lipshitz, S.P., and Vanderkooy, J.: 'Why professional 1-bit sigma-delta conversion is a bad idea'. 109th AES Convention, Los Angeles, September 2000, preprint 5188
18 Lipshitz, S.P., and Vanderkooy, J.: 'Towards a better understanding of 1-bit sigma-delta modulators'. 110th AES Convention, Amsterdam, May 2001, paper 5395
19 Lipshitz, S.P., and Vanderkooy, J.: 'Towards a better understanding of 1-bit sigma-delta modulators – Part 2'. 111th AES Convention, New York, December 2001, paper 5477
20 Lipshitz, S.P., and Vanderkooy, J.: 'Towards a better understanding of 1-bit sigma-delta modulators – Part 3'. 112th AES Convention, Munich, May 2002, paper 5620
21 Lipshitz, S.P., and Vanderkooy, J.: 'Towards a better understanding of 1-bit sigma-delta modulators – Part 4'. 116th AES Convention, Berlin, May 2004, paper 6093
22 Hawksford, M.O.J.: 'Time-quantized frequency modulation, time-domain dither, dispersive codes, and parametrically controlled noise shaping in SDM', *J. Audio Eng. Soc.*, 2004, **52**, (6), pp. 587–617
23 Flood, J.E., and Hawksford, M.O.J.: 'Adaptive delta-sigma modulation using pulse grouping techniques'. Proc. IERE Conf. on Digital Processing of Signals in Communications, Loughborough University, No. 23, April 1972, pp. 445–462
24 Flood, J.E., and Hawksford, M.J.: 'Exact model for deltamodulation processes', *Proc. IEE*, 1971, **118**, pp. 1155–1161
25 Hawksford, M.J.: 'Unified theory of digital modulation', *Proc. IEE*, 1974, **121**, (2), pp. 109–115

9 Appendix: Loop filter coefficients defining three example NSTFs

LPF and BPF bi-quadratic parameters optimised for $f_{sdm} = 2.8224\,\text{MHz}$ are defined:

order,	number of principal integrators
zorder,	number of parametric stages
$f[1{:}zorder]$,	vector defines resonant frequencies of parametric stages 1 to *zorder*
$k[1{:}zorder]$,	vector defines damping factors of parametric stages 1 to *zorder*
$g1[1{:}zorder]$,	vector defines BPF gain coefficients of parametric stages 1 to *zorder*
$g2[1{:}zorder]$,	vector defines LPF gain coefficients of parametric stages 1 to *zorder*

Integrator scale factor vector $p[1{:}zorder]$ is given by, $p[1:zorder] = \frac{2\pi f[1:zorder]}{f_{sdm}}$

9.1 'Aggressive' noise shaping filter parameters

$order = 4$ $zorder = 13$

$f(1) = 39000;$	$k(1) = .35;$	$g2(1) = 1.1;$	$g1(1) = 0$
$f(2) = 25000;$	$k(2) = .18;$	$g2(2) = 1.3;$	$g1(2) = 0$
$f(3) = 25000;$	$k(3) = .18;$	$g2(3) = 1.3;$	$g1(3) = 0$
$f(4) = 25000;$	$k(4) = .18;$	$g2(4) = 1.3;$	$g1(4) = 0$
$f(5) = 25000;$	$k(5) = .18;$	$g2(5) = 1.3;$	$g1(5) = 0$
$f(6) = 30000;$	$k(6) = .36;$	$g2(6) = 1.3;$	$g1(6) = 0$
$f(7) = 30000;$	$k(7) = .52;$	$g2(7) = 1.3;$	$g1(7) = 0$
$f(8) = 30000;$	$k(8) = .52;$	$g2(8) = 1.3;$	$g1(8) = 0$
$f(9) = 30000;$	$k(9) = .52;$	$g2(9) = 1.3;$	$g1(9) = 0$
$f(10) = 30000;$	$k(10) = .01;$	$g2(10) = 0;$	$g1(10) = 100$
$f(11) = 33000;$	$k(11) = .01;$	$g2(11) = 0;$	$g1(11) = 100$
$f(12) = 55000;$	$k(12) = .15;$	$g2(12) = 0;$	$g1(12) = 2.5$
$f(13) = 20000;$	$k(13) = .36;$	$g2(13) = 1.3;$	$g1(13) = 0$

9.2 'More relaxed' noise shaping filter parameters

$order = 4$ $zorder = 10$

$f(1) = 38000;$	$k(1) = .38;$	$g2(1) = 1.3;$	$g1(1) = 0$
$f(2) = 27000;$	$k(2) = .18;$	$g2(2) = 1.3;$	$g1(2) = 0$
$f(3) = 24300;$	$k(3) = .18;$	$g2(3) = 1.3;$	$g1(3) = 0$
$f(4) = 24300;$	$k(4) = .18;$	$g2(4) = 1.3;$	$g1(4) = 0$
$f(5) = 24300;$	$k(5) = .18;$	$g2(5) = 1.3;$	$g1(5) = 0$
$f(6) = 21000;$	$k(6) = .36;$	$g2(6) = 1.3;$	$g1(6) = 0$
$f(7) = 21000;$	$k(7) = .52;$	$g2(7) = 1.3;$	$g1(7) = 0$
$f(8) = 21000;$	$k(8) = .52;$	$g2(8) = 1.3;$	$g1(8) = 0$
$f(9) = 30000;$	$k(9) = .005;$	$g2(9) = 0;$	$g1(9) = 100$
$f(10) = 28000;$	$k(10) = .01;$	$g2(10) = 0;$	$g1(10) = 100$

9.3 'Relaxed' noise shaping filter parameters

$order = 5$ $zorder = 5$

$f(1) = 20000;$	$k(1) = .8;$	$g2(1) = 1.5;$	$g1(1) = 0$
$f(2) = 20000;$	$k(2) = .8;$	$g2(2) = 1.5;$	$g1(2) = 0$
$f(3) = 20000;$	$k(3) = .8;$	$g2(3) = 1.5;$	$g1(3) = 0$
$f(4) = 20000;$	$k(4) = .8;$	$g2(4) = 3;$	$g1(4) = 0$
$f(5) = 20000;$	$k(5) = .8;$	$g2(5) = 3;$	$g1(5) = 0$

PAPERS

Dynamic Model-Based Linearization of Quantized Pulse-Width Modulation for Applications in Digital-to-Analog Conversion and Digital Power Amplifier Systems*

M. O. J. HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, CO4 3SQ, UK

Quantized pulse-width modulation (PWM) offers an efficient means of converting digital data to analog either at low signal levels for DAC systems or at higher levels in power amplification. However, although a number of techniques exist, they are flawed by varying degrees of dynamic nonlinearity inherent to the conversion process, especially at lower sampling rates. The basic nonlinear mechanisms of PWM are described and a family of model-based solutions to the linearization problem is presented that retains the advantage of a uniform sampled digital format. As such it is possible to design PWM converters that exhibit vanishing levels of distortion even under broad-band high-level signal excitation.

0 INTRODUCTION

Pulse-width modulation (PWM) was invented by A. H. Reeves, who also invented pulse-code modulation (PCM) [1]–[3] and the capacitor microphone. The technique of PWM has had a long history of theoretical and practical development, where it has found wide application in power-efficient amplifiers and power-supply systems [4], [5]. Indeed, it is somewhat appropriate that both PCM and PWM originate from the same inventor, as this paper seeks to find an optimum solution to combining the two systems within the digital domain. Recently PWM has been identified as a means of achieving digital-to-analog conversion (DAC) and has been used in MASH conversion systems in association with noise shaping and oversampling. Such systems require the process of generating PWM to be performed in the digital domain, where an early study was undertaken by Sandler [6], with evolutionary extensions subsequently reported in numerous related papers [7]–[10].

One approach is to mimic the analog methods of implementing PWM using a digital ramp function and comparator. However, if the technique of natural sampling is implemented, then there is significant computational complexity in calculating the intersection of ramp and finely time-quantized, oversampled, and band-limited data as well as extremely high clock rates to time the pulse transitions. More recently it has been recognized that oversampling and noise shaping with multilevel quantization can be used to reduce the resolution required of the PWM signal, where an outline system was proposed [11, sec. 8.3] by the author in 1985. Later work by Sander and coworkers [12], [13] has developed this theme and supported the feasibility of more practical switching rates in association with acceptable signal-to-noise ratios. Theoretically the target performance of PWM is enhanced by an increase in the sampling rate, although the design of the switching output stage is then more problematic and efficiency degrades due to the extra signal transitions where both voltage and current occur simultaneously in the output transistors. Also edge slewing, jitter, and power-supply compliance and its associated transient response degrade performance. For high efficiency the sampling rate should be low, but for low distortion and ease of signal reconstruction using low-pass filtering, the rate should be high. However, it will also be shown that an increase in sampling rate requires an increase in system accuracy,

* Manuscript received 1991 September 19; revised 1991 December 14.

thus negating some of the reported advantage of noise shaping and excessive oversampling.

The principal thrust in modern PWM systems is to convert uniformly time-sampled and amplitude-quantized audio data directly into an efficient PWM code that exhibits a low level of nonlinearity. Ideally, only modest oversampling should be used, which is sufficient to gain a signal recovery and possible noise-shaping advantage, but limited so as not to degrade efficiency. Also, it is advantageous if the bulk of signal processing occurs at the uniform sampling rate to avoid the need for excessive oversampling ratios to mimic the process of natural sampling.

In this paper we address directly the principal distortion-generating mechanisms of PWM and demonstrate that by use of dynamic (or time-varying) filtering, model-based linearization can be achieved to a high degree of accuracy that generates a low residue, even in the presence of high-level and energetic signals. The process is first demonstrated using a symmetrical nonrecursive filter architecture with constrained recursive coefficient adaptation forward and backward of the present output sample, which is suitable for finely resolved PWM. The technique is then applied to more coarsely quantized structures, which also enables a noise-shaping advantage to lower the resolution of the PWM samples.

Two recent papers by Mellor et al. [14], [15] have demonstrated significant reductions in PWM distortion by modifying the uniform sampling process. As it is well documented that naturally sampled PWM offers enhanced linearity over uniform sampling, Mellor proposes using linear interpolation operation between adjacent input samples to estimate the PWM pulse transitions so as to approximate those of a naturally sampled system. This is a fundamental proposition in the evolution of digital PWM that paves the way to a low-distortion, all-digital power amplifier system. However, research now suggests that there exist more optimal strategies for linearizing uniformly sampled digital PWM, and in this paper an approach argued from the spectral domain is presented.

1 NONLINEAR MECHANISMS IN PWM

The discussion in this paper is restricted to a particular class of PWM, where the input data are assumed to be uniformly sampled and the pulse width is calculated at each sampling instance. This technique simplifies signal processing compared with natural sampling, where in digital architectures there is a significant complication in calculating the intersection of a finely interpolated audio signal with that of a linear staircase function, as illustrated in Fig. 1.

Initially we shall assume audio data in four times oversampling format, although alternative oversampling ratios can also offer advantage in association with noise shaping. The uniformly sampled PWM process maps each data sample to an equivalent-area rectangular pulse of constant amplitude but variable width. We consider a symmetrical pulse distribution, as shown in Fig. 2. The width of the pulse τ is then determined as

$$\tau = \frac{x(n)}{A} T_s \tag{1}$$

where A is the pulse amplitude, $x(n)$ the amplitude of the nth input sample, and T_s the sampling period of

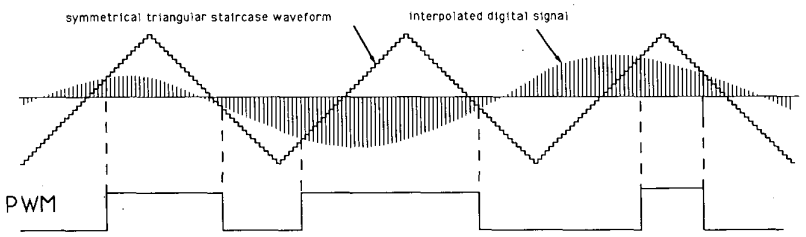


Fig. 1. Natural sampling showing interpolated digital signal and problem of calculating intersections.

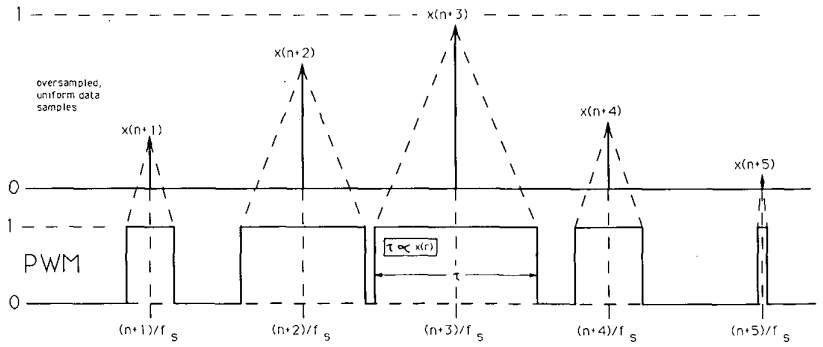


Fig. 2. Oversampled symmetrical PWM using uniform sampling where pulse width is directly proportional to sample amplitude.

the PWM sequence. In this example $0 < x(n) < A$ and corresponds directly to a pulse width range of $0 < \tau < T_s$. To accommodate a bipolar PWM format, a dc level of $A/2$ is subtracted from the square wave and a corresponding dc offset of $A/2$ is added to the input sequence, transforming $x(n)$ to a range of $-A/2$ to $A/2$.

The nonlinear distortion inherent in this class of PWM process arises from two elements.

1) The spectrum of the input sequence is replicated about the sampling frequency and its harmonics. Consequently the baseband signal must adhere to conventional uniform sampling theory to prevent aliasing distortion.

2) Each sample is then replaced by a constant-amplitude rectangular pulse with an area proportional to the sample amplitude $x(n)$. Although this guarantees exact dc coding, the broad spectrum of each individual pulse is dynamically modified as a function of the pulse width. Hence following subsequent summation over all pulses, it is this dynamic spectral modulation that is the root of nonlinearity in PWM.

To examine the mechanism of dynamic spectral modulation, the Fourier transform $F_n(f)$ of the nth pulse, illustrated in Fig. 2, is expressed as

$$F_n(f) = \frac{\tau A}{T_s}\left[\frac{\sin(\pi f\tau)}{\pi f\tau}\right] e^{-j2\pi nfT_s} \qquad (2a)$$

where τ is calculated from Eq (1). For a low-PWM modulation index, $\pi f\tau << \pi/2$,

$$F_n(f) \approx \left[\frac{\tau A}{T_s}\right] e^{-j2\pi nfT_s} \qquad (2b)$$

and the modulation process is approximately linear. However for higher modulation indices the $\sin(\pi f\tau)/\pi f\tau$ factor modifies the spectrum of each individual sample, introducing a frequency-dependent gain modulation that generates nonlinear distortion products over a broad spectrum. In essence, each pulse width has a unique Fourier transform that exhibits differential gain errors with increasing frequency. In Fig. 3 $\sin(\pi f\tau)/\pi f\tau$ is plotted as a function of f for a family of τ, where gain modulation is evident. To illustrate this distortion mechanism, Fig. 4(a) shows the output spectrum of a symmetrical uniformly sampled PWM system for a periodic input sequence $x(n)$ calculated over 1024 samples. Here

$$x(n) = A[0.5 + 0.15\sin(k_0 n) + 0.15\sin(k_1 n) + 0.15\sin(k_2 n)] \qquad (3)$$

The sampling rate is 176 kHz and the three equiamplitude superimposed input signal frequencies are 171.875 Hz, 17.1875 kHz, and 20.45313 kHz. However, to expose the intermodulation distortion in the

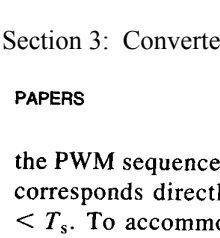


Fig. 3. Family of sinc(x) functions representing dynamic modulation of pulse spectrum in PWM.

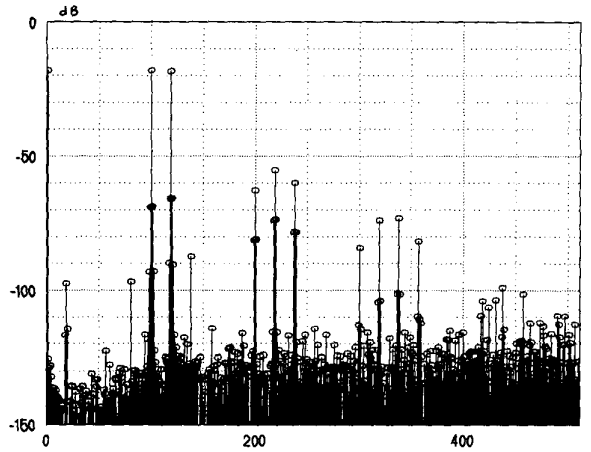


Fig. 4. (a) Noncorrected PWM spectrum computed over 1024 samples. Sampling rate 176 kHz; correction band 0–22 kHz; input frerquencies 171.875 Hz, 17.1875 kHz, and 20.45313 kHz. (b) Spectrum of (a) with expanded frequency scale to show intermodulation products.

PWM output spectrum, an expanded-frequency-scale plot is shown in Fig. 4(b). The resultant spectrum $F_p(f)$ is computed over 511 samples and corresponds to

$$F_p = \frac{1}{NS} \sum_{n=0}^{NS-1} F_n(f) \ . \tag{4}$$

It is observed that dynamic spectral modulation occurs after the sampling function. Hence the higher frequency distortion products do not undergo aliasing into the baseband, which enables an opportunity for linearization.

2 INTERLEAVED AND DYNAMIC FIR CORRECTION OF THE PWM PROCESS

To attain linearization, the uniformly sampled digital signal must be modified to compensate for the change in spectral shape with pulse width, although the frequency range over which this is achieved can be limited to the 0–22-kHz passband. The main constraint to be imposed upon this process is the restriction to uniformly spaced samples for the modified sequence, whereas the pulse width takes, by its nature, intersample values. The method of correction is to effectively time-interleave multiple finite-impulse-response (FIR) filters between the oversampled PCM data and the pulse-width modulator whose frequency responses are individually matched to the inverse of each corresponding pulse $x(n)$ in the PWM sequence. The overall filtering process is dynamic as the transfer function of each pulse is a function of pulse width, which can vary between zero and the sampling period, although while an individual filter is contributing to an output pulse, its local coefficients are held constant. In practice, the filtering process can be considered as a single filter, but where the coefficients shift with the input data. Thus a coefficient contributing to the current output is also uniquely linked to its corresponding input data sample. However, the exact structure and operation of the correction process become clearer to comprehend when the simulation program segments in Secs. 2.1 and 2.2 are examined.

To introduce the corrective procedure, a symmetrical five-tap FIR filter is shown in Fig. 5, where the incremental time delay T_s is matched to the system sampling

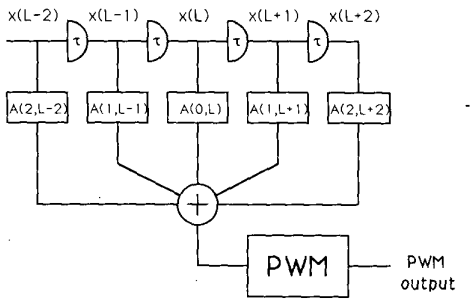


Fig. 5. Five-coefficient FIR filter.

period $1/f_s$ s. The input to the filter is a sampled-data sequence $x(n), 0 < x(n) < 1$, corresponding to a PWM sample width τ, $0 < \tau < 1/f_s$, and the PWM sample is assumed symmetrical about the sample instant. This symmetrical distribution is critical to the present system as it enables a symmetrical FIR filter to be used and therefore prevents complications arising from dynamic phase modulation.

The five coefficients of an FIR filter associated with the current sample n are defined as a_2, a_1, a_0, a_1, a_2. To express the dynamic form of the filter, the corresponding coefficients uniquely linked to individual input samples are represented as $A(2, n-2), A(1, n-1), A(0, n), A(1, n+1), A(2, n+2)$, where n is the current output sample. Thus the notation $A(x, y)$ means coefficient a_x associated with sample $x(y)$.

The aim, therefore, is to associate a unique FIR filter with each PWM mapped input sample $x(n)$ and to compute corresponding filter coefficient sets $\{A(0, n), A(1, n), A(2, n)\}$, where $a_0 = A(0, n)$, $a_1 = A(1, n)$, and $a_2 = A(2, n)$, such that the transfer function of the FIR filter $E_{x(n)}(f)$ and the Fourier transform of the corresponding PWM sample $F_{x(n)}(f)$ closely approximate a (constant) target function $T(f)$ that is matched over the passband 0 to f_u Hz, where f_u is typically 22 kHz for a digital audio system.

However, to maintain constancy of pulse area, the sum of the FIR filter coefficients is unity, where the central coefficient a_0 is given by

$$a_0 = 1 - 2 \sum_{r=1}^{\lambda} a_r \ . \tag{5}$$

That is, where the number of independent coefficients $\lambda = 2$, it is required to calculate two coefficients for a five-tap filter.

Although the target function $T(f)$ could be unity, we choose here the transfer function of a 50% PWM sample corresponding to the quiescent state, $x(n) = 0.5$,

$$T(f) = \frac{\sin(0.5\pi f T_s)}{0.5\pi f T_s} \tag{6}$$

where $T(f)$ is normalized to unity at dc.

The transfer function $E_{x(n)}(f)$ of the five-tap FIR equalization filter associated with the sample $x(n)$, as shown in Fig. 5, can be expressed as

$$E_{x(n)} = a_0 + 2a_1 \cos(2\pi f T_s) + 2a_2 \cos(4\pi f T_s) \tag{7}$$

and the normalized transform $F_{x(n)}(f)$ of a PWM sample corresponding to $x(n)$ is

$$F_{x(n)}(f) = \frac{\sin[x(n)\pi f T_s]}{x(n)\pi f T_s} \ . \tag{8}$$

Consequently the matching task that approximates the

product of the transform of the nth PWM sample and the transform of the nth associated correction FIR filter to the target transform, performed over a frequency band 0 to f_u Hz, can be stated as

$$E_{x(n)}(f)F_{x(n)}(f) \rightarrow T(f)|_{0 \text{ to } f_u} . \quad (9)$$

Substituting and rearranging,

$$a_1[\cos(2z) - 1] + a_2[\cos(4z) - 1] \rightarrow \frac{x(n)\sin(z/2)}{\sin[x(n)z]} - 0.5|_{0 \text{ to } f_u} \quad (10)$$

where

$$z = \frac{\pi f}{f_s} . \quad (11)$$

This matching task could be solved by an optmization procedure to minimize an error function within the 0 to f_u Hz frequency band. However, because the functions exhibit similar curvature over the lower frequency region, a simpler approach (for a_1 and a_2) is to force equality in the function of Eq. (10) at two frequencies, where we select f_u Hz and $0.5f_u$ Hz. Hence solving two simultaneous equations based on Eq. (10) and modifying the functional form to prevent errors due to small differences, the solution can be stated as

$$a_1 = -\frac{S_1 - 4S_2 + 4\sin^2(z)(S_2 - 1) + 3}{D_1}$$

$$a_2 = \frac{S_1 - 4S_2 + 4\sin^2(z/2)(S_2 - 1) + 3}{D_2}$$

where

$$D_1 = 4\left\{\sin^2(z) - 4\sin^2\left(\frac{z}{2}\right)[1 - \sin^2(z)]\right\}$$

$$D_2 = 4\cos^2\left(\frac{z}{2}\right) D_1$$

$$S_1 = \frac{2m\sin(z/2)}{\sin(mz)}$$

$$S_2 = \frac{2m\sin(z/4)}{\sin(mz/2)} . \quad (12)$$

Here m is the normalized input data, $0 < m < 1$, on which the coefficient estimation is based, and the solution is referred to as equation set (12).

To demonstrate the effectiveness of this transform matching process, two sets of Fourier transforms are shown in Fig. 6. The sets include:

1) The target transform corresponding to a pulse, where $m = 0.5$

2) Two extreme case example transforms of pulses, where $m = 0$ and 1

3) The corresponding compensated transforms for $m = 0$ and 1.

In these examples the sampling rate f_s was 176 kHz and the coefficients were calculated for exact matching at 22 and 11 kHz, respectively. The results clearly show how the compensated and target curves are closely matched over the band 0–22 kHz, even though the PWM sample widths vary over the extreme range from 0 to $1/f_s$. A detailed comparison of tabulated data also confirms equality at both 22 and 11 kHz.

In a pratical implementation of this system, a number of techniques can be adopted for determining a_1 and a_2 for each normalized value of m. First, a direct solution of equation set (12) can be implemented for each new sample m. Second, a discrete look-up table can be computed for a range of m and a ROM used to store coefficient values. Finally, a polynomial approximation can be formed for a_1 and a_2 as a function of m, which offers the advantage of efficient signal interpolation. For systems with fine quantization of data (approxi-

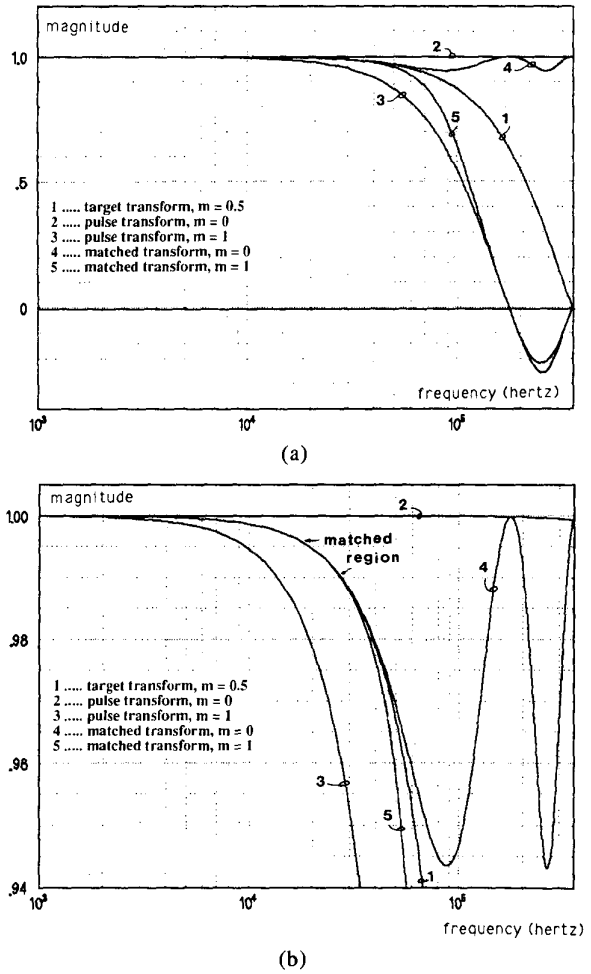


Fig. 6. PWM transform matching functions for $m = 0$ and $m = 1$. Sampling rate 176 kHz; matching frequencies 11 and 22 kHz. (a) Full-range vertical scale. (b) Expanded vertical scale.

mating a continuous function), the polynomial method is the simplest. However, in more coarsely quantized systems with, say, 1024 values of m, the ROM technique using predetermined values is the most efficient. Thus knowing the oversampling ratio and modulation index, the appropriate coefficients can be loaded into the FIR filter, thus making the filter response dynamic or sample-value dependent.

To demonstrate the variation of coefficients with modulation index m, Fig. 7 shows plots of a_1 and a_2 against m for a PWM system with four times oversampling (that is, f_s = 176 kHz). Approximate polynomial matched generating functions for a_1 and a_2 are also given as

$$a_1 = -0.00598x^4 + 0.000692x^3 - 0.0554x^2 + 0.0000680x + 0.0141$$

$$a_2 = 0.00151x^4 - 0.000180x^3 + 0.00345x^2 - 0.0000158x - 0.000925 .$$

However, there is a further level of complexity in the correction process that arises from pulse-width-dependent distortion and the associated dispersive response of the dynamic FIR filter. When the coefficients a_1 and a_2 are superimposed on adjacent samples in a more general $x(n)$ sequence and subsequently mapped into a PWM format, they, by their nature, each modify the width of the four adjacent PWM samples, which in turn modify the correction coefficients already calculated and associated with these samples. This nonlinear interdependence of coefficients is a recursive process, which because of the symmetric and two-sided form of the FIR impulse response, propagates both forward and backward of the current sample, although the number of iterations can be constrained.

There are a number of methods for calculating the coefficient set for a specific data sequence, and we present here two examples. Both methods rely on an interactive process to converge to a final solution of coefficients. Hence in the initial discussion we consider samples to be approximately continuous within computer precision.

Fig. 8 shows the general system architecture of the dynamic FIR filter used to determine the output sample sequence $y(L)$ that drives directly the pulse-width modulator. Each sample is associated with five coefficients, which also interact with their adjacent samples and where each coefficient set is calculated from a nonlinear dependence on both the central and the surrounding samples. Consequently as the data samples are shifted at the uniform PWM sampling rate, the coefficients having influence on the output sample also change, resulting in a dynamic FIR filter.

2.1 Zigzag Algorithm

In this example the input data $x(L)$ propagate through a 30-stage shift register, where the current output is taken at L = 18 and the clock rate $f_s = 1/T_s$ Hz. The input data sequence $x(L)$ is transformed to the output sequence $y(L)$ via a coefficient matrix $A(r, L)$,

$A(0, L)$ coefficient a_0 associated with sample L
$A(1, L)$ coefficient a_1 associated with sample L
$A(2, L)$ coefficient a_2 associated with sample L.

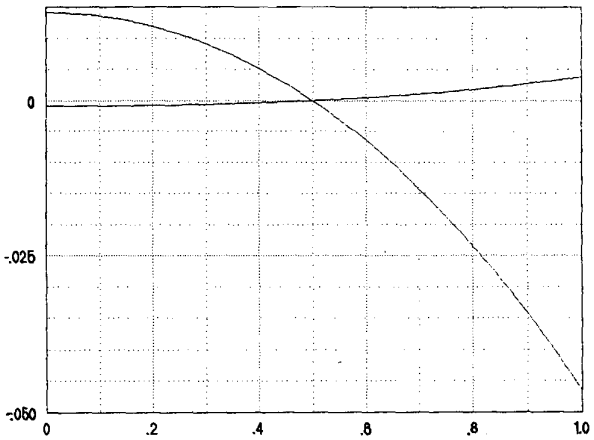


Fig. 7. Plots of coefficients a_1 and a_2 against m for f_s = 176 kHz and f_{opt} = 0–22 kHz.

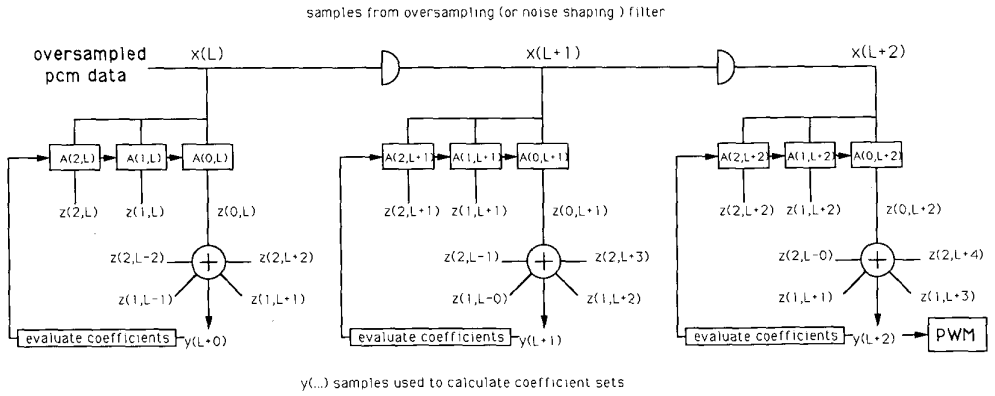


Fig. 8. Interconnected coefficient sets, where each set is calculated from knowledge of surrounding samples.

From Eq. (5),

$$A(0, L) = 1 - 2[A(1, L) + A(2, L)] . \quad (13)$$

As $a_2 << a_1 << a_0$, the influence of a sample on its surrounding samples decays rapidly, so only a limited number of samples need be considered either side of the output sample $L = 18$. In the present example, an input vector spans $L = 1$ to 30, where the zigzag procedure operates as follows:

1) All coefficients $A(0, L)$ are set to 1 and coefficients $A(1, L)$, $A(2, L)$ are set to zero.

2) The coefficient set associated with $L = 18$ is calculated.

3) The coefficient set associated with $L = 17$ is calculated using $A(1, 18)$.

4) The coefficient set associated with $L = 19$ is calculated using $A(1, 18)$ and $A(2, 17)$.

5) The procedure is repeated following a zigzag trajectory either side of $L = 18$, moving progressively outward, and calculating the output matrix $y(L)$ using the latest coefficient estimates via the expression

$$\begin{aligned} y(L) = {} & x(L-2) * A(2, L-2) + x(L-1) \\ & * A(1, L-1) + x(L) * A(0, L) \\ & + x(L+1) * A(1, L+1) \\ & + x(L+2) * A(2, L+2). \end{aligned} \quad (14)$$

6) Steps 2) to 5) are then repeated, say four times, allowing convergence to the final coefficient set.

An important attribute of this algorithm is that the input samples remain unchanged and the output samples $y(L)$ are used only to determine the latest coefficient sets. The recursive procedure that runs both forward and backward of the output sample allows a convergence of the coefficients surrounding $L = 18$, hence forming an estimate of the signal required to drive the pulse-width modulator.

The process is encapsulated in the following subroutines, where the coefficients are estimated using equation set (12), with appropriate constants determined by the oversampling ratio:

```
REM coefficient optimization using zigzag algorithm
FOR L = 1 TO 20
A(L, 0) = 1 : A(L, 1) = 0 : A(L, 2) = 0
NEXT
FOR Q=1 TO 6
FOR P=0 TO 7
L = 10 - P : Y(L) = X(L-2)*A(L-2, 2) + X(L-1)*A(L-1, 1) +
X(L)*A(L, 0) + X(L+1)*A(L+1, 1) + X(L+2)*A(L+2, 2)
GOSUB 10000: REM subroutine to evaluate A(L, 0), A(L, 1) and
A(L, 2) on Y(L)
L = 10 + P : Y(L) = X(L-2)*A(L-2, 2) + X(L-1)*A(L-1, 1) +
X(L)*A(L, 0) + X(L+1)*A(L+1, 1) + X(L+2)*A(L+2, 2)
GOSUB 10000: REM subroutine to evalute A(L, 0), A(L, 1) and
A(L, 2) on Y(L)
NEXT : NEXT
FOR L = 20 TO 1 STEP -1
X(L) = X(L-1)
NEXT
REM last evaluation of Y(18) forms PWM drive
RETURN

10000 : REM direct coefficient evaluation subroutine based on
equation set (12)
E1 = 2*Y(L)*SIN(.5*Z)/SIN(Y(L)*Z) : E2 = 2*Y(L)*SIN(.25*Z)/
SIN(.5*Y(L)*Z)
A(L, 1) = (-E1 + 4*E2 - 4*(E2 - 1)*SIN(Z)^2 - 3)/Z1
A(L, 2) = (E1 - 4*E2 + 4*(E2 - 1)*SIN(Z/2)^2 + 3)/Z2
A(L, 0) = 1 - 2*(A(L, 1) + A(L, 2))
RETURN
```

2.2 Single-Sided Unidirectional Shifting Algorithm

A more efficient procedure uses a single unidirectional scan of the data $x(L)$ to determine coefficients, but shifts both data and coefficients on the system clock. Consequently when the data are scanned, the new coefficient estimates can use the past coefficient estimates alongside their corresponding input sample values. After each shift function, the data are scanned from $L = 1$ to $L = 16$, calculating new coefficients via Eq. (12). The new data samples cause the latest input coefficients to adapt rapidly, but the changes become progressively smaller at higher sample values, as effectively the 16th sample has been through 16 iterations. The output is again taken from $L = 18$, but now the coefficient sets linked with samples $L = 16-20$ do not change, meaning that the pulse area contribution of each input sample is preserved precisely through Eq. (13). As processing is always based on the actual input samples $x(n)$ and Eq. (13) is applied rigorously after final convergence, effects of computational errors are noncumulative. The procedure is described succinctly in the following subroutine.

```
REM coefficient optimization using single-sided shifting algorithm
FOR L = 20 TO 1 STEP -1
X(L) = X(L-1) : A(L, 0) = A(L-1, 0) : A(L, 1) = A(L-1, 1): A(L,
2) = A(L-1, 2)
NEXT
XX = A(0, 0)*X(0) + A(1, 1)*X(1) + A(2, 2)*X(2)
GOSUB 10000: REM subroutine to evaluate A(L, 0), A(L, 1), and
A(L, 2) on XX
XX = A(0, 1)*X(0) + A(1, 0)*X(1) + A(2, 1)*X(2) + A(3, 2)*X(3)
GOSUB 10000 : REM subroutine to evaluate A(L, 0), A(L, 1), and
A(L, 2) on XX
FOR L = 2 TO 18
XX = A(L-2, 2)*X(L-2) + A(L-1, 1)*X(L-1) + A(L, 0)*X(L) +
A(L+1, 1)*X(L+1) + A(L+2, 2)*X(L+2)
GOSUB 10000 : REM subroutine to evaluate A(L, 0), A(L, 1), and
A(L, 2) on XX
NEXT
REM last value of XX forms PWM drive
RETURN
10000 : REM direct coefficient subroutine based on equation set (12)
E1 = 2*XX*SIN(.5*Z)/SIN(XX*Z) : E2 = 2*XX*SIN(.25*Z)/
SIN(.5*XX*Z)
A(L, 1) = (-E1 + 4*E2 - 4*(E2 - 1)*SIN(Z)^2 - 3)/Z1
A(L, 2) = (E1 - 4*E2 + 4*(E2 - 1)*SIN(Z/2)^2 + 3)/Z2
A(L, 0) = 1 - 2*(A(L, 1) + A(L, 2))
RETURN
```

2.3 Response of Linearization Process to Impulsive Data

Table 1 shows the system response to a single impulsive input. First the linearization algorithm is conditioned by an input sequence $x(L) = 0.5$, which sets coefficients a_1 and a_2 to zero. A single sample is then entered where $x(L) = 0.83$. (Note that a unity modulation index corresponds to a maximum pulse duration of T_s.) Ten samples on either side of the impulse are computed, and the distribution of coefficients is also

shown to demonstrate the bilateral dispersive response of the linearization algorithm.

2.4 Spectral Response of Linearization Process

Consider a PWM sequence that is periodic over NS samples and where the modulation index of sample r is $m(r)$ and is normalized to a range of $0 < m(r) < 1$ corresponding to a pulse width $0 < \tau < T_s$. Noting the spectral weighting function described by Eqs. (2) and (4), the Fourier transform of this sequence is given by

$$F(f) = \frac{1}{NS} \sum_{r=0}^{NS-1} m(r) \left\{ \frac{\sin[m(r)\pi fT]}{m(r)\pi fT} \right\} e^{-j2\pi rfT} .$$

Harmonic X of the fundamental frequency is $f_x = (X/NS)f_s$, and since $T_s = 1/f_s$, then $f_xT_s = X/NS$, whereby the transform becomes

$$F\left(\frac{Xf_s}{NS}\right) = \sum_{r=0}^{NS-1} \frac{\sin[\pi X m(r)]}{\pi X} e^{-j2\pi rX/NS} . \quad (15)$$

Eq. (15) can be used to compute the Fourier transform of the pulse sequence $m(r)$ that includes the PWM distortion, resulting from sample reconstruction using rectangular pulses of constant amplitude but variable width, that are symmetrically arranged about their sampling instants. However, if the data sequence $m(r)$ has been predistorted by the linearization algorithm, the effects of nonlinearity should be virtually negated. A data sequence, identical to that described in Sec. 1, has been processed by the algorithm described in Sec. 2.2, where the output spectrum [computed using eq. (10)] is shown in Fig. 9. Comparing this result with the noncorrected transform shown in Fig. 4 reveals a significant reduction in intermodulation products and thus demonstrates the effectiveness of the linearization process.

2.5 Optimum Estimation of Equalizer Target Function

Eq. (6) showed a target function normalized to $m = 0.5$, while in Fig. 7 an example set of coefficients was shown for $f_s = 176$ kHz. These results reveal asymmetry in the distribution of a_1 and a_2, where the moduli are greater for $m = 1$ compared with $m = 0$. An alternative strategy can shift the target response away from $m = 0.5$ in order to achieve greater coefficient equality at $m = 0$ and $m = 1$, respectively, and thus minimize the maximum contribution from the corrective signal. This can be achieved by modifying S_1 and S_2 in equation set (12) to

$$S_1 = \frac{m \sin(\sigma z)}{\sigma \sin(mz)} \qquad S_2 = \frac{m \sin(\sigma z/2)}{\sigma \sin(mz/2)}$$

and searching σ to give near equality in a_1 and a_2 at the extremes of m. As an example for $f_s = 176$ kHz and $f_u = 22$ kHz, the best match is achieved at $\sigma = 0.72373$, giving peak values of $a_1 = 0.0292264$, $a_2 = -0.001855974$ at $m = 0$ and $a_1 = -0.0292264$, $a_2 = 0.002353529$ at $m = 1$, respectively, which can be compared against Fig. 7 for $\sigma = 0.5$. Also, to demonstrate that the modified target transfer function represents a valid solution, a spectral plot is shown in Fig. 10 that should be compared with Fig. 9. Although the magnitudes of the correction coefficients are now reduced overall, their values at low signal level (such as $m \approx 0.5$) are actually greater than in the case where $\sigma = 0.5$. To regain symmetry in the processing of signals in the ranges 0 to 0.5 and 0.5 to 1, a differential topology can be used where two modulators, including linearization, now operate with input signals of the form $1 + m(r)/2$ and $1 - m(r)/2$, respectively. The output is then formed by either a difference stage or a

Table 1. Response of linearization algorithm to single impulsive input.

		Coefficients		
Input sample $x(n)$	Output sample $y(n)$	a_0	a_1	a_2
0.5	0.5	1.0	0.0	0.0
0.5	0.5	1.0	0.0	0.0
0.5	0.5	1.0	0.0	0.0
0.5	0.4999995	1.0	0.0	0.0
0.5	0.5	1.0	0.0	0.0
0.5	0.4999995	1.0	0.0	0.0
0.5	0.5000148	1.0	0.0	0.0
0.5	0.4997598	0.9999622	2.167327E-05	−2.816349E-06
0.5	0.5049456	1.000537	−2.889769E-04	2.065322E-05
0.5	0.4555999	0.9954436	2.449079E-03	−1.708585E-04
0.83	0.9943724	1.08405	−4.578839E-02	3.763581E-03
0.5	0.4555992	0.9954542	2.441855E-03	−1.689809E-04
0.5	0.504934	1.000537	−2.889769E-04	2.065322E-05
0.5	0.4997598	0.9999622	2.167327E-05	−2.816349E-06
0.5	0.5000229	1.0	0.0	0.0
0.5	0.4999959	0.9999946	3.612211E-06	−9.387829E-07
0.5	0.5000018	1.0	0.0	0.0
0.5	0.4999995	1.0	0.0	0.0
0.5	0.5	1.0	0.0	0.0
0.5	0.5	1.0	0.0	0.0
0.5	0.5	1.0	0.0	0.0

bridge, as shown in Fig. 11. The technique is applicable for both DAC and power-amplifier applications, where the latter could use the differential output stage. Also symmetry can be regained irrespective of the chosen target alignment via the σ factor, so it is applicable for both $\sigma = 0.5$ and $\sigma = 0.72373$.

3 NOISE SHAPING AND OVERSAMPLING IN PWM

3.1 Single-Stage Noise Shaper

The example presented in Sec. 2 used input data in an essentially nonquantized format, whereby extremely fine resolution is required to position the edges of the PWM transitions. However, if input data are uniformly quantized (such as 16 bit, 44 kHz), then the process should attempt to accommodate quantized data in an efficient way. Since oversampling (typically 4 to 32 times) is commonly used in PWM, there is an opportunity, as previously suggested [11], to use noise shaping to quantize more coarsely each rectangular pulse to below the resolution of the input data. A basic scheme is shown in Fig. 12. It consists of an oversampling filter, a noise shaper possibly with an order of up to 6, and a linearization algorithm.

The noise-shaping process is essentially that presented in an earlier paper [16], where the following subroutine describes the algorithm:

```
REM PWM noise-shaping routine
REM S(T%) input data; X(T%) output data; RN% noise-shaper
order; U0% clock
S(T%) = DC + S0*SIN(K0*U0%) + S1*SIN(K1*U0%) +
S2*SIN(K2*U0%)
B(0) = S(T%) - DO : DI = 0
```

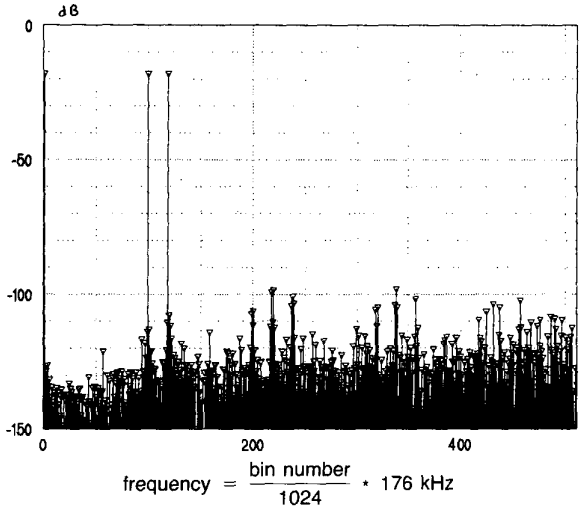


Fig. 9. Corrected PWM spectrum, computed over 1024 samples. Sampling rate 176 kHz; correction band 0–22 kHz; input frequencies 171.875 Hz, 17.1875 kHz, and 20.45313 kHz. (Compare with Fig. 4.)

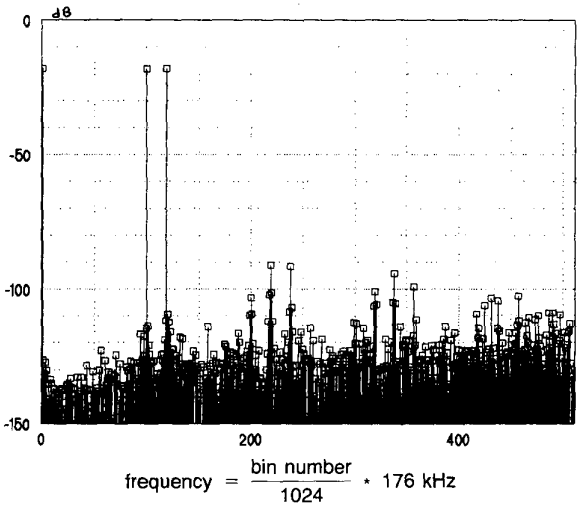


Fig. 10. As Fig. 9, but with optimized σ for a_1 symmetry at $m = 0$ and $m = 1$.

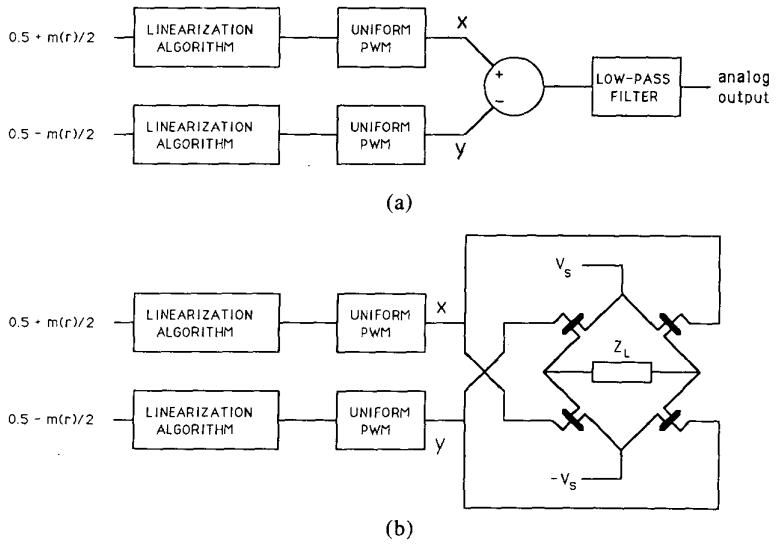


Fig. 11. Differential linearized PWM systems to enhance symmetrical operation for use in high-resolution applications. (a) Configured as DAC. (b) Configured as power amplifier with bridge output stage.

```
FOR L% = 1 TO RN%
B(L%) = B(L%-1) + C(L%) : C(L%) = B(L%)
DI = DI + C(L%)
NEXT
X(T%) = INT(DI) + .5
RETURN
```

The output of the noise shaper, where the word length has been reduced while maintaining an adequate signal-to-noise ratio, can be applied directly to the linearization algorithm described in Sec. 2. However, because the output data of the correction algorithm are quantized more finely than the input data, a requantization process may be used where it is recommended that to reduce additional noise, the output sequence be resolved to either 2 or 3 bit greater than the input sequence, possibly with a further stage of noise shaping, as described in Sec. 3.2.

Alternatively, the PWM signal can be formed using a coarse–fine modulator where the output of the linearization algorithm is initially quantized to the same resolution as the noise-shaper output. The quantization error is then used to fine-tune the edge timing by selecting appropriate taps on a precision delay circuit, as shown in Fig. 13, where for example the clock phase can be controlled to time indirectly the PWM output. In this scheme, if $\Delta\phi$ corresponds to the clock phase determined by the quantization error, then in region A the phase is, say, $\Delta\phi/2$, while in region B it is $-\Delta\phi/2$, thus maintaining symmetry about the sampling instant.

Results of a computer simulation of a fourth-order noise-shaper and linearization system are shown in Figs. 14–17, corresponding to 4, 8, 16, and 32 times oversampling, respectively. In each case the PWM code is quantized to 13 bit and the noise shaper to 10 bit, and the a_1 and a_2 coefficient generators are described by Eq. (12), respectively. As a benchmark, spectral plots are shown for the PWM system, including the noise shaper, both without and with the linearization algorithm, where the noncorrected plots are taken directly from the pulse-width modulator driven by a noise shaper with 10-bit output resolution. The results demonstrate that although noise shaping yields a progressive reduction in intermodulation distortion at higher oversampling ratios, the linearization process is the more effective, although with direct truncation there is a noise penalty.

In computing the Fourier transform of baseband noise-shaped systems over a finite number of samples, the low-level spectral detail at low frequency can be lost. To illustrate this point, the spectral output of a fourth-order, four times oversampled noise-shaped PWM system is shown in Fig. 18, where spectral flattening is evident. The distortion results from the mean level of the finite-analysis sequence being nonzero. Therefore to compensate, one sample is modified so as to force the dc average to zero (or in the special case of our PWM system, to 0.5). In practice, this single sample modifier is small, but it is sufficient to subtract a constant-level spectral distortion, which otherwise corrupts low-level spectral components particularly evident at low frequency with a baseband noise shaper. (Note that a single sample produces a constant spectral response.) The same noise-shaped spectrum, but with a single modified sample to correct the mean value of the sequence, is shown in Fig. 14(a), where the low-level noise-shaped characteristic is now clearly portrayed. This compensation procedure is applied to all the spectral domain plots in this paper, where the justification is that the noise-shaper loop gain is infinite at dc, meaning that an infinite sequence would have a dc value equal to that of the input sequence.

3.2 Two-Stage Noise Shaping

The problem of data-truncation postlinearization can be improved using a second stage of low-order noise shaping to reduce the requantization distortion, where the scheme is shown in Fig. 19. Here the output of the linearization algorithm is requantized using a quantizer enclosed within a first order noise-shaping loop to aid maintenance of resolution prior to driving the PWM.

There is a balance to be made between linearity and additional low-level noise. If the order of the second-stage noise shaper is increased, chaotic behavior [16] makes the output sequence deviate substantially from that of the input sequence, and therefore invalidates the linearization procedure, as the coefficient sets are now no longer accurately matched to the pulse pattern.

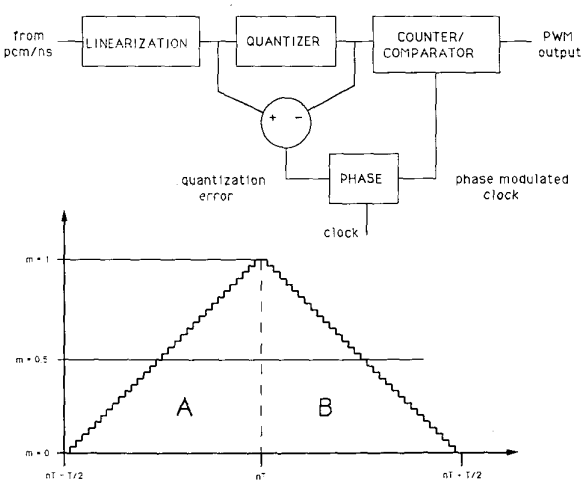


Fig. 13. Double-edge modulation using counter/comparator.

Fig. 12. Uniform sampled PWM with linearization, noise shaping, and data truncation.

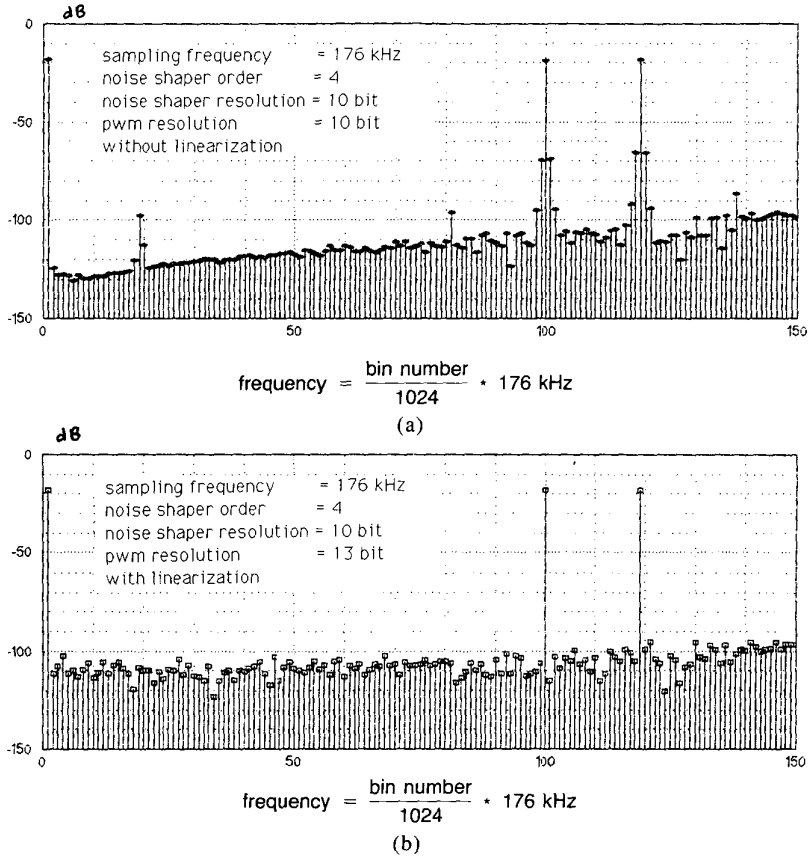


Fig. 14. Single-stage noise shaping and PWM (a) Without linearization. (b) With linearization and direct truncation.

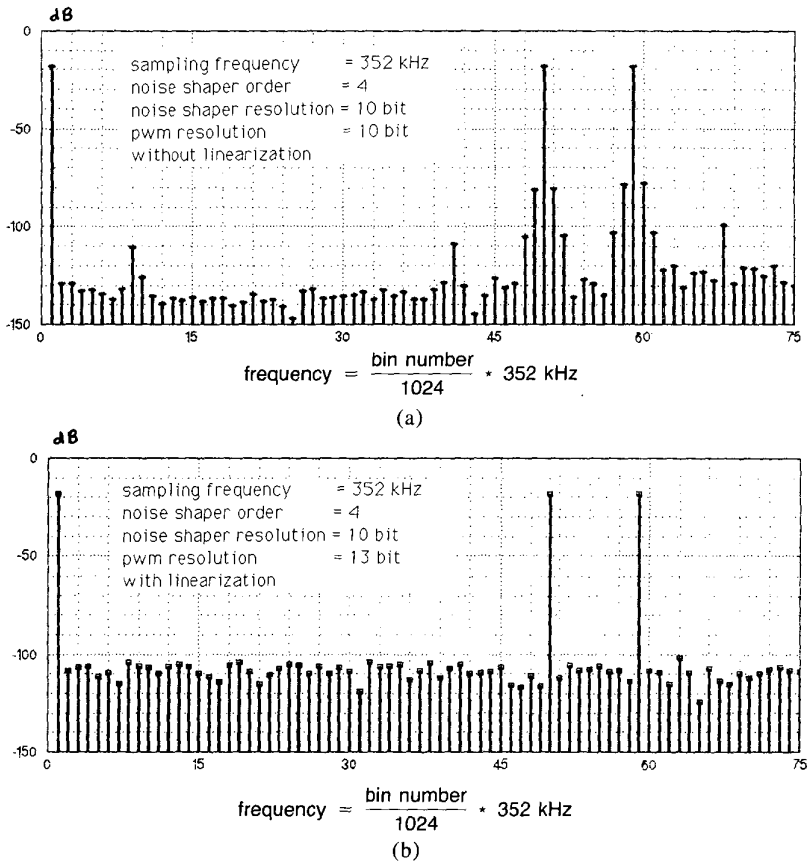
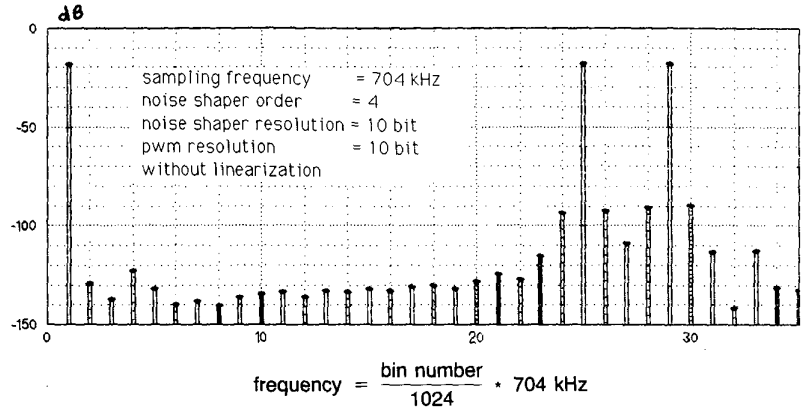
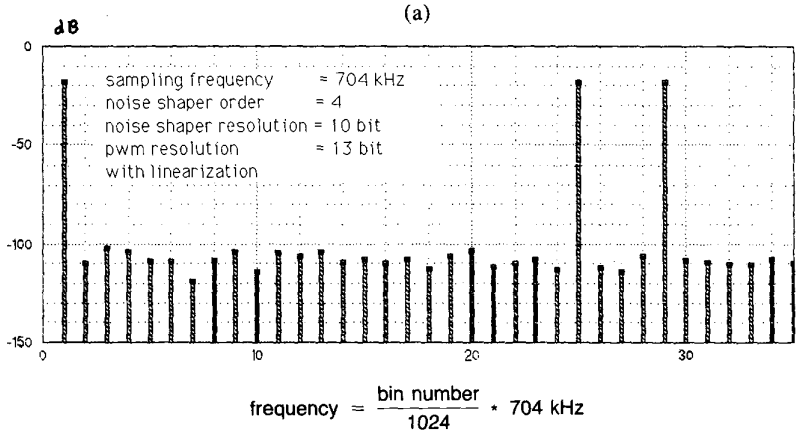


Fig. 15. Single-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and direct truncation.

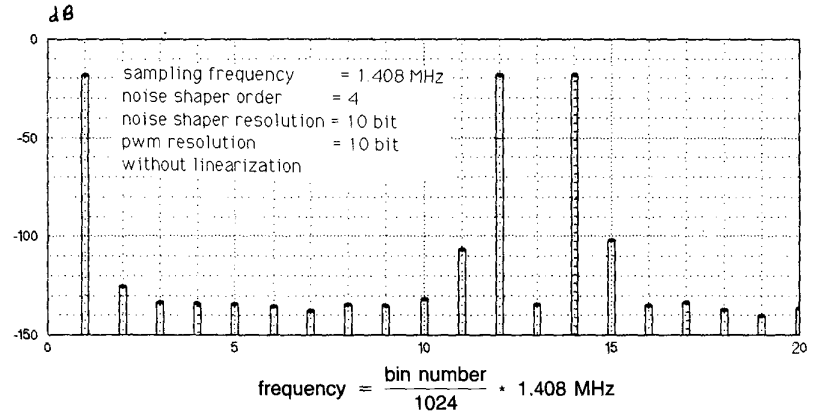


(a)

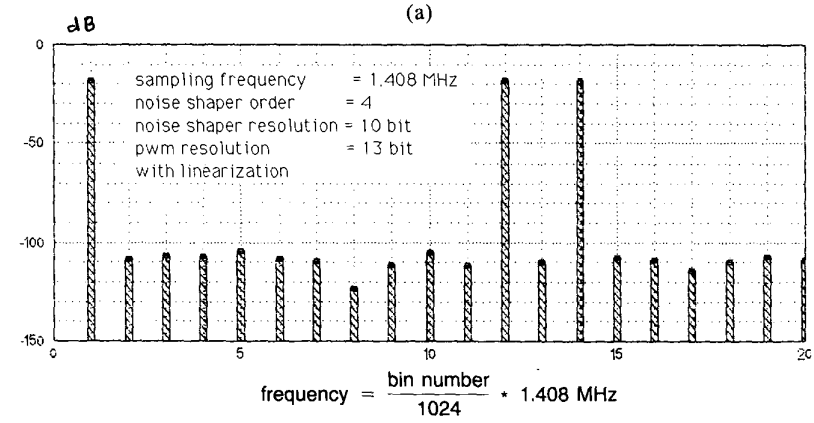


(b)

Fig. 16. Single-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and direct truncation.



(a)



(b)

Fig. 17. Single-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and direct truncation.

To compare performance, a set of results of simulations is presented in Figs. 20–23, corresponding to the system data listed in Table 2.

3.3 Pulse Construction Accuracy in PWM Systems

It has been shown that noise shaping can reduce the resolution requirement of the PWM samples, thus easing the problem of pulse construction by requiring coarser steps in pulse width. However, there remains the question of absolute accuracy for each pulse to attain an acceptable overall performance commensurate with, say, a 16-bit 44.1-kHz PCM specification. Such considerations are independent of signal processing prior to PWM modulation and relate to errors introduced after theoretical ideal conversion, where system-dependent error sources can be attributed to the following:

1) Edge jitter and timing errors
2) Slew limiting on pulse edges (meaning pulse area is not proportional to pulse width)
3) Amplitude errors (producing both scale and dynamic error)
4) Processing errors (adding a noise source to the output samples).

These sources contribute to pulse area error which, via dynamic modulation of the spectrum of each pulse including a dc term, produce output noise and distortion. However, in a system where these contributions are small and random we can, to a good approximation, consider an impulse error sequence at the uniform sampling rate f_s Hz, where each impulse weighting equates with an area error in the output pulse. As a method of performance estimation, a system resolution clock f_r Hz is defined, which represents the accuracy to which pulses are required to be timed. Thus if A is the amplitude of a PWM sample, then the pulse area error, hence error impulse weighting, ranges from $-Af_s/2f_r$ to $Af_s/2f_r$, where the PWM sampling rate f_s Hz also represents the error sequence repetition frequency. The error sequence appears as a final system level of requantization distortion where, if treated as a random noise source with a uniform probability distribution function, it contributes an extra noise source of $(Af_s/f_r)^2/12$, distributed over the half-sampling band 0 to $f_s/2$ Hz. Hence, assuming a uniform spectral power density, the noise power N_a within the half-band 0 to $f_n/2$ Hz (where f_n is the Nyquist sampling frequency, $f_n = 44.1$ kHz) is

$$N_a = \frac{A^2}{12}\frac{f_s f_n}{f_r^2} . \tag{16}$$

As a benchmark, consider an N-bit PWM system sampled at the Nyquist rate f_n Hz, where the system clock $f_r = f_n 2^N$. Then the reference in-band noise N_{ref} is

$$N_{ref} = \frac{A^2}{12}\frac{f_n^2}{(f_n 2^N)^2} = \frac{A^2}{12.2^{2N}}$$

whereby

$$\frac{N_a}{N_{ref}} = \frac{f_s f_n}{f_r^2}\, 2^{2N} .$$

To maintain a constant audio-band noise performance

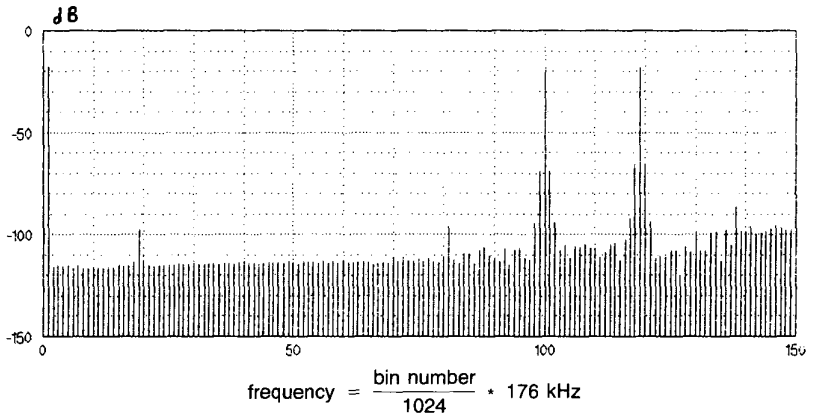


Fig. 18. PWM/noise-shaped spectrum illustrating spectral flattening. [See also Fig. 14(a).]

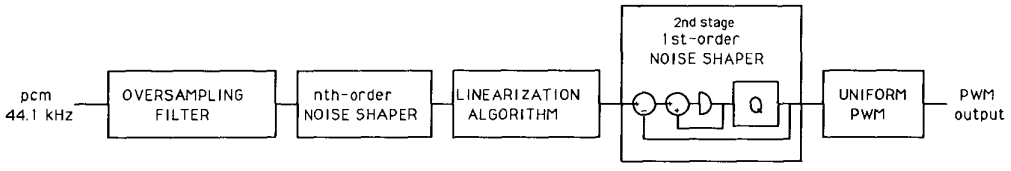


Fig. 19. Uniform sampled PWM with linearization and two stages of noise shaping.

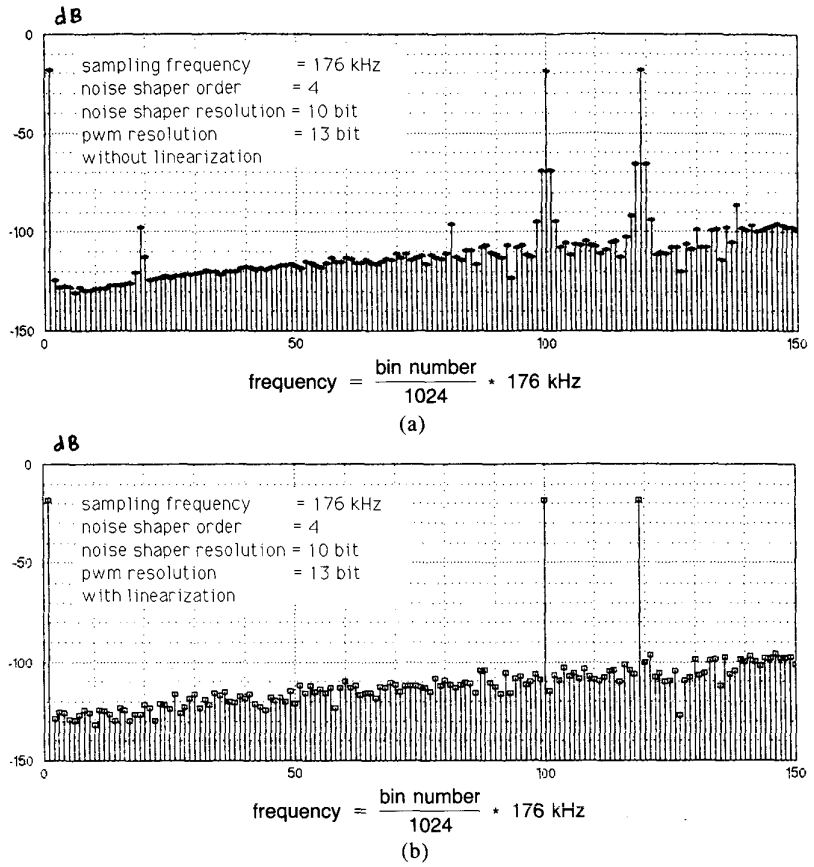


Fig. 20. Two-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and first-order noise-shaped truncation prior to PWM.

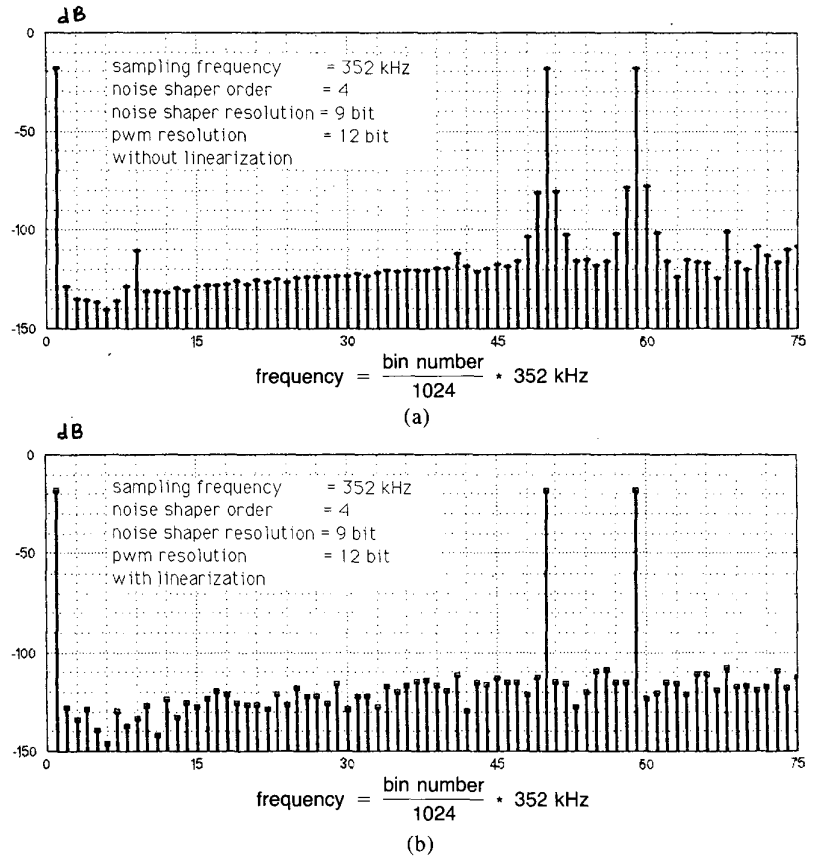
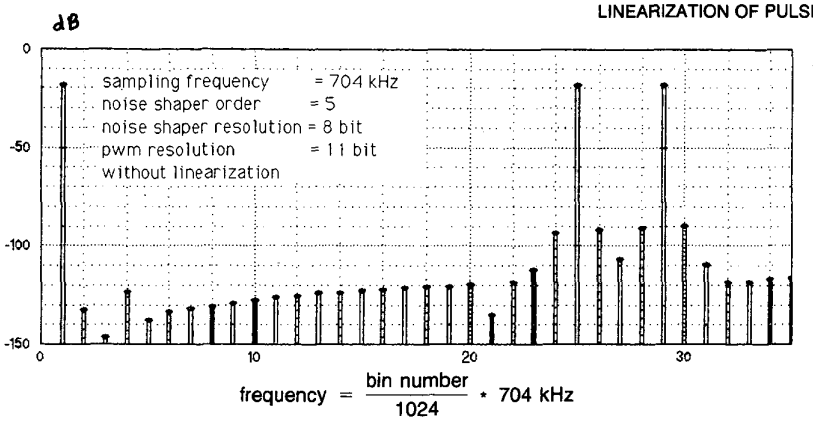
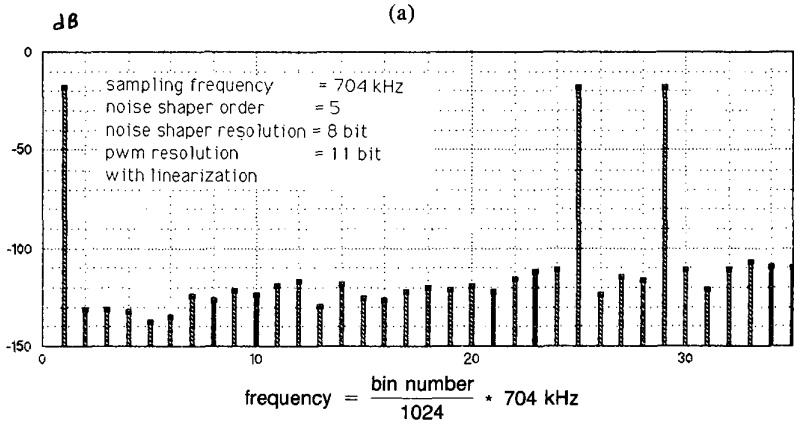


Fig. 21. Two-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and first-order noise-shaped truncation prior to PWM.

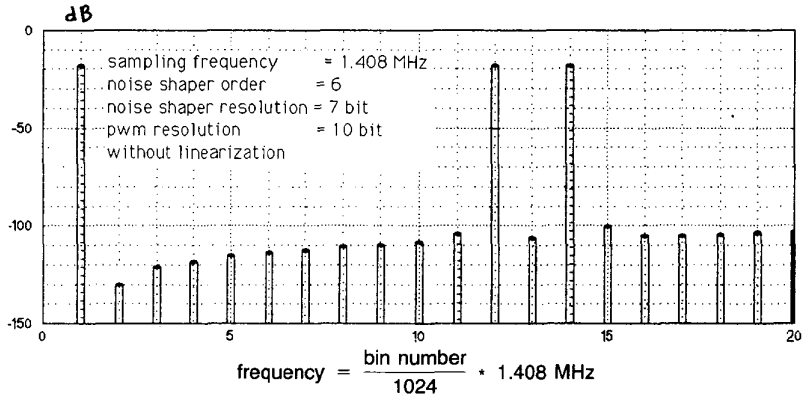


(a)

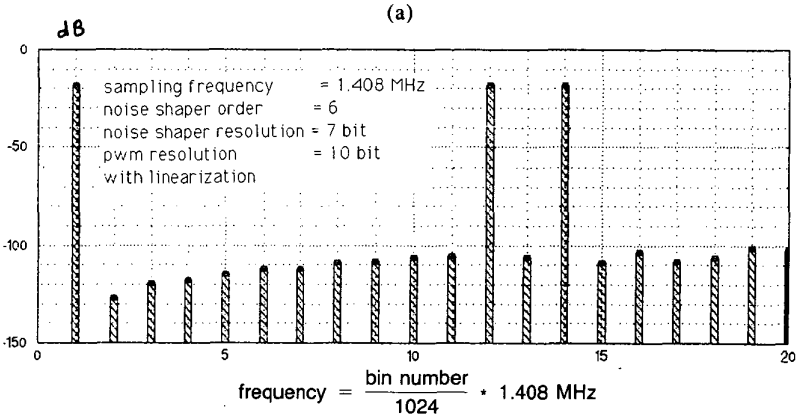


(b)

Fig. 22. Two-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and first-order noise-shaped truncation prior to PWM.



(a)



(b)

Fig. 23. Two-stage noise shaping and PWM. (a) Without linearization. (b) With linearization and first-order noise-shaped truncation prior to PWM.

with sampling rate, make $N_a/N_{ref} = 1$, that is,

$$f_r = 2^N(f_s f_n)^{1/2} . \quad (17)$$

Eq. (17) suggests that if the system dynamic range is to be maintained, even if noise shaping is used, to reduce pulse resolution, then system accuracy must actually increase in proportion to $(f_s)^{1/2}$. Consequently noise shaping can allow a more coarse quantization of pulse-width resolution, but the absolute accuracy of the timing and amplitude stability must be maintained or, indeed, increased. This is an important system design consideration, especially where the linearization algorithm described in Sec. 2 allows enhanced linearity at lower sampling rates.

4 CONCLUSION

A method of linearizing uniform sampled PWM has been described, where excellent linearity can be achieved at moderate sampling rates. Although numerous results can be computed to show the variation of distortion (for example, % total harmonic distortion) as a function of signal level and frequency, we recognize that for PWM, high-frequency, high-amplitude input signals are a particularly severe test, a fact adequately demonstrated in much of the supporting literature. Consequently the results presented here have been restricted to to a high-level three-tone equiamplitude intermodulation test, where simulations demonstrate the correction procedure to result in almost zero distortion residues within the audio band, including both intermodulation components and harmonics of the fundamental.

The nonlinear, model-based process was then extended to include noise shaping in order to reduce the resolution of the PWM code and therefore facilitate PWM modulator implementation. However, for an effective noise-shaping advantage, the sampling rate needs to be a minimum of four times, with a preference for higher rates. Also the linearization algorithm can lead to requantization, which must be controlled if degradation in dynamic range is to be minimized.

A number of computer simulations were presented that demonstrate the effectiveness of linearization for various system parameters. It was also shown that provided the optimization band is maintained at 0–22 kHz, the magnitude of the dispersive, corrective signal diminishes as the sampling rate is increased where there was greater opportunity for noise shaping to reduce the resolution of the output code, although with higher sampling rates there is some advantage in extending the optimization frequency band. However, an increase in sampling rate also implies more pulse transitions per second, reflecting a reduction in amplifier power efficiency. Also, pulse jitter, slew rate, and amplitude-related distortions, including the effects of nonideal power supplies, will limit the resolution of each pulse and thus degrade the dynamic range, implying that the promise of a very high signal-to-noise ratio with low distortion may prove illusive.

An interesting comparison therefore emerges between a noise-shaped system with a low pulse resolution requirement but higher sampling rate and a non-noise-shaped system using lower sampling rates but with a higher sample resolution requirement. It is important to observe that noise shaping was shown in Sec. 3.3 not to be a method of reducing accuracy, only resolution, where the accuracy requirement at a given sampling rate is similar for both a noise-shaped and a non-noise-shaped system, with actually a more stringent specification required at the higher sampling rates. The linearization algorithm means that uniform sampling can be used, which yields a simpler implementation of the pulse-width modulator. Although a high clock rate is required for pulse timing, the sampling rate remains modest, leading to good power efficiency. In earlier systems much higher sampling rates were reported with the uniform sampling process in order to achieve an adequate high-frequency intermodulation distortion performance.

The results presented in this paper suggest that although noise-shaped systems offer advantage, a direct-conversion PWM system using linearization is also a serious candidate, provided the pulse timing can be accommodated, and should therefore find applications in both power amplification and precision DAC systems. Clearly for high-power PWM amplifier applications there are a number of associated problems to be addressed in terms of EMC and output-stage design, although regular advances are reported and the potential for reduced sampling rates through linearization are of significance when considering output-stage efficiency. However, the output stage itself can introduce distortion due to edge jitter and edge rise time, and these in association with power-supply compliance can still generate distortion, although this is now an instrumental problem rather than a fundamental limitation. The development of highly linear digital power amplifiers is of major significance and will have wide application in digital audio systems, including digital and active loudspeaker technology.

Table 2. System data.

Figure	Loop order	NS resolution	PWM resolution	PWM sampling frequency (kHz)
20	4	10	13	176
21	4	9	12	352
22	5	8	11	704
23	6	7	10	1408

However, it is also suggested that the low-level PWM system is an attractive alternative to other forms of DAC, especially as with correct pulse edge timing it can exhibit perfect low-level linearity (commensurate with a quantized data sequence). Once linearized, PWM can be considered the dual of multilevel, oversampled DAC technology, where multilevel constant width (that is, PCM) is translated to constant-level, multiwidth coding or, indeed, other intermediate combinations of multilevel and multiwidth formats. It may well prove easier to achieve optimum resolution using pulse-edge timing with a two-level signal than the inherent complexities of designing multilevel DACs with their accuracy problems and myriad of interlevel signal combinations and associated slew-rate and pulse-area modulation errors.

The PWM DAC with reduced sampling rate is also a viable alternative to the widely used bit stream converter [17] and earlier reported look-up table techniques [18]. Although edge resolution is now more stringent, the considerable reduction in the number of edge transitions per second, typically a factor of 64, suggests lower jitter noise and associated high-frequency spuriae. There is also the potential for enhanced low-level resolution with no idle-channel artifacts. Furthermore, as a point of circuit detail, linearized PWM converters do not require either transimpedance (current-to-voltage) amplifiers, particularly significant in oversampled systems, or switched-capacitor filters, either of which can represent additional sources of noise and nonlinear distortion. It is therefore proposed that the linearized PWM DAC can represent an optimum conversion strategy that should also prove well matched to signals that include psychoacoustic noise shaping [19] to enhance overall system dynamic range.

Future research will consider the feasibility of greater reductions in the sampling rate to merge more closely the duality between PWM and PCM. Also methods of implementing the iterative linearization procedures using digital processors will be considered in order to enable real-time performance and subjective measures to be achieved. However, the extension to a fully functional power amplifier system is the main application, even though EMC criteria are severe and the need to control output-stage distortion is crucial to realizing the optimum performance of digital audio signals.

5 REFERENCES

[1] A. H. Reeves, French patent 852, 183 (1938).

[2] A. H. Reeves, British patent 535, 860 (1939).

[3] A. H. Reeves, U.S. patent 2,272,070 (1942).

[4] S. Kashiwagi, "A High Frequency Audio Power Amplifier Using a Self-Oscillating Switching Regulator," 0093-9994/85/0700-0906 *IEEE Int. Appliance Ind. Conf.* (1984 May).

[5] J. Hancock, "A Class D Amplifier Using MOSFETs with Reduced Minority Carrier Lifetime," *J. Audio Eng. Soc.*, vol. 39, pp. 650–662 (1991 Sept.).

[6] M. B. Sandler, "Digital Power Amplifier Design," Ph.D. dissertation, University of Essex, Colchester, UK (1983).

[7] M. B. Sandler, "Toward a Digital Power Amplifier," presented at the 76th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 32, p. 1009 (1984 Dec.), preprint 2135.

[8] M. B. Sandler, "Progress toward a Digital Power Amplifier," presented at the 80th Convention o the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 378 (1986 May), preprint 2361.

[9] M. B. Sandler, "Techniques for Digital Power Amplification," *Proc. Inst. Acoustics* (Reproduced Sound 3, Windermere, 1987).

[10] J. Goldberg and M. B. Sandler, "Comparison of PWM Modulation Techniques for Digital Power Amplifiers," *Proc. Inst. Acoustics* (Reproduced Sound 6), vol. 12, pt. 8, pp. 57–65 (1990 Nov.).

[11] M. O. J. Hawksford, "*N*th-Order Recursive Sigma-ADC Machinery at the Analog-Digital Gateway," presented at the 78th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, pp. 58–87 (1985 July/Aug.), preprint 2248.

[12] R. G. Bowman, R. E. Hiorns, and M. B. Sandler, "Design and Performance of a Noise Shaping, Pulse Width Modulated, Digital to Analogue Converter," presented at the IEE Colloquium on Digital Audio Signal Processing, London (1991 May).

[13] J. M. Goldberg and M. B. Sandler, "Noise Shaping and Pulse-Width Modulation for an All-Digital Audio Power Amplifier," *J. Audio Eng. Soc.*, vol. 39, pp. 449–460 (1991 June).

[14] P. H. Mellor, S. P. Leigh, and B. M. G. Cheetham, "The Implementation and Performance Enhancement of a Completely Digital Power Amplifier," *Proc. Inst. Acoustics* (Reproduced Sound 6), vol. 12, pt. 8, pp. 67–75 (1990 Nov.).

[15] P. H. Mellor, S. P. Leigh, and B. M. G. Cheetham, "Improved Sampling Process for a Digital Pulse-Width Modulated Class D Power Amplifier," presented at the IEE Colloquium on Digiital Audio Signal Processing, London (1991 May).

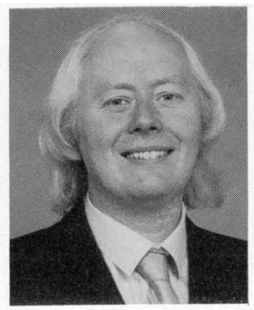
[16] M. O. J. Hawksford, "Chaos, Oversampling, and Noise Shaping in Digital-to-Analog Conversion," *J. Audio Eng. Soc.*, vol. 37, pp. 980–1001 (1989 Dec.).

[17] P. J. A. Naus, E. C. Dijkmans, E. F. Stikvoort, A. J. McKnight, D. J. Holland, and W. Bradinal, "A CMOS Stereo 16 Bit D/A Converter for Digital Audio," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 390–394 (1987 June).

[18] M. O. J. Hawksford, "Multi-Level to 1 Bit Transformations for Applications in Digital-to-Analogue Converters Using Oversampling and Noise Shaping," *Proc. Inst. Acoustics*, vol. 10, pp. 129–143 (1988 Nov.).

[19] R. A. Wannamaker, "Psycho-Acoustically Optimal Noise Shaping," presented at the 89th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, p. 871 (1990 Nov.), preprint 2965.

THE AUTHOR



Malcolm Omar Hawksford is a reader in the Department of Electronic Systems Engineering at the University of Essex, where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honors B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship, where the field of study was the application of delta modulation to color television and the development of a time compression/time multiplex system for combining luminance and chrominance signals. Since his employment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing, and loudspeaker systems has been undertaken. Since 1982 research into digital crossover systems has begun within the group and, more recently, oversampling and noise shaping investigated as a means of analog-to-digital/digital-to-analog conversion.

Dr. Hawksford has had several AES publications that include topics on error correction in amplifiers and oversampling techniques. His supplementary activities include writing articles for *Hi-Fi News* and designing commercial audio equipment. He is a member of the IEE, a chartered engineer, a fellow of the AES and of the Institute of Acoustics, and a member of the review board of the *AES Journal*. He is also a technical adviser for *HFN* and *RR*.

PAPERS

Linearization of Multilevel, Multiwidth Digital PWM with Applications in Digital-to-Analog Conversion*

MALCOLM OMAR HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

The process of uniformly sampled pulse-width modulation is investigated as a means of high-performance digital-to-analog conversion where comparisons of duality are made with signals in standard modulated format. Emphasis is placed on using linearization algorithms to suppress nonlinear distortion products, and a new digital-to-analog conversion architecture is proposed that employs digital linearization together with a multiwidth and multiamplitude signal format configured within a current-steering autocalibration system topology.

0 INTRODUCTION

Research [1]–[3] has shown that partial linearization of uniformly sampled pulse-width modulation (PWM) can be achieved by introducing a modified data sequence that takes account of the dynamic distortion inherent within the spectral domain when the width of a pulse is modulated. An oversampling factor of times 4 was used both to relax the design of the dynamic bidirectional recursive compensation and to allow latitude for low-order analog filters to achieve signal reconstruction in digital-to-analog conversion.

In this paper we explore the possibility of using lower orders of oversampling with the aim of implementing a low-distortion PWM converter that can operate at twice the Nyquist rate of a typical digital audio system. The advantage of lower sampling rates is a reduction in the number of pulse transitions per second and a corresponding lowering of the system clock rate in situations that do not use noise shaping. Also, a near-Nyquist sampled PWM system can be considered the dual of the near-Nyquist sampled pulse-code modulated (PCM) data. The former maintains constant amplitude but variable width, whereas the latter has constant width and variable amplitude.

The method of linearizing near-Nyquist sampled PWM is extended to a more general pulse format using both multilevel and multiwidth structures. This both reduces the resolution required of the PWM component and lowers the pulse timing clock rate. The technique of combining multiwidth and multilevel pulses enables a parallel digital-to-analog conversion (DAC) architecture to be implemented, a system we designate a flash DAC (FDAC). It is shown that an FDAC can be synthesized from an array of identical fast current-switching cells, and because of the low number of pulse transitions per Nyquist sample, the system offers low jitter noise and slew-rate distortion. Finally a calibration procedure is introduced for the FDAC, which, in association with a stable low-jitter clock source and fast logic circuits, offers a further candidate for a high-quality DAC for use in digital audio systems.

1 DESIGN OF STATIC COMPENSATION FILTER

It has been shown [3], [4] that nonlinearity in uniformly sampled PWM arises because of variations in the spectral shape within the audio band when the width of the PWM samples changes. Thus instead of the overall spectrum being scaled in proportion to the area under a pulse, modulation of pulse shape means that only the dc term is proportional to the pulse area while other regions of the spectrum undergo varying gain changes. A fundamental requirement of any DAC is that the transfer function of all samples remain in exact proportion and only undergo uniform scaling with the signal level, a condition readily met by rectangular PCM samples of constant width but varying level. This is a theoretical ideal where in practice slew-rate limiting, pulse jitter, and a nonconstant pulse crest can lead to dynamic spectral errors and are factors that contribute to nonideality in the DAC process.

* Presented at the 97th Convention of the Audio engineering Society, San Francisco, CA, 1994 November 10–13.

In Fig. 1 a rectangular pulse is shown centered on a sampling instant where the overall width can vary from 0 to $1/f_{ns}$, f_{ns} being the Nyquist sampling frequency in hertz. Also shown is a normalized family of spectral domain plots for a rectangular pulse of varying pulse width computed over the band of 0 to f_{ns} Hz. This result is computed directly from the Fourier transform of a normalized rectangular pulse $F_p(f)$ of width m/f_{ns}, where m is the modulation depth, $0 < m < 1$,

$$F_p(f) = \frac{\sin(m\pi f/f_{ns})}{m\pi f/f_{ns}}. \quad (1)$$

The dc component of $F_p(f)$ is normalized to unity for all m.

Next we consider the design of a uniformly sampled digital finite impulse response (FIR) filter that can compensate for the spectral error within the band of 0 to $f_{ns}/2$ Hz. However, unlike the examples considered earlier [3], the correction band now extends to approximately one-half of the sampling frequency. Therefore to achieve similar correction accuracy, an increase in the number of taps in the FIR filter is required. Also, to reduce the magnitude of the correction coefficients, a target transfer function $\{\sin(m_t\pi f/f_{ns})/(m_t\pi f/f_{ns})\}$ is used, which, in the first instance, corresponds to a rectangular pulse of duration $1/2f_{ns}$. Here the target modulation index $m_t = 0.5$, although for the multilevel case a value of $m_t = 1.0$ is preferred as this width is dominant at higher signal levels. The equalization transfer function $E(m, f)$ over the band of 0 to $f_{ns}/2$ Hz follows from Eq. (1),

$$E(m,f) = \frac{m}{m_t}\frac{\sin(m_t\pi f/f_{ns})}{\sin(m\pi f/f_{ns})} \quad \text{matched over 0 to } f_{ns}/2 \text{ Hz}. \quad (2)$$

The impulse response of the FIR correction filter can be obtained from the inverse Fourier transform of $E(m, f)$. However, this process is approximate as the number of coefficients in the FIR filter is finite and because spectral replication causes a peak of $f_{ns}/2$ Hz, as shown in Fig. 2. Spectral replication about the sampling frequency and its harmonics means that correction cannot be applied to frequencies $\geq f_{ns}/2$. Also, modulation products are reflected back into the baseband and prevent a total cancellation of distortion.

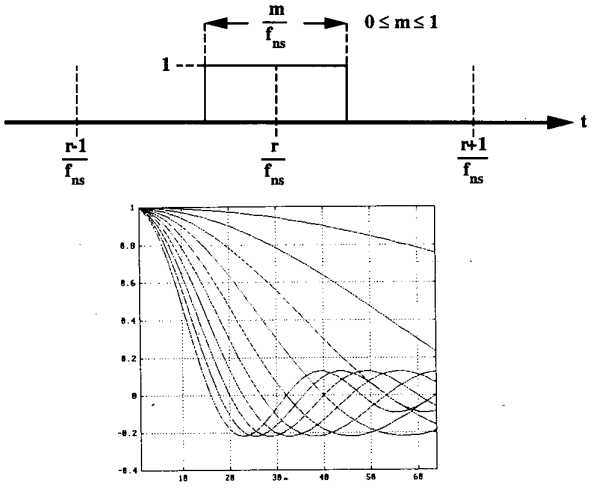


Fig. 1. Pulse of normalized width 1 together with family of corresponding transforms.

1.1 Window Function

The first proposal for the approximation process is to introduce a guard band of width f_g Hz, consisting of a frequency and amplitude scaled, reflected image of the response within the acceptance band to form a smooth overall function with no discontinuities in the first derivative. This strategy reduces the higher order terms of the signal within the transform signal space, which in this case yields less time dispersion in the nonrecursive equalization filter. The construction of the overall function is shown in Fig. 3, where f_0 Hz is the transition frequency defining the boundary between acceptance band and guard band and is defined as

$$f_0 = \frac{f_{ns}}{2} - f_g. \quad (3)$$

The function within the acceptance band of 0 to f_0 Hz is given by Eq. (2). Its value A_0 at $f = f_0$ follows as

$$A_0 = \frac{m}{m_t}\frac{\sin[m_t\pi(0.5 - f_g/f_{ns})]}{\sin[m\pi(0.5 - f_g/f_{ns})]}. \quad (4)$$

To form the scaled and reflected function $A(m, f)$ in the guard band that smoothly interfaces with $E(m, f)$, the function for $E(m, f)$ is scaled by a factor k, then offset and reflected about $f_{ns}/2$ to give

$$A(m,f) = \frac{1}{k}\left\{A_0(1+k) - \frac{m}{m_t}\frac{\sin[m_t\pi k(0.5 - f/f_{ns})]}{\sin[m\pi k(0.5 - f/f_{ns})]}\right\}. \quad (5)$$

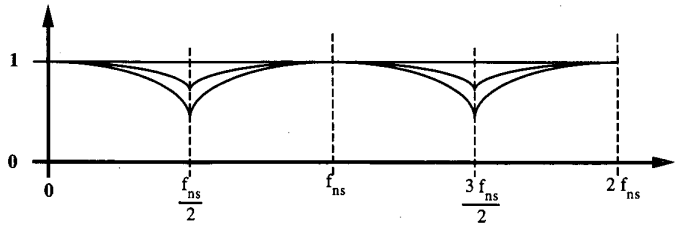


Fig. 2. Transfer function $E(f)$ of digital FIR correction filter for $m = 0$, 0.5, and 1.

The scaling factor k is determined by equating the arguments of the sine functions in Eqs. (2) and (5) at $f = f_0$,

$$k = \frac{f_{ns}}{2f_g} - 1 . \quad (6)$$

Examination of Eq. (5) shows that $A(m, f_0) = A_0$ and that differentiation of $E(m, f)$ and $A(m, f)$ reveals a continuous functional form over the transition frequency f_0.

Consider a symmetrical transversal filter with coefficients $\{a_0, a_1, \ldots, a_n\}$, where the transfer function $F_c(f)$ is given by

$$F_c(f) = a_0 + 2\sum_{r=1}^{n} a_r \cos\left(\frac{2\pi rf}{f_{ns}}\right) . \quad (7)$$

The FIR filter design requires a transfer function that is matched to the composite equalization filter response $E_c(f)$, where

$$E_c(m,f) = E(m,f) + A(m,f) . \quad (8)$$

Here $E(m, f)$ is valid for $0 < f < f_0$, whereas $A(m, f)$ is valid for $f_0 < f < 0.5f_{ns}$.

The coefficient set $\{a\}$ can be determined by forming a set of simultaneous equations to match the windowed Fourier transform of a rectangular pulse corresponding to a modulation depth m (where $0 < m < 1$) with that of the finite data sequence of the FIR filter. Assuming NC independent coefficients in a symmetrical FIR filter, then using a uniform frequency distribution where $f = if_{ns}/2NC$ and defining a matrix element $Y(r, i)$, where

$$Y(r, i) = 2\cos\left(\frac{\pi ri}{NC}\right) \quad (9a)$$

the matrix equation is defined as

$$[E_c] = [Y][a] . \quad (9b)$$

Inversion of Eq. (9b) produces the NC coefficients where, to preserve the dc content of the input data sequence, the central coefficient a_0 is given by

$$a_0 = 1 - 2\sum_{r=1}^{n} a_r . \quad (10)$$

For the special case where the sampling rate is $2f_{ns}$, the guard band f_g can be set to $f_{ns}/2$, which relaxes the filter design. By following the procedure discussed in Section 1.1, the scaling factor $k = 1$ and the target response show odd symmetry about $f_{ns}/4$ Hz. Consequently, even-order terms in the equalization filter are zero, and the filter can be compared with a half-band filter where alternate samples are zero.

1.2 Scaling of Frequency-Domain Sampling

An alternative approach to the estimation of coefficients is to concentrate the frequency domain sampling over a subset of the Nyquist band. Hence if there are NC unknown coefficients, the NC samples are focused into the relevant frequency space, thus improving the approximation in that region.

The scaling is achieved by introducing a factor μ into the arguments of Eqs. (2) and (9a),

$$E(m,f) = \frac{m}{m_t}\,\frac{\sin(m_t\mu\pi f/f_{ns})}{\sin(m\mu\pi f/f_{ns})}$$

matched over 0 to $\mu f_{ns}/2$ Hz (11)

$$Y(r, i) = 2\cos\left(\frac{\mu\pi ri}{NC}\right) . \quad (12)$$

A similar analysis using Eq. (9b) is then used to determine the transversal filter coefficients $\{a\}$.

Fig. 4 shows example approximation functions for filters using four coefficients for $\mu = 0.5$ and 0.9. Equalization characteristics are computed for discrete m in the range $0 \leq m \leq 1$ in steps of 0.1. The results show that by reducing the upper frequency limit of equalization, improved in-band approximations result as the frequency domain sampling is concentrated into a narrower bandwidth. For $\mu = 0.9$ the in-band approximation is less accurate.

Although coefficient sets can be calculated for each value of m, this is computationally inefficient. Thus two methods are proposed:

1) A look-up table is computed for coefficient values against m and linear interpolation performed for intermediate values. Eq. (10) is always used to ensure that there is no overall loss of pulse area.

2) Polynomial approximations are estimated from the discrete m values following a similar procedure reported in [3].

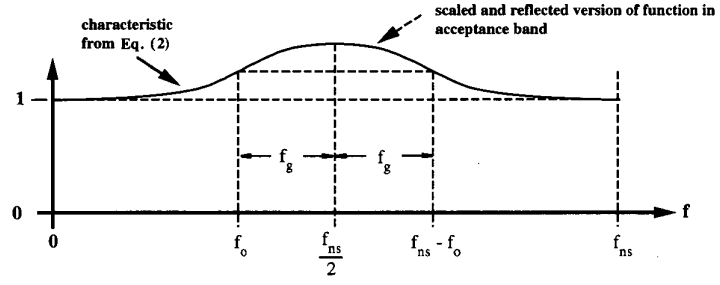


Fig. 3. Modified correction characteristic to eliminate discontinuities in first derivative.

2 DYNAMIC FILTER SIMULATION FOR BINARY PWM

The filter design procedures presented in Sections 1.1 and 1.2 show how a filter coefficient set can be calculated for a particular m value of input data. However, in a general PCM sequence ip(L), each pulse can take any value within the data range and requires individual correction for the PCM-to-PWM mapping. Effectively, a unique FIR filter is required for each input data sequence, where for sample L a coefficient set (of NC independent coefficients) is determined and is represented using the notation $a(x, y)$ as follows:

Coefficient Set for Sample L

Sample L	$a(L, \text{NC}), a(L, \text{NC}-1), \ldots a(L, 1), \ldots a(L, \text{NC}-1), a(L, \text{NC})$

The filter structure is shown in Fig. 5, where the coefficients are not constant but are uniquely associated with each sample value. Thus as the samples are shifted through the FIR filter, the coefficients are also shifted through the individual coefficient sets. In Fig. 6 the filter process is extended to show the computation of three adjacent output samples op(L), op($L + 1$), and op($L + 2$), corresponding to input samples ip(L), ip($L + 1$), and ip($L + 2$).

As reported earlier [3], the calculation of coefficients requires iterative adaptation as the output samples are dependent on both the present sample and contributions from adjacent FIR filters. In this paper the "single-sided, unidirectional shifting algorithm" is modified for use with NC coefficients and is illustrated using matrix notation. For NC = 2,

$$\begin{bmatrix} a(1,3) & 0 & 0 & 0 & 0 & 0 \\ a(1,2) & a(2,3) & 0 & 0 & 0 & 0 \\ a(1,1) & a(2,2) & a(3,3) & 0 & 0 & 0 \\ a(1,2) & a(2,1) & a(3,2) & a(4,3) & 0 & 0 \\ a(1,3) & a(2,2) & a(3,1) & a(4,2) & a(5,3) & 0 \\ 0 & a(2,3) & a(3,2) & a(4,1) & a(5,2) & a(6,3) \\ 0 & 0 & a(3,3) & a(4,2) & a(5,1) & a(6,2) \\ 0 & 0 & 0 & a(4,3) & a(5,2) & a(6,1) \\ 0 & 0 & 0 & 0 & a(5,3) & a(6,2) \\ 0 & 0 & 0 & 0 & 0 & a(6,3) \end{bmatrix} \begin{bmatrix} \text{ip}(1) \\ \text{ip}(2) \\ \text{ip}(3) \\ \text{ip}(4) \\ \text{ip}(5) \\ \text{ip}(6) \\ \text{ip}(7) \\ \text{ip}(8) \\ \text{ip}(9) \\ \text{ip}(10) \end{bmatrix} = \begin{bmatrix} \text{op}(1) \\ \text{op}(2) \\ \text{op}(3) \\ \text{op}(4) \\ \text{op}(5) \\ \text{op}(6) \\ \text{op}(7) \\ \text{op}(8) \\ \text{op}(9) \\ \text{op}(10) \end{bmatrix}.$$

Matrix [a] is initialized such that all coefficients are zero except for elements $a(x, 1) = 1$. Computation starts by calculating op(3) and the corresponding coefficient set $\{a(3, 1), a(3, 2), a(3, 3)\}$, as described previously, where upon the new elements take their places in matrix [a]. The procedure is repeated for the next row and so on throughout the matrix, forming, for this example, an output op(7). The input vector [ip] is then shifted and the [a] matrix diagonally shifted by $a(x + 1, y + 1) = a(x, y)$ and $a(1, 1) = 1$. The procedure then repeats to determine the next output sample. The iterative procedure accommodates the nonlinear interaction between samples due to the PWM process and allows the coefficient values to converge to a stable solution, although in practice more rows in the [a] matrix may be desirable.

To validate the procedure for calculating coefficients for use with binary PWM, a simulation program was run over NS = 2048 samples, where $m_t = 1$ and the input signal is dc biased at $x = 0.5$,

$$x(L) = 0.5 + 0.3\,[\sin(k_0 f_{ns} t) + \sin(k_1 f_{ns} t) + \sin(k_2 f_{ns} t)] + \text{rnd} * 2^{-17}. \quad (13)$$

Here $k_0 = 2\pi/\text{NS}$, $k_1 = 300k_0$, $k_2 = 310k_0$ and rnd has a uniform probability density function from -0.5 to 0.5 and $t = L/f_{ns}$ for $\leq L \leq$ NS. Fig. 7 shows computed output spectra for the PWM process with and without correction for $\mu = 0.5$ and NC = 6.

3 MULTILEVEL DAC WITH PWM INTERPOLATION

Direct DAC using PWM requires a high clock rate to time the pulse transitions. For example, 16-bit resolution at a 44.1-kHz sampling rate corresponds to a PWM clock of 2.89 GHz. Although this rate is high, when the timing and jitter performance of a high-quality DAC are considered, this rate is not unreasonable, and fast-counting circuits suggest that economic realization will be feasible with clocks offering subnanosecond jitter.

However, although techniques now exist using oversampling and noise shaping which can substantially reduce the timing clock rate, alternative systems that combine multilevel and PWM are also attractive. For example, if we assume a 7-bit multilevel DAC (128 levels), then for 20-bit resolution, a PWM edge timing clock rate of 44.1 kHz $* 2^{(20-7)} = 361$ MHz appears feasible.

Two desirable characteristics of an audio DAC are accurate conversion of low-level signals and a monotonic coding characteristic. It is here that the PWM technique offers significant advantage as a progressive change in pulse width is synonymous with monotonic coding of small signals. Of course, once a multilevel format is employed, there is an additional requirement to match

the levels to a high degree of accuracy. However, provided the number of levels is not too great, this can be achieved by a parallel array of identical cells that can be progressively switched in and out of the circuit as required. It is proposed that 7-bit (128-level) multilevel coding can be handled by a parallel architecture and that economic automatic calibration is feasible within a VLSI system.

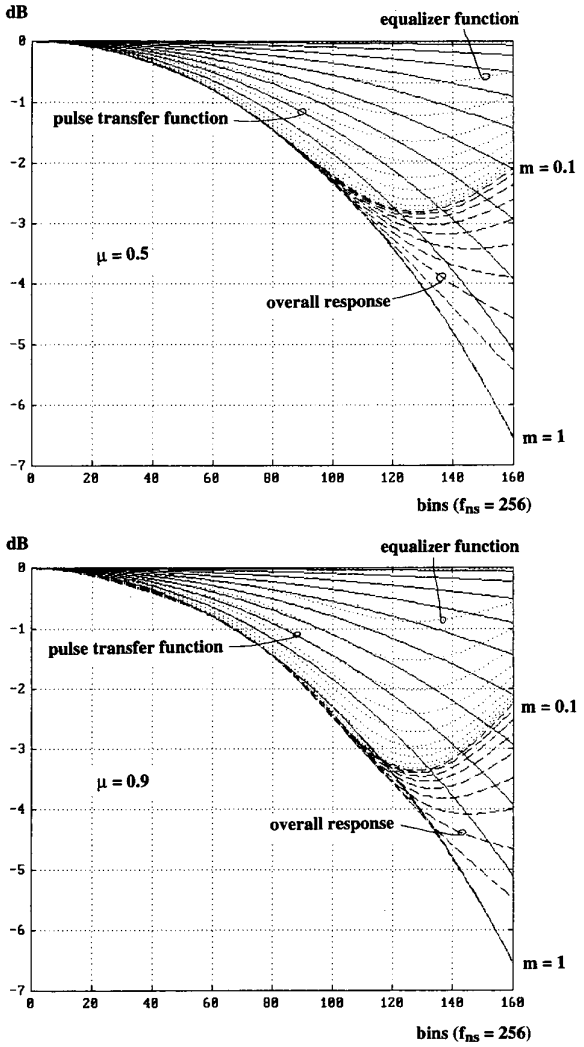


Fig. 4. Approximate equalization characteristics for $\mu = 0.5$ and $\mu = 0.9$.

Before considering PWM interpolation with linearization, a possible FDAC architecture is outlined based on a current-steering topology. The system is shown in Fig. 8 and uses emitter-coupled logic (ECL) bistables to drive differential current switches. The use of ECL allows fast pulse transitions, which are necessary to give precise edge definition that is compatible with low-jitter clock sources. Also, it is imperative that the current sources be matched to a defined reference current, although this matching must also include the effect of current loss, for example, in the bases of the switching transistors. A possible calibration procedure nominates one current source as reference and then compares individually its value against the remaining sources. This can be done by alternate switching and adjusting the current level until no perceived ac component appears on the output. The technique theoretically has a high accuracy because of the following attributes:

1) Observing the error as an ac signal allows a high-gain error amplifier without dc drift problems.

2) Switching rates at the DAC output are kept low, reducing effects of slew distortion and jitter.

3) Fine adjustment current values for each cell can be stored in a local memory and DAC.

4) Current steering is fast and with appropriate capacitive elements on the output, slew-induced distortion is eliminated as there are no rapid dv/dt effects or feedback amplifiers being momentarily operated near open loop [8].

Once the multilevel DAC current sources are calibrated, PWM can be used to interpolate between output levels with guaranteed monotonicity and potentially high

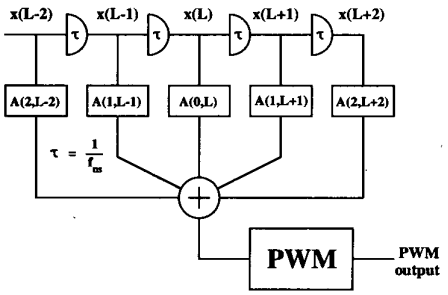


Fig. 5. Five-coefficient FIR filter.

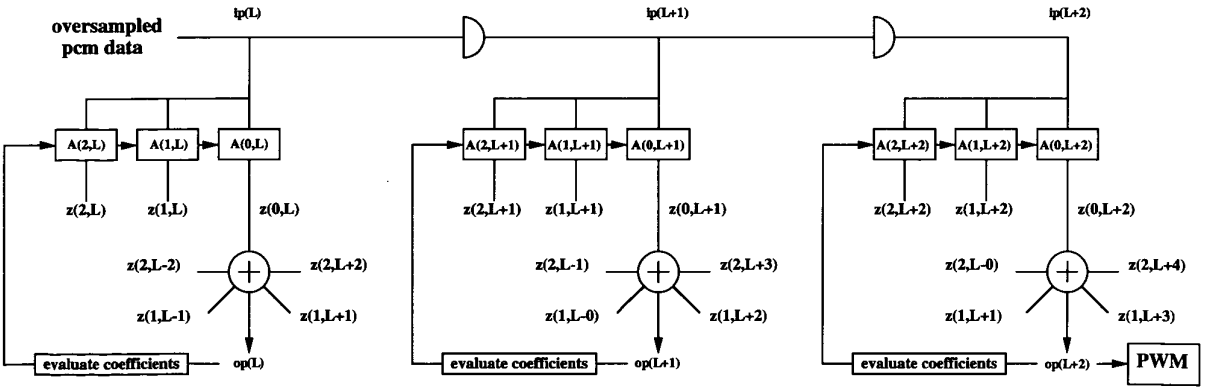


Fig. 6. Interconnected coefficient sets where each set is calculated from knowledge of surrounding samples.

linearity, where the proposed form of PWM interpolation can take a variety of formats. However, to reduce switching rates and to have a monotonic change in pulse area with progressive changes in input data, the two formats illustrated in Fig. 9 have been selected and designated types 1 and 2.

Type 1, as shown in Fig. 9(a), offers reduced switching transitions but has a step change in pulse width at each multilevel boundary, which may add broad-band distortion within the correction process. However, type 2 has symmetry about even quanta and maintains similar pulse widths either side of a multilevel boundary. These differences can be seen by examining the progressive pulse structures in Fig. 9(a) and (b).

In type 1, for an increasing signal level the pulse width always grows from the center switching between two adjacent output DAC levels. When the pulse width equals the sample period, a new pulse emerges at the center switching up to the next quantization level. However, in type 2 the pulse alternates between outward growth from the center and inward growth from the pulse edges. The symmetrical pulse structure guarantees zero dynamic phase distortion, and the PWM format yields a smooth transition between adjacent DAC levels with no repeated codes. In the DAC interpolation process only one cell is switched at a time, and, between adjacent

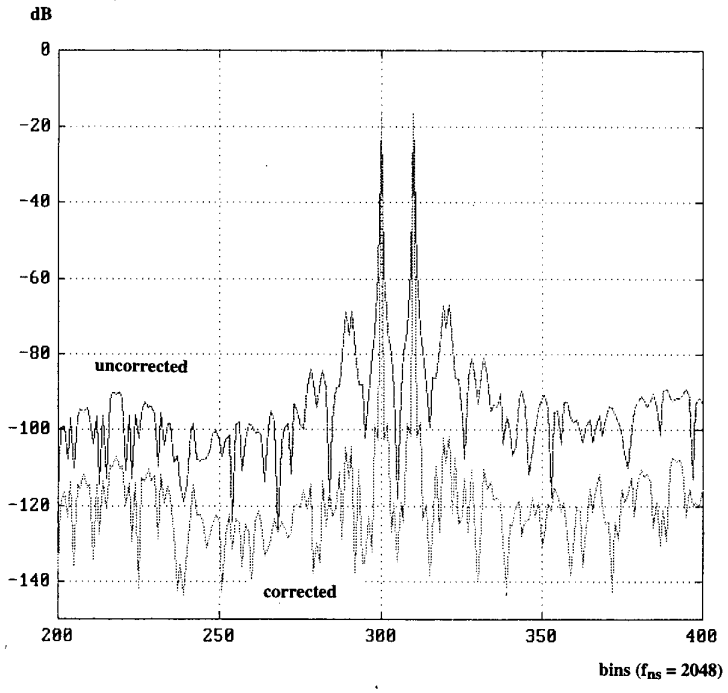


Fig. 7. Output spectra for binary PWM for μ = 0.5 and NC = 6.

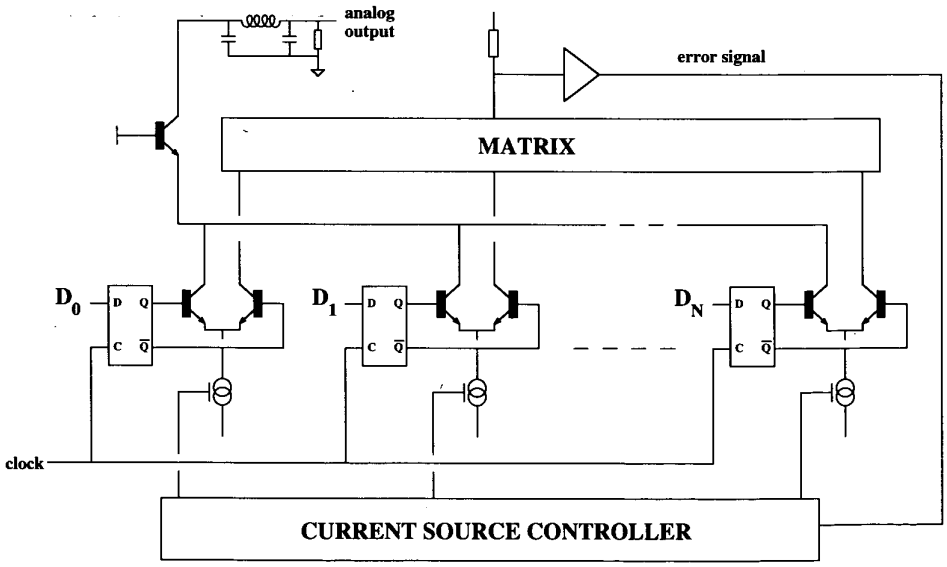


Fig. 8. Basic FDAC system topology.

samples, current-switching cells become active in a controlled and progressive sequence.

For the type 1 sample format, consider a sample $x(L)$ of amplitude $\{N + m\}$, where N is an integer and $0 < m < 1$, that is,

$$x(L) = N + m . \tag{14}$$

N corresponds to the Nth quantization level of the FDAC, which has a width $1/f_{ns}$, whereas m determines the intermediate sample width m/f_{ns} (see example pulse in Fig. 9(a)) of PWM interpolation between levels N and $(N + 1)$. The dc normalized transfer function $F_{np}(f)$ of this composite pulse is

$$F_{np}(f) = \left[\frac{\sin(m\pi f/f_{ns})}{\pi f/f_{ns}} + N\frac{\sin(\pi f/f_{ns})}{\pi f/f_{ns}}\right]\frac{1}{N+m} .$$

If the normalized target transfer function is $\sin(\pi f/f_{ns})/(\pi f/f_{ns})$, corresponding to $m_t = 1$, then the equalizer transfer function $E_m(m, f)$ follows as

$$E_m(m,f) = \frac{1}{F_{np}(f)}\frac{\sin(\pi f/f_{ns})}{\pi f/f_{ns}}$$

whereby

$$E_m(m,f) = \frac{N+m}{N + \sin(m\pi f/f_{ns}/\sin(\pi f/f_{ns})} . \tag{15}$$

For $N = 0$ this reduces to

$$E_m(f,m)\Big|_{N=0} = \frac{m\sin(\pi f/f_{ns})}{\sin(m\pi f/f_{ns})} \tag{16}$$

which is equivalent to Eq. (2) for $m_t = 1$.

The reason for selecting $m_t = 1$ is that for $N >> 1$, $E_m(m, f) \approx 1$, which reduces the equalizer coefficient values, hence dependence on correction, and improves

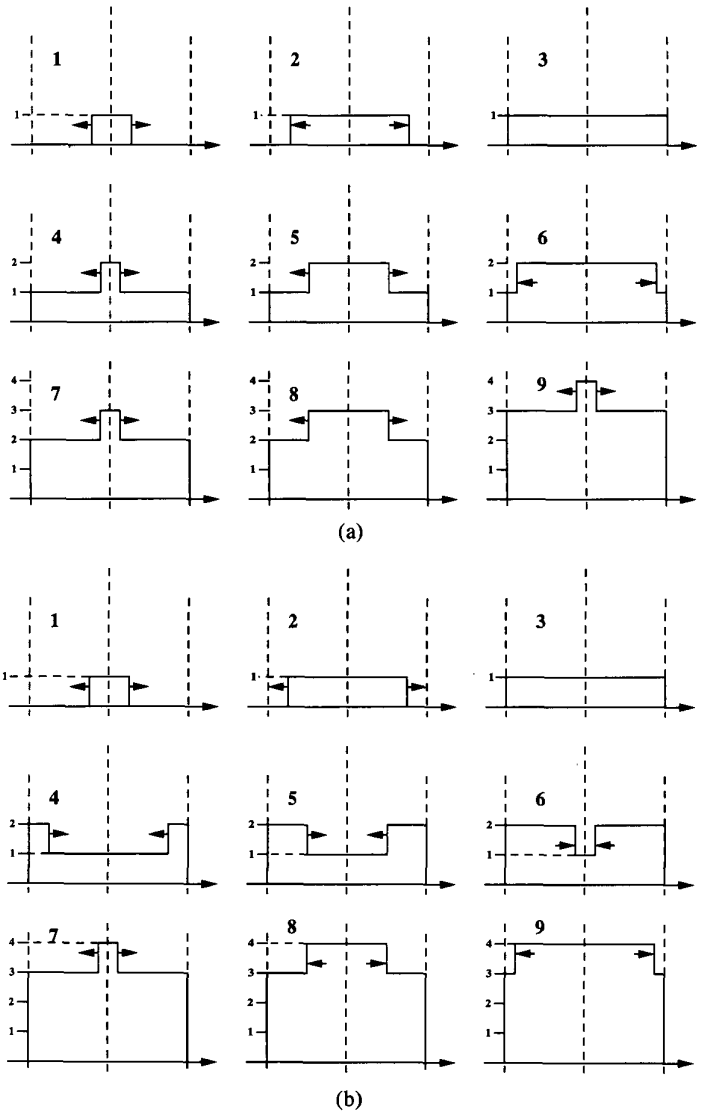


Fig. 9. Simultaneous pulse-amplitude and pulse-width modulation. (a) Type 1. (b) Type 2.

the accuracy of the matching functions.

If a type 2 pulse format is used, the process is similar, except that N in Eqs. (14) and (15) is rounded to even integer values and m takes bipolar values, where $-1 < m < 1$.

Using Eq. (15), the optimum equalization characteristics are plotted in Fig. 10 as functions of m and f for both type 1 and 2 pulse formats, where type 1 reveals more ripples at the integer boundaries.

Because of the PWM format, a dynamic correction filter is required to operate in a mode similar to that of two-level PWM. However, observation of Fig. 10 shows that the form of correction differs in detail such that as the pulse amplitude is increased, there is a corresponding dilution of spectral modulation. In fact spectral modulation is greatest for the lowest levels, where fortunately absolute distortion levels are lowest. The equalization function $E_m(m, f)$ for the multilevel code is the reciprocal of the transfer function of the composite pulse multipled by the target transfer function corresponding to $m_t = 1$.

To determine coefficients for the FIR filter, a procedure similar to that described in Section 1.1 can be followed by introducing a guard band of f_g Hz and forming a function $A_c(m, f)$ which takes a scaled and reflected form of the function in the acceptance band. At the transition band edge f_0 let $E_c(m, f) = A_c$,

$$A_c = \left(1 + \frac{m}{N}\right) \frac{\pi(0.5 - f_g/f_{ns})}{\sin[\pi(0.5 - f_g/f_{ns})]} \left\{1 + \frac{1}{N} \frac{\sin[m\pi(0.5 - f_g/f_{ns})]}{\sin[\pi(0.5 - f_g/f_{ns})]}\right\}^{-1}. \quad (17)$$

The function $A_c(m, f)$ in the guard band is then given by

$$A_c(m,f) = \frac{1}{k}\left\{A_c(1 + k) - \frac{\pi k(0.5 - f/f_{ns})}{\sin[\pi k(0.5 - f/f_{ns})]} \frac{1 + \frac{m}{N}}{1 + \frac{1}{N} \frac{\sin[m\pi k(0.5 - f/f_{ns})]}{\sin[\pi k(0.5 - f/f_{ns})]}}\right\} \quad (18)$$

where the scaling factor k again follows from Eq. (6). Using Fourier analysis, a set of coefficients can then be determined directly. Alternatively, the technique described in Section 1.2 can be used, where the equalization bandwidth is set by selecting the factor μ.

As $x(L)$ increases above unity, the variations in the transfer function become progressively reduced, which relaxes the iterative evaluations of the coefficients. Also, the coefficients can be estimated using interpolation, which simplifies overall computation. In Fig. 11 a coefficient map for NC = 6 is presented for signal levels corresponding to a range of $m = 0$ to ~7, where the general trends can be observed. Further simplification also results as positive and negative signals of equal magnitude can be assigned the same coefficient values.

To demonstrate the performance of the multilevel DAC, results are plotted in Figs. 12 to 15 for type 1 and type 2 pulse formats, with $\mu = 0.5$ and 0.9, respectively. For these examples, the input excitation is defined by

$$x(L) = 50[\sin(k_0 f_{ns} t) + \sin(k_1 f_{ns} t) + \sin(k_2 f_{ns} t)] + \text{rnd} * 2^{-17}. \quad (19)$$

Again, $k_0 = 2\pi/\text{NS}$, $k_1 = 300k_0$, $k_2 = 310k_0$, and rnd has a uniform probability density function from -0.5 to 0.5 and $t = L/f_{ns}$ for $1 \leq L \leq \text{NS}$ where Ns = 2048. Data relating to each computation are presented in Table 1.

The results show that, because of the multilevel structure of the signal, once the higher quantization levels are excited, the distortions become noiselike rather than discrete frequencies. The error-correction process reduces this distortion significantly and also shows that there is little advantage gained in using more than NC = 6.

4 CONCLUSION

This paper has considered the application of linearization to the PWM process approached from the spectral domain, where improved linearization of uniformly sampled PWM was achieved. The work extends earlier results by achieving better spectral matching, which allowed a reduction in the system sampling rate close to the Nyquist frequency.

Although the technique is directly applicable to two-level PWM, a multilevel converter was also explored as a means of enhancing linearity and also for reducing the internal clock rates necessary for timing the PWM edges.

A feature of the converter was the monotonicity of pulse areas as a function of the input sample level together with dynamic linearization to compensate for spectral domain modulation with the signal level. It was demonstrated that low-level signals show less distortion in the corrected FDAC together with a general reduction in PWM-related distortion at higher levels.

An advantage of combining a calibrated FDAC with PWM interpolation is to reduce switching rates at the digital–analog gateway. Fast switching of pulses using current steering without associated high dv/dt is an important attribute in minimizing distortion. Also low logic propagation times and stable clock sources lower jitter-related errors.

In comparison with bit-stream technology, the number of signal transitions at the output of the DAC is reduced greatly, which offers the potential of lower jitter sensitivity. It is also suggested that the problems of RF circuit

design and layout, commonly associated with interconnecting bit-stream integrated circuits, are reduced, even though high clock rates remain necessary for timing the width of the PWM pulses. The technique overcomes the need to use excessive oversampling ratios and orders of noise shaping to achieve an acceptable clock rate. By their nature such processes imply a greater sensitivity to jitter and slew rate artifacts as the number of significant signal transitions is considerably higher.

This paper has presented an alternative DAC for high-performance digital audio. The technique offers a fast and readily calibrated architecture, where known nonlinearities can be reduced to low levels. It does not require excessive oversampling ratios or the use of noise shaping and therefore moves against the trend of many bit-stream DACs. However, for power DAC applications [5], the use of binary PWM remains attractive where alternative error-correction strategies have been reported [6], [7].

5 REFERENCES

[1] P. H. Mellor, S. P. Leigh, and B. M. G. Cheetham, "Improved Sampling Process for a Digital Pulse Width Modulated Class D Power Amplifier," presented at the IEE Colloquium on Digital Audio Signal Processing, London, UK (1991 May).

[2] P. Craven, "Toward the 24-bit DAC: Novel Noise-Shaping Topologies Incorporating Correction for the Nonlinearity in a PWM Output Stage," *J. Audio Eng. Soc.*, vol. 41, pp. 291–313 (1993 May).

[3] M. O. J. Hawksford, "Dynamic Model-Based Linearization of Quantized Pulse-Width Modulation for

Table 1. Computational information for type 1 and 2 multilevel PWM spectra.

Fig. 12: Type 1 pulse format	
μ = 0.5, NS = 2048, NC = 2, 4, 6, and 8	
Frequency range 1 to 1000 bins,	amplitude range −150 dB to 40 dB
Fig. 13: Type 1 pulse format	
μ = 0.9, NS = 2048, NC = 2, 4, 6, and 8	
Frequency range 1 to 1000 bins,	amplitude range −150 dB to 40 dB
Fig. 14: Type 2 pulse format	
μ = 0.5, NS = 2048, NC = 2, 4, 6, and 8	
Frequency range 1 to 1000 bins,	amplitude range −150 dB to 40 dB
Fig. 15: Type 2 pulse format	
μ = 0.9, NS = 2048, NC = 2, 4, 6, and 8	
Frequency range 1 to 1000 bins,	amplitude range −150 dB to 40 dB

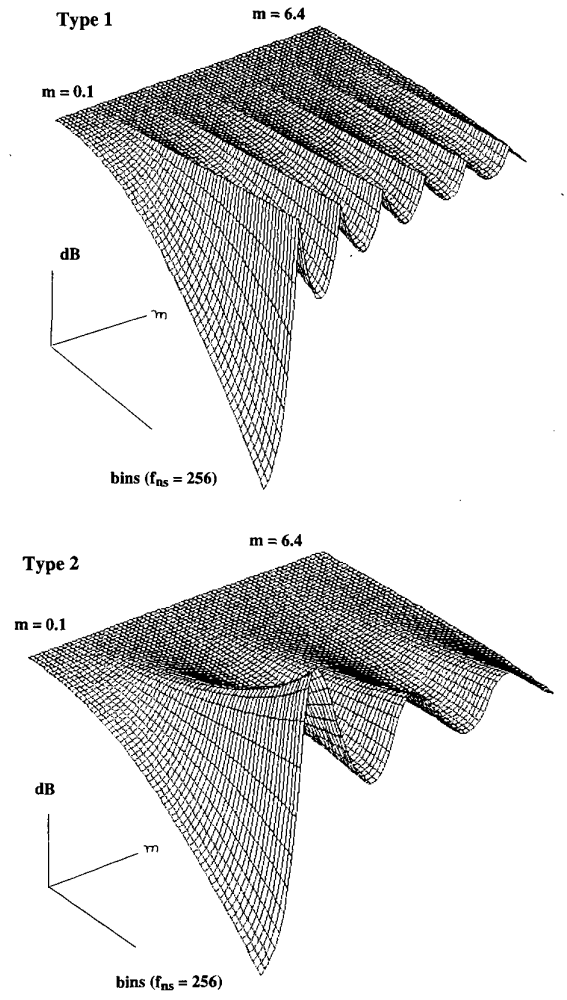


Fig. 10. Optimum equalization $E_m(m, f)$ as a function of amplitude and frequency for type 1 and 2 pulse formats.

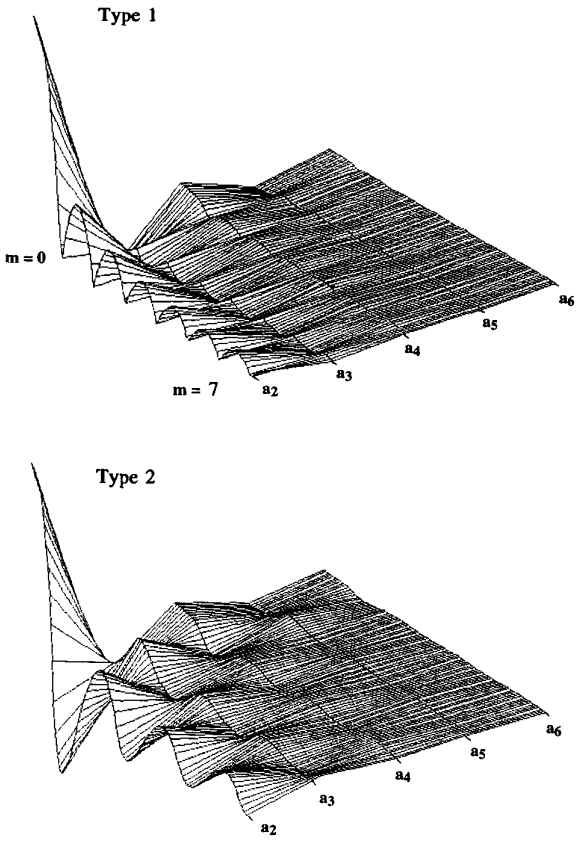


Fig. 11. Coefficient maps for NC = 6 (a_1 not shown) and m = 0 to 7.

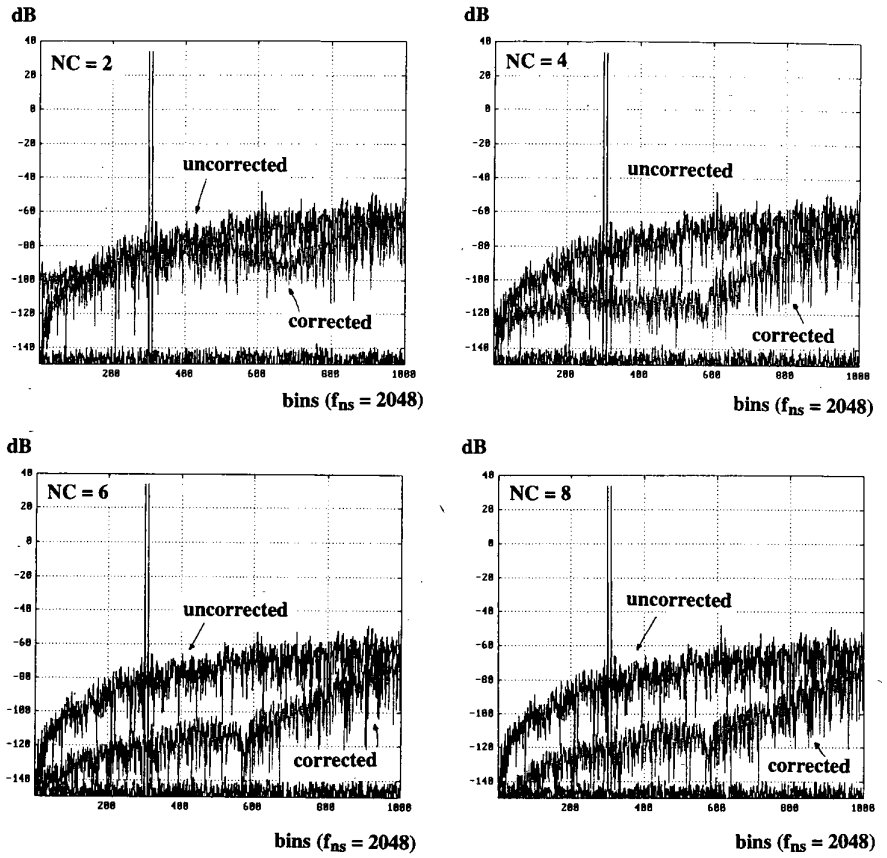


Fig. 12. Output spectra of multilevel DAC. Type 1, μ = 0.5.

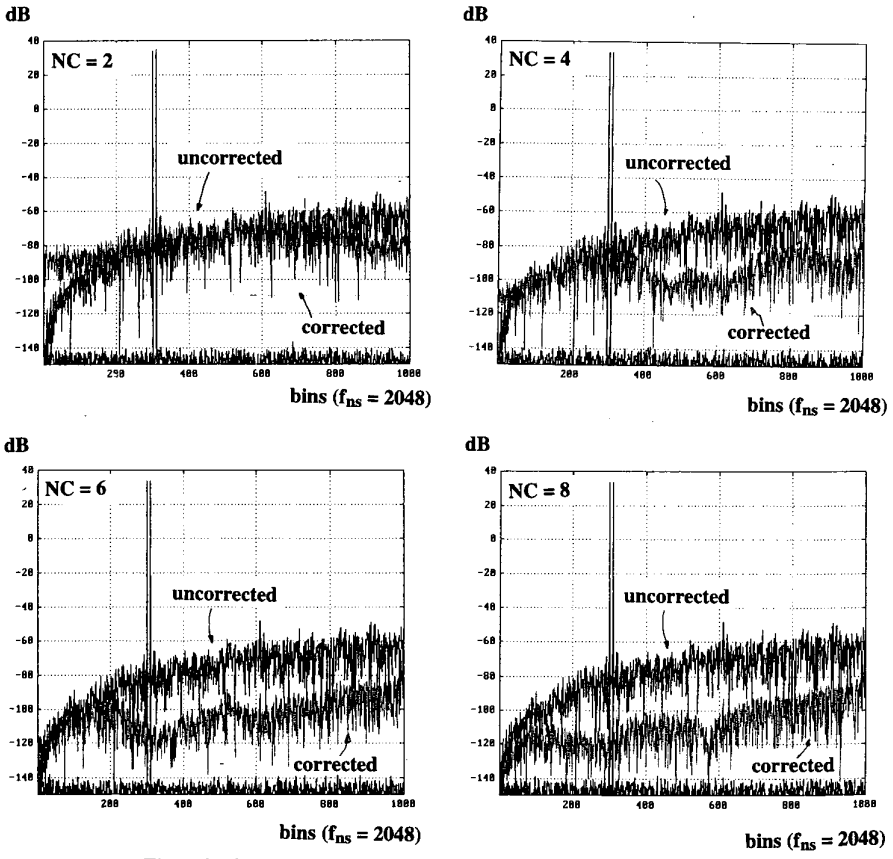


Fig. 13. Output spectra of multilevel DAC. Type 1, μ = 0.9.

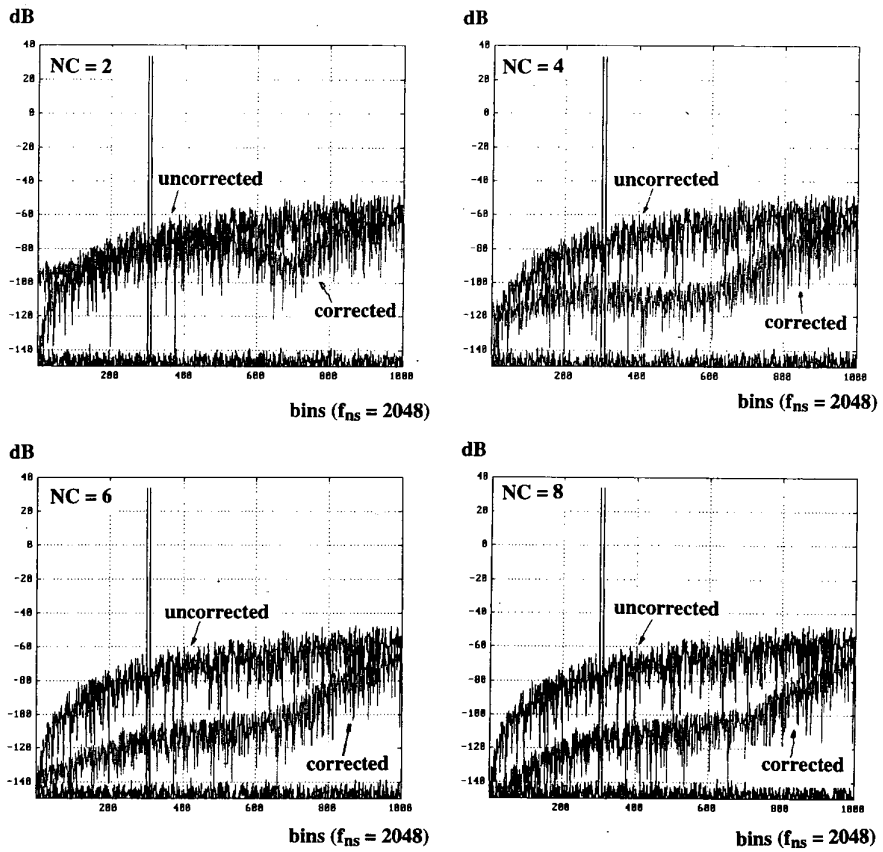


Fig. 14. Output spectra of multilevel DAC. Type 2, μ = 0.5.

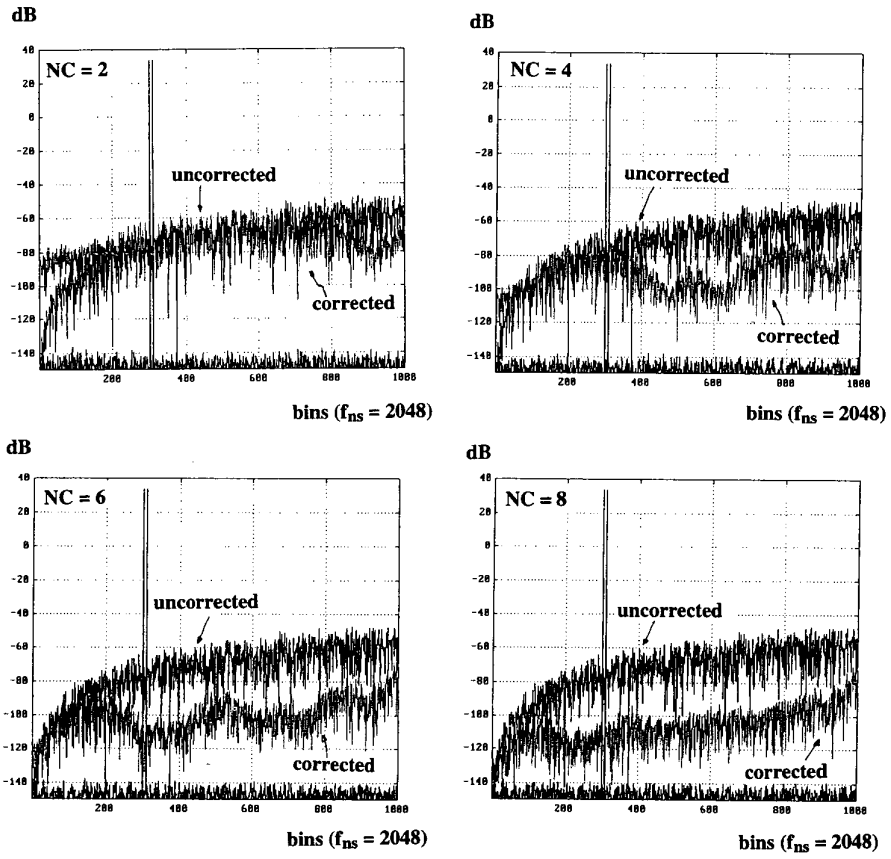


Fig. 15. Output spectra of multilevel DAC. Type 2, μ = 0.9.

Applications in Digital-to-Analog Conversion and Digital Power Amplifier Systems," *J. Audio Eng. Soc.*, vol. 40, pp. 235–252 (1992 Apr.).

[4] M. O. J. Hawksford, "Multi-Level to 1 bit Transformations for Applications in Digital-to-Analogue Converters Using Oversampling and Noise Shaping," *Proc. Inst. Acoust.*, vol. 10, pp. 129–143 (1988 Nov.).

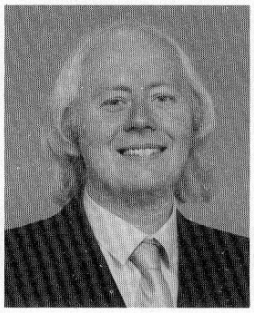
[5] J. M. Goldberg and M. B. Sandler, "Noise Shaping and Pulse-Width Modulation for an All-Digital Audio Power Amplifier," *J. Audio Eng. Soc.*, vol. 39, pp. 449–460 (1991 June).

[6] S. Logan, "Linearization of Class D Output Stages for High Performance Audio Power Amplifiers," M.Sc. dissertation, Essex University, Colchester, UK (1993 Autumn).

[7] M. O. J. Hawksford and S. Logan, "Linearization of Class D Output Stages for High-Performance Audio Power Amplifiers," presented at the 2nd International IEE Conference on Analogue-to-Digital and Digital-to-Analogue Conversion, Cambridge University, Cambridge, UK (1994 July).

[8] M. O. J. Hawksford, "Digital-to-Analog Converter with Low Intersample Transition Distortion and Low Sensitivity to Sample Jitter and Transresistance Amplifier Slew Rate," *J. Audio Eng. Soc.*, vol. 42, pp. 901–917 (1994 Nov.).

THE AUTHOR



Malcolm Omar Hawksford is director of the Centre for Audio Research and Engineering, a professor in the Department of Electronic Systems Engineering at the University of Essex, and Scheme Director for the M.Sc. in Audio Systems Engineering, where his interests encompass audio engineering, electronic circuit design, and signal processing. Professor Hawksford studied electrical engineering at the University of Aston in Birmingham where he gained a First Class Honours BSc and Ph.D. The Ph.D. program, sponsored by the BBC via a Research Scholarship, investigated the application of delta modulation for color television and the development of a time-compression/time-multiplex system for combining luminance and chrominance signals. While at Essex University, he has undertaken research principally in the fields of analog amplifiers, digital signal processing, and loudspeaker systems. Since 1982 research into digital crossover networks and equalization for loudspeakers has been pursued, which has culminated in an advanced digital and active loudspeaker system being designed within the university. Research topics have also encompassed oversampling and noise shaping techniques applied to analog-to-digital and digital-to-analog conversion and the linearization of PWM encoders.

Professor Hawksford has published in the *Journal of the Audio Engineering Society* on topics that include error correction in amplifiers, oversampling techniques, and MLS techniques. His supplementary activities include writing for *Hi-Fi News and Record Review* and designing high-end analog and digital audio equipment. He is a chartered engineer and is a fellow of the AES, the Institution of Electrical Engineers, and the Institute of Acoustics. He is a member of the technical committee of Acoustic Renaissance for Audio (ARA), a group currently promoting a system for storing multichannel, high-definition audio signals on high-capacity DVD optical disks. He is also technical adviser to *HFN and Record Review* and a technical consultant to LFD Audio, UK.

IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, VOL. 51, NO. 9, SEPTEMBER 2004

An Oversampled Digital PWM Linearization Technique for Digital-to-Analog Conversion

Jin-Whi Jung and Malcolm J. Hawksford

***Abstract*—An algorithmic-based linearization process for uniformly sampled digital pulsewidth modulation (PWM) is described. It is shown that linearization of the intrinsic distortion resulting in uniformly sampled PWM can be achieved by using a fractional delay digital filter embedded within a noise shaping re-quantizer. A technique termed direct PWM mapping is proposed as a pre-compensation filter scheme for applications in high-resolution digital-to-analog conversion.**

***Index Terms*—Digital-to-analog (D/A) conversion, pulsewidth modulation (PWM), sigma–delta modualtion (SDM).**

I. Introduction

TIME-DISCRETE pulsewidth modulation (PWM) has a natural synergy with digital circuitry used within both digital systems and more general very large-scale integrated (VLSI) devices. Since MASH [6], the amalgamation of low-bit uniformly sampled PWM and noise shaping re-quantization has invited considerable research, not least because it relaxes the problem of quantizer saturation encountered in two-level quantization sigma–delta modulatoion (SDM). However, the intrinsic nonlinearity of uniformly sampled PWM is a well-known problem that can produce significant levels of in-band distortion whereas naturally sampled PWM, as commonly used in analog PWM systems, is linear within the baseband frequency range. Consequently, uniformly sampled PWM is subject to spectral distortion implying the filtered output signal is partially corrupted, see Craven [3], Rowe [11], and Leigh [12] for detailed discussions.

Several research papers that address the subject of linearization in PWM have been reported. Mellor *et al.*, [5] and Leigh [12] introduced a method using time-domain interpolation, while Goldberg [2] and Goldberg and Sandler [4] presented a more refined interpolation technique to approximate uniformly sampled PWM to natural sampling. Hawksford [1] developed a novel uniformly sampled PWM distortion compensation technique for uniformly sampled PWM that was later adapted to embrace multilevel, multiwidth PWM [10]. Craven [3] then proposed a similar compensation technique but where error correction was implemented using negative feedback incorporated within a noise shaping loop.

Manuscript received January 4, 2002; revised March 28, 2004. This paper was recommended by Associate Editor U.-K. Moon.

J.-W. Jung is with the Department of Electronic and Electric Engineering, University College London, London WC1E 6BT, U.K. (e-mail: j.jung@ucl.ac.uk).

M. J. Hawksford is with Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: mjh@essex.ac.uk).

Digital Object Identifier 10.1109/TCSI.2004.834487

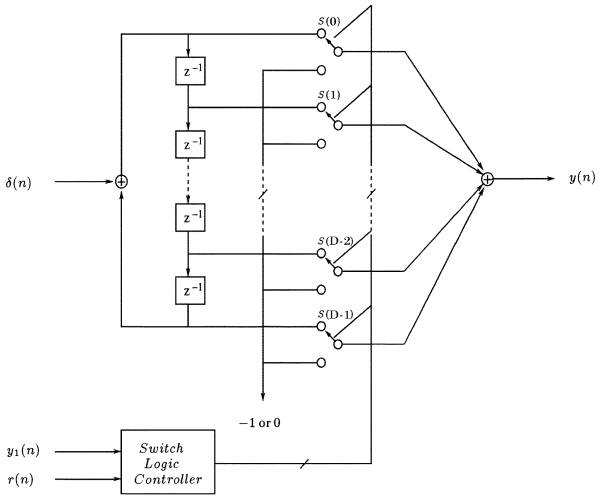


Fig. 1. Implementation of digital PWM in which 2-level quantizers s_i are connected within circular formation and gated by the switch logic controller according to α, where D is the PWM resolution bits, $\delta(n)$ is the unit impulse, $r(n)$ is the PWM reference signal, $y_1(n)$ is the output of the noise shaping re-quantizer, and $y(n)$ PWM output.

This paper presents a fractional sample interpolation scheme to reduce the intrinsic distortion resulting from the uniformly sampled PWM. The technique designated direct PWM mapping (DPM), exploits a Farrow structure fractional delay finite-impulse response (FIR) digital filter employing third-order Lagrange interpolation. The principal feature of the DPM scheme is that the normalized fractional sample grid of the Farrow structure directly generates the digital PWM time base defining the PWM output-wave transitions. It is shown here that a 3-tap Farrow structure FIR filter is sufficient for linearization and as such contributes to the simplicity of the overall PWM system.

II. Digital PWM and Intrinsic Distortion

Fig. 1 shows the proposed digital PWM where the modulating signal α is derived from a logic combination of both the quantized output signal $y_1(n)$ and the PWM reference signal $r(n)$. Let α be scaled to an integer value i so that $i \in (0, D-1)$ where $i \in \mathbb{Z}$ and D is derived from the pulse-repetition period $T_D = DT_s$, i.e., integer multiples D of the sampling period $T_s = 1/f_s$. The signal α addresses a switch state control circuit where s_1 and s_0 denote the on-state and off-state switches, respectively. The PWM time-domain sequence $h(\alpha, n)$ is therefore given as

$$h(\alpha,n) = h(\alpha, n-D) + s_1(0)\cdot\delta(n) + s_1(1)\cdot\delta(n-1) + \cdots + s_1(\alpha-1)\cdot\delta(n-\alpha+1)$$

$$+ s_1(\alpha)\cdot\delta(n-\alpha) + s_0(\alpha+1)\cdot\delta(n-\alpha-1) + \cdots + s_0(D-1)\cdot\delta(n-D+1). \quad (1)$$

Alternatively, in z domain, $h(\alpha, n)$ is expressed as

$$H(\alpha, z) = \frac{1}{1-z^{-D}} \cdot \left(\sum_{i=0}^{\alpha} s_1(i)\cdot z^{-i} + \sum_{i=\alpha+1}^{D-1} s_0(i)\cdot z^{-i} \right). \quad (2)$$

The last term of (2) is required for zero padding in order to represent the logic state 0 within a pulse-repetition period for unipolar leading-edge PWM; also, if s_0 is set to -1 it represents negative magnitude padding in bipolar leading-edge PWM.

Expressions (1) and (2) provide an intuitive understanding of digital PWM.

1) The sampled data z^{-i} is gated by corresponding 2-level quantizers and connected within a circular formation.
2) While the sampled data $z^{-\alpha}$ relates to pulse-position modulation (PPM) where the summation of impulses from 0 to α constitutes a rectangular pulse having pulsewidth α.

In this manner, the denominator $R(z) = 1/(1-z^{-D})$ determines the PWM pulse repetition with D roots.

For the purpose of illustration, if the allocation of α and D are chosen arbitrarily, noting that all the logic states of the switches $s_1(i)$ are asserted 1, then the periodic impulse response sequence will appear as

$$h = \begin{bmatrix} s_1(0) & s_1(1) & \cdots & s_1(\alpha) & 0 & \cdots & 0 \\ s_1(0) & s_1(1) & s_1(2) & \cdots & s_1(\alpha) & \cdots & 0 \\ s_1(0) & \cdots & s_1(\alpha) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ s_1(0) & s_1(\alpha) & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 & \cdots & 0 \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}. \quad (3)$$

Note also that the transfer function of the digital PWM can be expanded in an infinite geometric series

$$H(\alpha, z) = (1 + z^{-1} + \cdots + z^{-\alpha} + 0 + \cdots + 0) \cdot (1 + z^{-D} + z^{-2D} + \cdots). \quad (4)$$

The pulse-repetition period D has a base of $1-z^{-D}$, hence there exist D frequencies within the Nyquist interval with D roots of unity. In this notation each PWM bit corresponds to the delays at each sampling instant. Defining $\omega_i = (2\pi f_i/f_s) = (2\pi i/D)$, the roots of $H(\alpha, z)$ are obtained by solving $z^D = 1$ for D solutions, that is D roots of unity in the digital PWM system. The zeros of the denominator $R(z)$ that constitute dips in spectral response form D roots of unity on the unit circle, where if D corresponds to the PWM bit resolution, then, generally, $D = 2^M$ where M is even integer.

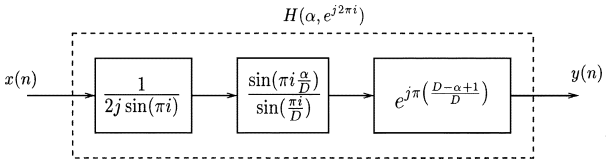


Fig. 2. Uncompensated PWM system represented as a nonlinear magnitude response and a variational delay response, derived from (5).

Hence, from (2), the transfer function $H(\alpha, \exp[j2\pi i])$ may be derived as

$$H(\alpha, e^{j2\pi i}) = \frac{1}{1-e^{-j2\pi i}} \cdot \frac{1-e^{-j2\pi i\frac{\alpha}{D}}}{1-e^{-j\frac{2\pi i}{D}}} = \frac{1}{2j\sin(\pi i)} \cdot \frac{\sin\left(\pi i\frac{\alpha}{D}\right)}{\sin\left(\frac{\pi i}{D}\right)} \cdot e^{j\pi i\left(\frac{D-\alpha+1}{D}\right)}. \quad (5)$$

The transfer function (5) represents conventional uncompensated digital PWM that is characterized in the frequency domain by a nonlinear magnitude response, see Fig. 2. As such, this process can be compared to a similar result given previously by Hawksford [1], where the linear magnitude component of a PWM pulse is multiplied by a nonlinear transfer function, i.e., $H(\alpha, \exp[j2\pi i]) = \alpha \cdot H(\exp[j2\pi i])$ that results from modulating the pulsewidth, also [10] extended this discussion to include multilevel digital PWM.

For large values of D, the transfer function of $H(\alpha, \exp[j2\pi i])$ in (5) becomes

$$H(\alpha, e^{j2\pi i}) = \frac{1}{2j\sin(\pi i)} \cdot \frac{\sin\left(\pi i\frac{\alpha}{D}\right)}{\frac{\pi i}{D}} \cdot e^{j\pi i\left(\frac{D-\alpha+1}{D}\right)} \quad (6)$$

where (6) shows that transfer function $H(\alpha, \exp[j2\pi i])$ has an α-dependent nonlinearity of the form $\sin(\alpha x)/x$ that describes deterministically the nonlinear modulation process inherent in uniformly sampled digital PWM. From (6), it follows that for sufficiently small α, the transfer function $H(\alpha, \exp[j2\pi i])$ approximates to

$$H(\alpha, e^{j2\pi i}) = \frac{\alpha}{2j\sin(\pi i)} \cdot e^{j\pi i\left(\frac{D-\alpha+1}{D}\right)}. \quad (7)$$

The spectral distortion described by (5) can be calculated directly using computer simulation, where over a bandwidth of 192 kHz, Fig. 3 reveals a typical spectrum for uniformly sampled PWM. The intermodulation components that occur at the multiples of sampling frequency of 48 kHz can also be calculated using the mathematical results in [11], they are observed for each harmonic of the pulse-repetition frequency, where their magnitudes become lower and spectra broaden as frequency increases. Also, reflected distortion components can be seen within the baseband close to the fundamental frequency of 750 Hz. Both these classes of distortion are exploited in this study to assess the performance of PWM linearization and are estimated by simulation.

1) *Harmonic distortion:* With an input signal of frequency $f = 6$ kHz sampled initially at 48 kHz and using four times upsampling, PWM for both 6-bit leading-edge sampling Fig. 4(a) and 7-bit double edge sampling Fig. 4(b) are simulated to observe the reflected spectral distortion appearing within baseband. Harmonics $2f$ and $3f$ are the

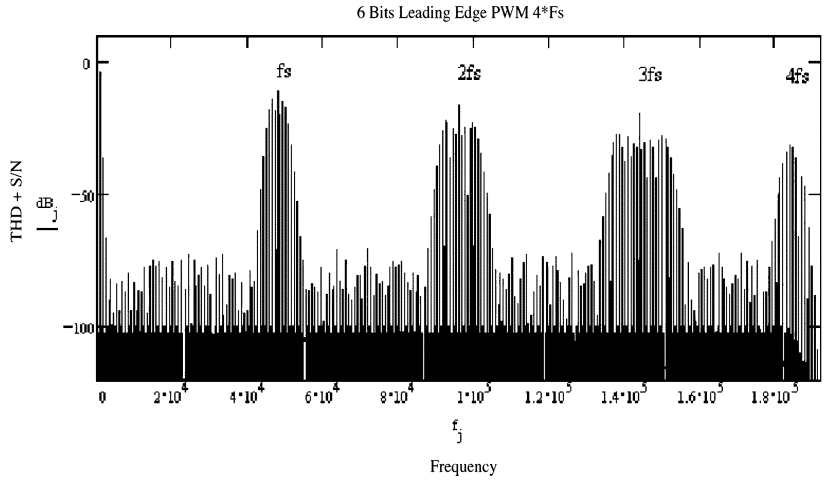


Fig. 3. Typical spectra of uniformly sampled PWM; intermodulation harmonic distortions at each multiples of PWM pulse-repetition frequency 48 kHz and reflected harmonic distortions in baseband, where the input frequency is 750-Hz single tone.

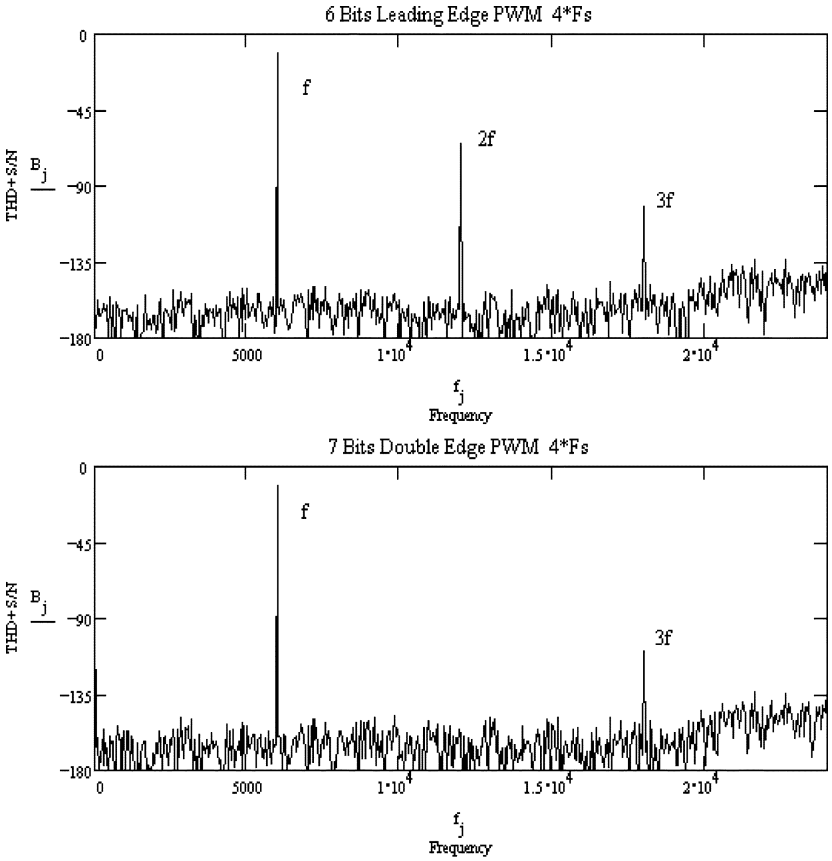


Fig. 4. PWM distortion in audio frequency band. (a) A 6-bit leading-edge sampling. (b) A 7-bit double edge sampling PWM systems at 4 times $f_s = 48$ kHz, input signal is 6 kHz.

reflected distortion components with respect to the 6-kHz input signal. The results confirm that double edge sampling achieves even order harmonic cancellation where for example the component $2f = 12$ kHz is absent, while it remains for leading-edge PWM.

2) *Intermodulation distortion:* The intermodulation results are shown in Fig. 5 where three superimposed input frequencies $f_1 = 0.75$ kHz, $f_2 = 6$ kHz, and $f_3 = 9$ kHz drive a 6-bit $4f_s$ PWM and hence the sampling rate is 192 kHz. The observed intermodulation distortion components are located both sides of the fundamental frequencies f_1, f_2, and f_3 and their multiples.

III. DPM

A. Windowing and Transfer Functions

In the following discussion, only leading-edge sampling digital PWM is presented although DPM can also be implemented for use with double-edge sampling PWM.

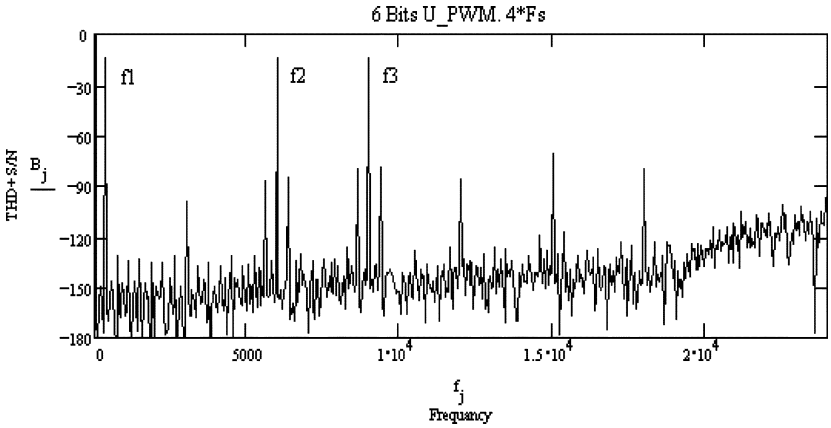


Fig. 5. Intermodulation distortion simulation results, the input frequencies are $f_1 = 0.75$ kHz, $f_2 = 6$ kHz, and $f_3 = 9$ kHz.

DPM is a dynamic switch operation map applied to digital PWM that is composed of a circular structure of 2-level quantizers whose delay length varies with α where, $0 \le \alpha < D$, consequently, D is the pulse-repetition period. Let a time sequence $\{\cdots, k = D, k+1 = 2D, k+2 = 3D, \cdots\}$ be defined to determine each PWM pulse edge sampling instant where each has the sampled grid $i = 0, 1, 2, \cdots, D-1$ used to upsample in between two consecutive sequences. For notational convenience, in this section, we normalize this time sequence so as D is set at 1, and hence, α is correspondingly fractional.

Based upon this iterative definition, (1) can be rewritten as a difference equation

$$y(\alpha, k+1) = y(\alpha, k) + \sum_{i=0}^{\alpha} \frac{\sin(\pi(k-i))}{\pi(k-i)} + \sum_{i=\alpha+1}^{D-1} \Xi(i) \quad (8)$$

where $\Xi(i) = 0$ or $\Xi(i) = -1$ by the PWM definition in (2). Further, let the impulse function in (8) be defined as

$$h_{\text{pwm}}(\alpha, k) = \sum_{i=0}^{\alpha} \frac{\sin(\pi(k-i))}{\pi(k-i)}. \quad (9)$$

The last term $\sin(\pi(k-\alpha))/\pi(k-\alpha)$ represents the transition edge, where the summation from 0 to α for the pulses defines the method of PWM.

In order to apply a windowing function to (9) by taking $N+1$ consecutive samples in k iterations, samples are denoted as integer $j \in (0, N)$ where they are composed of an odd-N Lagrange interpolation FIR filter on an iterative scale of k. Hence, the property can be applied that impulse responses h_{pwm} for $N - j$ are the same as those for j but placed in reverse order, that is,

$$h_{\text{pwm}}(\alpha, j) = \sum_{i=0}^{\alpha} \frac{\sin(\pi(j-i))}{\pi(j-i)} = (-1)^{N-j} \sum_{i=0}^{\alpha} \frac{\sin(\pi i)}{\pi(j-i)}. \quad (10)$$

In our example where $N = 3$, the sign of $h_{\text{pwm}}(\alpha, j)$ in (10) alternates as $\{-1, 1, -1\}$. Therefore, the PWM pulses that determine the transition edges should shift one repetition period to the right in order to prevent the sign of impulses changing during leading-edge sampling (see the dotted line in Fig. 6).

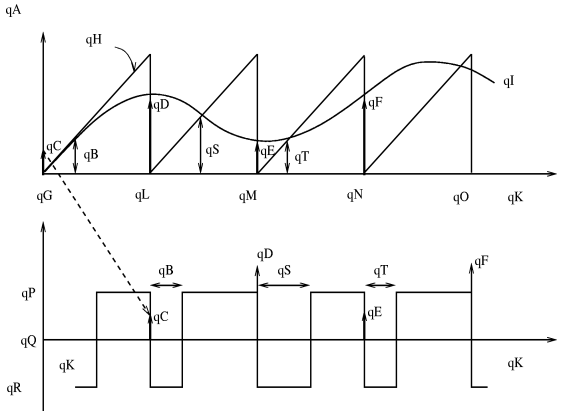


Fig. 6. Mapping is accomplished by an Lagrangian interpolation where the magnitude of modulating signal data α in each frames are linearly converted into time-domain data in leading-edge modulation.

Note that $h(\alpha, j)$ corresponds to the impulse response of DPM defined by the window and the Lagrangian interpolation FIR filter coefficient for Nth order polynomials, which has a classic form

$$h(\alpha, j) = \prod_{k=0,\ k \ne j}^{N} \frac{\alpha - k}{j - k}, \qquad \text{for } j \in \{0, 1, 2, \cdots, N\} \quad (11)$$

where $h(\alpha, j)$ has the Kronecker Delta function property

$$h(\alpha, j) = \begin{cases} 1, & \text{if } j = k \\ 0, & \text{if } j \ne k. \end{cases} \quad (12)$$

Once the Lagrange coefficient polynomial (11) is derived, it is unnecessary to calculate equations simultaneously as the new upsampled data fall on the grid obtained from the summation of products of each coefficient and the output values of the re-quantizer. In order to derive an explicit transfer function expression, take a binomial coefficient for (11) so that it determines the number of ways of choosing sampling time.

For a generic Lagrangian interpolation of (11), one should refer to [7] where several results for the binomial coefficients are introduced. When N is odd, it follows

$$h(\alpha, j) = (-1)^{N-j}\binom{\alpha}{j}\binom{\alpha-j-1}{N-j} = (-1)^{N-j}\binom{\alpha}{L}\binom{N}{j}\frac{L}{\alpha-j} \tag{13}$$

since the following relations are satisfied for $N, j \in \mathbb{Z}$ where

$$\binom{\alpha-j-1}{N-j} = \binom{\alpha-j}{N-j}\frac{1}{\alpha-j} \tag{14}$$

and

$$\binom{\alpha}{j}\binom{\alpha-j}{N-j} = \binom{\alpha}{N}\binom{N}{j}. \tag{15}$$

During the upsampling operation that generates the digital PWM clocks, the $sinc$ function under the window of length $L = N + 1$ for the DPM Lagrangian FIR filter is

$$h(\alpha, j) = (-1)^{N-j}\sum_{i=0}^{\alpha}\left\{\frac{\sin(\pi i)}{\pi(j-i)}\binom{\alpha}{L}\binom{N}{j}\frac{\pi L}{\sin(\pi i)}\right\} \tag{16}$$

in which the two terms, $W_b(j)$ and $C_b(\alpha)$ are defined as

$$W_b(j) = \binom{N}{j} \tag{17}$$

$$C_b(\alpha) = \sum_{i=0}^{\alpha}\left\{\binom{\alpha}{L}\frac{\pi L}{\sin(\pi i)}\right\} \tag{18}$$

is the scaling coefficient and forms an intrinsic equalization function with regard to the digital PWM, which eliminates the intrinsic spectral distortions. This equalization process can be viewed as a pre-processing stage formed by a FIR digital filter whose impulse response compensates for the frequency response of h_{ppm}. Fig. 6(b) depicts frequency-domain magnitude weighting, which can be interpreted as a pre-compensation filter to the digital PWM.

B. Mapping and Implementation

The implementation of (16) for DPM is accomplished using a third-order Lagrangian interpolator. A limit needs to be imposed on the interpolation filter whose normalized gain must not exceed 1 to prevent additional distortion by over modulation resulting from an excessive modulation index. The Farrow structure fractional delay FIR filter (see [8] and [14] for details) is selected for the Lagrangian interpolation filters for DPM in which the coefficients are dynamically adjusted by the reference signal $r(i)$ stored in ROM. Fig. 7 shows the direct mapping principle in which the PWM reference signal $r(i)$ consists of unit-tread staircase $y_1(i)$ is the re-quantized signal at the upsampling instants of the Lagrangian Farrow structure interpolator for which there must be only a single solution to detect the position of the sampling instant for α.

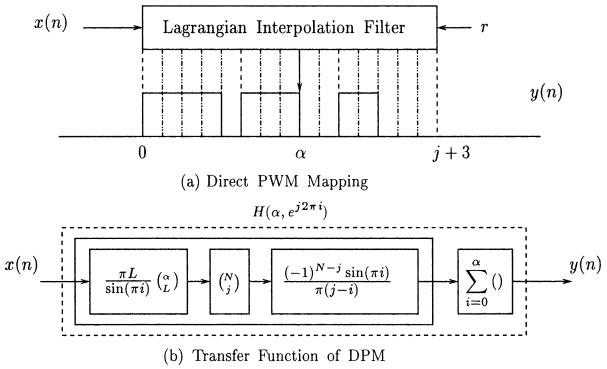


Fig. 7. Direct PWM mapping system represented as a pre-compensation filter and PWM. (a) The Lagrangian interpolation FIR filter consists of the PWM frame where each PPM signal is summed over α. (b) Depicts the transfer function diagram derived from (16).

TABLE I
SAMPLED EXAMPLE OF PWM REFERENCE SIGNAL

t_k	t_{k+1}	t_{k+2}	t_{k+3}
0.4627151	0.6008112	0.7565712	0.8775633

TABLE II
4-BIT DIRECT PWM MAP EXAMPLE BY TABLE I

PWM bit	Interpolation	N.S output	PWM output
16	0.7565712	0.6875	1
15	0.7473108	0.6875	1
14	0.7379273	0.8125	1
13	0.7284337	0.75	1
12	0.7188426	0.8125	1
11	0.7091669	0.625	1
10	0.6994194	0.625	1
9	0.6896129	0.8125	-1, 0
8	0.6797602	0.6875	-1, 0
7	0.6698741	0.5625	-1, 0
6	0.6599674	0.625	-1, 0
5	0.6500529	0.6875	-1, 0
4	0.6401433	0.6875	-1, 0
3	0.6302516	0.6875	-1, 0
2	0.6203905	0.5625	-1, 0
1	0.6105727	0.625	-1, 0

For example, the points at each sample periods before interpolation are shown in Table I, which is taken from a simulation result of the schematic in Fig. 8. α is uniformly sampled and each value provides information on the pulse transition instant of the PWM output signal after one sample delay of the pulse-repetition period. A leading-edge example is illustrated in Fig. 7 where the rising pulse sampling instant is passed to the DPM output signal. Both Tables I and II present examples of 4-bit DPM where t_k, t_{k+1}, t_{k+2}, and t_{k+3} are internal state values of 3-tapped delay line for the Lagrangian Farrow structure FIR filter. In this sequence, the interpolated samples correspond to each digitized 2^4 PWM values. Looking at the period between t_{k+1} and t_{k+2}, the upsampled 16 values can be obtained which are displayed in the Interpolation column in Table II.

The complete digital-to-analog (D/A) conversion system for an audio application is shown in Fig. 8. The pre-compensation FIR filter has a 3-tap delay line and is synchronously controlled

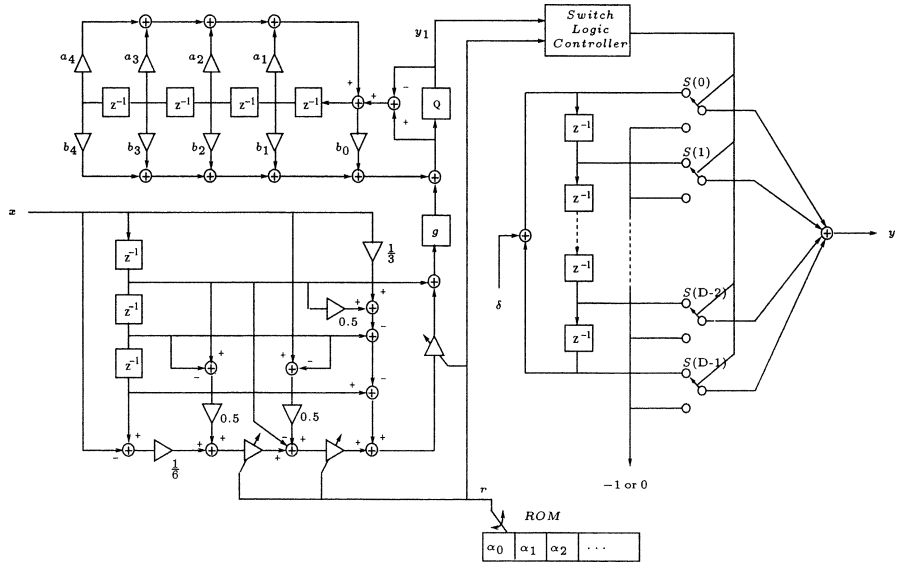


Fig. 8. Direct PWM mapping schematic for $4 \cdot f_s$ input 4-bit PWM D/A converter example.

by r which is stored in the ROM. The output of the Lagrange interpolation filter is scaled by a gain g in order to match the input range of the noise shaping re-quantizer and its output y_1 drives the digital PWM. This re-quantization from the PCM signal, using the *floor* rounding function that returns the greatest integer from the quantizer, has a D-bit range and is given by

$$Q(x) = \frac{\text{floor}(D \cdot x + 0.5)}{D}. \tag{19}$$

The coefficients a_n and b_n in the noise-shaping re-quantization block are set according to the noise shaping system design requirements; a_n should have negative values while b_n positive, etc. Due to the fourth-order noise shaping loop used in this example, which goes unstable unless the overall loop gain is kept to less than 0.5, the coefficient calculation of the noise shaper must be carefully chosen. Although a fourth-order noise shaping re-quantizer is used here, the DPM can take noise-shaping schemes of any type and order. The main idea of using noise-shaping re-quantization is to provide compression of the input PCM signal for the digital PWM. However, the performance of the noise shaping re-quantization has a direct effect on the output signal-to-noise (S/N) ratio.

TABLE III
COMPARISON OF ERROR CORRECTION METHODS OF FOUR PWM DACs

	Linearization Methods	Implementation Methods
PA	Interpolation to approximate analogue PWM - Time Domain	Interpolation and digital signal processing for Newton-Rapson algorithm
MFIR	Frequency moment matching - Frequency Domain	5 tap FIR filter in form of Toeplitz matrix
PNS	Frequency moment matching -Frequency Domain	3 tap FIR filter(ROM) in feedback loop and look-ahead scheme
DPM	Interpolation and windowing -Time and Frequency both	Farrow structure 3 tap FIR filter and digital PWM mapping

TABLE IV
COMPARISON OF SPECIFICATIONS OF FOUR PWM DACs

	PA	MFIR	PNS	DPM	SDM
Sampling Frequency	48kHz	48kHz	48kHz	48kHz	48kHz
PWM Bits	8	10	4	6	1
Oversampling	$8{\cdot}f_s$	$4{\cdot}f_s$	$64{\cdot}f_s$	$4{\cdot}f_s$	$256{\cdot}f_s$
Clock Rate (MHz)	98.304	196.608	49.152	12.288	12.288
THD+S/N	-125dB	-110dB	-160dB	-120dB	-112dB
NS order	5th	(2+2)th	5th	4th	2nd

C. Features and Comparison

The DPM operation can be summarized as D-bit PWM whose 2^D times upsampled values are located within the pulse-repetition period after filtering the input signal by the 3-tap Farrow structure FIR filter. The output of the Farrow structure FIR filter is fed to a noise shaping re-quantizer in order for the input signal to be compressed to D-bit PWM. DPM may be compared with earlier PWM linearization techniques that have been developed previously using time and frequency-domain methods. Time-domain PWM linearization methods have appeared in several studies [2], [4], [5], [12]. These papers demonstrated the typical error found in uniformly sampled PWM and introduced an improved sampling process for digital PWM power amplifiers [5], [12]; to conclude, possible remedies were suggested using upsampling introduced between two consecutive samples to locate the digital PWM edges. However, for finer accuracy, [2], [4] raised the concept of using the polynomial approximation method. On the other hand, the frequency-domain methods are studied in [1], [3], [10] and are based on the concept of moment matching.

In Tables III and IV, outline principles and example results are summarized that correspond to the respective papers, [1]–[3]. However, it should be noted that the clock rates given in Table IV are not related directly to the error correction method and corresponding performance but to the implementation differences. For brevity, the following terms are used; moment matching FIR

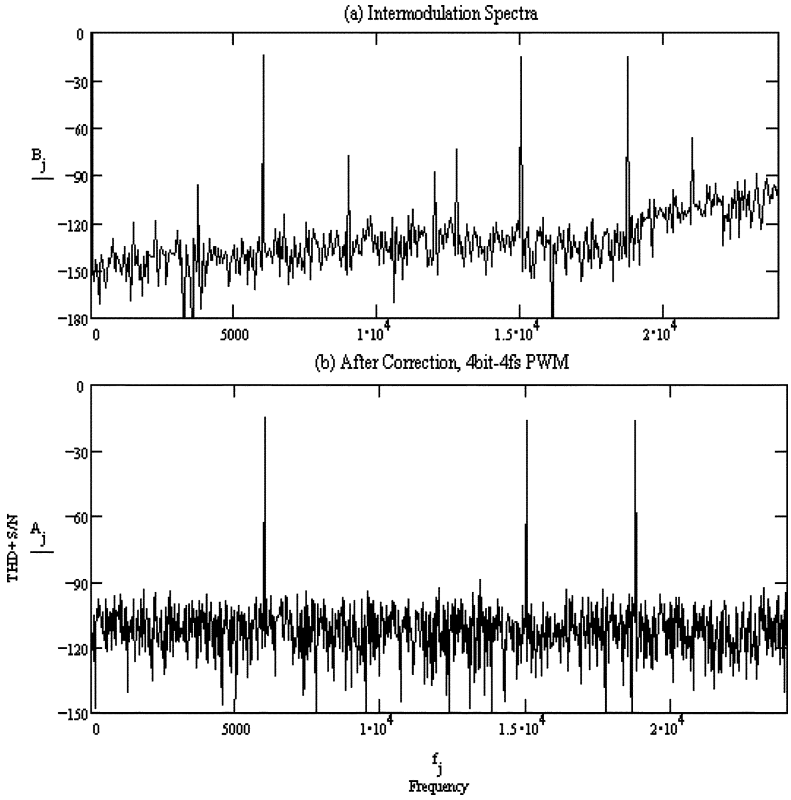


Fig. 9. Simulation results of 4-bit $4f_s$ PWM system (a) uncorrected (b) after correction.

filtering (MFIR) for the paper [1], polynomial approximation (PA) [2], procrastinate noise shaping (PNS) [3], and SDM.

The principal features of the DPM method are as follows.

1) In the time domain, DPM has better computational efficiency since it does not require iterative calculation steps for cross points or roots finding of two polynomials which is used in PA method.
2) In the frequency domain, its implementation is simpler because a 1-tap or 3-tap FIR filter is sufficient for the cascaded pre-compensation FIR filter, compared with MFIR and PNS methods.
3) Unlike SDM adopting 2-level quantization, PWM-based D/A converters (DACs) are known to have a distinctive feature in that the quantization overload problem can be relaxed due to higher bit quantization, [3], this is important in real-time applications although recent developments in Trellis SDM should be observed [17], [18].

The concept behind DPM can possibly be adapted to all forms of PWM's. For example, for double-edge modulation, it should use an even-order (2∼4 tap) Farrow structure Lagrangian interpolation FIR filter. A series of simulations were performed to test the error correction performance of DPM. Taking into account typical DACs; input signals are upsampled by 4 to 8 times.

IV. Simulation Results

In initial investigations, PWM with 4∼8 bit resolution were tested using the processes shown in Fig. 8. Fig. 9(a) shows simulation result of 4-bit, $4f_s$ PWM system with a four times oversampled input signal, where $f_s = 48$ kHz. This uncorrected PWM has an overall system clock rate 3.072 MHz, which is significantly lower than that of the SDM in Table IV. To observe the harmonic and intermodulation distortion in the base-band three superimposed input signals of 6, 15, and 18.75 kHz were used. After DPM operation, Fig. 9(b) reveals that signal purity is preserved more accurately. Following a series of investigations, it was concluded that 3.072 MHz is the minimum system clock frequency required to facilitate appropriate correction for the system; 4-bit PWM fed by $4f_s$ pre-oversampled input signal where $f_s = 48$ kHz.

Next, Fig. 10 presents results where the overall system bit rates are increased to 6.144 and 12.288 MHz, respectively, as the PWM bit resolution is increased from 4 to 5 and 6 bits. With the same superimposed input signal as the earlier example, the PWM intrinsic error is corrected more completely. In Fig. 10 (a), the residual noise and distortion is marked around -120 dB, hence this fact leads us to conclude that the DPM operated at 5-bit $4f_s$ can satisfy the required S/N ratio for high-quality audio applications. The 6-bit $4f_s$ DPM DAC shown in the subfigure (b) has the same system clock rate 12.288 MHz as the commercialized SDM DAC, which is the system proposed in this paper.

Finally, broad-band spectra are considered. Fig. 11 shows two PWM modes 6-bit $4f_s$ and 8-bit $4f_s$, where $f_s = 48$ kHz. The shaped quantization noise and distortion in the high frequency region are chosen so as not to be emphasized significantly where even the highest noise level is lowered to around -100 dB, see the subfigure (b). This is due to one of the distinctive features of the SDM+PWM structure compared to SDM only and can be a desirable characteristic for high resolution DAC applications.

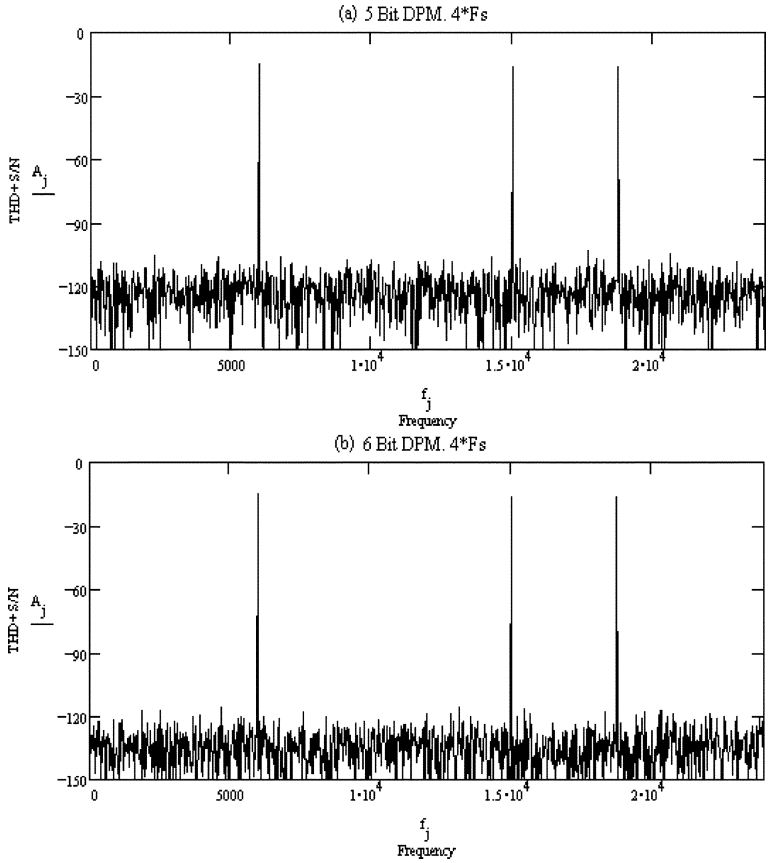


Fig. 10. Simulation results of (a) 5-bit $4f_s$ and (b) 6-bit $4f_s$ PWM systems after error correction by the DPM.

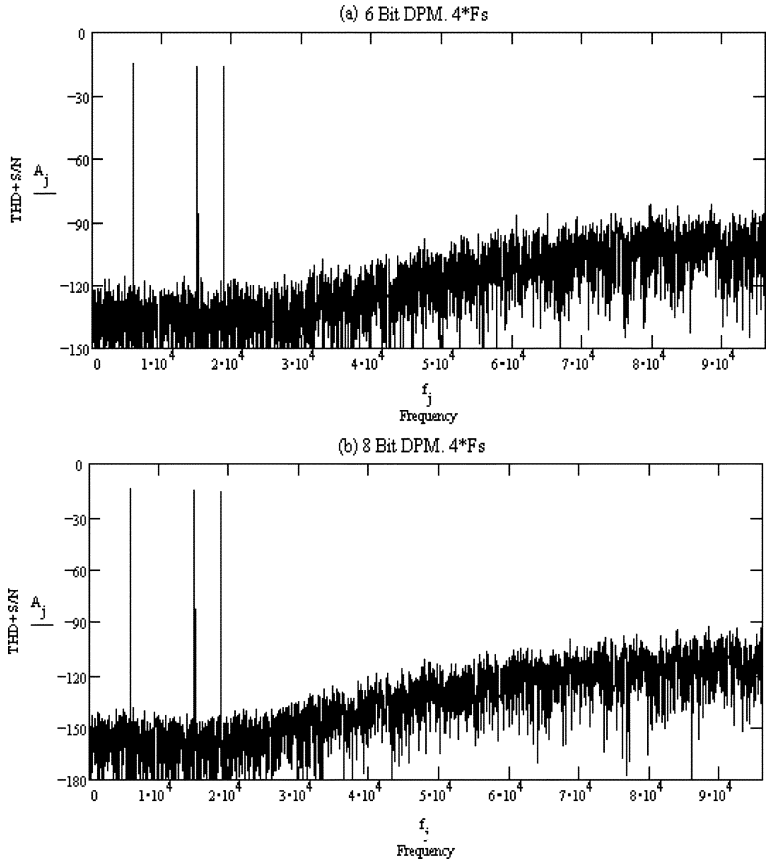


Fig. 11. Simulation results of broad-band spectra of (a) 6-bit $4f_s$ and (b) 8-bit $4f_s$ PWM systems after error correction by the DPM.

V. Summary

A novel error correction scheme for digital PWM has been investigated for use in high-resolution D/A conversion. The main discussion focused on the intrinsic error cancellation achieved using a window function with Lagrangian interpolation in a third-order Farrow structure fractional FIR filter synchronously combined with a digital PWM configured in circular formation. This method is shown to offer advantages against other previously proposed PWM linearization methods.

Although performance estimates are derived from a series of computer simulations, it is shown that relatively high-bit resolution digital PWM with a proper pre-compensation algorithm can outperform SDM; assuming similar system operating environments. Higher resolution D/A conversion can be achieved due not only to the relaxation of the 2-level quantization overload problem but also the significant reduction of noise and distortion generation in high frequency region. In addition, no high-order upsampling digital filter (typically 128 to 256 times) is required compared to SDM, since the third-order Farrow structure fractional FIR filter performs both pre-compensation and upsampling functions in one structure.

References

[1] M. J. Hawksford, "Dynamic model-based linearization of quantized pulsewidth modulation for applications in digital-to-analog conversion and digital power amplifier systems," *J. Audio Eng. Soc.*, vol. 40, no. 4, pp. 235–252, 1992.

[2] J. M. Goldberg, "Signal processing for high resolution pulsewidth modulation based digital-to-analog conversion," Dept. Elect. Eng., Ph.D. dissertation, King's College London, London, U.K, 1992.

[3] P. G. Craven, "Toward the 24 bit DAC: Novel noise-shaping topologies incorporating correction for the nonlinearity in a PWM output stage," *J. Audio Eng. Soc.*, vol. 41, no. 5, pp. 291–313, 1993.

[4] J. Goldberg and M. B. Sandler, "Comparison of PWM modulation techniques for digital amplifiers," in *Proc. Inst. Acoust.*, vol. 12, 1990, pp. 57–65.

[5] P. H. Mellor, S. P. Leigh, and B. M. G. Cheetham, "Improved sampling process for a digital pulsewidth modulated Class D power amplifier," in *Proc. IEE Colloq. Digital Audio Signal Processing*, London, U.K., 1991.

[6] K. Uchimura *et al.*, "VLSI A to D and D to A converter with multistage noise shaping," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'86)*, Apr. 1986, pp. 1545–1548.

[7] P. J. Kootsookos and R. C. Williamson, "FIR approximation of fractional sample delay systems," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 269–271, Mar. 1996.

[8] T. I. Laakso, V. Valimaki, M. Karjalinen, and U. K. Laine, "Splitting the unit delay," *IEEE Signal Processing Mag.*, vol. 13, pp. 30–60, Jan. 1996.

[9] A. C. Paul, "A cathedral-2 implementation of a pre-compensation algorithm for use in a PWM DAC," King's College London, London, U.K., Internal Rep., 1993.

[10] M. J. Hawksford, "Linearization of multilevel, multiwidth digital PWM with applications in digital-to-analog conversion," *J. Audio Eng. Soc.*, vol. 43, no. 10, pp. 787–798, 1995.

[11] H. E. Rowe, *Signals and Noise in Communication Systems*. London, U.K.: D. Van Nostrand, 1965.

[12] S. P. Leigh, "Pulsewidth modulation sampling process for digital Class D amplification," Ph.D. thesis, University of Liverpool, Liverpool, U.K., 1991.

[13] V. Valimaki, "A new filter implementation strategy for Lagrange interpolation," in *Proc. IEEE Int. Symp. Circuits Systems(ISCAS'95)*, vol. 1, 1995, pp. 362–364.

[14] C. W. Farrow, "A continuously variable digital delay element," in *Proc. IEEE Int. Symp. Circuits Systems (ISCAS'88)*, vol. 3, June 1988, pp. 2641–2645.

[15] G. D. Cain, N. P. Murphy, and A. Tarczynski, "Evaluation of several FIR fractional—sample delay filters," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'94)*, vol. 3, Apr. 1994, pp. 621–624.

[16] T. I. Laakso, V. Valmaki, and J. Henrikson, "Tunable downsampling using fractional delay filters with applications to digital TV transmission," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'95)*, vol. 2, May 1995, pp. 1304–1307.

[17] H. Kato, "Trellis noise-shaping converters and 1-bit digital audio," in *Proc. 112th AES Convention*, Munich, Germany, Mar. 10–13, 2002, paper 5615.

[18] E. Janssen and D. Reefman, "Advances in Trellis based SDM structures," in *Proc. 115th AES Convention*, New York, Oct. 10–13, 2003.

Jin-Whi Jung received the B.Sc. degree in electronics engineering from Dong-A University, Busan, South Korea, in 1985, the M.Sc. degree (with dissertation) in electronic systems engineering from the University of Essex, Colchester, U.K., in 1997. He is currently working toward the Ph.D. degree in nonlinear dynamics at the University College London, London, U.K.

His research interests include digital telecommunication circuits and systems, audio engineering, and application-specified integrated circuit designs.

Malcolm J. Hawksford received the B.Sc. degree with first class honors and the Ph.D. degree from the University of Aston, Birmingham, U.K., in 1968 and 1972, respectively.

He is the Director of the Centre for Audio Research and Engineering, and a Professor in the Department of Electronic Systems Engineering, Essex University, Colchester, U.K., where his research and teaching interests encompass audio engineering, electronic circuit design, and signal processing. While studying for his doctoral degree, he worked on delta modulation and sigma–delta modulation (SDM) for color television applications, and invented a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system. His research covers both analog and digital systems with a strong emphasis on audio systems including signal processing and loudspeaker technology. Since 1982, his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. A "first" in 1986 was for a prototype system to be demonstrated at the Canon Research Centre, Tokyo, Japan, research that had been sponsored by a research contract from Canon. Much of this work has appeared in the *Audio Engineering Society (AES) Journal* together with a substantial number of contributions at AES conventions. His research has encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with special emphasis on SDM and its application to super audio compact disc (SACD) technology. In addition, his research has included the linearization of pulsewidth-modulated encoders, diffuse loudspeaker technology, array loudspeaker systems and three-dimensional spatial audio and telepresence including scalable multichannel sound reproduction. He is a Technical Consultant for NXT, U.K., and for LFD Audio, U.K.

Dr. Hawksford was a recipient of the BBC Research Scholarship for his Ph.D. studies, is a recipient of the publication award of the AES for his paper entitled "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," for the best contribution by an author of any age to volumes 45 and 46 of the *AES Journal*. He is currently chairman of the AES Technical Committee on High-Resolution Audio and is a Founder Member of the Acoustic Renaissance for Audio (ARA), and a Technical Adviser for *Hi-Fi News and Record Review*. He is a Chartered Engineer, a Fellow of the AES, a Fellow of the Institution of Electrical Engineers, U.K., and a Fellow of the Institute of Acoustics (IOA).

PAPERS

Modulation and System Techniques in PWM and SDM Switching Amplifiers*

MALCOLM HAWKSFORD, *AES Fellow*
(mjh@essex.ac.uk)

University of Essex, Department of Electronic Systems Engineering, Colchester, Essex, CO4 3SQ, UK

Continuous- and discrete-time switching audio power amplifiers are studied both with and without feedback. Pulse-width modulation (PWM) and sigma–delta modulation (SDM) amplifier configurations are simulated and their interrelationship is described using linear phase modulation (LPM) and linear frequency modulation (LFM). Distortion generation encountered when applying negative feedback to PWM is demonstrated and strategies to improve linearity are presented. Recent innovations in SDM coding and output-stage topologies using pulse-shaping techniques are reviewed with emphasis on stable, low-distortion operation, especially under high-level signal excitation. A simplified low-latency variant of predictive SDM with step back is introduced, which together with dynamic compression of the state variables extends stable operation to a modulation depth of unity, thus allowing SDM to compete with PWM power amplifiers in terms of peak signal capability.

0 INTRODUCTION

There is a growing awareness of the role of high-efficiency topologies in power amplifier applications and especially the strategic significance in their application to a wide range of audio products [1], [2]. This is already evident in the digital theater market, where smaller size yet high performance products continue to emerge. There is also the opportunity to lower overall power dissipation, which when mapped into volume production is a most critical issue in terms of environmental factors. Key advantages stem from reduced size and heat loss. However, there are also more fundamental philosophical reasons for adopting class-D switching in the amplification process. (Hereafter switching amplifier infers "class-D" switching amplifier.) Switching amplifiers configured specifically for use with digital signals take on the mantle of a power digital-to-analog converter. As such there is reduced analog processing as the digital signals are, in a figurative sense, brought into closer proximity with the loudspeaker. Consequently there is opportunity to maintain better signal integrity and to achieve a more transparent overall performance, commensurate with appropriate design and physical implementation. The reduction in analog-related artifacts such as dynamic modulation of the closed-loop transfer function through device nonlinearity, including active device transconductance and internal capacitance modulation, implies less amplifier-dependent signal coloration and should lead to a more neutral and consistent sound quality. However, switching output stages offer potentially better control of a loudspeaker, as the source impedance of the amplifier remains low and almost resistive even under overload. This is contrary to most analog amplifiers, which are highly dependent on negative feedback to lower distortion and output impedance. The output impedance of a switching amplifier is determined principally by the actual on resistance of the output switching devices, moderated only by negative feedback, when used, and by the passive low-pass filter necessary to limit the extreme high-frequency signal components resulting as a consequence of the type of modulation scheme selected. Also depending upon both topology and whether or not feedback is used, the characteristics of the power supply can be critical.

Until recently most switching power amplifiers for audio were based only on a paradigm of PWM, which offers a number of attractive features. Classical PWM generates a binary output signal "switched" between two voltage levels, where the width of each pulse is modulated in proportion to the input signal such that the short-term average of the square-wave output follows the instantaneous amplitude of the input signal. Since the output power devices switch rapidly between two states, they only dissi-

*Presented at the 118th Convention of the *Audio Engineering Society,* Barcelona, Spain, 2005 May 28–31, under the title "SDM versus PWM Power Digital-to-Analog Converters (PDACs) in High-Resolution Digital Audio Applications." Manuscript received 2005 May 11; revised 2005 November 11 and 2006 January 11.

pate significant power in the pulse transition regions, enabling the output stage to achieve high efficiency. However, with the application of SDM used in both analog-to-digital converters (ADC) and digtal-to-analog converters (DAC) and also in Super Audio CD (SACD) [3], SDM has become a viable alternative binary modulation method for switching power amplifier systems. Unlike PWM, where the pulse width is modulated, SDM forms a sequence of equal-area binary pulses such that the relative densities of positive and negative pulses are complementary and track the instantaneous amplitude of the input signal. Traditionally SDM forms a digital signal where the pulse instants are assigned to discrete time slots. But just as PWM has both continuous-time (analog) and discrete-time (digital) variants, so can SDM. However, where the source signal is digital, it is prudent to maintain signals within the digital domain and not to impose additional cascaded stages of ADCs and DACs. Also because uniformly sampled PWM associated with digital source signals is inherently nonlinear together with the generation of requantization noise due to the output pulses being constrained in time, additional signal processing is required both to linearize the modulator and to shape the requantization noise spectrally. Nevertheless, even for a digital switching amplifier, when the output stage and the power-supply rejection requirements are considered together with the role of negative feedback, the principal design factors revert back to those of an analog system. Consequently analog feedback techniques can apply either to analog or to digitally (see Section 6) derived PWM.

Digital PWM requires two clocks that ultimately bound its performance. First there is the sampling frequency that determines the repetition rate of pulse transitions, and second there is the much higher frequency clock, which defines the discrete-time locations of each pulse transition. However, with the advent of the direct-stream digital[1] (DSD) format based on SDM [3], the effective sampling rate of the system has been elevated to 2.8224 MHz where, unlike in PWM, the sampling rate and the pulse repetition rates are the same. Also SDM can readily be implemented so that within the audio band the modulator is virtually linear and does not require an additional linearization processor to achieve acceptable levels of distortion. The increase in sampling rate also facilitates a simpler low-pass filter with the potential for reduced signal losses because such output filters have to handle the full output current of the amplifier. However, offsetting this advantage, the high number of pulse transitions per second imply potentially lower power efficiency, thus demanding the use of high-speed switching transistors and possibly zero-voltage switching, as discussed in Section 7.

This paper commences in Section 1 with a study of both PWM and SDM directed at switching amplifiers with an emphasis on modulation and the exploitation of negative feedback. The approach taken considers the relationship between PWM and SDM using modulation models based on linear phase modulation (LPM) and linear frequency modulation (LFM). Analytical modeling establishes the native linearity of each modulator class and enables the results to benchmark the performance of amplifier variants that incorporate combinations of negative feedback and linearization strategies. It is desirable to use negative feedback in a switching amplifier in order to reduce distortion dependence on both output-stage nonidealities and power-supply variations. However, it is shown in Section 3 that for PWM the introduction of output-voltage-derived feedback, although achieving anticipated performance gains, can introduce additional distortion products that are not observed in optimized open-loop PWM. Consequently corrective means are required to reduce these undesirable artifacts where a number of amplifier topologies are presented. This aspect of the study exploits a high precision Matlab[2] simulator of a PWM negative-feedback amplifier, where a spectral resolution in excess of 200 dB reveals all significant distortion products. The simulations show a noise-shaping advantage for induced jitter, the generation of intermodulation distortion for a multitone excitation, and the resulting effects on both noise and distortion when the loop gain is changed.

The study concludes with Section 7 by describing an SDM power amplifier topology [4] designed both to attenuate switching components above the SDM pulse repetition rate and to lower switching losses by forcing the output transistors to commutate only when switching voltages are zero. To complement this amplifier a digital SDM encoder algorithm is presented, which combines predictive look-ahead with dynamic compression of state variables that together enable stable coding up to a modulation index of unity, an important factor in achieving the full output signal capability of an SDM-based switching amplifier.

1 SDM–PWM ANALYTICAL COMPARISON

It is well known that naturally sampled, non-time-domain quantized PWM implemented using a symmetrical sawtooth waveform and a binary comparator offers low intrinsic distortion provided the bandwidth of the input signal is suitably constrained to prevent reflected components about the sampling frequency from falling within the audio band [5]. On the first encounter it may appear unusual for a signal that is processed by a two-level amplitude quantizer to have low intrinsic distortion. However, in this section it is shown that the modulation process has a fundamental affinity with LPM and is therefore intimately related to the LFM model proposed for SDM [6]–[9]. Consequently to establish a proper mathematical framework, models for both SDM and PWM are developed and their interrelationship is established.

1.1 SDM Modeling

Fig. 1 shows an SDM model using two differentially driven (or complementary) LFMs to produce a non-time-domain quantized pulse train of density-modulated posi-

[1]DSD was proposed by Sony to describe sigma–delta modulation for high-resolution audio.

[2]Matlab is a trade name of The MathWorks, Inc.

tive and negative pulses. The principal features of this structure have been reported in previous papers [6]–[7], [9]. LFM is a technique where a sinusoidal carrier is modulated such that its instantaneous carrier frequency is proportional to the amplitude of the input signal. In the SDM model based upon LFM, common reference points are designated on each cycle of the carrier where in this study the positive-slope zero crossings (PSZCs) are chosen for their convenience of definition. Constant-area (such as Dirac) impulses are then located at each PSZC to form the output pulse sequence for non-time-domain quantized SDM. However, because SDM generates two time-interleaved sequences of positive and negative pulses, as illustrated in Fig. 1, then carriers $s_1(t)$ and $s_2(t)$ with complementary LFM are used to form pulse density modulated sequences $P_1(t_{1r})$ and $P_2(t_{2r})$, where $\{t_{1r}, t_{2r}\}$ are the respective pulse time coordinates. Both carrier center frequencies (that is, the frequency when the input signal is zero) are set to f_{lfm}. Thus because of modulation symmetry, the combined pulse repetition rate of positive and negative output pulses always remains constant at $2f_{lfm}$ pulses per second, where the interleaved SDM output stream is $P_1(t_{1r}) - P_2(t_{2r})$. Observing this process it is evident that, say for a positive input signal, the frequency of positive pulses increases while simultaneously the frequency of negative pulses falls by an equal amount. Hence the short-term average of the composite output tracks the input signal where, if time-domain quantization is ignored, this process shows similar behavior to SDM. In the following analysis each LFM is realized as a cascade of signal integration and LPM that yields low intrinsic distortion [7], [8] provided the output pulse time coordinates are unconstrained so as to adopt their natural sampling instants. This modulation process is summarized in the following to facilitate comparisons between SDM and PWM.

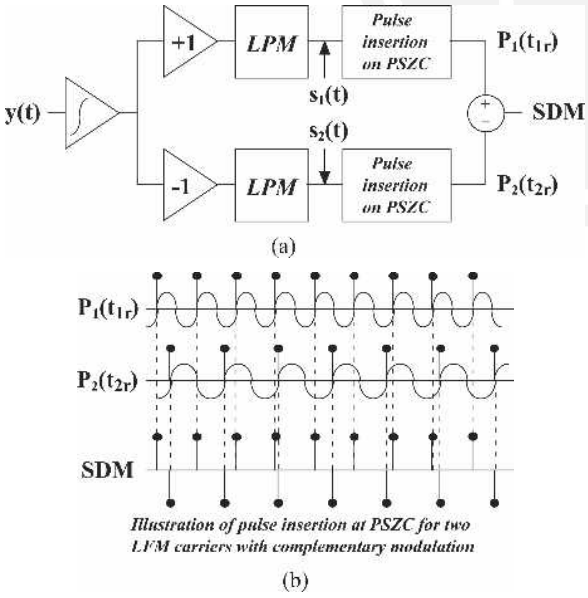


Fig. 1. (a) Non-time-quantized SDM using LFM. (b) Pulse insertion at PSZC for two LFM carriers with complementary modulation.

Assume the input signal $y(t)$, where t is time, is normalized by $\hat{Y}$ to limit the respective instantaneous frequencies $f_1(t)$ and $f_2(t)$ of the two complementary LFM carriers to a range of 0 Hz $< f_{1,2}(t) < 2f_{lfm}$ Hz, which are defined in terms of $y(t)$, $\hat{Y}$, and f_{lfm} as

$$f_1(t) = f_{lfm}\left[1 + \frac{y(t)}{\hat{Y}}\right] \quad \text{and} \quad f_2(t) = f_{lfm}\left[1 - \frac{y(t)}{\hat{Y}}\right]. \tag{1}$$

Instantaneous LFM carrier phases $\theta_1(t)$, $\theta_2(t)$ expressed as functions of $f_1(t)$, $f_2(t)$ are

$$\theta_1(t) = 2\pi \int_{u=0}^{t} f_1(u)\, du \quad \text{and}$$

$$\theta_2(t) = 2\pi \int_{u=0}^{t} f_2(u)\, du. \tag{2}$$

Hence the two carriers $s_1(t)$ and $s_2(t)$ with complementary LFM become

$$s_1(t) = A\cos[\theta_1(t)] = A\cos\left[2\pi f_{lfm}t + \frac{\pi}{2} + \frac{2\pi f_{lfm}}{\hat{Y}}\int_{u=0}^{t} y(u)\, du\right] \tag{3}$$

$$s_2(t) = A\cos[\theta_2(t)] + A\cos\left[2\pi f_{lfm}t - \frac{\pi}{2} - \frac{2\pi f_{lfm}}{\hat{Y}}\int_{u=0}^{t} y(u)\, du\right]. \tag{4}$$

Note in Eqs. (3) and (4) that complementary phase shifts of 0.5π and -0.5π are included, so that under quiescent conditions when $y(t) = 0$, positive and negative output pulses idle with an alternating symmetry.

To derive non-time-domain quantized SDM using the LFM model, Fig. 1 shows positive pulses located at the PSZC of $s_1(t)$ (that is, at times t_{1r}) and negative pulses located at the PSZC of $s_2(t)$ (that is, times t_{2r}), where solutions for t_{xr} occur when $\theta(t_{xr}) = 2\pi r$ for integer r, that is,

$$t_{1r} + \frac{1}{\hat{Y}}\int_{t=0}^{t_r} y(t)\, dt = \frac{r - 0.25}{f_{lfm}} \tag{5}$$

$$t_{2r} - \frac{1}{\hat{Y}}\int_{t=0}^{t_r} y(t)\, dt = \frac{r + 0.25}{f_{lfm}}. \tag{6}$$

To solve Eqs. (5) and (6) in terms of t_{xr} a four-stage methodology is adopted [9].

1) LFM signals $s_1(t)$, $s_2(t)$ are sampled at uniformly spaced time instants at the rate of $\cdot(f_{lfm})$ Hz, where the oversampling factor of is greater than 1, in order to give a finer time resolution (for example, of = 16).

2) Sampled carriers are scanned to identify negative to positive polarity transitions to identify all PSZCs.

3) For each detected PSZC, linear interpolation between adjacent samples improves on $\{t_{1r}, t_{2r}\}$ estimation.

4) An iterative error-driven procedure then seeks optimum solutions for $\{t_{1r}, t_{2r}\}$.

The SDM output is formed by subtracting the two complementary LFM-derived pulse streams. The corre-

sponding spectrum $\text{out}_{\text{sdm}}(f)$ is then calculated [9] by a summation over N, with $r = N$ being the Nth rotation of 2π in $\theta(t_{xr}) = 2\pi r$, as

$$\text{out}_{\text{sdm}}(f) = \sum_{r=1}^{N} (e^{-j2\pi f t_{1r}} - e^{-j2\pi f t_{2r}}). \quad (7)$$

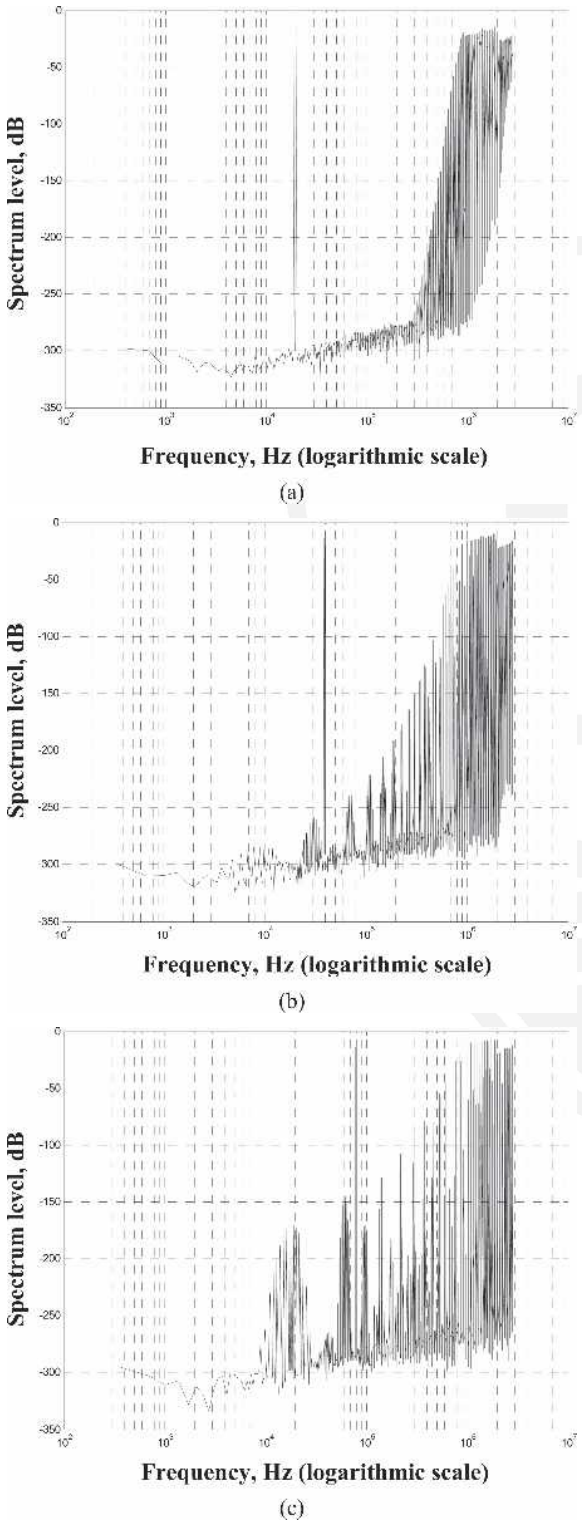


Fig. 2. SDM output spectra. (a) Example 1. (b) Example 2. (c) Example 3.

By way of illustration, Fig. 2 shows three examples of SDM output spectra using two-tone input signals with the following data. All SDM examples have $f_{\text{sdm}} = 64(44.1)$ kHz.

Example 1:
[A1 = 0.2, f1 = 19 kHz]
[A2 = 0.2, f2 = 20 kHz]
Example 2:
[A1 = 0.2, f1 = 39 kHz]
[A2 = 0.2, f2 = 40 kHz]
Example 3:
[A1 = 0.2, f1 = 79 kHz]
[A2 = 0.2, f2 = 80 kHz].

All computations show extremely low in-band distortion, which confirms a high degree of linearity. Distortion only becomes problematic when both input signal frequency and modulation depth are sufficiently high for the sidebands centered about the carrier frequency to migrate toward the audio band.

1.2 PWM Modeling

Having examined SDM, a similar model is now constructed for non-time-domain quantized PWM. Fig. 3 shows open-loop, naturally sampled PWM based on binary amplitude quantization with a symmetrical triangular wave added to the input signal. On first encounter it ap-

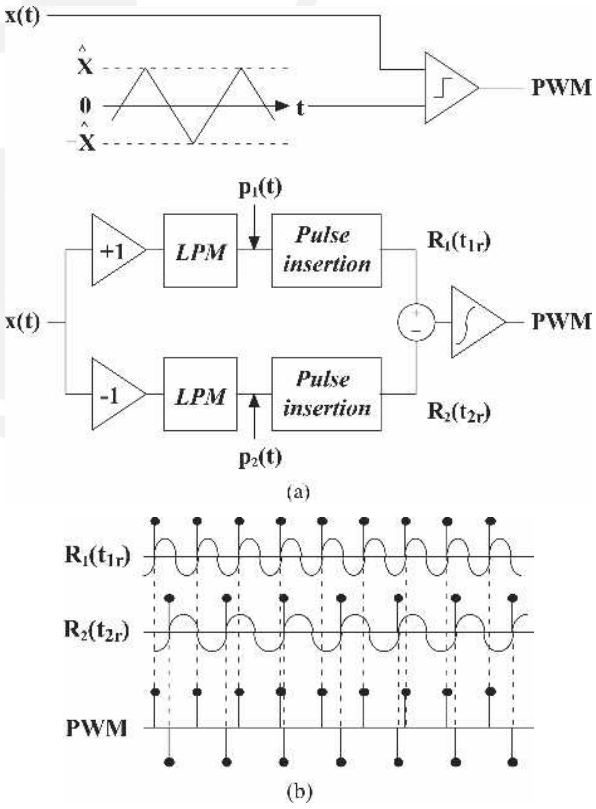


Fig. 3. (a) Non-time-quantized PWM using LPM. (b) Pulse insertion at PSZC for two LPM carriers with complementary modulation.

pears unlikely that such a simple structure can pass a signal with low intrinsic distortion. Following earlier work on uniformly sampled PWM [10], [11], it could be argued that as the input signal modulates the pulse width then, because the rth rectangular pulse of width τ_r transforms to the frequency domain with magnitude response $\tau_r \sin(\pi f \tau_r)/(\pi f \tau_r)$, the resulting nonlinear frequency response as a function of pulse width produces distortion. Therefore PWM appears to exhibit an intrinsic dynamic spectral modulation although, as will be shown, this does not occur with natural sampling.

The key to understanding naturally sampled PWM is the observation that each output pulse transition, both −1 to +1 and +1 to −1, undergoes linear modulation in time (provided the input amplitude does not exceed the peak-to-peak amplitude of the triangular wave) as a function of the instantaneous amplitude of the input signal. This differs from the SDM process described in Section 1.1 as there it was the frequency of the output pulses that was proportional to the input signal, as shown by Eqs. (3) and (4), which describe classic LFM. However, in PWM the process is closer to LPM, although the relationship between LFM and LPM is only the inclusion of linear integration of the input signal.

The proposed model for naturally sampled PWM, illustrated in Fig. 3, is described as follows. It is recognized that in naturally sampled PWM the positive and negative transitions of the PWM output are individually associated with the input signal and that there is no coupling between edges other than indirectly through filtering applied to the input signal. Hence in the non-time-domain quantized PWM model the pulse transitions can be determined individually by two independent linear-phase modulators. However, because for a given change in input signal amplitude, PWM edges move in opposite directions, these modulators are driven differentially by the input signal, mirroring the complementary LFM process used for modeling SDM. Also, because for a zero input signal the two output pulse sequences must be offset by a half-cycle to form a symmetrical interlaced sequence, the complementary dc signals are added to the phase modulator inputs to shift the respective phases of the carriers by $\pi/2$ and $-\pi/2$, as shown in Eqs. (8) and (9). Hence if the PWM carrier frequency is f_{pwm} Hz and the input signal $x(t)$ is normalized by $\hat{X}$, then the two phase modulator signals $p_1(t)$ and $p_2(t)$ are given by

$$p_1(t) = A \cos\left[2\pi f_{\text{pwm}} t + \frac{\pi}{2}\left(1 + \frac{x(t)}{\hat{X}} \right) \right] \tag{8}$$

$$p_2(t) = A \cos\left[2\pi f_{\text{pwm}} t - \frac{\pi}{2}\left(1 + \frac{x(t)}{\hat{X}} \right) \right]. \tag{9}$$

As with the LFM-SDM model, the PSZC time coordinates $\{t_{1\text{r}}, t_{2\text{r}}\}$ form natural time sampling instants derived from the two LPM carriers $p_1(t)$, $p_2(t)$. Unit-area pulses (such as Dirac) $R_1(t_{1\text{r}})$, $R_2(t_{2\text{r}})$ are then located at $\{t_{1\text{r}}, t_{2\text{r}}\}$, whereby $R_1(t_{1\text{r}}) - R_2(t_{2\text{r}})$ forms the composite data stream of positive and negative pulses. Finally the composite stream is integrated, which translates each Dirac pulse into a unit step function. However, because of the interleaved nature of the pulse sequence and since on average the number of positive and negative pulses must remain the same, a binary PWM square-wave output is formed, provided $|x(t)| < \hat{X}$. If this limit is exceeded, then the output pulse sequence becomes multilevel as it is possible to have two or more sequential pulses of the same sign. This compares directly with replacing the two-level quantizer normally used in PWM with a uniform multilevel quantizer. It should be noted that in order to achieve an amplitude-symmetric bipolar sequence typical of a practical PWM amplifier, a constant dc offset is subtracted from the integrated output equal to one-half the weight of the first Dirac pulse in the composite pulse sequence, thus achieving a long-term average of zero. Also, because of the relationship between phase and frequency, as stated in Eq. (2), it follows that LPM acts as a differentiator and therefore requires integration to correct the overall frequency response. In the LFM-SDM model the integrator precedes LPM, whereas for the LPM-PWM model it is located after LPM. The location of integration gives insight into modulator linearity as conventional wisdom describes PWM akin to a sample-and-hold function, where the hold period is modulated by the input signal, which when mapped into the frequency domain suggests a mechanism for dynamic spectral modulation. However, in the PWM model it can now be seen that the output square wave is actually formed by integration of two interleaved phase-modulated pulse sequences. As such there is no finite-duration hold function being used; it is an illusion. This is a critical observation, which reveals that system linearity depends solely on the characteristics of the two differentially driven LPMs, where any distortion is just frequency shaped by an integrator.

A system simulation was performed where for both LPM carriers every PSZC instant $\{t_{1\text{r}}, t_{2\text{r}}\}$ was determined using in iterative procedure similar to that used for the SDM model. The output spectrum of the time-integrated output pulse sequence was then calculated using a complex exponential method similar to that described by Eq. (7),

$$\text{out}_{\text{pwm}}(f) = \sum_{r=1}^{N} \left(\frac{e^{-j2\pi f t_{1\text{r}}} - e^{-j2\pi f t_{2\text{r}}}}{j2\pi f T} \right) \tag{10}$$

where the integration time constant $T = 1$ second such that a Dirac pulse maps to a unit step. A principal advantage of this approach is that multiple-sine-wave input signals can be generated so that the linearity of the overall process can be explored in detail in a way that can be problematic using pure analytical techniques. Also, the comparison between SDM and PWM using non-time-domain quantized models reveals that the principal differences are the location of the signal integration function and that SDM normally operates with a higher sampling rate. The higher sampling rate for SDM adopted in digital applications is a result of differing quantization strategies, as when the PWM time coordinates $\{t_{1\text{r}}, t_{2\text{r}}\}$ are quantized

in time, a finer quantization interval must be used compared to that set by the PWM natural sampling rate, that is, it is the higher PWM quantization related clock rate that should be compared to the SDM pulse repetition rate. To show the formation of the PWM output waveform, Fig. 4 illustrates phase-modulated Dirac pulses derived from the PSZC of the complementary phase modulators both before and after linear integration and including the constant dc offset required to achieve an amplitude-symmetric pulse distribution.

Examples of output spectra derived using the LPM model are shown in Fig. 5 for three sets of input signals with progressively higher signal frequency. As with the SDM simulations, the input signals adopt the same two tones to facilitate comparison, although the center frequencies of LPM and LFM differ. The example spectra pre-

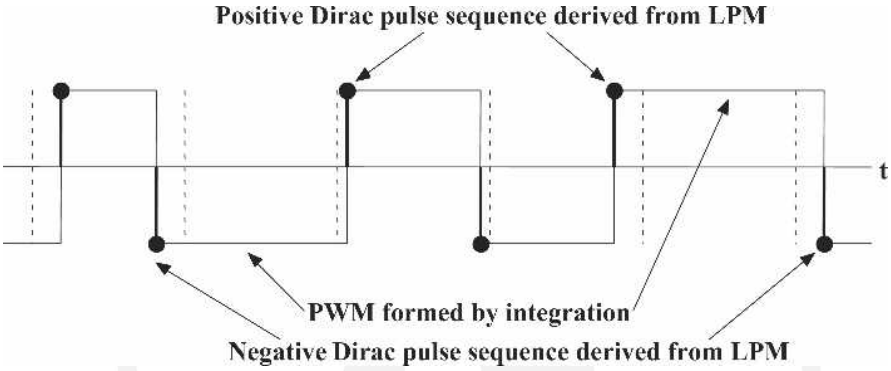


Fig. 4. PWM derived by integration of LPM.

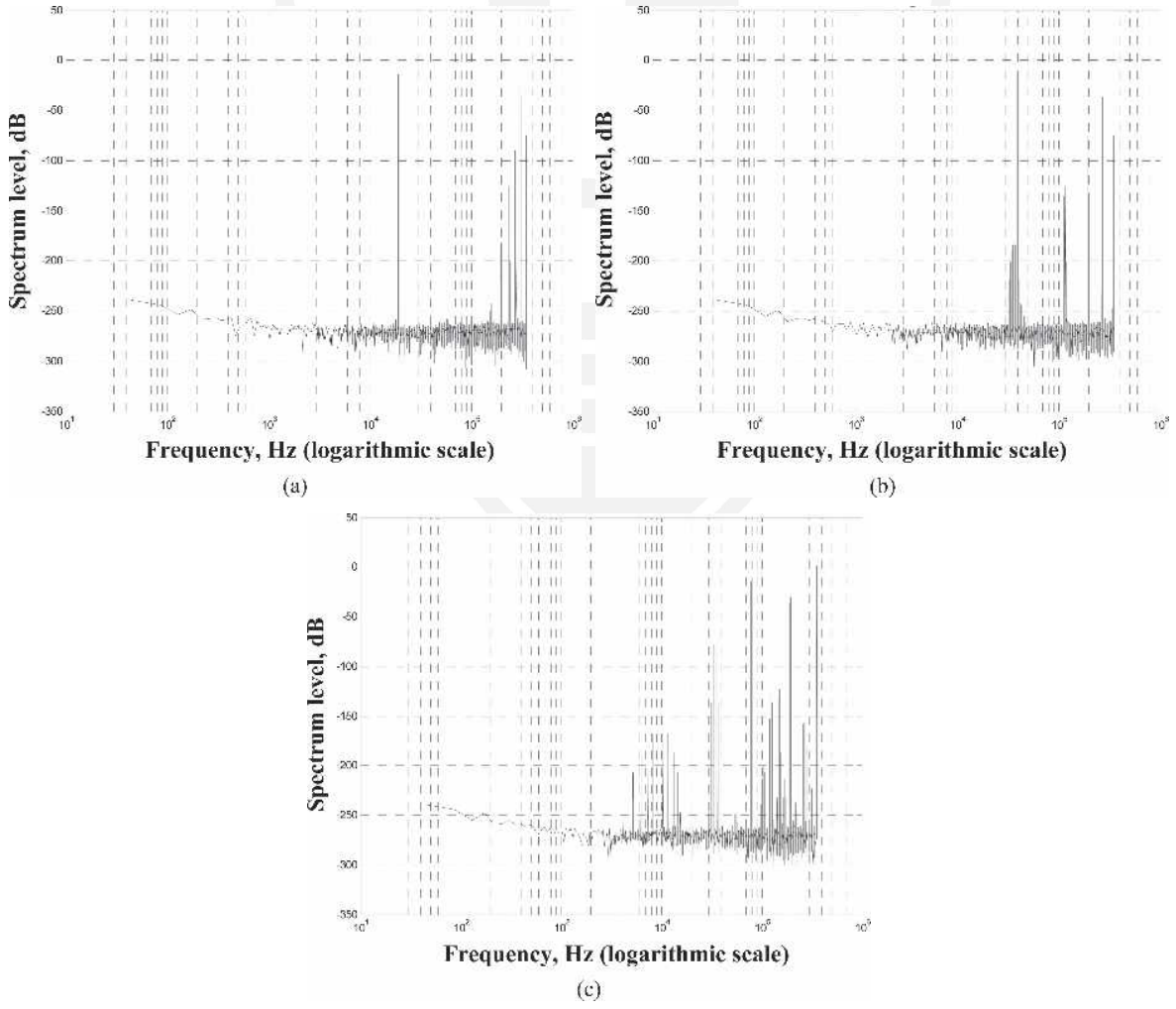


Fig. 5. PWM output spectra. (a) Example 1. (b) Example 2. (c) Example 3.

sented confirm the low inherent distortion achievable with LPM, although as the frequency of the input signal is increased, the migration of aliased frequency components toward the audio band can be observed. All PWM examples have $f_{pwm} = 8(44.1)$ kHz.

Example 1:
[A1 = 0.2, f1 = 19 kHz]
[A2 = 0.2, f2 = 20 kHz]
Example 2:
[A1 = 0.2, f1 = 39 kHz]
[A2 = 0.2, f2 = 40 kHz]
Example 3:
[A1 = 0.2, f1 = 79 kHz]
[A2 = 0.2, f2 = 80 kHz].

1.3 Analytical Derivation of Dirac Pulse Sequences

The method described for determining PSZC is well matched to the task of simulation. However, it is also possible to adapt analytically both LFM and LPM models such that short-duration pulses with appropriate polarity are formed at PSZC. Consider the uniform quantizer characteristic shown in Fig. 6 with quantum 2π, where the input function is $\phi(t)$ and the quantized output is $Q[\phi(t)]$. Adopting the procedure reported previously [9] but putting quantum $\delta \to 2\pi$, a series representation of the quantization process follows,

$$Q[\phi(t)]|_{\delta=2\pi} = \phi(t) + \frac{\delta}{\pi}\sum_{r=1}^{\infty}\left[\frac{1}{r}\sin\left(2\pi r\frac{\phi(t)}{\delta}\right)\right]$$

$$= \phi(t) + 2\sum_{r=1}^{\infty}\left[\frac{1}{r}\sin(r\phi(t))\right]. \quad (11)$$

Differentiating $Q[\phi(t)]$ with respect to time forms a time-domain series of pulses that are located at each quantizer transition,

$$\delta[\phi(t)]|_{R\to\infty} = \frac{d\phi(t)}{dt}\left\{1 + 2\sum_{r=1}^{R}[\cos(r\phi(t))]\right\}. \quad (12)$$

Fig. 6 illustrates the uniform quantizer expressed as a function of $\phi(t)$ whereas Fig. 7 shows by way of example the synthesis of pulses located at PSZC for 8, 32, and 256 harmonics and where $d\phi(t)/dt = 1$. The example for 256 harmonics reveals a narrow synthesized pulse width compared to the sequence period, although for the Dirac pulse to be formally represented $R \to \infty$. Hence applying Eqs.

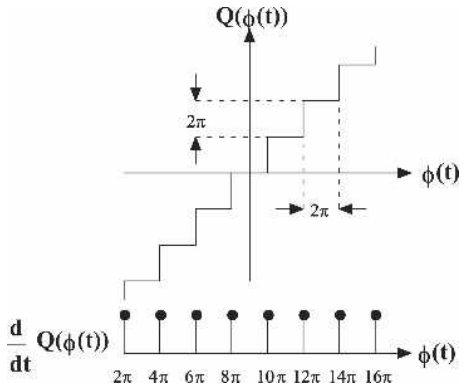


Fig. 6. Dirac pulse formation using differentiation of quantization characteristic.

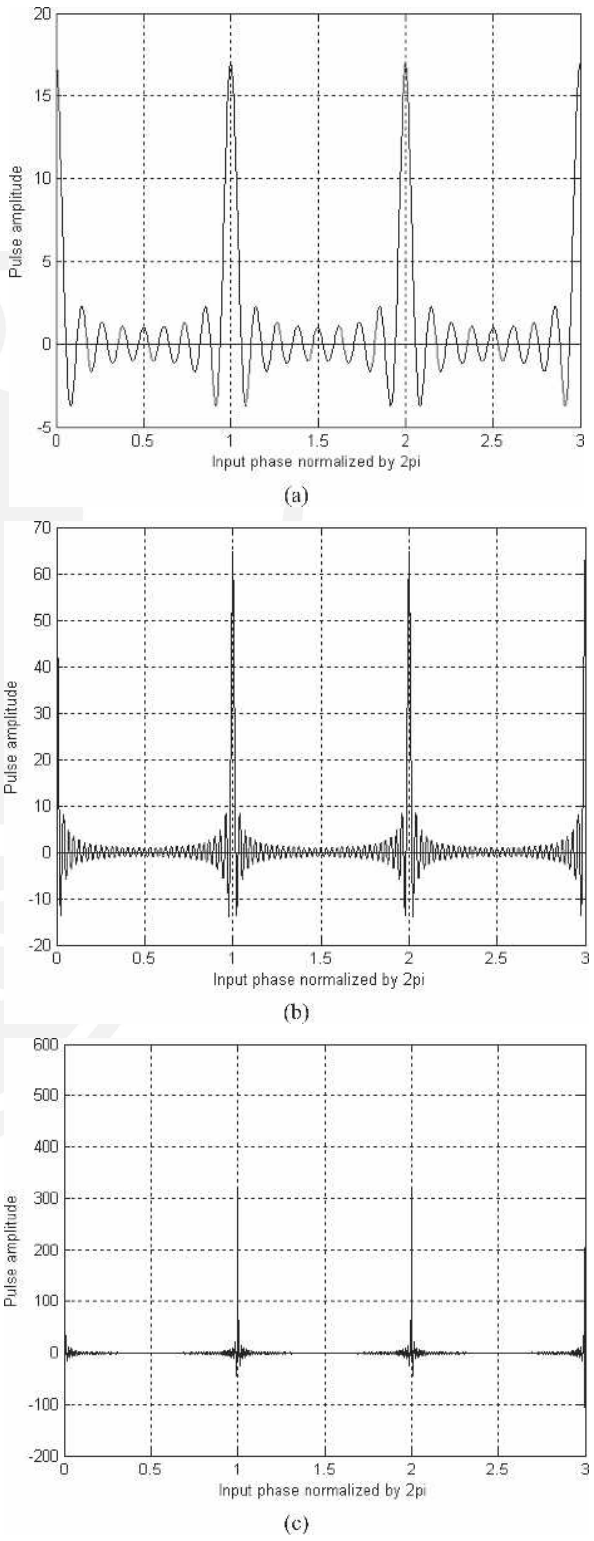


Fig. 7. Pulse synthesis. (a) 8 harmonics. (b) 32 harmonics. (c) 256 harmonics.

(3) and (4) to Eq. (12), the SDM pulse sequence δ_{sdm} follows,

$$\delta_{sdm} = \frac{y(t)}{\hat{Y}} + \left[1 + \frac{y(t)}{\hat{Y}}\right] \sum_{r=1}^{\infty} \left\{ \cos\left[r\left(2\pi f_{lfm} t + \frac{\pi}{2} + \frac{2\pi f_{lfm}}{\hat{Y}} \int_{u=0}^{t} y(u)\, du \right) \right] \right\} - \left[1 - \frac{y(t)}{\hat{Y}}\right] \sum_{r=1}^{\infty} \left\{ \cos\left[r\left(2\pi f_{lfm} t - \frac{\pi}{2} - \frac{2\pi f_{lfm}}{\hat{Y}} \int_{u=0}^{t} y(u)\, du \right) \right] \right\}. \quad (13)$$

Applying Eqs. (8) and (9) to Eq. (12) and integrating the output, the non-time-quantized PWM pulse sequence δ_{pwm} is given,

$$\delta_{pwm} = \frac{1}{\hat{X}} \frac{dx(t)}{dt} + 2\left[2f_{pwm} + \frac{1}{2\hat{X}} \frac{dx(t)}{dt} \right] \sum_{r=1}^{R} \left\{ \cos\left[r\left(2\pi f_{pwm} t + \frac{\pi}{2}\left(1 + \frac{x(t)}{\hat{X}}\right)\right)\right]\right\} - 2\left[2f_{pwm} - \frac{1}{2\hat{X}} \frac{dx(t)}{dt} \right] \sum_{r=1}^{R} \left\{ \cos\left[r\left(2\pi f_{pwm} t - \frac{\pi}{2}\left(1 + \frac{x(t)}{\hat{X}}\right)\right)\right]\right\}. \quad (14)$$

2 JITTER SENSITIVITY OF SDM AND PWM

From the respective non-time-quantized models of SDM and PWM described in Sections 1.1 and 1.2 it is straightforward to predict the relative sensitivity of each system to pulse jitter. In making this comparison the key difference can be derived from the position of the integrator, where it was shown that for SDM the input to the LPM is integrated, whereas for PWM the output pulse sequence is integrated. Jitter is represented in terms of output pulse time displacement with the conversion rule that the pulse area must remain invariant, a requirement normally met in SDM using switched-capacitor circuits [12].

Consider a general time-modulated impulse sequence of M samples $\sum_{r=1}^{M}\{\alpha_r \delta(t - t_r)\}$, where $\alpha_r = 1$ or -1, depending on the pulse polarity, derived from either the LFM or the LPM models and located at time t_r, displaced in time by instantaneous jitter Δt_r such that its actual time coordinate is $t_r + \Delta t_r$. The resulting instantaneous time-domain error sequence $\Gamma(t)$ is

$$\Gamma(t) = \sum_{r=1}^{M} \alpha_r \{\delta(t - t_r) - \delta(t - t_r - \Delta t_r)\}. \quad (15)$$

For SDM the corresponding spectral error Esdm(f) follows directly from Equation (15),

$$\mathrm{Esdm}(f) = \sum_{r=1}^{M} \alpha_r [e^{-j2\pi f t_r} - e^{-j2\pi f (t_r + \Delta t_r)}]$$

that is,

$$\mathrm{Esdm}(f) = j2 \sum_{r=1}^{M} [\alpha_r e^{-j2\pi f (t_r + 0.5\Delta t_r)} \sin(\pi f \Delta t_r)]. \quad (16)$$

Eq. (16) shows that in the lower frequency region the SDM spectral error is proportional to frequency since $\sin(\pi f \Delta t_r) \approx \pi f \Delta t_r$. However, for PWM the output pulses derived from LPM are spectrally weighed by an integrator with time constant T second, giving a PWM jitter spectrum Epwm(f) in terms of Esdm(f),

$$\mathrm{Epwm}(f) = \frac{\mathrm{Esdm}(f)}{j2\pi f T}. \quad (17)$$

Hence substituting for Esdm(f) from Eq. (16) and introducing a sinc function,

$$\mathrm{Epwm}(f) = \sum_{r=1}^{M} \left[\alpha_r e^{-j2\pi f (t_r + 0.5\Delta t_r)} \frac{\Delta t_r}{T} \mathrm{sinc}(\pi f \Delta t_r) \right]. \quad (18)$$

Eq. (18) reveals that because of integration, the PWM jitter spectrum is virtually constant with frequency and proportional to Δt_r since at low frequency $\mathrm{sinc}(\pi f \Delta t_r) \approx 1$. Exploiting Eqs. (16) and (18), two simulations were performed for SDM and PWM to reveal the output spectral error due only to jitter. The LPM center frequency for PWM was 8(44.1) kHz whereas the LFM center frequency for SDM was 64(44.1) kHz. Also both simulations used RPDF[3] jitter noise of 2 ns peak to peak. For both SDM and PWM the input signal consisted of 19- and 20-kHz sine waves, each with peak amplitude 0.1. This allowed both spectral shape and distribution about the respective carrier frequencies to be compared. The results are shown in Figs. 8 and 9. Overall the PWM results reveal greater sensitivity to jitter and also confirm the predicted spectral distributions. However, in making comparisons the different numbers of pulse transitions per unit time for SDM and PWM should be noted as the sampling rates were selected to reflect typical amplifier applications.

3 NEGATIVE-FEEDBACK-DEPENDENT DISTORTION IN PWM AMPLIFIERS

This section considers the problem of applying negative feedback in PWM power amplifiers to lower output-stage distortion and dependence on power-supply variations, both of which are major performance-limiting factors, especially in open-loop designs. Section 1 has confirmed the linearity of naturally sampled idealized PWM and has shown that an LPM model can generate equivalent non-time-quantized PWM signals. Here the only distortion fundamental to LPM occurs when the input signal has high-

[3]Rectangular probability distribution function.

frequency, high-amplitude content so nonlinear aliased components migrate downward from the sampling frequency, as illustrated in Fig. 5. The key observation is that the presence of high-frequency input signal components causes an increase in distortion. In a naturally sampled PWM amplifier that uses output-voltage-derived negative feedback, a critical performance factor is the influence on modulator linearity of the output switching signal residue following its subsequent filtering by the forward-path amplifier. The presence of these high-frequency switching components inevitably increases sideband generation in LPM and thus causes the distortion performance to degrade over what might otherwise be anticipated from the use of negative feedback. Fig. 10 shows a basic PWM modulator, which includes a negative feedback loop with feedback factor B and a forward-path amplifier with transfer function A. In practice an output reconstruction filter is required, although this is omitted here as the unfiltered output signal is to be analyzed. Also included in the loop is an analog output stage N or power switch, which is susceptible both to switching distortion and to power-supply voltage, shown here to generate an instantaneous error voltage V_ε.

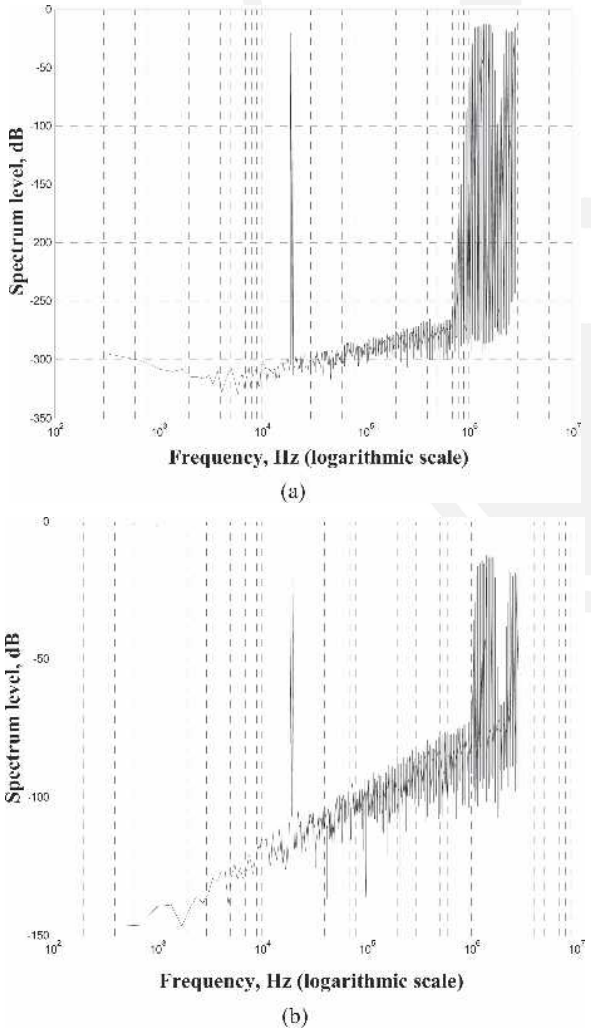


Fig. 8. SDM. (a) Without jitter. (b) With 2-ns peak-to-peak jitter.

In PWM the output voltage is a square wave and normally switches over almost the full range dictated by the power supply. It is therefore a high-amplitude signal containing high-frequency switching components. Following attenuation by B, the fed back signal is applied to the forward-path amplifier, where the transfer function A approximates

$$A = \frac{A_0}{1 + \mathrm{j}f/f_0} \Rightarrow \frac{A_0 f_0}{\mathrm{j}f} = \frac{f_T}{\mathrm{j}f} = \frac{1}{\mathrm{j}2\pi f T_A} \tag{19}$$

where A_0 is the dc gain, f_0 the 3-dB break frequency (dominant pole), $f_T = A_0 f_0$ the unity-gain bandwidth or the

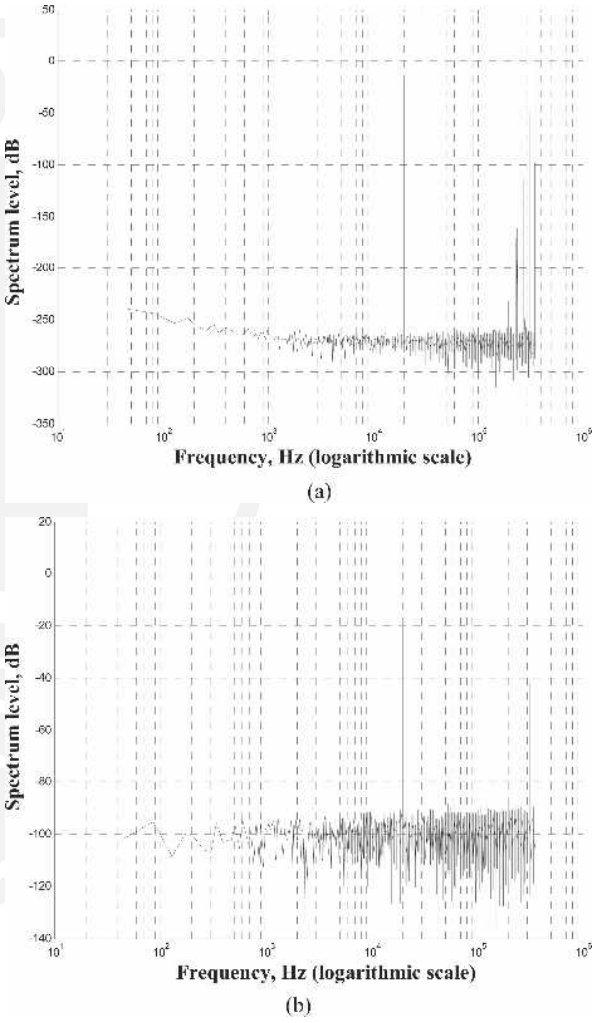


Fig. 9. Natural sampling PWM. (a) Without jitter. (b) With 2-ns peak-to-peak jitter.

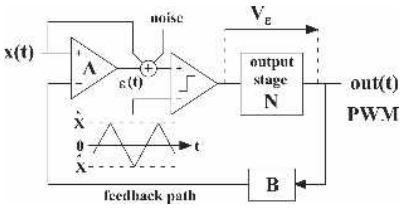


Fig. 10. Natural sampling analog PWM with feedback.

gain–bandwidth product, and T_A the corresponding time constant of the integrator (that is, amplifier A). T_A and f_T are then related as

$$f_T = \frac{1}{2\pi T_A}. \tag{20}$$

Because from Eq. (19) A tends to a first-order integrator when $A_0 \to \infty$, then $\varepsilon(t)$, the time-domain output of A, includes a triangular component produced by integration of the PWM square-wave output. Hence $\varepsilon(t)$ is a function of the PWM peak-to-peak output signal, feedback factor B, unity-gain frequency f_T, and PWM sampling rate f_{pwm}. Consequently there are high-frequency switching components present at the PWM modulator input, which can induce distortion by perturbing the natural PWM sampling instants.

In practice there are two principal factors that determine the maximum unity-gain frequency f_T of amplifier A. The first is related to the phase margin, which must include all phase shift within the amplifier loop, whereas the second, and more fundamental, is the triangular signal resulting from integration by amplifier A of the PWM square-wave output. To establish the maximum unity-gain frequency $\hat{f_T}$ for amplifier A in a simple negative-feedback PWM amplifier, assume in the forward path a unity-gain PWM stage with output power stage gain N, where both the triangular wave applied to the comparator and the PWM output square wave span −1 V to 1 V. Then allowing a 1-V margin for the signal headroom, the peak-to-peak range of $\varepsilon(t)$ is set to 1 V, that is,

$$\varepsilon(t) = \frac{NB}{T_A}\int_{t=0}^{0.5/f_{\text{pwm}}} \mathrm{d}t = 1$$

whereby the integrator time constant T_A is limited to

$$T_A \geq \frac{NB}{2f_{\text{pwm}}}. \tag{21}$$

Hence from Eqs. (20) and (21), and also confirmed by simulation to be a realistic estimate, the maximum unity-gain frequency $\hat{f_T}$ for amplifier A is

$$\hat{f_T} = \frac{f_{\text{pwm}}}{\pi NB}. \tag{22}$$

To gain further insight into PWM with negative feedback, observe that open-loop, naturally sampled PWM produces no discernable distortion in the absence of high-frequency input signals. It can therefore be said to produce an optimum PWM signal. Consequently if a closed-loop PWM amplifier were to correct completely for internal loop deficiencies then, in the absence of jitter, it should yield an output PWM waveform identical to that of the naturally sampled open-loop modulator. Any pulse relocation (other than pure delay) would represent degradation. However, because of the additional pulse-edge modulation caused by switching components present in $\varepsilon(t)$, even if the output stage N is perfect, then fundamentally a negative-feedback PWM amplifier as presented in Fig. 10, although improving on some performance aspects such as reduced power-supply sensitivity and output-stage dependence, could be anticipated to introduce distortion not present in open-loop PWM. Solutions to this problem therefore require signal processing targeted to eliminate (or significantly reduce) undesirable pulse-edge modulation.

Before presenting techniques to improve PWM closed-loop linearity, a precision simulation is presented to demonstrate distortion generation due to switching artifacts present within a PWM amplifier with overall negative feedback. A Matlab program was written based on the topology shown in Fig. 10 where, to minimize complexity, $NB = 1$. In the simulation amplifier A was modeled as an ideal z-domain integrator specified just in terms of its unity-gain frequency. Each PWM sample period was subdivided into 2^{14} increments (selected by experiment to reduce the simulation noise close to that set by finite-word-length artifacts in Matlab). For each computational increment an iterative procedure evaluated the state variables and estimated the time coordinate of each PWM transition. Time resolution was further enhanced by applying linear interpolation between computational increments. Since the topology included feedforward from the input to the PWM stage (see Fig. 10), changing the unity-gain frequency allowed the simulation to model amplifiers ranging from zero feedback, thus becoming an open-loop PWM amplifier, to maximum feedback where, From Eq. (22), $\hat{f_T} = f_{\text{pwm}}/\pi$. The simulation calculated the time coordinates $\{t_{1\text{r}}\}$ for the negative-to-positive transitions and $\{t_{2\text{r}}\}$ for the positive-to-negative transitions, where the output spectrum was evaluated using Eq. (10). To reveal noise shaping together with distortion generation as a function of feedback-loop unity-gain frequency, a noise source was added at the input to the comparator (see Fig. 10) to deliberately induce a small level of output jitter. In all other respects the amplifier had no other imperfections as the aim here was to expose only fundamental distortion mechanisms due entirely to the encapsulation of ideal, naturally sampled PWM with negative feedback.

All simulations used an input signal consisting of two sine waves, each of normalized amplitude 0.45, with respective frequencies of 17 and 20 kHz; the PWM sampling rate was set at $f_{\text{pwm}} = 16(44.1)$ kHz. The highest loop gain selected used $\hat{f_T} = f_{\text{pwm}}/\pi$, with subsequent gains reduced in increments of 20 dB up to a maximum of 60-dB attenuation. Also two simulations were performed with the amplifier gain set to zero (open-loop case), both with and without jitter noise, in order to benchmark the simulations with feedback and to observe the native resolution of the simulation. Computations were taken over 2^{14} PWM sample periods, with each period subdivided into a further 2^{14} increments. The open-loop spectral results are shown in Fig. 11 together with a histogram of the output jitter, indicating that noise induction produced a peak jitter of about 0.5 ps. The jitter level was deliberately kept low so as not to mask low-level distortion products. Spectral results for the case with feedback are shown in Fig. 12 for loop-gain attenuations of 0, −20, −40, and −60 dB. By way of comparison, Fig. 13 presents the results for uniform sampling PWM with loop gains of −20 and −60 dB.

The same simulator was used, but the a sample-and-hold was applied to the input signal with a sampling rate of $2f_{\text{pwm}}$ Hz.

The output spectra confirm that for zero-loop gain (that is, open-loop naturally sampled PWM) there is no evidence of in-audio-band intermodulation products. Also demonstrated is the extremely high resolution of the simulations with a spectral noise floor in excess of −250 dB. The inclusion of low-level jitter noise can also be seen where the histogram allows a direct observation of the jitter level that can be linked to the spectral noise level for calibration purposes. The level of time-domain jitter and the resulting spectral noise are in line with that discussed for PWM in Section 2. When feedback is applied, the

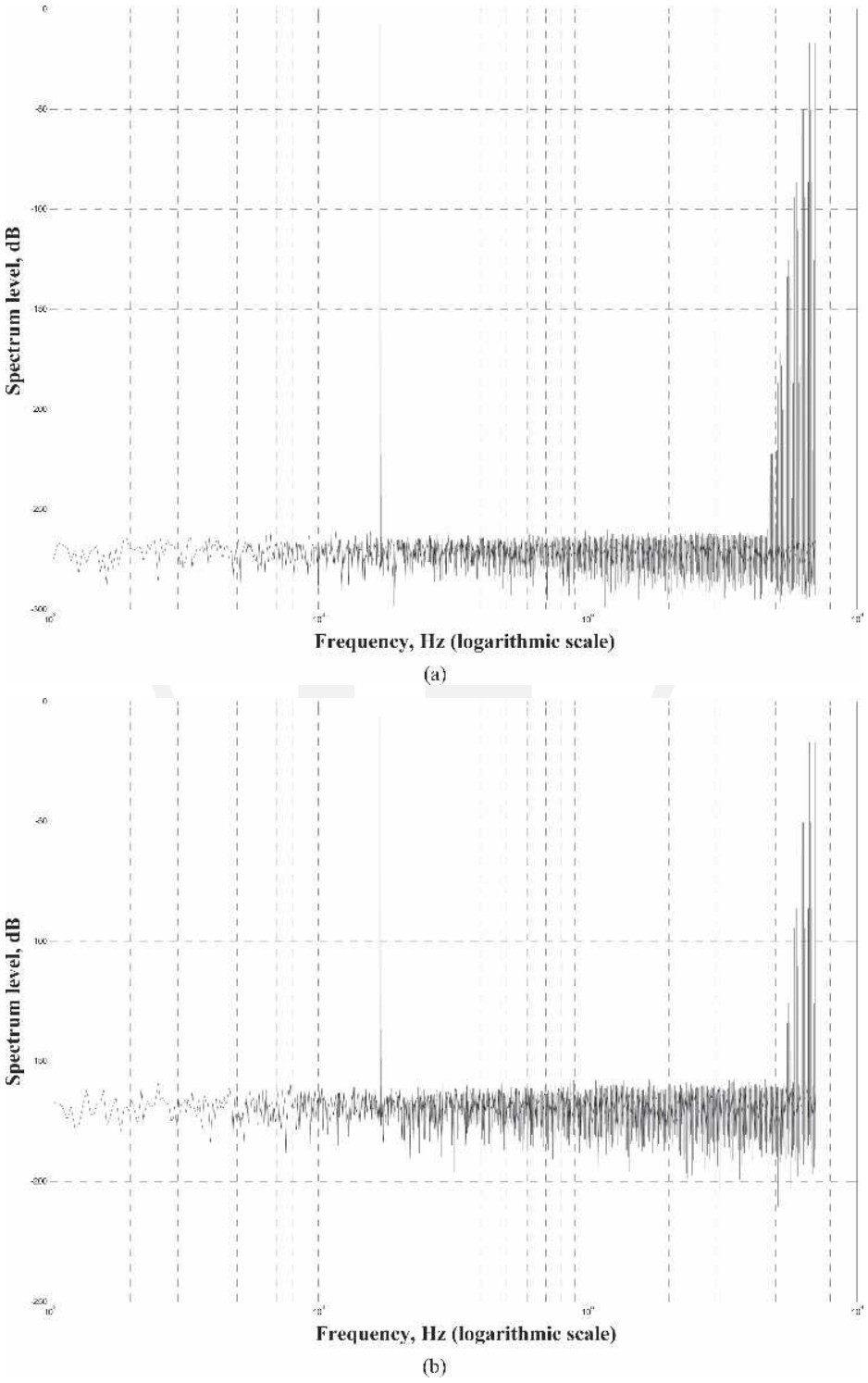


Fig. 11. Natural sampling open-loop PWM output spectra. (a) Without jitter. (b) With jitter. (c) Histogram of jitter present in PWM output.

spectral effect on jitter noise can be observed in terms of first-order noise shaping and follows theoretical predictions. However, of concern is the generation of intermodulation distortion, which for high-amplitude input signals is considered unacceptable in the context of a high-resolution amplifier. Also, contrary to normal feedback behavior and of specific interest in this paper, the distortion level is seen to rise progressively as the loop gain of the feedback amplifier is increased. Observe, however, that the results in Fig. 13 for uniform sampling PWM reveal a distortion similar to that induced by feedback, but here the distortion remains when the loop gain is reduced, a consequence of the input signal not being sampled at times coincident with the pulse transitions in the PWM output.

4 NODAL TRANSITION FILTERS (NTF) IN FEEDBACK PWM

In this section a technique using a loop filter to attenuate switching components is described to improve the linearity of feedback PWM. However, the inclusion of a filter introduces additional phase shift, which in the context of negative feedback presents problems of stability. The solution to stability has been solved here using a "constant-voltage" crossover filter [13], [14], which enables effectively the node in the circuit from which feedback is derived to be changed in a controlled frequency-dependent manner so as to bypass the internal PWM stage at high frequency. Constant-voltage filters are a specific class of loudspeaker crossover filters, where, for example, in a two-way crossover the high- and low-pass filter transfer functions sum to unity, thus exhibiting zero phase shift, a characteristic critical for stability when such filters are introduced within a negative-feedback loop. Because of the node-shifting property of a crossover filter, the process is called a nodal transition filter (NTF). Fig. 14 illustrates a negative-feedback PWM amplifier similar to that shown in Fig. 10, but using an NTF with a low-pass filter $\gamma(f)$ and a high-pass filter $1 - \gamma(f)$, where the constant-voltage properties follow from $\gamma(f) + [1 - \gamma(f)] \equiv 1$.

Fig. 15 presents an equivalent but more efficient topology than Fig. 14, where the filter $1 - \gamma(f)$ is derived from $\gamma(f)$. To demonstrate equivalence let the internal PWM and output stages have a nonlinear transfer function N. Referring to Fig. 14, the signal V_f fed back to the inverting input of A is

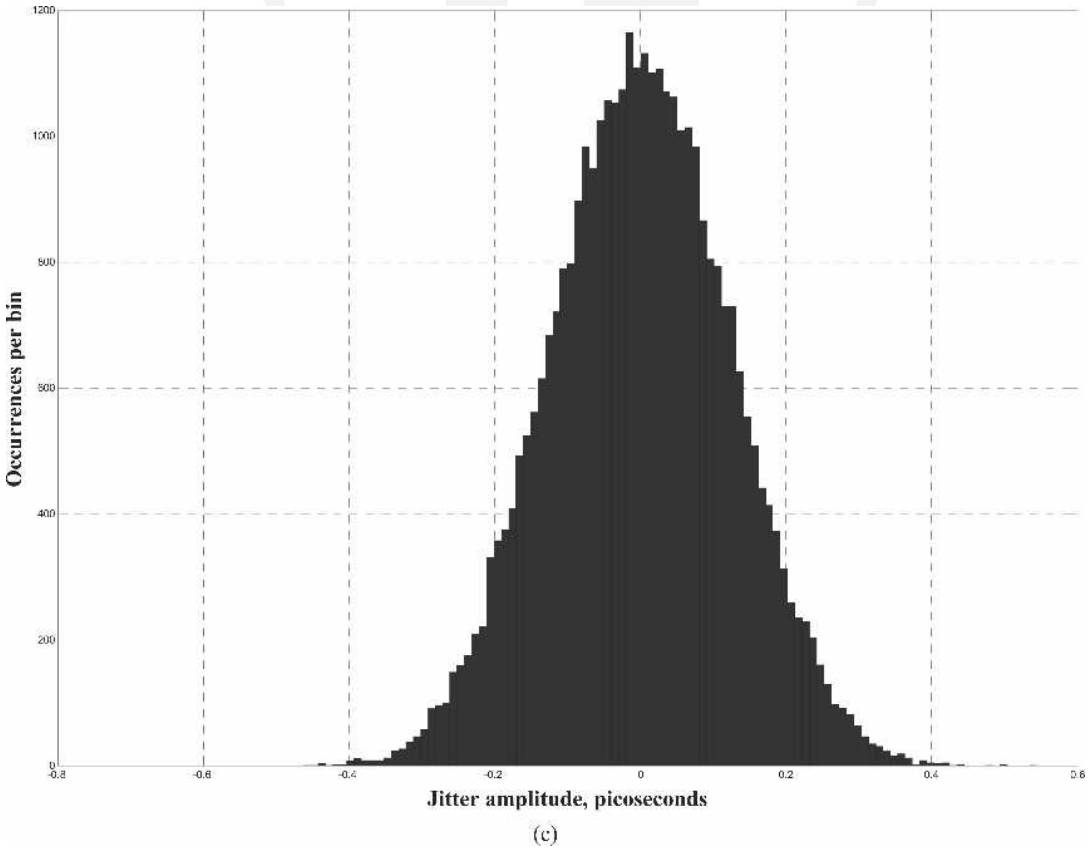
$$V_f = \mathrm{OUT}(f)\left[\gamma(f) + \frac{(1 - \gamma(f))}{N}\right]B. \tag{23}$$

Similarly, analyzing the topology of Fig. 15,

$$V_f = [\mathrm{OUT}(f) - V_\varepsilon + \gamma(f)V_\varepsilon]B$$

$$V_\varepsilon = \mathrm{OUT}(f)\left(1 - \frac{1}{N}\right).$$

Eliminating V_ε, equivalence is confirmed since V_f is again given by Eq. (23).



(c)

Fig. 11. *Continued*

To determine the properties of an NTF enhanced amplifier, the topology shown in Fig. 15 is analyzed to derive the closed-loop transfer function $G_{\mathrm{NTF}}(f)$,

$$G_{\mathrm{NTF}}(f) = \frac{AN}{1 + AB[1 + \gamma(f)(N-1)]}. \tag{24}$$

To observe NTF operation note that $\gamma(f)$ is a low-pass filter, where $\gamma(f \to 0) \to 1$ and $\gamma(f \to \infty) \to 0$. Thus

$$G_{\mathrm{NTF}}(f)|_{f\to 0} = \frac{AN}{1+ABN} \to \left.\frac{1}{B}\right|_{AB\gg 1} \tag{25}$$

and

$$G_{\mathrm{NTF}}(f)|_{f\to \infty} = \frac{AN}{1+AB} \to \left.\frac{N}{B}\right|_{AB\gg 1}. \tag{26}$$

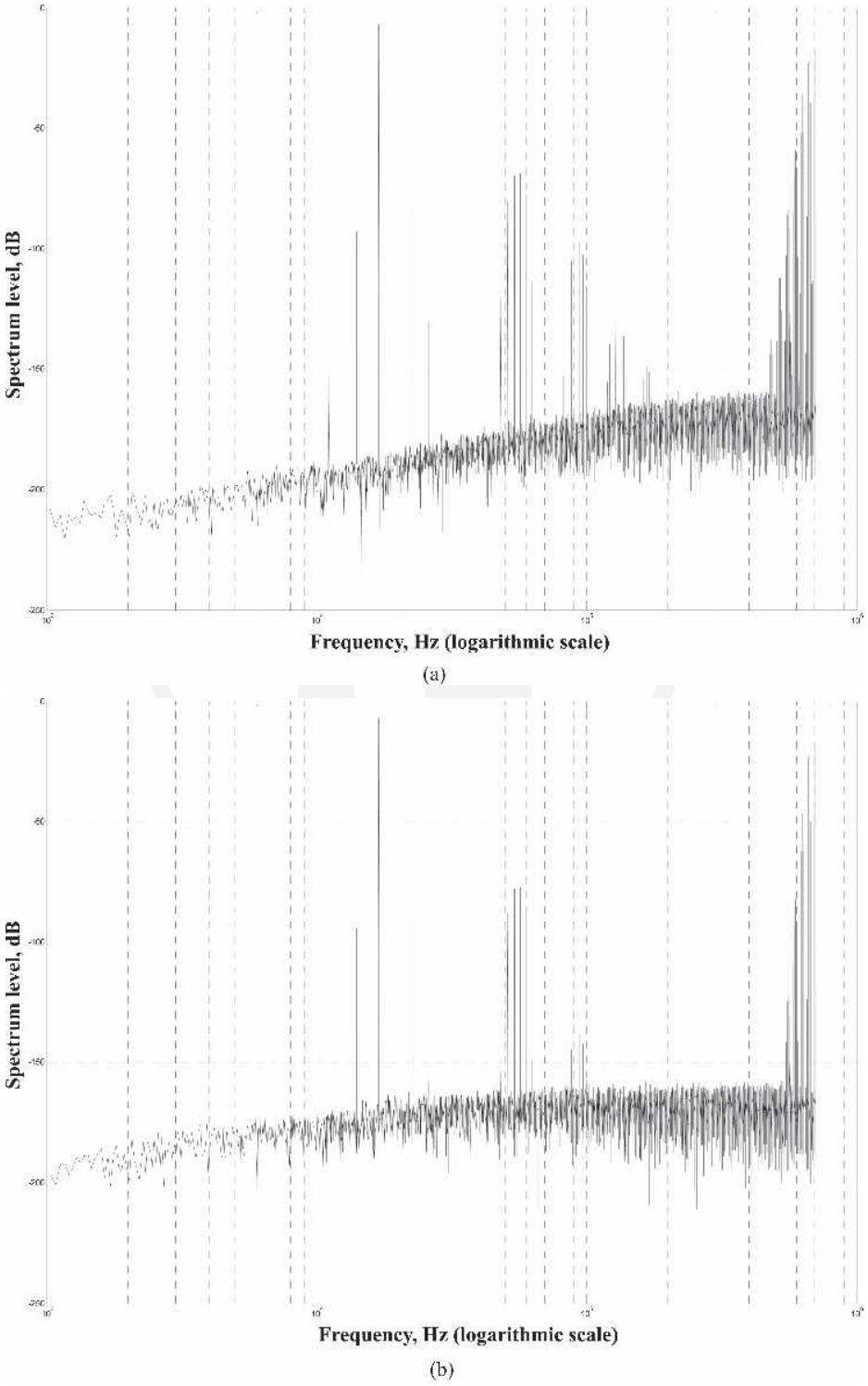
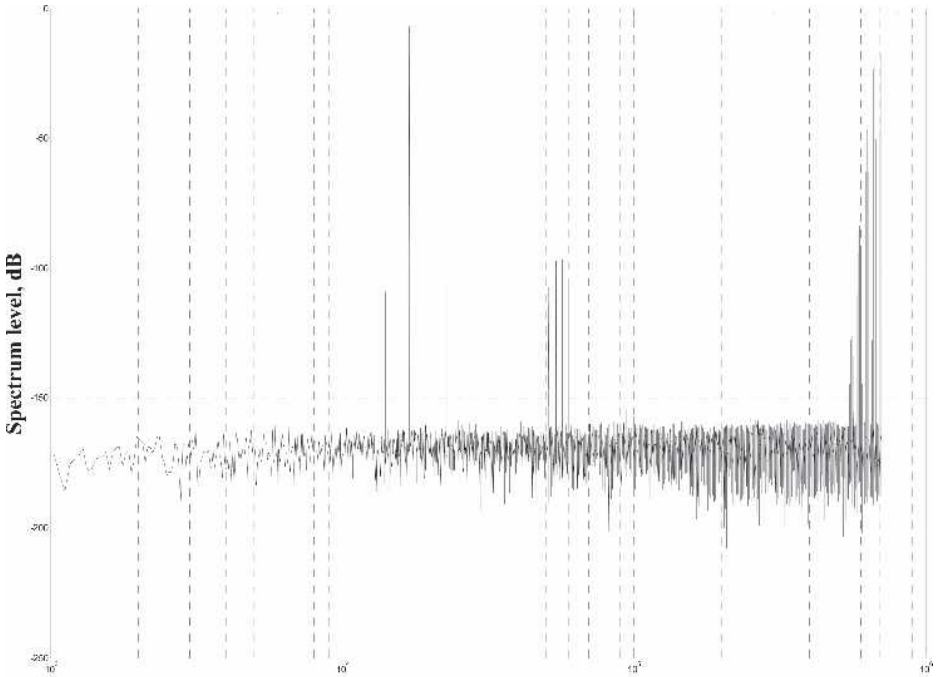


Fig. 12. Natural sampling closed-loop PWM output spectra. (a) With jitter, maximum loop gain. (b) With jitter, loop gain −20 dB below maximum. (c) With jitter, loop gain −40 dB below maximum. (d) With jitter, loop gain −60 dB below maximum.

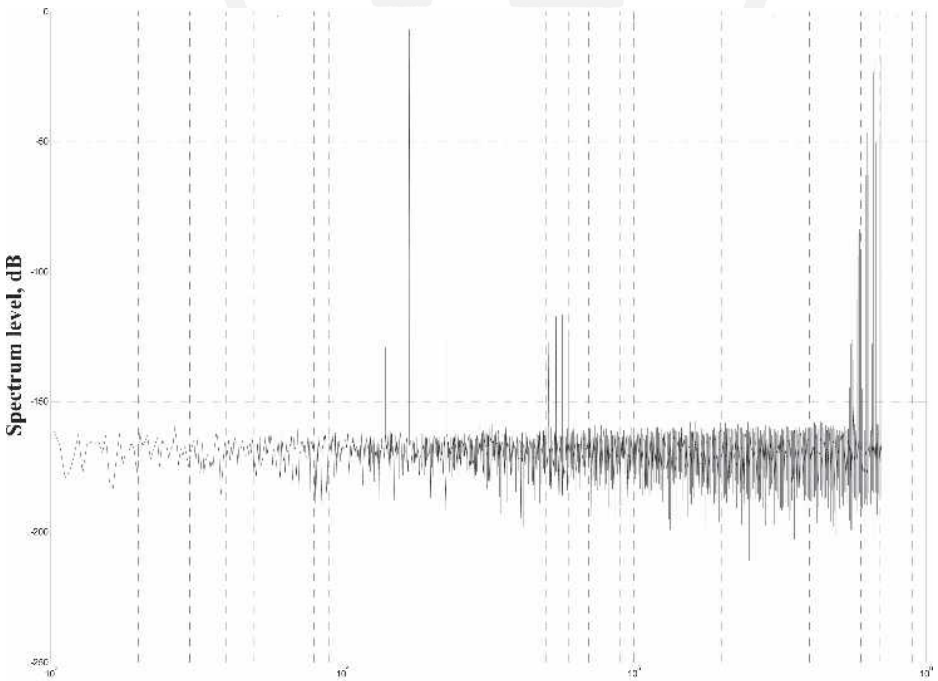
These limiting cases reveal that at lower frequencies the modulator and the output stage N are located within the overall feedback loop whereas at high frequency the feedback path is progressively transferred between nodes until feedback is derived from the output of the amplifier A. Thus PWM and the power output stages are excluded and appear open loop. Consequently with appropriate NTF design, $\gamma(f)$ can filter most of the switching components produced by the PWM stage while $1 - \gamma(f)$ maintains closed-loop stability by allowing feedback directly from the output of A at high frequency.

To investigate the distortion reduction performance of an NTF-enhanced PWM amplifier, let

$$\gamma(f) = \frac{1}{1 + \sum_{r=1}^{R} a_r (\mathrm{j}2\pi f)^r}. \tag{27a}$$



(c)



(d)

Fig. 12. *Continued*

Here $\gamma(f)$ is a low-pass filter of order R with filter coefficients $\{a_1, \ldots, a_R\}$, where at high frequency

$$\gamma(f)|_{f\to\infty} \to \frac{1}{a_R(\mathrm{j}2\pi f)^R}. \tag{27b}$$

To determine the sensitivity of $G_{\mathrm{NTF}}(f)$ to N, an error function $E_{\mathrm{NTF}}(f)$ is defined [15],

$$E_{\mathrm{NTF}}(f) = 1 - \frac{G_{\mathrm{NTF}}(f)}{G_{\mathrm{NTF}}(f)|_{\mathrm{target}}}. \tag{28}$$

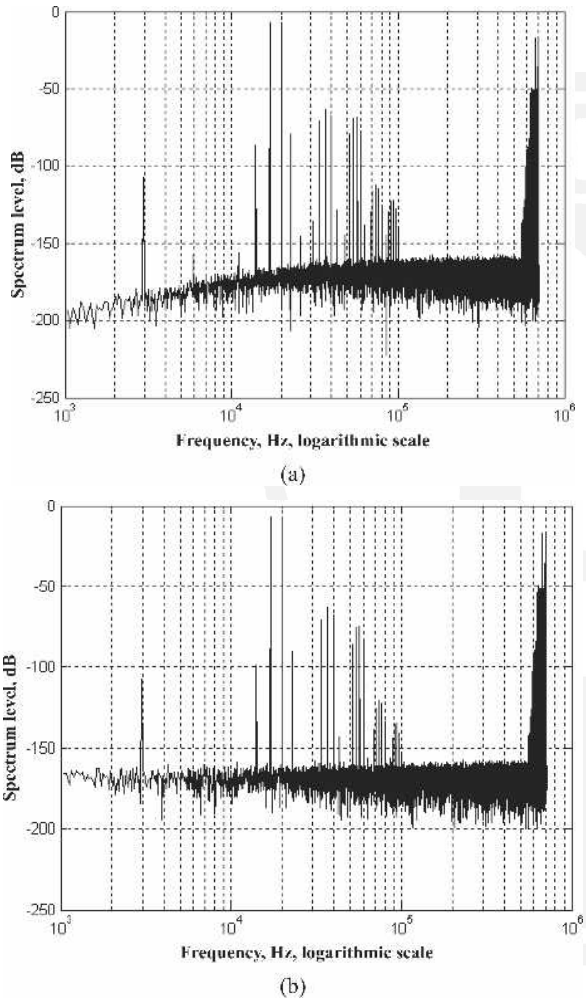


Fig. 13. Uniform sampling closed-loop PWM output spectra. (a) With jitter, loop gain −20 dB below maximum. (b) With jitter, loop gain −60 dB below maximum.

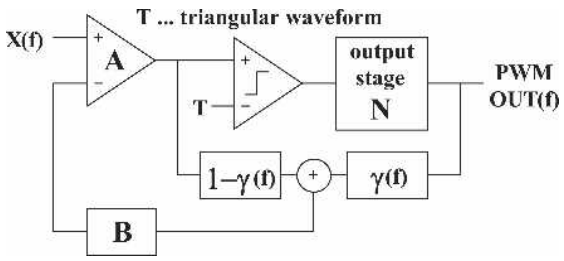


Fig. 14. Analog NTF feedback PWM.

The target transfer function $G_{\mathrm{NTF}}(f)|_{\mathrm{target}} = G_{\mathrm{NTF}}(f)$ for $N = 1$, and from Eq. (24),

$$E_{\mathrm{NTF}}(f) = \frac{(1-N)\{1+AB[1-\gamma(f)]\}}{1+AB[1+\gamma(f)(N-1)]}. \tag{29a}$$

Assuming $N \approx 1$ and $AB \gg 1$, $E_{\mathrm{NTF}}(f)$ then simplifies to

$$E_{\mathrm{NTF}}(f) \approx \left.\frac{(1-N)[1-\gamma(f)]}{1+\gamma(f)(N-1)}\right|_{AB\gg 1} \approx (1-N)[1-\gamma(f)]|_{N\approx 1}. \tag{29b}$$

Eq. (29b) reveals that the output-stage error is shaped in frequency by the high-pass filter $1 - \gamma(f)$ and not the loop gain AB. However, because the NTF is a constant-voltage crossover filter when high- and low-pass transfer functions sum to unity, the high-pass filter is limited to (pseudo) first order even if the low-pass filter has a high rate of attenuation. This constraint on the filter order is demonstrated in the following.

From Eq. (27a) the derived high-pass filter $1 - \gamma(f)$ follows,

$$1-\gamma(f) = \frac{\sum_{r=1}^{R} a_r(\mathrm{j}2\pi f)^r}{1+\sum_{r=1}^{R} a_r(\mathrm{j}2\pi f)^r}. \tag{30a}$$

Eq. (30a) shows that at high frequency $[1 - \gamma(f)]|_{f\to\infty} \to 1$, but at low frequency the asymptotic rate of attenuation is dictated by the term with the lowest power in the numerator, namely, as $f \to 0$,

$$[1-\gamma(f)]|_{f\to 0} = \frac{\sum_{r=1}^{R} a_r(\mathrm{j}2\pi f)^r}{1+\sum_{r=1}^{R} a_r(\mathrm{j}2\pi f)^r} \to a_1\mathrm{j}2\pi f. \tag{30b}$$

A limit on the asymptotic slope applies also to the derived high-pass transfer function when the high- and low-pass filter orders are interchanged, that is, for a high-pass filter of order R,

$$[1-\gamma(f)]|_{f\to 0} = \frac{a_R(\mathrm{j}2\pi f)^R}{1+\sum_{r=1}^{R} a_r(\mathrm{j}2\pi f)^r} \to a_R(\mathrm{j}2\pi f)^R. \tag{31a}$$

Then the derived low-pass filter high-frequency asymptotic response is

$$\gamma(f)|_{f\to\infty} = \frac{1+\sum_{r=1}^{R-1} a_r(\mathrm{j}2\pi f)^r}{1+\sum_{r=1}^{R} a_r(\mathrm{j}2\pi f)^r} \to \frac{a_{R-1}}{a_R\mathrm{j}2\pi f}. \tag{31b}$$

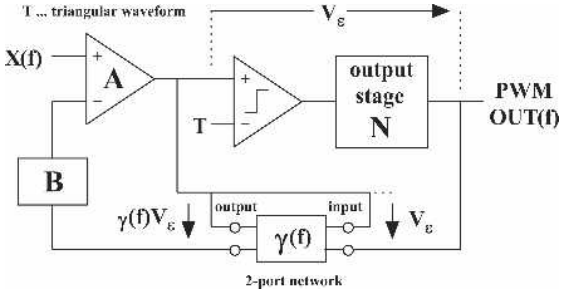


Fig. 15. Equivalent NTF feedback PWM.

Although a high-order high-pass filter increases the noise-shaping advantage within the bound of loop gain, Eq. (31b) shows that there is less attenuation of the switching components due to the first-order response of the derived low-pass filter. Conversely, even with a high-order low-pass filter and with $AB \gg 1$, the derived filter cannot exceed a first-order response. Here the low-frequency asymptotic noise-shaped error is derived from Eqs. (29b) and (30b),

$$E_{\mathrm{NTF}}(f)|_{f\to 0} \approx a_1 \mathrm{j}2\pi f(1-N). \tag{32}$$

To conclude this section the error function $E_{\mathrm{NTF}}(f)$ is compared against the noise-shaping transfer function (NSTF) used, for example, in the study of SDM. Fig. 16 shows the same NTF amplifier topology, but where the PWM output stage has been replaced with a linear gain stage N followed by an additive noise source V_{d}. Analyzing this topology, the output signal OUT(f) expressed as a function of input $X(f)$ and noise source V_{d} then follows,

$$\mathrm{OUT}(f) = \frac{1-AB[1-\gamma(f)]}{1+AB[1+N\gamma(f)-\gamma(f)]}V_{\mathrm{d}} + \frac{NA}{1+AB[1+N\gamma(f)-\gamma(f)]}X(f). \tag{33}$$

If the output stage gain is set to $N = 1$ to match the optimum conditions in the NTF amplifier, Eq. (33) reduces to

$$\mathrm{OUT}(f) = \frac{1}{1+AB}V_{\mathrm{d}} - \frac{AB}{1+AB}[1-\gamma(f)]V_{\mathrm{d}} + \frac{A}{1+AB}X(f) \tag{34a}$$

that is, when the loop gain is large,

$$\mathrm{OUT}(f)|_{AB\gg 1} \to \frac{1}{AB}V_{\mathrm{d}} - [1-\gamma(f)]V_{\mathrm{d}} + \frac{1}{B}X(f). \tag{34b}$$

Alternatively, expressed in terms of the closed-loop transfer function $H(f)$ and the noise-shaping transfer functions $D_{f_1}(f)$, $D_{f_2}(f)$,

$$\mathrm{OUT}(f) = D_{f_1}(f)V_{\mathrm{d}} - D_{f_2}(f)V_{\mathrm{d}} + H(f)X(f). \tag{34c}$$

Eqs. (34) reveal the closed-loop transfer function $H(f)$ as the standard canonic expression for a negative-feedback amplifier, whereas the additive output noise is shaped by two distinct NSTFs. $D_{f_1}(f)$ follows normal feedback theory and shows the noise shaped by the inverse of the amplifier loop gain; however, the second function, $D_{f_2}(f)$ [see Eq. (34b)], tends to $1-\gamma(f)$ when the loop gain is large. In practice noise shaping is dominated by $D_{f_2}(f)$. Thus the NSTF is equivalent to the high-pass transfer function of the NTF and compares to $E_{\mathrm{NTF}}(f)$ defined in Eq. (29b). This analysis confirms that increasing the loop gain AB offers little advantage with respect to the noise-shaping output error as the NTF limits the NSTF to a first-order response.

To demonstrate the validity of the NTF methodology proposed in this section, a circuit simulation was performed on the PWM amplifier presented in Fig. 17, which is derived from the conceptual topology shown in Fig. 15. In this example a six-stage RC ladder network was used, where $R = 1\ \mathrm{k}\Omega$ and $C = 1$ nF. Results are presented both with and without NTF, so changes in waveform detail can be observed. In both simulations the PWM output and the output of amplifier A were computed. The results shown in Fig. 18 are for a simple feedback amplifier where high-frequency signal components are present within the feedback loop, whereas in Fig. 19 the inclusion of the NTF suppresses these elements to reveal a waveform segment virtually free of high-frequency artifacts. Fig. 20 presents an alternative NTF using a second-order LCR filter, and Fig. 21 shows a simulation to demonstrate high-frequency suppression. In Fig. 22 a variant of the NTF-enhanced PWM amplifier is illustrated. Here the low-pass filter $\gamma(f)$

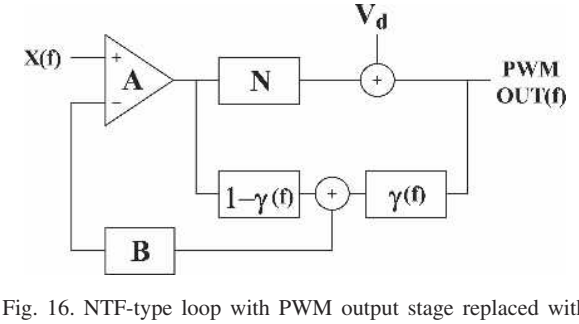


Fig. 16. NTF-type loop with PWM output stage replaced with noise source.

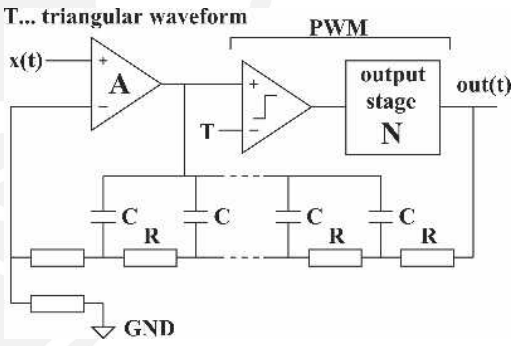


Fig. 17. Simulation of NTF PWM amplifier.

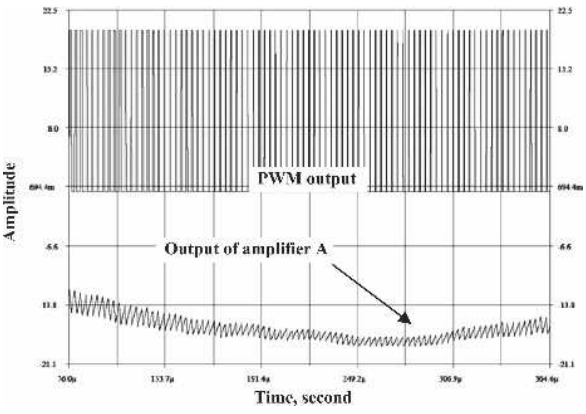


Fig. 18. PWM amplifier without NTF.

is now positioned after amplifier *A* to attenuate switching artifacts prior to the PWM stage and the high-pass filter $1 - \gamma(f)$ is placed in a feedforward path to the output where the composite signal forms the feedback signal.

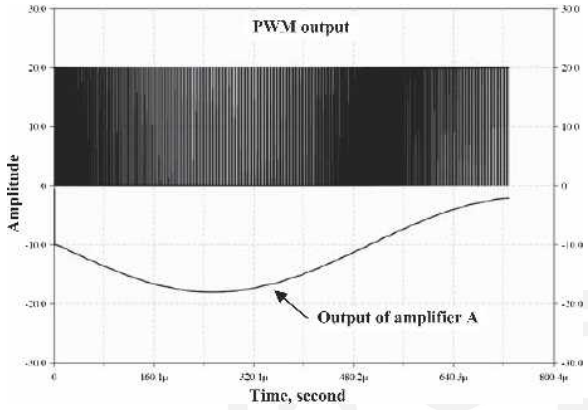


Fig. 19. PWM amplifier with NTF.

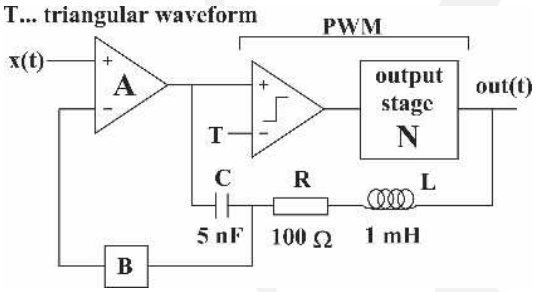


Fig. 20. PWM amplifier with *LCR* NTF.

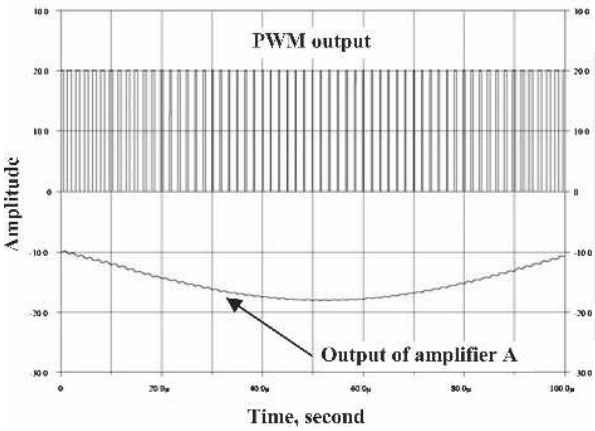


Fig. 21. Signal segment for PWM amplifier with *LCR* NTF.

To summarize this section, the technique of using an NTF within the closed loop of a PWM feedback amplifier achieves two principal performance enhancements:

- The low-pass filter $\gamma(f)$ of unconstrained order suppresses high-frequency switching artifacts present within the feedback loop and thus lowers distortion as the input signal to the modulator has reduced high-frequency content.
- Feedback achieves first-order distortion reduction as described by Eqs. (29), (32), and (34), similar to a conventional feedback amplifier incorporating a first-order forward-path amplifier.

5 INTERNAL SWITCHING COMPENSATION IN NEGATIVE-FEEDBACK PWM AMPLIFIERS

Section 3 analyzed the problem of distortion resulting from switching components within a PWM feedback amplifier and in Section 4 a method of filtering within the loop was discussed. As simulation revealed, the suppression of high-frequency signals within the feedback loop enables PWM to realize its full linearity potential within the bounds of comparator switching performance, output signal level control and both power supply and output-stage-induced switching artifacts. In this section an alternative approach is presented where a reference (namely, very low distortion) naturally sampled PWM side chain is located within the amplifier feedback loop and used to cancel the switching components of the main PWM stage. In this scheme filtering of the PWM signal fed back from the output of the amplifier can be omitted so that full-bandwidth closed-loop control is retained. The proposed scheme is shown in Fig. 23, where the loop contains two modulators, both driven from the output of amplifier *A*.

In the side chain the output of the reference PWM is subtracted from the output of amplifier *A* to form the idealized PWM error signal, which under optimum alignment contains no low-frequency information and thus carries only high-frequency switching distortion. This signal can be filtered, although there are constraints imposed by phase distortion which may affect adversely the process of compensation. The high-frequency error signal is then added to the fed back signal, where the summation process is designed so that, ideally, the switching signal in the main PWM output and the derived side chain PWM error cancel. As such the fed back signal matches closely the output of amplifier *A,* necessary to maintain proper loop behavior, but remains sensitive to the error signal between the main PWM output stage and the side chain reference

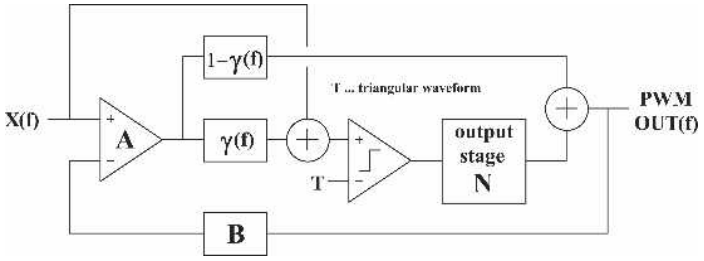


Fig. 22. NTF PWM amplifier reconfigured with feedforward path.

PWM, the latter having to be optimized for low distortion. It is critical for PWM signal levels to be matched by scaling as in practical amplifiers there will be differences between the main PWM output and the low-level reference PWM output. It is therefore desirable to calibrate the amplifier to achieve maximum switching cancellation.

To demonstrate the effectiveness of compensation, circuit simulation was performed for an optimally aligned PWM feedback amplifier. Fig. 24 presents an example output waveform for amplifier *A,* where an absence of switching-induced ripple is revealed. The corresponding PWM output waveform is also shown. However, for a nonideal output stage (and that includes gain error) there can be additional distortion resulting from unsuppressed high-frequency switching artifacts degrading the PWM process. Nevertheless, switching compensation can still offer substantial improvement, even if cancellation is imperfect. To summarize,

- Nonideality in the output stage adds distortion to the output voltage, as occurs in conventional analog amplifiers.
- Nonideality including simple gain error also implies error between the main PWM output stage and the reference PWM stage. This results in high-frequency switching ripple being added to the fed back signal, which can degrade modulator linearity.

These factors need to be considered when feedback is applied to PWM, otherwise some of the distortion reduction capability of feedback is impaired. For a practical amplifier design the following strategy is suggested:

1) Introduce an NTF to further reduce residual switching artifacts, as shown in Fig. 25. However, because this filter no longer has to suppress the whole of the switching distortion of the PWM, it can be less invasive and designed to facilitate an increase in loop gain. For example, it was shown in Section 4 [see Eqs. (31)] that if an NTF high-pass filter is high order then the derived low-pass filter is first order. Hence if switching suppression has already been implemented, a first-order low-pass filter allows some additional switching attenuation whereas a high-order high-pass filter enables improved noise shaping according to Eqs. (34). In practice the NTF would be selected using experimental data, including the phase margin and taking into account the PWM sampling rate to achieve the best overall performance compromise.

2) It is recommended that the side-chain process has an embedded controller (see also Section 6) to optimize the gain of the output stage so that, on average, switching artifacts are minimized. Fig. 25 shows a possible basic system topology incorporating variable gain and controller.

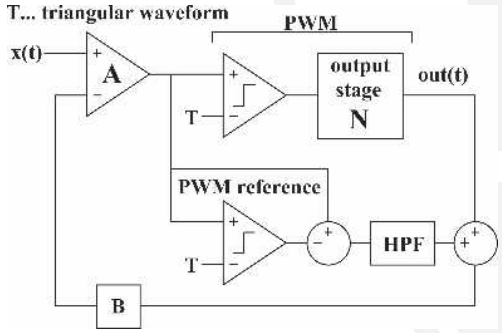


Fig. 23. PWM amplifier including reference PWM configured for switching compensation.

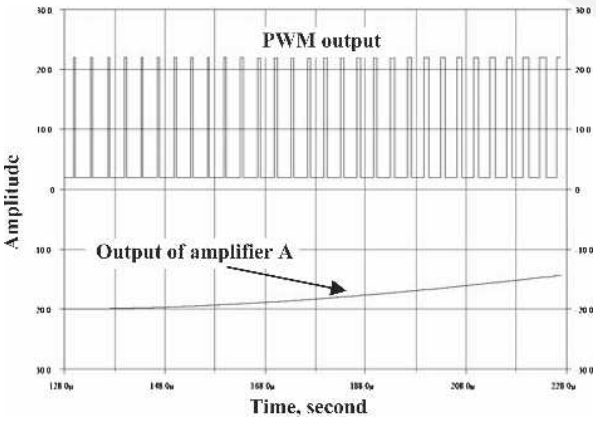


Fig. 24. Signal segment for PWM amplifier with side-chain switching compensation.

6 PREDICTIVE SWITCHING COMPENSATION IN NEGATIVE-FEEDBACK PWM AMPLIFIERS

An alternative approach to correct for switching distortion in PWM feedback amplifiers is to use a predictive side-chain based on open-loop PWM located in the input path to the main amplifier. In this scheme high-precision open-loop PWM first generates a naturally sampled PWM signal that predicts the optimum output of the feedback PWM amplifier. This signal is then subtracted from the input to produce a prediction of the switching error, which if performed accurately carries only the high-frequency signal components that normally reside well above the audio band. Now if the main feedback PWM amplifier were to output optimum naturally sampled PWM, then by subtraction, using the predicted switching error now present in the input signal to the feedback amplifier, the switching components in the amplifier feedback path are canceled, resulting in the desired low-ripple signal at the output of amplifier *A*. This scheme is shown in Fig. 26, where it is mandatory for both predictive PWM and inter-

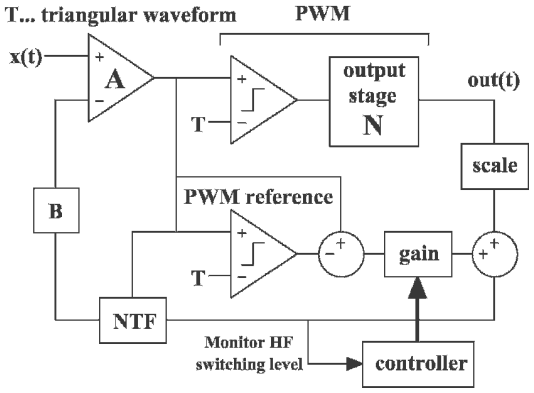


Fig. 25. PWM with NTF and auto balance.

nal-loop PWM to use the same reference triangular wave to match the gains and to time synchronize the two modulators. To confirm the efficacy of this technique, a circuit-level simulation achieved near-perfect ripple suppression using a relatively high loop gain established by a forward path amplifier with a time constant of 0.4 μs, a feedback factor of unity, and a PWM sampling rate set at 300 kHz.

Inspection of the signal flow in Fig. 26 reveals that the predictive process can be simplified to a cascaded open-loop PWM stage, as shown in Fig. 27, a result that initially may not have been anticipated. For this system to yield optimum switching waveform suppression requires identical comparators with matched output voltage ranges, consequently defining N_i as the gain of the predictive input PWM stage and N as the voltage gain of the internal PWM stage used to scale signals to the full output voltage swing. Then

$$NB = N_i. \tag{35}$$

Hence if the feedback factor B is constant in order to establish a well-defined closed-loop gain, N should be programmable so that loop conditions can be optimized dynamically according to Eq. (35). Thus, for example, changes in power-supply voltage and component tolerances can be compensated. A practical means to vary N is to modulate its power-supply voltage as in PWM this has a multiplicative function. Fig. 28 illustrates a forward control loop that monitors the short-term (say rms) power of the error $\varepsilon(t)$. This information can then be used to control the output-stage gain N by modulating the output-stage power-supply voltage with the aim of minimizing the power of $\varepsilon(t)$, where this implies $NB = N_i$. Since this strategy minimizes the switching error, the generation of secondary distortion as discussed in Section 3 is also minimized. Consequently if the PWM output stage and feedback path together have a gain equal to 1, assuming $N_i = 1$, then the loop gain of the negative-feedback amplifier is A. This process can be viewed as a secondary error correction strategy compensating for nonideal gain in the power output stage and also allowing a degree of regulation for the associated power supply. Such a procedure then integrates power-supply regulation within the PWM output stage rather than as an independently controlled subsystem that has to be calibrated. However, in systems where the power-supply voltage is regulated independently an alternative method to control the output stage gain is to modulate the amplitude of the triangular wave used in the forward path PWM.

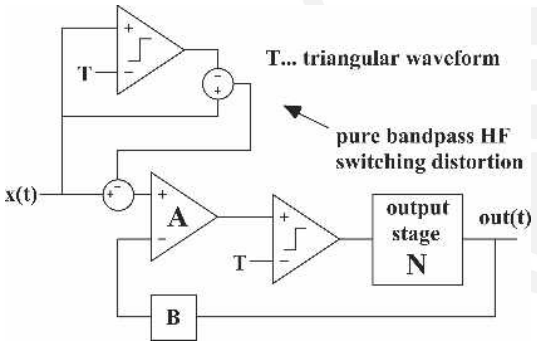


Fig. 26. PWM feedback amplifier with feedforward switching compensation.

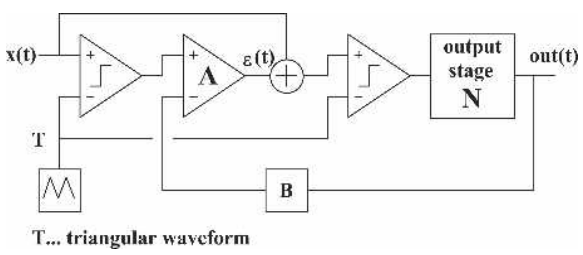


Fig. 27. Simplified PWM feedback amplifier with feedforward switching compensation.

The following analysis describes the distortion reduction characteristics of the PWM amplifier shown in Fig. 28 as well as the sensitivity to the magnitude of the error signal $\varepsilon(t)$. The following Fourier transforms are assumed:

$$\varepsilon(t) \Rightarrow E(f), \qquad x(t) \Rightarrow X(f), \qquad \text{out}(t) \Rightarrow \text{OUT}(f).$$

Also let input and output PWM stages have transfer functions N_i and N, respectively, where the optimum alignment is $NB = N_i$, with $N_i \to 1$. In this system the function of the feedforward path from amplifier input to comparator input should be observed since it makes $\varepsilon(t)$ a true error signal that is zero under optimum conditions. Hence if $A = 0$, the structure reverts to open-loop, naturally sampled PWM. From the PWM topology in Fig. 28 $E(f)$ can be expressed as

$$E(f) = A[X(f)N_i - \text{OUT}(f)B]$$

with $\text{OUT}(f)$ given via the output stage by

$$\text{OUT}(f) = N[X(f) + E(f)].$$

Substituting for $E(f)$ the input–output transfer function then becomes

$$\frac{\text{OUT}(f)}{X(f)} = N\left(\frac{1 + N_i A}{1 + NAB}\right) \to \left.\frac{N_i}{B}\right|_{NAB \gg 1}. \tag{36}$$

Similarly $E(f)$ can be determined by eliminating $\text{OUT}(f)$,

$$\frac{E(f)}{X(f)} = \left(\frac{NAB}{1 + NAB}\right)\left(\frac{N_i}{NB} - 1\right) \to \left.\left(\frac{N_i}{NB} - 1\right)\right|_{NAB \gg 1}. \tag{37}$$

Eq. (37) confirms that $E(f)$ is zero for $A = 0$ and for $NB = N_i$, as implied by Eq. (35). Hence amplifier A only

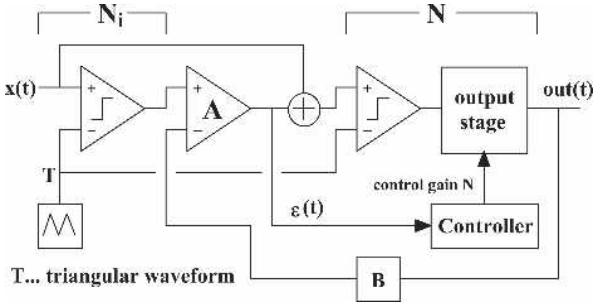


Fig. 28. PWM feedback amplifier with feedforward switching compensation and dynamic loop gain control.

produces a finite output when the input PWM and output PWM stages differ in their performance due to gain errors and output-stage imperfections. The analysis also shows that the ultimate performance of this type of PWM feedback topology is bounded by the linearity of the open-loop input PWM stage, where Eq. (36) confirms that even for finite gain A, with optimum alignment $NB = N_i$, then the overall transfer function is N_i/B. Also, it should be observed that the closed-loop gain performance is not sensitive to the condition stated in Eq. (35). Thus even if a secondary control loop modulates N to seek optimum loop conditions, this will not strongly affect the output other than to fine-tune the distortion performance by suppressing switching artifacts present in $\varepsilon(t)$.

6.1 Output Stage for Digital PWM

As a corollary to this section, because the amplifier with predictive compensation requires a PWM input signal, it follows that it can be used not only with a naturally sampled PWM code but also with a code derived from uniformly sampled, digitally derived PWM where the output is not directly amenable to analog control to take into account output-stage imperfections. Digital PWM normally uses linearization and noise shaping [10], [11] with pulse transitions calculated algorithmically. Consequently the pulses can be applied directly via a two-level DAC to the analog feedback power output stage because digital PWM forms the predictive stage required to reduce switching ripple. To implement a practical power amplifier there are signal-processing factors to consider. Fig. 29 shows a conceptual system together with a simplified signal flow diagram. In interpreting the functionality of this system, its mixed signal architecture must be considered as some processes are digital while others are analog. The basic function of subsystems N_1, N_2, N_3, N_4 are summarized as follows: N_1 represents uniformly sampled digital PWM, where the output is a time-domain quantized PWM signal sampled at a high clock frequency so that noise shaping realizes an acceptable signal-to-noise ratio (SNR). Including time-delay compensation T_x, an error signal is derived from across N_1 and fed forward to the amplifier input via a high-order digital low-pass filter N_2 designed to attenuate switching components. Significantly the filtered error signal is of low level; consequently the DAC and the analog reconstruction filter in this path require high accuracy but only low resolution. N_3 also includes a high-order, low-pass digital filter and DAC to form the feedforward signal summed with the PWM comparator input. Finally N_4 represents the transfer function of the PWM output stage. In addition there are two delay networks T_x and T_y to compensate for digital filter and process time delays. Analyzing this system the overall transfer function G_d of the simplified digital PWM amplifier is

$$G_d = \frac{N_4 A}{1 + N_4 AB}\left[\frac{N_1 N_3}{A} + N_1(e^{-j\omega T_y} - N_2) + N_2 e^{-j\omega T_x}\right]. \quad (38)$$

Eq. (38) shows that when $AB \gg 1$, then there is low sensitivity to N_3 and N_4, implying that the DAC performance in the feedforward path is noncritical and that output-stage distortion is reduced by feedback. Assuming high loop gain, Eq. (38) reduces to

$$G_d|_{AB\gg 1} \rightarrow N_1(e^{-j\omega T_y} - N_2) + N_2 e^{-j\omega T_x}. \quad (39)$$

If within the low-frequency pass band $N_2 \Rightarrow e^{-j\omega T_y}$, then Eq. (39) reduces further to just a pure time delay of $T_x + T_y$, revealing the most critical process is the low-pass filter N_2 associated with feedforward error correction about the uniformly sampled PWM stage, noting that T_x is chosen to minimize the level of error in this path. However, if N_1 includes linearization then even this condition for N_2 is noncritical. To summarize, the following features are highlighted:

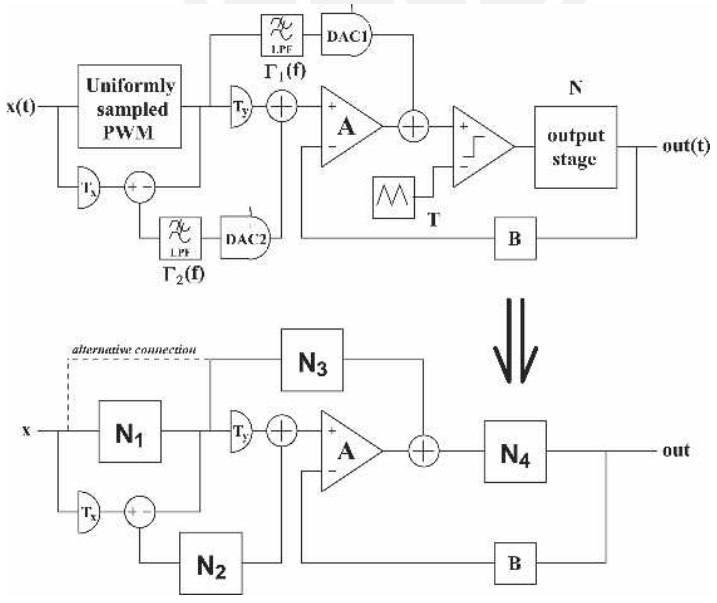


Fig. 29. Conceptual amplifier configuration for uniformly sampled digital PWM.

- The digital PWM stage N_1 normally includes linearization; therefore it produces low distortion.
- N_2 is critical to desensitizing system dependence on N_1, but because it processes only very low-level signals and is defined mainly by the passband of a digital low-pass filter, N_2 can be extremely accurate.
- N_3 is noncritical where any associated distortion is lowered because of feedback.
- If N_2 is optimized then the system has low sensitivity to the linearization process in the digital PWM stage N_1; possibly this feedforward correction procedure renders linearization unnecessary.
- The PWM output of N_1 drives the power amplifier directly, forming a predictive switching signal that facilitates the lowering of switching ripple in the feedback loop of the power amplifier.

This summary concludes the discussion on PWM. In the next section SDM digital power amplification is considered, using a highly stable coder and an output stage with low commutation losses.

7 SDM POWER AMPLIFIER SYSTEMS

Earlier work [4] has presented a discussion on switching power amplifiers based on a quantized SDM code. In this scheme the output of an SDM directly controls a power switch which, as with a PWM power amplifier, drives the loudspeaker via an analog low-pass filter. Variants of the output stage topology are reviewed in Section 7.2, whereas Sections 7.1.1 and 7.1.2 revisit SDM loop design as there are the following specific requirements.

- Adequate audio band SNR determined by the choice of NSTF.
- High-level input signal coding required as conventional high-order SDM normally limits maximum modulation depth, thus inhibiting maximum output from being achieved for a given power-supply voltage.
- Robust stability [16], that is, a low probability of instability at high input signal levels, ideally allowing a modulation index of unity.

7.1 Robust Loop Stability in SDM with High-Amplitude Input Signals

A conceptual digitally addressed SDM power amplifier scheme is shown in Fig. 30. Here, because of its wide adoption as a professional SDM coding algorithm and by way of illustration, the Sony FF SDM[4] topology [17] is included as the front-end coder. In practical SDM systems the sampling rate conversion is required to match the source signal to the SDM sampling rate, although for clarity that is not included in Fig. 30. The output stage is configured as a standard H bridge, where a method for circuit efficient ac-coupled interfacing with self dc restoration has been reported [4]. The design concept can use a resonant-mode power supply where the overall amplifier gain is modulated by scaling the H-bridge supply voltage. The use of a resonant supply locked in frequency and phase to the SDM clock can allow the H bridge to commutate during zero voltage transitions and thus reduce both switching loss and EMC interference significantly.

7.1.1 One-Sample SDM Look-Ahead

Theoretically the Sony FF SDM coder can yield more than sufficient SNR for power amplifier applications, especially in the context of the additional signal distortion inherent in power switches and also power supply noise. However, a weakness of the standard Sony topology is stability, especially under high levels of input signal,

[4]Sony FF SDM is a proprietary coding algorithm using local feedforward and local feedback paths; see [16] for a description.

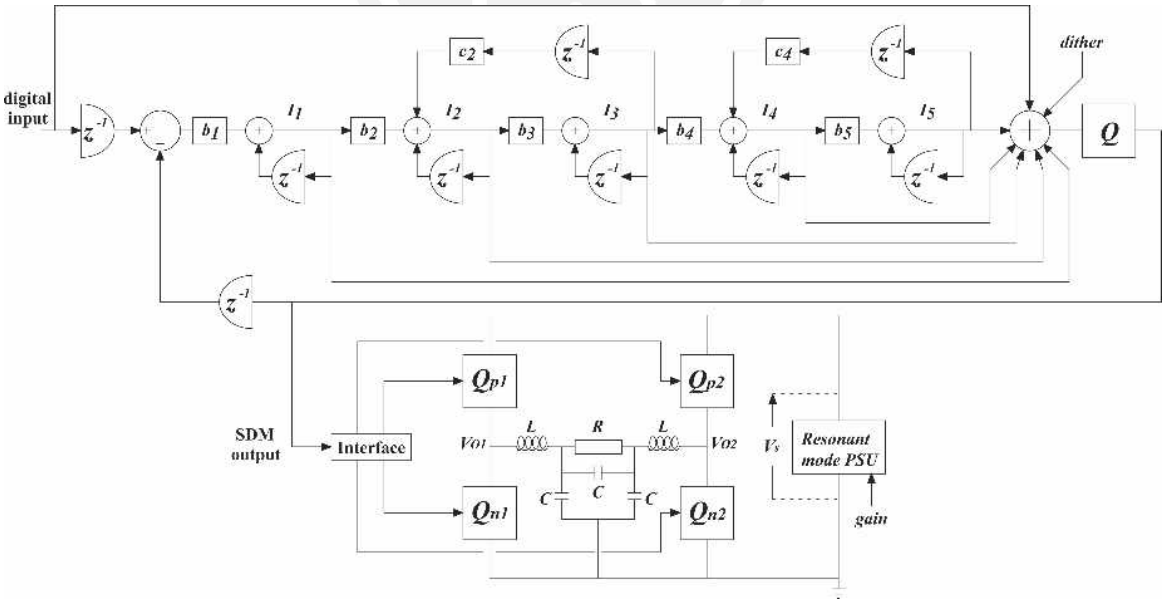


Fig. 30. Sony FF SDM with Type 2 bridge output stage.

which ultimately limits the maximum achievable modulation index. In this section a modified algorithm is explored based on the look-ahead principle [18] and an energy-balancing comparator. It is shown that with modest look-ahead significant gains in stability robustness can be achieved together with a low and constant coding latency. A further extension compresses under high-level excitation a subset of the state variables and is shown to allow stable operation up a modulation depth of unity. To demonstrate the probability of instability of the standard Sony FF SDM, results are derived using an earlier reported coding scheme [19], which stabilized the loop using a step-back-in-time procedure. A virtue of this method is that the step-back activity is an inverse measure of stability, that is, the more robust the loop, the lower the step-back activity. The simulation used the standard Super Audio CD (SACD) [3], [17] sampling rate f_{DSD} of 2.8224 MHz together with a 1-kHz input signal of amplitude 0.5. The corresponding SDM output spectrum [including a 24-bit, 88.2-kHz linear pulse-code modulation (LPCM) reference spectrum], quantizer input, and related amplitude histogram plots are shown in Fig. 31 with individual step-back activity events indicated by asterisks in Fig. 31(b). The step-back activity is relatively frequent, and if the input is increased further in level, catastrophic failure occurs.

In the standard feedback SDM algorithm a simple threshold decision is made on sample n as to whether the corresponding output sdm(n) is 1 or −1. Using Matlab notation, a typical threshold decision statement takes the following form:

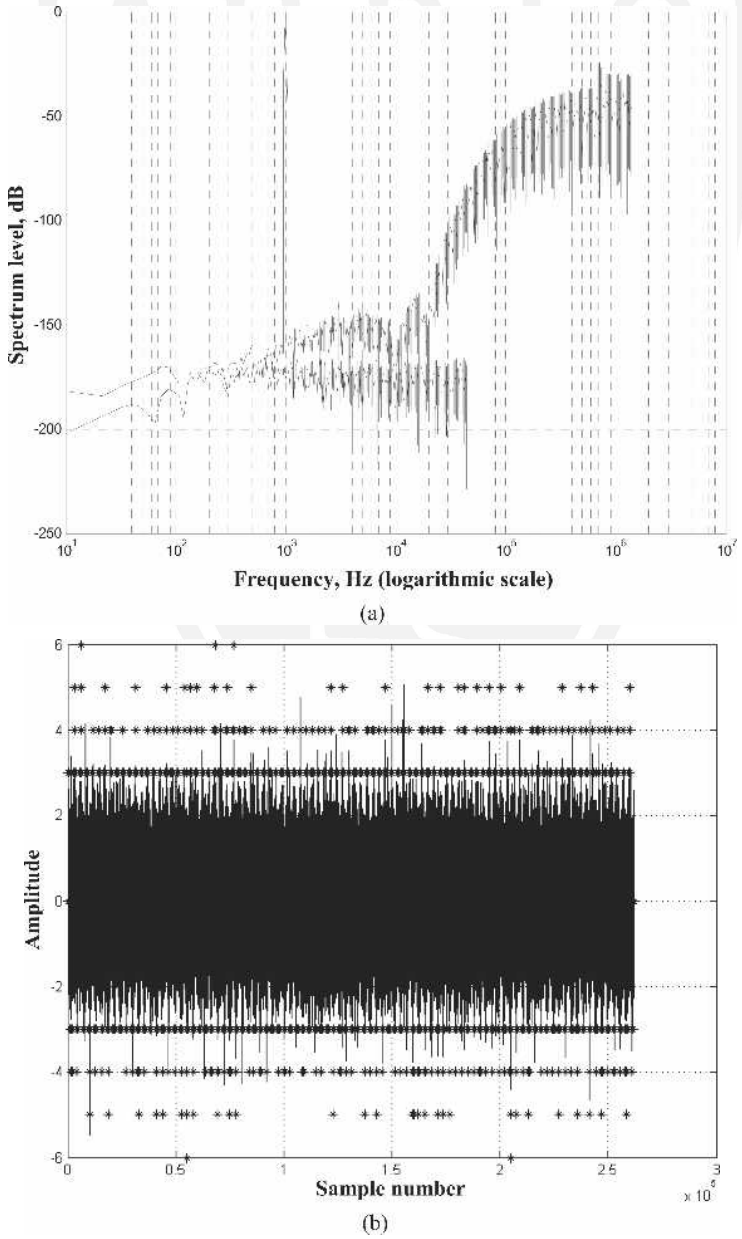


Fig. 31. Standard Sony FF SDM with step-back-in-time correction; input 0.50. (a) SDM output and 24 bit @ 88.2 kHz LPCM reference spectra. (b) Quantizer input with step back. (c) Quantizer input histogram.

```
% calculate input to quantizer
ss(n) = sum(I(1:order))+ax(n)+0.35*(rand(1,1)-.5);
% form SDM output using 2-level quantization
sdm(n) = sign(ss(n));
```

The first statement calculates the quantizer input ss(n) which includes a summation of the integrator outputs I(1:order), as shown in Fig. 30. It also includes RPDF dither, with a peak-to-peak amplitude range of 0.35 and a sampled input signal $ax(n)$. The second statement then performs a binary threshold comparison about a level of zero and generates the required SDM binary code.

A modification of the standard Sony FF SDM loop was investigated, where initially a one-sample look-ahead was implemented. In this scheme at sample n, the two pathways corresponding to sdm(n) = 1 and sdm(n) = −1 were computed and the corresponding states for sample n + 1 evaluated for the two options. The look-ahead algorithm is summarized as follows: Taking the present integrator states as I(1:5), the two possible state updates are calculated as $I1x, \ldots, I5x$ for sdm(n) = 1 and $I1y, \ldots, I5y$ for sdm(n) = −1, where, in Matlab notation,

```
% integrator update for sdm(n) = 1
I1x = I(1)+ax(n)-1;
I2x = I(2)+b2*I1x+c2*I(3);
I3x = I(3)+b3*I2x;
I4x = I(4)+b4*I3x+c4*I(5);
I5x = I(5)+b5*I4x;

% integrator update for sdm(n) = -1
I1y = I(1)+ax(n)+1;
I2y = I(2)+b2*I1y+c2*I(3);
I3y = I(3)+b3*I2y;
I4y = I(4)+b4*I3y+c4*I(5);
I5y = I(5)+b5*I4y;
```

Next, for each pathway signals PX and PY are calculated, representing the total energy of the five integrators states, although for integrator 5 this must include the summation of integrators 1 to 4 to represent the mandatory feedforward paths shown in Fig. 30 of the fifth-order Sony FF encoder, namely,

```
% look-ahead decision based on energy balance
```

$$PX = (I1x + I2x + I3x + I4x + I5x)^2 + I1x^2 + I2x^2 + I3x^2 + I4x^2 \quad (40)$$

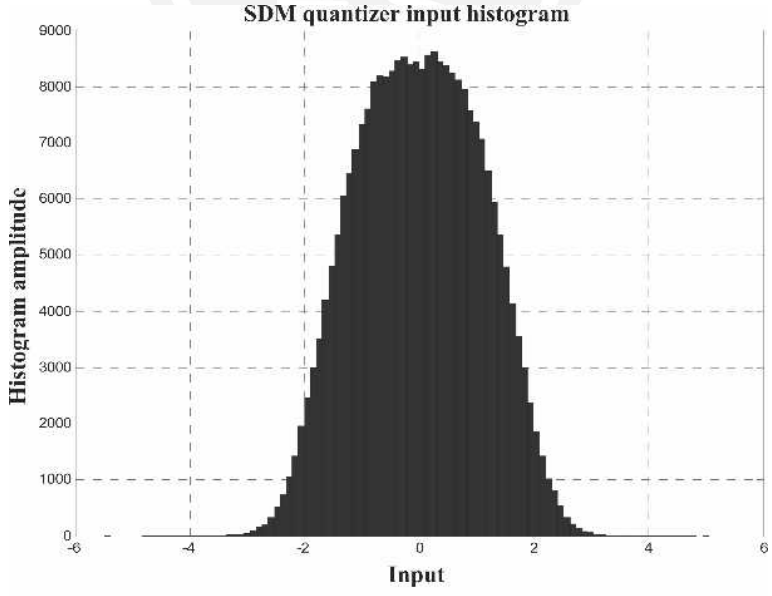
$$PY = (I1y + I2y + I3y + I4y + I5y)^2 + I1y^2 + I2y^2 + I3y^2 + I4y^2. \quad (41)$$

Having calculated the state energy estimates corresponding to the two pathways, an energy balance that includes dither rd(n) is used to select the actual SDM output sdm(n) for sample n,

$$\text{sdm}(n) = \text{sign}(\text{rd}(n) + PY - PX). \quad (42)$$

The loop integrators are then updated using the SDM output calculated by Eq. (42) in preparation for the next computational cycle. In addition, to help control stability using this modified algorithm, the loop (in this example) retained a step-back-in-time procedure [19] instigated by a simplified amplitude threshold comparator operating on the quantizer input.

A similar set of simulations was performed as for the standard Sony-FF SDM but using elevated peak input levels of 0.70 and 0.73. The corresponding results are shown in Figs. 32 and 33. Close inspection of the two output spectra reveals a slight overall reduction in SNR although the spectral shapes are almost identical. However, the quantizer input waveforms, and in particular the step-back activity, are changed. Although the input level has been



(c)

Fig. 31. *Continued*

increased from 0.5 to 0.7, activity is actually reduced with only two step-back incidents recorded in a window of 2^{18} samples. Increasing the input level to 0.73 caused the step-back activity to increase further, as might be anticipated for such a high-level signal, although it still falls within acceptable bounds and allowed the loop to remain stable. Consequently the energy-balancing equation with one-sample look-ahead achieves more robust SDM coding, especially with higher level input signals, a characteristic better matched to digital power amplifier applications.

7.1.2 Two-Sample SDM Look-Ahead

The look-ahead procedure was then extended from one to two samples to ascertain whether a further coding advantage is possible. The algorithm first calculated the state variables for a one-sample look-ahead for both a one- and a zero-output decision, as described in Section 7.1.1. Then from these two decisions two further sets of state variables were calculated, giving four sets for paths [1 0], [1 1], [0 0], and [0 1], as follows:

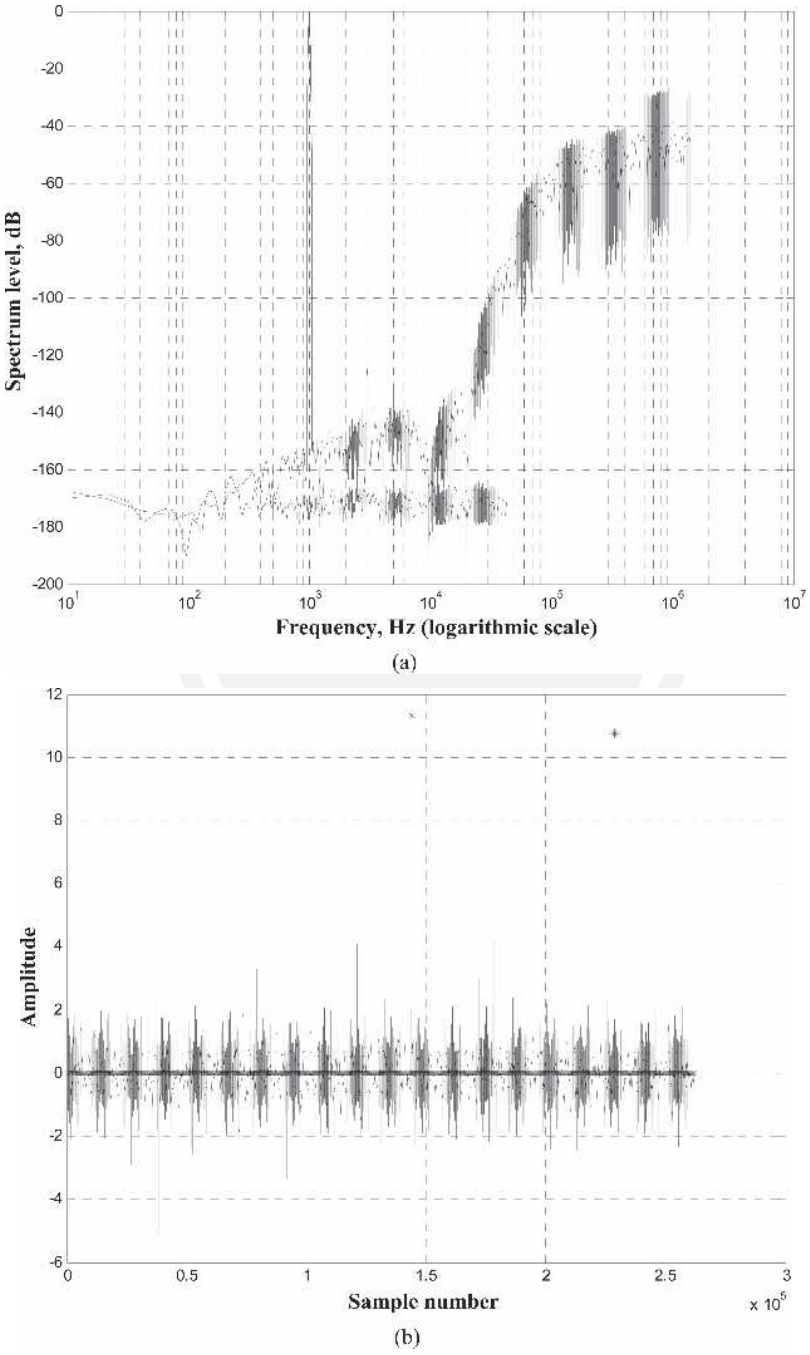


Fig. 32. Sony FF SDM with one-sample look-ahead; input 0.70. (a) SDM output and 24 bit @ 88.2 kHz LPCM reference spectra. (b) Quantizer input with step back. (c) Quantizer input histogram.

{I1×0, I2×0, I3×0, I4×0, I5×0} for path [1 0]

{I1×1, I2×1, I3×1, I4×1, I5×1} for path [1 1]

{I1y0, I2y0, I3y0, I4y0, I5y0} for path [0 0]

{I1y1, I2y1, I3y1, I4y1, I5y1} for path [0 1].

Energy estimates $mx0$, $mx1$, $my0$, and $my1$ were then determined for each of the four sets of state variables,

$$mx0 = (I1x0 + I2x0 + I3x0 + I4x0 + I5x0)^2 + I1x0^2 + I2x0^2 + I3x0^2 + I4x0^2 \quad (43)$$

$$mx1 = (I1x1 + I2x1 + I3x1 + I4x1 + I5x1)^2 + I1x1^2 + I2x1^2 + I3x1^2 + I4x1^2 \quad (44)$$

$$my0 = (I1y0 + I2y0 + I3y0 + I4y0 + I5y0)^2 + I1y0^2 + I2y0^2 + I3y0^2 + I4y0^2 \quad (45)$$

$$my1 = (I1y1 + I2y1 + I3y1 + I4y1 + I5y1)^2 + I1y1^2 + I2y1^2 + I3y1^2 + I4y1^2. \quad (46)$$

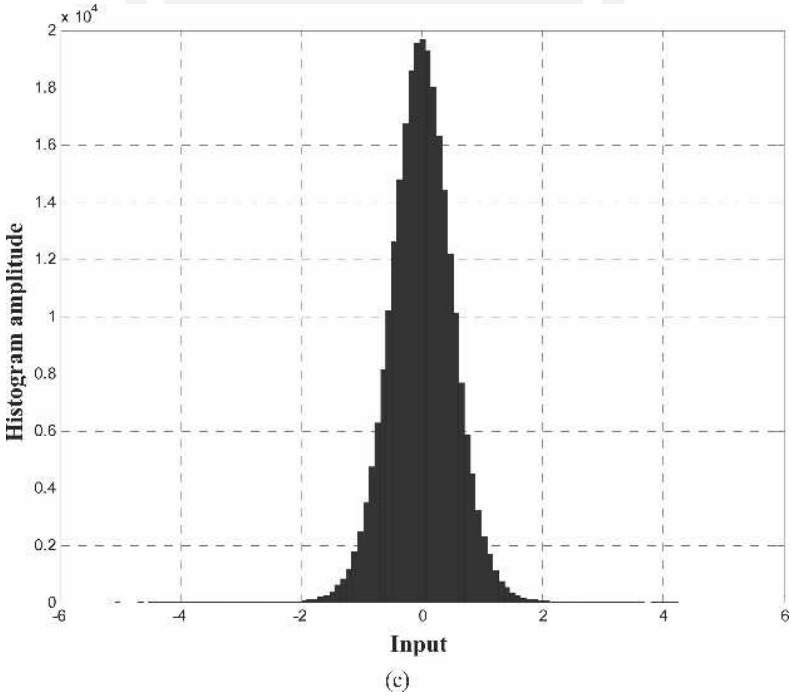
Note that x corresponds to a +1 output and y to a −1 on the first sample look-ahead. Finally, the SDM output decision sdm(n) was made based on the four look-ahead energy states by extracting initially the minimum of [$my0$ $my1$] and the minimum of [$mx0$ $mx1$] and then forming an energy difference equation,

$$\mathrm{sdm}(n) = \mathrm{sign}(rd(n) + \min([my0 \quad my1]) - \min([mx0 \quad mx1])). \quad (47)$$

Note that in Eq. (47) a dither source rd(n) is again included in the decision process. To show the complete procedure, a Matlab SDM simulation program without step-back correction for the two-sample look-ahead coder is presented in the Appendix. Simulation results confirm further improvement, such that with a peak input signal amplitude of 0.75 no step-back events were recorded. Also the simulation (using the code in the Appendix) without step-back correction remained stable over 2^{20} samples. This was repeated several times without problem other than mild nonlinear distortion similar to that displayed in Figs. 32(a) and 33(a). However, this simulation, although stable, proved to be close to the overload threshold such that increasing the peak input signal to 0.77 caused failure.

7.1.3 Dynamic State-Variable Compression with Two-Sample Look-Ahead SDM

A further modification is to incorporate dynamic compression of the state variables to enable two-sample look-ahead SDM to remain stable up to the maximum modulation index of unity. The method also removes the need for step-back correction so latency is no longer variable. Loop activity is monitored by observing the signal ss(n) described in Section 7.1.1, which in non-look-ahead SDM forms the input to the comparator Q shown in Fig. 30. If the magnitude of ss(n) exceeds a predetermined threshold, then the accumulated outputs of selected integrators $I_1(n)$, $I_2(n)$, . . . , $I_5(n)$ are replaced by compressed values $\beta_1 I_1(n)$, $\beta_2 I_2(n)$, . . . , $\beta_n I_5(n)$, where β_1, . . . , β_5 are the five integrator compression coefficients. The threshold level was determined by experiment but was set to be greater than the maximum signal normally encountered when the two-sample look-ahead loop was operating with an input of 0.75, that is, just within its stable regime as described in Section 7.1.2. This ensured that noise-shaping performance remained undisturbed for input signals up to a level of at least 0.75. Two variants of dynamic compression were tested with similar results as follows: Define T_h



(c)

Fig. 32. *Continued*

as the threshold level and D_p as damping. The compression algorithm then takes the generic form of a conditional loop. If abs(ss(n)) > T_h, then

$$I_1(n) \Rightarrow \beta_1 I_1(n), \quad I_2(n) \Rightarrow \beta_2 I_2(n), \quad I_3(n) \Rightarrow \beta_3 I_3(n),$$
$$I_4(n) \Rightarrow \beta_4 I_4(n), \quad \text{and} \quad I_5(n) \Rightarrow \beta_5 I_5(n)$$

else if abs(ss(n)) $\leq T_h$, then no change to the integrator outputs.

Variant 1: fixed compression
$\{\beta_1 = \beta_2 = 1$ and $\beta_3 = \beta_4 = \beta_5 = 0.5\}$

Variant 2: variable compression
$\{\beta_1 = \beta_2 = 1$ and $\beta_3 = \beta_4 = \beta_5 = f(\text{ss}(n), T_h, D_p)\}$

where the variable compression function is defined,

$$f(\text{ss}(n), T_h, D_p) = e^{-[|\text{ss}(n)|-T_h]D_p}.$$

Typical parameters found by experiment are $T_h = 20$ and $D_p = 0.1$. In the first variant a simple fixed substitution is made for the integrator outputs when the threshold is exceeded whereas in the second variant the degree of compression is progressive, being controlled by an exponential law. These modified integrator output signals are retained and carried forward to the next step in the loop. If on the next cycle ss(n) still has a magnitude above the threshold, then the compression procedure is again implemented. This repeats until ss(n) falls within the detection window,

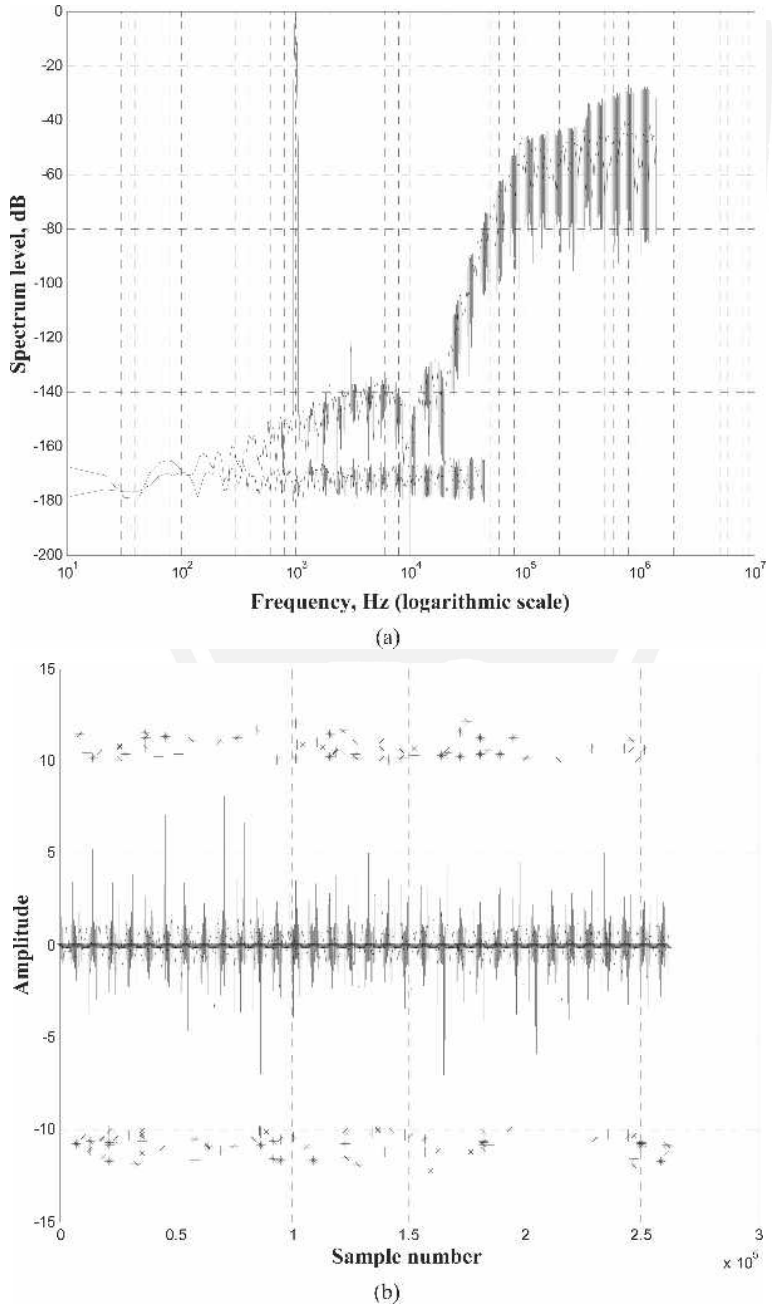
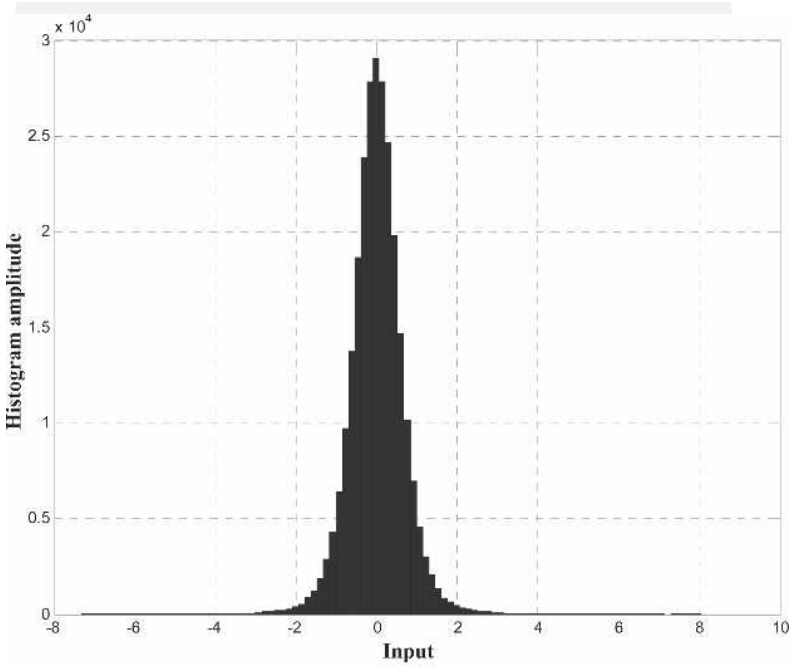


Fig. 33. Sony FF SDM with one-sample look-ahead; input 0.73. (a) Output spectrum. (b) Quantizer input with step back. (c) Quantizer input histogram.

whereupon normal loop behavior is resumed. Critical to the method is that the compression rate of the integrator outputs rises progressively with the magnitude of ss(n) and thus the input level. Also when the process is in a near permanent state of compression, then the first two integrator outputs remain unmodified while the other three are attenuated each loop cycle. Thus the loop tends to second-order behavior, which is known to be unconditionally stable, even with signals up to a modulation depth of unity. The dynamic modification of state variables is included in the SDM program presented in the Appendix. To demonstrate performance, the program was used to compute two example output spectra, shown in Fig. 34 for a 1-kHz input signal of amplitudes 0.78 and 1.0, respectively, where considering the high-input levels excellent noise shaping is revealed.

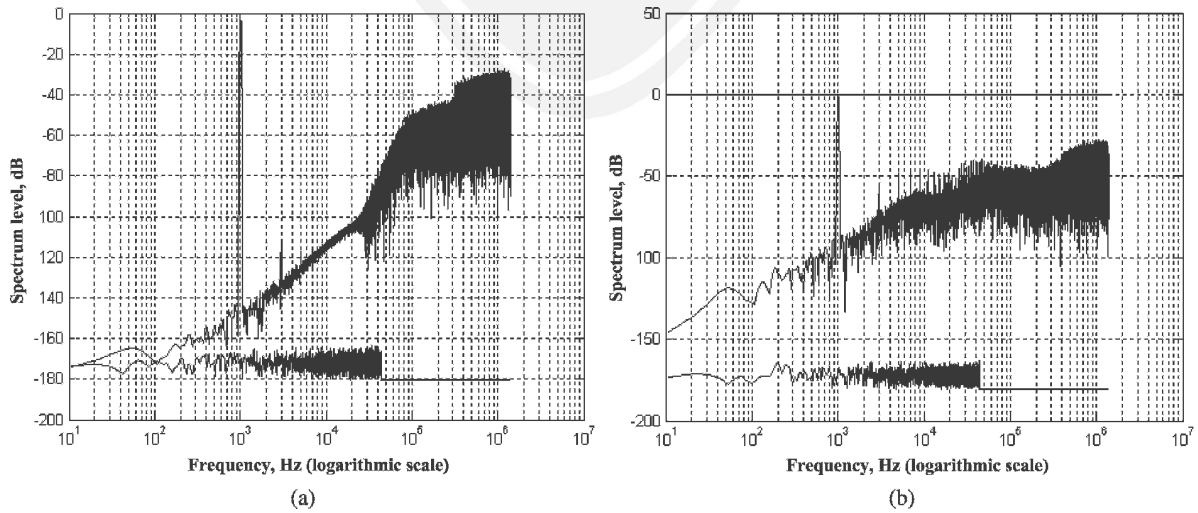
7.2 Review of SDM Power Amplifier Output-Stage Topologies

To conclude this section on switching amplifiers using SDM, three variants of an output stage topology [4] are discussed which reveal that a square-wave output is not mandatory and that reduced switching losses are possible compared to the standard H bridge with a constant voltage supply. The variants designated Types 1 to 3 offer progressive performance improvements in a number of critical areas, with Types 2 and 3 exploiting resonant-mode power supply techniques. The Type 1 stage uses an H-



(c)

Fig. 33. *Continued*



(a) (b)

Fig. 34. Sony FF with look-ahead and dynamic control of state variables. (a) Output spectrum; 0.78 input, 1 kHz. (b) Output spectrum; 1.00 input, 1 kHz.

bridge configuration powered by a constant-voltage power supply where this supply voltage determines the gain of the power amplifier, as changing the supply voltage directly modulates the amplitude of the output waveform [4]. However, Types 2 and 3 are new schemes that employ a resonant-mode power supply to produce a pure sinusoidal output superimposed on a constant-voltage component such that a raised-cosine repetitive waveform V_S of frequency f_{DSD} Hz is presented to the H-bridge stage, where

$$V_S = \text{gain}\left[\frac{1+\cos(2\pi f_{DSD}t)}{2}\right]. \quad (48)$$

The parameter "gain" defines the peak amplitude of the supply voltage and therefore determines the gain of the amplifier. Eq. (48) describes a voltage that swings between zero and "gain" volt, that is, in synchronism with the SDM sample clock. Also, V_S is phase locked so that the zero voltage instants are aligned precisely to the SDM sampling instants, thus virtually eliminating switching losses as all power transistors now switch at zero voltage. The Type 2 output stage is shown in Fig. 30, where the power supply uses the raised-cosine power supply V_S. Consequently the rectangular output pulses of the Type 1 amplifier are replaced with raised-cosine pulses whereby the output pulse stream is effectively prefiltered and has lower spectral content above the SDM sampling rate.

On first encounter it may appear that modifying the pulse waveform will introduce nonlinear distortion. However, this does not occur provided the modified output symbol shape is the same for all pulses in the data stream and where any further waveform changes resulting, for example, from low-pass filtering only affect the sample ensemble as a whole and do not give rise to intersample differences or pulse-sequence-dependent memory effects. To confirm this observation, simulations were performed for both Type 1 and Type 2 amplifier configurations. Normalized time-domain output waveforms are illustrated in Fig. 35 with the corresponding output spectra shown in Fig. 36. Using oversampling techniques, nonrectangular output pulses can be accommodated and output spectra calculated for frequencies in excess of the SDM sampling rate to show how using raised-cosine pulses reduces high-frequency content. Similar spectra to those presented in Figs. 32(a) and 33(a) can be observed with no evidence of additional in-band distortion resulting from waveform

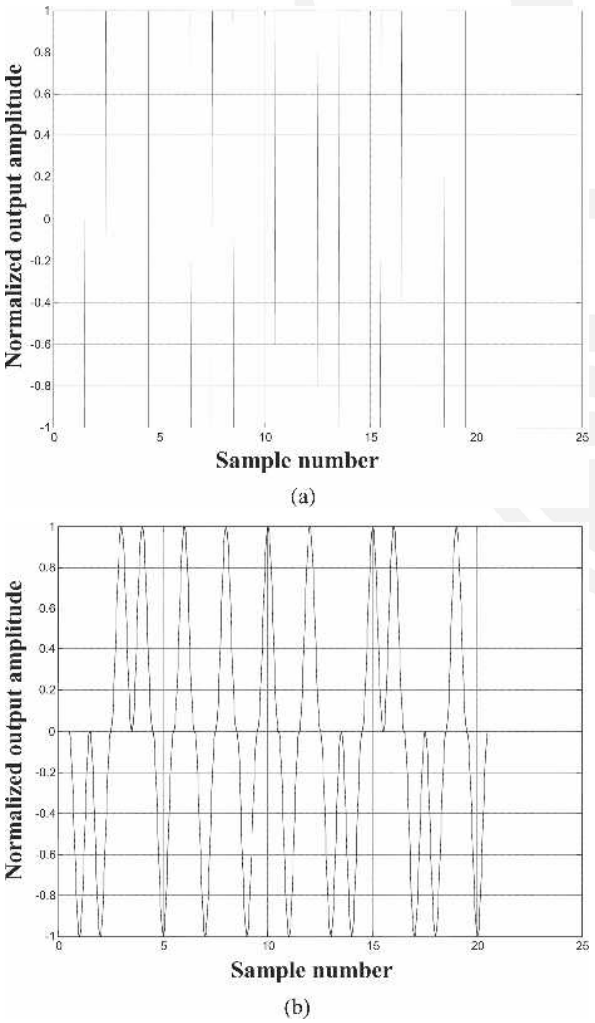


Fig. 35. SDM time-domain outputs. (a) Type 1. (b) Type 2.

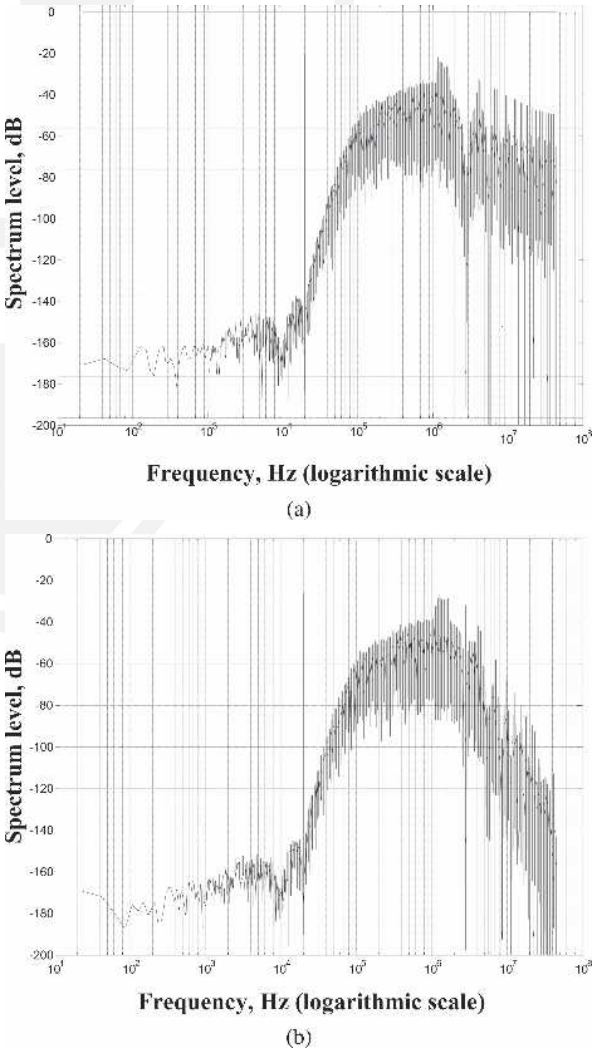


Fig. 36. SDM frequency-domain outputs. (a) Type 1. (b) Type 2.

modification. However, it is evident that significant attenuation of high-frequency components produced by the Type 2 amplifier output spectrum has been achieved. In these simulations a computation vector length of 2^{17} was elected and an additional oversampling factor of 32 applied to the SDM output code in order to accommodate the raised-cosine pulse shape.

Although the Type 2 amplifier topology addresses the problem of switching loss and to some extent alleviates the problem of EMC, it has the disadvantage in that the ratio of the low-pass-filtered output signal amplitude to the peak SDM output voltage is relatively poor compared to a PWM amplifier. This is exacerbated by the fact that even when a sequence of all-1 or all-0 pulses is generated, the differential output signal of the bridge stage always returns to zero between samples, as shown in Fig. 35(b), thus lowering the short-term average of the waveform. To overcome this deficiency a further modification to the amplifier is made and shown conceptually in Fig. 37. A similar raised-cosine power-supply voltage is used, but an additional constant-amplitude voltage source is introduced with its amplitude set precisely to that of the peak value of the raised-cosine waveform. As with the Type 1 and 2 amplifiers these waveforms are controlled by the parameter "gain" as the pulse amplitude modulation method is retained for gain control. Also in addition is a switch that can select either the constant voltage supply or the dynamic power supply. The positions of this switch are defined in Fig. 37 as positions 1 and 2, respectively. Fig. 37 includes an illustration of the relative levels of both the static and the dynamic power supply output waveforms with respect to the control parameter "gain."

The operation of the Type 3 amplifier is as follows: Normally when a change from 1 to 0 or 0 to 1 occurs then switch position 2 is selected and the amplifier operates as a Type 2 amplifier. However, if a burst of all-1 or all-0 pulses occurs in the SDM data stream then the switch changes to position 1 and the amplifier reverts effectively to a Type 1 configuration. This operation then forces the differential output signal of the H bridge to remain constant throughout the period of the burst. An example output sequence is shown in Fig. 38. The effect of this process is to retain constant-amplitude output pulses during a burst of like-valued SDM data. But where a data transition occurs then instead of a rectangular output, the waveform follows a smoothed path determined by the raised-cosine power supply. Consequently Type 3 is a hybrid of Type 1 and 2 amplifiers. Finally, to confirm that the output pulse processing used in the Type 3 amplifier does not introduce additional in-band distortion, a simulation was performed, and the output spectrum is shown in Fig. 39.

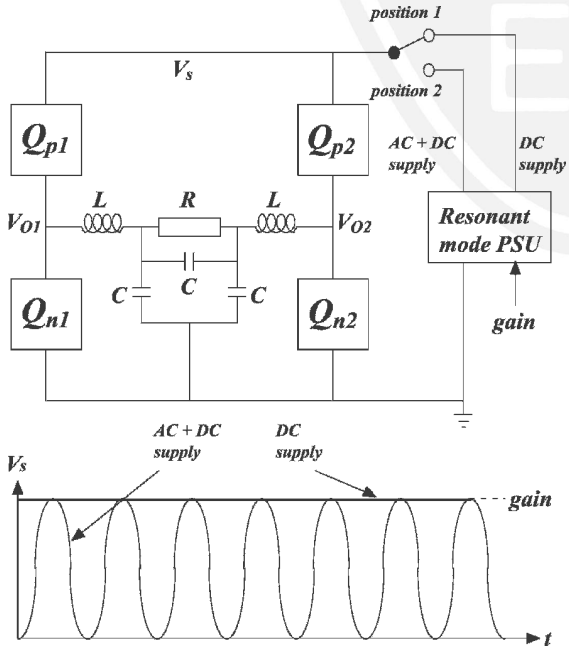


Fig. 37. Type 3 output stage powered by variable-output, resonant-mode power supply with ac and dc supply commutation.

8 CONCLUSIONS

This paper has endeavored to bring together topics on switching amplifiers that are relevant to both SDM and PWM power amplifier systems. It is evident that with the growing interest in SDM this type of modulation has ap-

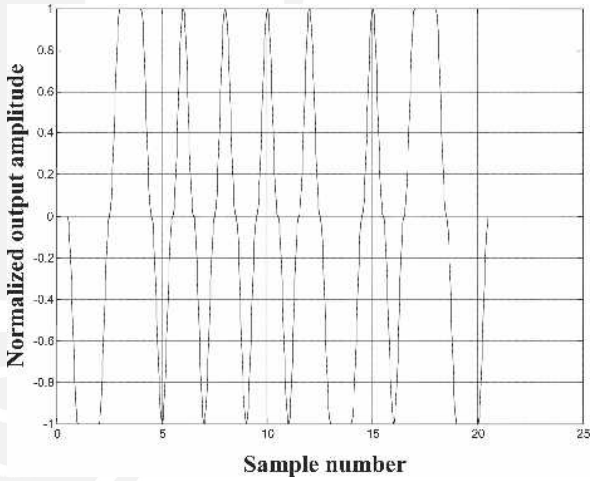


Fig. 38. Type 3 time-domain output.

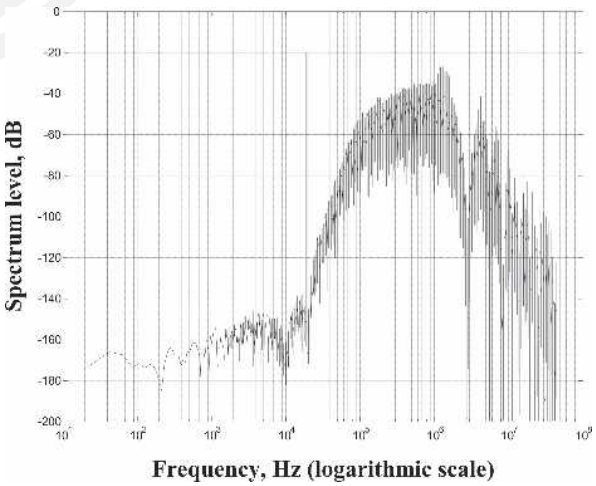


Fig. 39. Type 3 frequency-domain output.

plication, especially as it forms a natural digital system with a sampling rate that is compatible with available digital signal processors. Also, it offers a logical partner for DSD systems, especially as programmable supply voltages can be used to implement power-efficient gain control and thus eliminate the need for additional DSD signal processing.

Theoretical issues of natural sampling were presented, and the linkage between SDM and PWM was discussed for the case without quantization. Here it was shown that both systems may be modeled in terms of linear angle modulation, but the linear process of integration is repositioned from input to output to account for subtleties in waveform construction.

The application of negative feedback to improve PWM linearity was investigated. It was demonstrated by precision simulation that the presence of switching artifacts within the feedback loop causes high-frequency components to appear at the input of the PWM stage, degrading performance, and it was shown that these elements must be suppressed to render the PWM stage linear. As such the use of negative feedback cannot guarantee to reduce all distortion as its presence, without proper corrective procedures, can actually increase distortion as a function of loop gain. Three methods of suppressing switching components within the feedback loop were presented, namely, the NTF, a reference PWM stage within a feedback loop, and predictive correction using an open-loop PWM stage. All methods were shown capable of reducing switching distortion. The discussion of PWM was concluded by adapting the technique of predictive switching compensation for use with digital PWM applications, which also incorporated both feedback and feedforward error correction strategies.

In considering the SDM switched power amplifier it was recognized that robust SDM encoding is critical, especially with the requirement to encode high-amplitude signals. A look-ahead SDM coder was explored, which incorporated an energy-balancing binary decision threshold. This gave robust encryption up to a modulation index of about 0.75. However, by incorporating dynamic compression of the state variables which comes into operation only at high signal levels, stable operation up to the maximum modulation index of unity can be achieved. Simulation results showed that in this high-level operation region, excellent noise-shaping characteristics were retained whereas at lower levels there is no performance compromise. The technique is therefore ideal for power-amplifier applications. In this respect the use of SDM now competes with PWM in terms of peak signal handling. Finally, three configurations for SDM output stages were discussed. These revealed how the H-bridge topology can be adapted for SDM, and especially how switching losses could be lowered by incorporating both a resonant-mode power supply with a raised-cosine output voltage and a constant voltage supply, together with dynamic interpower supply switching related to the SDM code.

REFERENCES

[1] K. Nielsen, "High Fidelity PWM-Based Amplifier Concept for Active Loudspeaker Systems with Very Low Energy Consumption," *J. Audio Eng. Soc.,* vol. 45, pp. 554–570 (1997 July/Aug.).

[2] A. J. Magrath and M. B. Sandler, "Digital Power Amplification Using Sigma–Dalta Modulation and Bit Flipping," *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 45, pp. 476–487 (1997 June).

[3] J. Verbakel, L. van de Kerkhof, M. Maeda, and Y. Inazawa, "Super Audio CD Format," presented at the 104th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 46, p. 570 (1998 June), preprint 4705.

[4] F. M. Prime and M. O. J. Hawksford, "Digital Audio Power Amplifier for DSD Data Streams," presented at the 117th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 53, p. 118 (2005 Jan./Feb.), convention paper 6303.

[5] A. C. Floros and J. N. Mourjopoulos, "Analytical Derivation of Audio PWM Signals and Spectra," *J. Audio Eng. Soc.,* vol. 46, pp. 621–633 (1998 July/Aug.).

[6] J. E. Flood and M. J. Hawksford, "Exact Model for Deltamodulation Processes," *Proc. IEE,* vol. 118, pp. 1155–1161 (1971).

[7] M. J. Hawksford, "Unified Theory of Digital Modulation," *Proc. IEE,* vol. 121, pp. 109–115 (1974 Feb.).

[8] M. J. Hawksford, "Application of Delta-Modulation to Television Systems," Ph.D. Thesis, University of Aston, Birmingham (1972).

[9] M. O. J. Hawksford, "Time-Quantized Frequency Modulation, Time-Domain Dither, Dispersive Codes, and Parametrically Controlled Noise Shaping in SDM," *J. Audio Eng. Soc.,* vol. 52, pp. 587–617 (2004 June).

[10] M. O. J. Hawksford, "Dynamic Model-Based Linearization of Quantized Pulse-Width Modulation for Applications in Digital-to-Analog Conversion and Digital Power Amplifier Systems," *J. Audio Eng. Soc.,* vol. 40, pp. 235–252 (1992 Apr.).

[11] M. O. J. Hawksford, "Linearization of Multi-Level, Multi-Width Digital PWM with Applications in Digital-to-Analog Conversion," *J. Audio Eng. Soc.,* vol. 43, pp. 787–798 (1995 Oct.).

[12] D. J. Naus, E. C. Dijkmans, E. F. Stikvoort, A. J. McKnight, D. J. Holland, and W. Bradinal, "A CMOS Stereo 16-bit Converter for Digital Audio," *IEEE J. Solid-State Circuits,* vol. SC-22 (1987 June).

[13] R. H. Small, "Constant-Voltage Crossover Network Design," *J. Audio Eng. Soc.,* vol. 19, pp. 12–19 (1971 Jan.).

[14] W. M. Leach Jr., "Loudspeaker Driver Phase Response: The Neglected Factor in Crossover Network Design," *J. Audio Eng. Soc.,* vol. 28, pp. 410–421 (1980 June).

[15] M. O. J. Hawksford, "System Measurement and Identification Using Pseudorandom Filtered Noise and Music Sequences," *J. Audio Eng. Soc.,* vol. 52, pp. 275–296 (2005 Apr.).

[16] L. Risbo, "Σ–Δ Modulators—Stability Analysis and Optimization," Ph.D. Thesis, Technical University of Denmark (1994 June).

[17] H. Takahashi and A. Nishio, "Investigation of Practical 1-bit Delta–Sigma Conversion for Professional

Audio Applications," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 49, p. 544 (2001 June), convention paper 5392.

[18] J. A. S. Angus, "Tree Based Lookahead Sigma Delta Modulators," presented at the 114th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 51, p. 435 (2003 May), convention paper 5825.

[19] M. O. J. Hawksford, "Parametrically Controlled Noise Shaping in Variable State-Step-Back Pseudo-Trellis SDM," *Proc. IEE—VIS. Image Signal Process.,* vol. 152, pp. 87–96 (2005 Feb.).

APPENDIX
SONY FF MATLAB ENCODER WITH TWO-SAMPLE LOOK-AHEAD WITH IMPROVED STABILITY

```
% 5th-order Sony ff with 2-sample look-ahead configured for digital power amplifier
% AESDM
% 31.12.05
% 2-sample look-ahead with minimum energy state detection
% dynamic state variable compression to improve stability
% based on 5th order Sony FF SDM
% fsdm: SDM sampling rate
% fin1, fin2: input frequency of sinusoidal components
% a1, a2: input amplitude of sinusoidal components
% input quantized to "bit" resolution reference Nyquist sampling, then oversampled
% choice of output symbol, sigma sets samples per symbol
% NN-tap FIR output filter
close; home; clear; colordef white
fprintf('5th-order Sony FF with look-ahead configured for digital power amplification with output symbol options\n')

% set output pulse symbol shape
group = 3;
while group>2
fprintf('\nSelect type of output pulse symbol:\n')
fprintf('group = 0 standard rectangular symbol\n')
fprintf('group = 1 raised-cosine symbol\n')
fprintf('group = 2 raised-cosine symbol with hold applied between pulse groups\n')
group = input('Pulse symbol type: ');
end

% set input signal amplitude and frequency (2 sinewaves)
a1 = .4; a2 = .4; % input signal level
fin1 = 19000; fin2 = 2000; % input frequencies

% set data
fprintf('\nLoad input data\n')
sigma = 2^5; % samples per symbol of SDM output code (suggest 32)
NN = 32; % number of taps in output filter (32)
over = 64; % SDM oversample ratio above Nyquist (64 for DSD)
order = 5; % SDM order (5 for Sony FF)
L = 2^18; % computational vector length
bit = 24; % input signal quantization (24 bit)
nyquist = 44100; % Nyquist sampling rate (44100 Hz)
ditheramp = .35; % select dither amplitude (0.35)
thresh = 20; % dynamic state variables threshold (20)
damp = 10; % dynamic state variables damping (10)

% calculate constants, Blackman window and dither
fprintf('Calculate constants, window and dither\n\n')
fsdm = over*nyquist;
Lx = L/over; Ls = L*sigma;
f0 = nyquist/Lx;
w1 = round(fin1/f0)*2*pi/Lx; w2 = round(fin2/f0)*2*pi/Lx;
dither = 2*ditheramp*(rand(1,L)-1);
bl = blackman(Ls)'; bl = bl/mean(bl); % calculated over Ls samples
sdb = zeros(1,L);

% quantize input at 44.1 kHz
axx = a1*sin(w1*(1:Lx))+a2*sin(w2*(1:Lx));
axx = round(axx*2^(bit-1)+rand(1,Lx)+rand(1,Lx)-1)/2^(bit-1);

% interpolate input to a sampling rate of fsdm
ax = zeros(1,L);
for x = 1:Lx
ax((x-1)*over+1:x*over) = over*[axx(x) zeros(1,over-1)];
end
clear axx
win = [ones(1,Lx/2) zeros(1,(L-Lx)/2)];
win = [win 0 win(L/2:-1:2)];
ax = real(ifft(fft(ax).*win));

% set SDM initial conditions and constants
sdb = zeros(1,L); aa = 2*ditheramp;
I = zeros(1,5); ss = sdb;

% run Sony FF SDM loop with look-ahead
fprintf('Run Sony FF SDM: 2-sample look-ahead + dynamic state variable modulation\n\n')
b1 = 1; b2 = .5; b3 = .25; b4 = .125; b5 = .0625;
c2 = -.001953125; c4 = -.03125;
% dither sequence
rd = aa*(rand(1,L)-rand(1,L));
for n = 2:L-1
% update integrators
I(1) = I(1)+ax(n-1)-sdb(n-1);
I(2) = I(2)+b2*I(1)+c2*I(3);
I(3) = I(3)+b3*I(2);
I(4) = I(4)+b4*I(3)+c4*I(5);
I(5) = I(5)+b5*I(4);
% look-ahead 1 sample D(n) = 1
I1x = I(1)+ax(n)-1;
I2x = I(2)+b2*I1x+c2*I(3);
I3x = I(3)+b3*I2x;
I4x = I(4)+b4*I3x+c4*I(5);
I5x = I(5)+b5*I4x;
```

```
% look-ahead 2 sample D(n)=1 D(n+1)=1
I1x1=I1x+ax(n+1)-1;
I2x1=I2x+b2*I1x+c2*I3x;
I3x1=I3x+b3*I2x;
I4x1=I4x+b4*I3x+c4*I5x;
I5x1=I5x+b5*I4x;
% look-ahead 2 sample D(n)=1 D(n+1)=-1
I1x0=I1x+ax(n+1)+1;
I2xo=I2x+b2*I1x+c2*I3x;
I3x0=I3x+b3*I2x;
I4x0=I4x+b4*I3x+c4*I5x;
I5x0=I5x+b5*I4x;
% look-ahead 1 sample D(n)=-1
I1y=I(1)+ax(n)+1;
I2y=I(2)+b2*I1y+c2*I(3);
I3y=I(3)+b3*I2y;
I4y=I(4)+b4*I3y+c4*I(5);
I5y=I(5)+b5*I4y;
% look-ahead 2 sample D(n)=1 D(n+1)=1
I1y1=I1y+ax(n+1)-1;
I2y1=I2y+b2*I1y+c2*I3y;
I3y1=I3y+b3*I2y;
I4y1=I4y+b4*I3y+c4*I5y;
I5y1=I5y+b5*I4y;
% look-ahead 2 sample D(n)=1 D(n+1)=-1
Ily0=I1y+ax(n+1)+1;
I2y0=I2y+b2*I1y+c2*I3y;
I3y0=I3y+b3*I2y;
I4y0=I4y+b4*I3y+c4*I5y;
I5y0=I5y+b5*I4y;
ss(n)=sum(I(1:5))*(sdb(n)+1)-sum(I(1:5))*(sdb(n)-1); %
calculate input to quantizer for sdb(n)=1 or -1
if abs(ss(n))>thresh % attenuate state variables when
threshold exceeded
%I(3:5)=[.5 .5 .5].*I(3:5);
I(3:5)=I(3:5)*exp((-abs(ss(n))+thresh)/damp);
end
mx0=((I1x0+I2x0+I3x0+I4x0+I5x0)^2+I1x0^2+I2x0^2
+I3x0^2+I4x0^2);
mx1=((I1x1+I2x1+I3x1+I4x1+I5x1)^2+I1x1^2+I2x1^2
+I3x1^2+I4x1^2);
my0=((I1y0+I2y0+I3y0+I4y0+I5y0)^2+I1y0^2+I2y0^2
+I3y0^2+I4y0^2);
my1=((I1y1+I2y1+I3y1+I4y1+I5y1)^2+I1y1^2+I2y1^2
+I3y1^2+I4y1^2);
sdb(n)=sign(rd(n)+min([my0 my1])-min([mx0 mx1]));
% 2 step look-ahead minimum energy
n=n+1; % increment sample value
end % SDM loop end ****************************
*******************
plot(ss,'k')
title('Signal ss(n) applied to standard quantizer')
ylabel('Amplitude')
xlabel('Time')
grid; pause; close

% shape output pulses for digital amplifier application
using oversampling by sigma
sdbrc=zeros(1,Ls);
if group>0
fprintf('Shape SDM output pulses: raised cosine\n')
symb=(1-cos(2*pi*(0:sigma-1)/sigma))/2;
else
fprintf('Shape SDM output pulses: rectangular pulse
shape\n')
symb=ones(1,sigma);
end
for x=1:L
ss=(x-1)*sigma+1;
sdbrc(ss:ss+sigma-1)=sdb(x)*symb(1:sigma);
end

% detect groups of 2 or more pulses and hold pulse
amplitude at maximum
if group==2
fprintf('Shape SDM output pulses with hold function for
pulse groups\n')
for x=1:L-1
ss=(x-1)*sigma+1;
if sdb(x)-sdb(x+1)==0
sdbrc(ss+.5*sigma:ss+1.5*sigma-1)=sdb(x)*ones(1,
sigma);
%plot(sdbrc(ss:ss+2*sigma)); pause
end; end; end

% FFT routine
fprintf('Calculate output spectrum\n')
sdbf=abs(fft(sdbrc(1:Ls).*bl));
sdbf(1:Ls/2-1)=20*log10(10^-10+2*sdbf(2:Ls/2)/Ls);

% plot spectrum SDM to fsdm/2
fprintf('Plot SDM spectrum\n')
semilogx(f0*(1:L/2-1),zeros(1,L/2-1),'k')
hold
semilogx(f0*(1:L/2-1),sdbf(1:L/2-1),'k')
title('SDM output spectrum to fsdm/2')
xlabel ('Frequency, Hz (logarithmic scale)')
ylabel('Spectrum level, dB')
grid; pause; close

% plot spectrum SDM full
fprintf('Plot SDM spectrum\n')
semilogx(f0*(1:Ls/2-1),zeros(1,Ls/2-1),'k')
hold
semilogx(f0*(1:Ls/2-1),sdbf(1:Ls/2-1),'k')
title('SDM output spectrum to sigma*fsdm/2')
xlabel('Frequency, Hz (logarithmic scale)')
ylabel('Spectrum level, dB')
grid; pause; close

% plot example output time domain
kk=20;
plot((1:kk*sigma)/sigma+.5,sdbrc(1:kk*sigma),'k')
title('SDM output sequence')
xlabel('Sample number')
ylabel('Normalized output amplitude')
grid; pause; close

% NN-tap filtered output time domain plot
kk=Lx;
sdbrcav=sdbrc(1:kk*sigma);
for x=1:NN-1
```

```
sdbrcav = sdbrcav+sdbrc(1+x*sigma:(kk+x)*sigma);
end
sdbrcav = sdbrcav/NN;
plot((1:kk*sigma)/sigma+.5,sdbrcav(1:kk*sigma),'k')
title('Normalized filtered output over NN DSD
samples')
xlabel('Sample number')
ylabel('Normalized output amplitude')
grid; pause; close

% end of program
fprintf('\nProgram terminated\n')
return
% **********************************
```

THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and he studied delta modulation and sigma-delta modulation (SDM) for color television applications. During this period he also invented a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, Colchester, UK, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems, with a strong emphasis on audio systems including signal processing and loudspeaker technology. Since 1982 his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. The first one (believed to be unique at the time) was developed in 1986 for a prototype system to be demonstrated at the Canon Research Centre and was sponsored by a research contract from Canon. Much of this work has appeared in *JAES,* together with a substantial number of contributions at AES conventions. He is a recipient of the AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," for the best contribution by an author of any age for *JAES,* volumes 45 and 46.

Dr. Hawksford's research has encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with special emphasis on SDM and its application to SACD technology. In addition, his research has included the linearization of PWM encoders, diffuse loudspeaker technology, array loudspeaker systems, and three-dimensional spatial audio and telepresence including scalable multichannel sound reproduction.

Dr. Hawsford is a chartered engineer and a fellow of the AES, IEE, and IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA).

4 Loudspeaker systems

4-1 System theory

4-1 APPLICATION OF THE GEOMETRIC THEORY OF DIFFRACTION (GTD) TO DIFFRACTION AT THE EDGES OF LOUDSPEAKER BAFFLES, Bews, R.M. and Hawksford, M.J., *JAES*, vol.34, no.10, pp.771-779, October 1986

4-10 REDUCTION OF LOUDSPEAKER POLAR RESPONSE ABERRATIONS THROUGH THE APPLICATION OF PSYCHOACOUSTIC ERROR CONCEALMENT, Rimell, A. and Hawksford, M.O.J., IEE Proceedings on Vision Image Signal Processing, vol. 145, no 1, Feb. 1998, pp 11-18
[Awarded "The Associates Premium Award" by the IEE to Dr Rimell]

4-18 INTRODUCTION TO DISTRIBUTED MODE LOUDSPEAKERS (DML) WITH FIRST-ORDER BEHAVIOURAL MODELLING, Harris, N. and Hawksford, M.O.J., IEE Proc.-Circuits Devices Systems, Vol. 147, No. 3, pp 153-157, June 2000

4-2 Current drive

4-23 DISTORTION REDUCTION IN MOVING-COIL LOUDSPEAKER SYSTEMS USING CURRENT-DRIVE TECHNOLOGY, Mills, P.G.L., Hawksford, M.O.J., *JAES*, vol.37, no.3, pp.129-148, March 1989

4-3 Crossover and equalization systems

4-43 EFFICIENT FILTER DESIGN FOR LOUDSPEAKER EQUALIZATION, Hawksford, M.O.J. and Greenfield, R., *JAES*, vol. 39, no. 10, pp 739-751, November 1991

4-56 ASYMMETRIC ALL-PASS CROSSOVER ALIGNMENTS, Hawksford, M.O.J., *JAES*, vol. 41, no. 3, pp 123-134, March 1993

4-68 ON THE DITHER PERFORMANCE OF HIGH-ORDER DIGITAL EQUALIZATION FOR LOUDSPEAKER SYSTEMS, Greenfield, R.G. and Hawksford, M.O.J., *JAES*, vol. 43, no. 11, pp 908-915, November 1995

4-76 DIGITAL SIGNAL PROCESSING TOOLS FOR LOUDSPEAKER EVALUATION AND DISCRETE-TIME CROSSOVER DESIGN, Hawksford, M.O.J., *JAES*, vol. 45, no. 1/2, pp 37-62, Jan/Feb 1997. [*Awarded AES Publication prize for the best paper from an author of any age, from JAES volumes 45 and 46.]*

4-102 MATLAB PROGRAM FOR LOUDSPEAKER EQUALIZATION AND CROSSOVER DESIGN, Hawksford, M.O.J., *JAES*, vol. 47, no. 9, pp 707-719, September 1999

4-4 Digital and array loudspeakers

4-116 SMART DIGITAL LOUDSPEAKER ARRAYS, Hawksford, M. O. J., *JAES*, vol. 51, no. 12, pp 1133-1162, December 2003

4-146 SPATIAL DISTRIBUTION OF DISTORTION AND SPECTRALLY-SHAPED QUANTIZATION NOISE IN DIGITAL MICRO-ARRAY LOUDSPEAKERS, Hawksford, M.O.J. *JAES*, vol. 55, no. 1/2, pp. 1-27, January/February 2007

PAPERS

Application of the Geometric Theory of Diffraction (GTD) to Diffraction at the Edges of Loudspeaker Baffles*

R. M. BEWS AND M. J. HAWKSFORD

Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

The response of a loudspeaker system employing a baffle is modified by diffraction at the baffle edges. Numerical solutions are derived using a model based on the geometric theory of diffraction to examine the major features of the diffraction process. From the model it is shown that a smoother frequency and step response result, provided the drivers are placed at unequal distances from the sides of small regularly shaped baffles or if small irregularly shaped baffles are used.

0 INTRODUCTION

Most moving-coil loudspeakers use a finite baffle. Consequently an analysis of the problems caused by diffraction at the baffle edges needs consideration. This phenomenon is significant, since amplitude response fluctuations in excess of ±5 dB can occur for particular baffle shapes.

In the late 1950s Olson [1] presented experimental data and qualitatively predicted various diffraction effects within loudspeaker systems using the geometric theory of diffraction (GTD), a theory originally proposed by Keller in 1952 [2]. This work involved the placement of multiple diffraction point sources at the diffraction edge, which resulted in secondary radiation, which subsequently interfered with the direct driver output, giving rise to frequency and phase reference irregularities. Olson concluded that an asymmetrical placement of drive units on a curved or irregularly shaped baffle offered improvements due to the randomization of the diffraction signal path lengths. Unfortunately the calculations of amplitude and phase of these diffracted rays at the baffle edge have received minimal attention. In the present paper we offer numerical solutions to this specific problem, where to the authors' knowledge a more formal and quantitative treatment has not been discussed in the context of loudspeaker systems.

* Manuscript received 1985 May 9; revised 1986 May 16.

1 DIFFRACTION MODEL

A model of acoustic diffraction at the loudspeaker baffle edges is developed, and for simplicity the following will be assumed.

1) There is only one rebated drive unit present on the baffle. Two or more can be considered by applying the law of superposition.

2) The driver acts as a point source, later to be extended to drive units with cones of finite size.

3) The baffle is constructed out of acoustically reflective material such as wood.

The development of the model commences by calculating the sound pressure on the baffle edge at E, produced by the driver located at S, as shown in Fig. 1. The acoustical reciprocity theorem [1, pp. 24–26] is then applied, which enables the interchange of points E and S. This procedure allows a virtual source to represent the diffraction edge, where its response takes full account of the baffle-edge geometry.

At a microscopic level, the loudspeaker baffle, at point E, appears as a wedge with a solid angle γ (Fig. 2). From [3] it is found that the ratio of the sound pressure produced by a point source at the apex of a wedge to that of the point source in free air is inversely proportional to 4π minus the solid angle of the wedge. Thus the sound pressure at E, P_E, is given by

$$P_E = \frac{4\pi}{4\pi - \gamma} P_{fs}|_r \qquad (1)$$

where $P_{fs}|_r$ is the sound pressure at a distance r, produced by a point source in free air, that is $(A/r)\exp(-jkr)$.

Using [3] and noting that the driver is placed on a flat plane of solid angle 2π, the sound pressure P_S produced by the driver will then be

$$P_S = 2P_{fs} . \quad (2)$$

If the baffle were to be infinite, the sound pressure at E would be $2P_{fs}|_r$. However, the baffle ends here and the actual sound pressure is given by Eq. (1). Consequently a change in the sound pressure occurs at the edge. For example, if $\gamma = \pi$, there will be a pressure drop from $2P_{fs}|_r$ to $1.333P_{fs}|_r$ at E.

2 DIFFRACTION SOURCES

To produce the pressure change, a point source with a suitable amplitude and phase will be placed at E. This point source will be known from now on as a diffraction point source, where essentially the GTD is being applied. At the edge,

$$M_E \exp(j\theta_E) = M_I \exp(j\theta_I) + M_D \exp(j\theta_D)$$

where

$M_E \exp(j\theta_E)$ = resultant sound pressure at E
$M_I \exp(j\theta_I)$ = incident sound pressure at E, produced by driver
$M_D \exp(j\theta_D)$ = sound pressure produced by diffraction source.

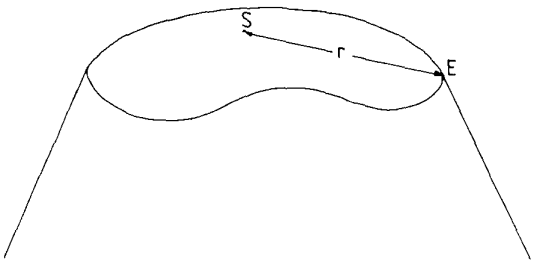


Fig. 1. Source locations on baffle.

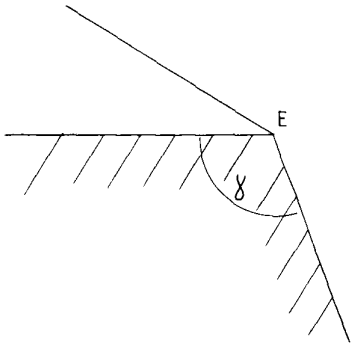


Fig. 2. Microscopic view of baffle at point E. γ—solid angle of wedge.

There is no phase difference between incident and diffracted waves, because a phase difference implies an energy loss at the edge, which cannot occur since the baffle is acoustically reflective. Most, if not all, commercially manufactured enclosures employ materials which are acoustically reflective. Hence $\theta_E = \theta_I = \theta_D$, and for convenience let them be zero. So

$$M_E = M_I + M_D . \quad (3)$$

If a numerical solution to this problem is to be sought, the infinite number of diffraction point sources required is impractical. As a result the total baffle edge will be quantized into N equally spaced sections, dx in length, which give rise to a finite number N of diffraction line sources. Now r becomes the average distance from the driver to a diffraction line source.

Substituting Eqs. (1) and (2) into Eq. (3) with M_D replaced by M_L as a diffraction line source is now being considered. Thus,

$$\left(\frac{4\pi}{4\pi - \gamma}\right) (\text{amplitude of } P_{fs}|_r) = 2(\text{amplitude of } P_{fs}|_r) + M_L$$

$$M_L = \left(\frac{4\pi}{4\pi - \gamma} - 2\right) (\text{amplitude of } P_{fs}|_r) .$$

Since the diffraction line source extends over a distance dx and the amplitude of the point source driver in free air falls as A/(radial distance), then

$$M_L = \left(\frac{4\pi}{4\pi - \gamma} - 2\right) \frac{dx}{2\pi r} A . \quad (4)$$

Provided dx is sufficiently small, it is acceptable to assume that the line source behaves as a point source. dx is chosen such that $(r_{max} - r_{min}) < r/1000$, since if the error $r_{max} - r_{min}$ is made even smaller, there is no detectable change in the amplitude and phase responses in the audio band from 20 Hz to 20 kHz.

3 ON-AXIS RESPONSE OF LOUDSPEAKER

3.1 Steady-State Response

Let the response of the driver at the point of observation be

$$\tilde{M}_p \exp(j\tilde{\theta}_p) .$$

where $\sim$ signifies on axis, and let the response of the kth diffraction line source at the point of observation be

$$\tilde{M}_k \exp(j\tilde{\theta}_k) .$$

The response on axis can then be calculated by considering the interference of all the outputs from the N

diffraction line sources and the driver.

The amplitude response on axis $M_{on}(\omega)$ is given by

$$M_{on}(\omega) = \left| \tilde{M}_p \exp(j\tilde{\theta}_p) + \sum_{k=1}^{N} \tilde{M}_k \exp(j\tilde{\theta}_k) \right|$$

$$= \left| \tilde{M}_p(\cos\tilde{\theta}_p + j\sin\tilde{\theta}_p) + \sum_{k=1}^{N} \tilde{M}_k(\cos\tilde{\theta}_k + j\sin\tilde{\theta}_k) \right|$$

$$= \sqrt{\left(\tilde{M}_p\cos\tilde{\theta}_p + \sum_{k=1}^{N}\tilde{M}_k\cos\tilde{\theta}_k\right)^2 + \left(\tilde{M}_p\sin\tilde{\theta}_p + \sum_{k=1}^{N}\tilde{M}_k\sin\tilde{\theta}_k\right)^2} . \quad (5)$$

The corresponding on-axis phase response $P_{on}(\omega)$ is

$$P_{on}(\omega) = \arctan\left(\frac{\tilde{M}_p\sin\tilde{\theta}_p + \sum_{k=1}^{N}\tilde{M}_k\sin\tilde{\theta}_k}{\tilde{M}_p\cos\tilde{\theta}_p + \sum_{k=1}^{N}\tilde{M}_k\cos\tilde{\theta}_k}\right) . \quad (6)$$

The diffraction line sources will in general be farther away from the observer than the driver. Denoting the observer-to-driver distance by OBD, then on axis the distance from a diffraction line source to the observer will be $\sqrt{OBD^2 + r_k^2}$ (Fig. 3). Using Eq. (4) and noting that the magnitude of the pressure is inversely proportional to the distance from the diffraction source, then at the observation point the magnitude of the response produced by the kth diffraction line source $\tilde{M}_k$ is

$$\tilde{M}_k = \frac{1}{\sqrt{OBD^2 + r_k^2}} \left(\frac{4\pi}{4\pi - \gamma} - 2\right) \frac{dx}{2\pi r_k} A .$$

If $\tilde{\theta}_p$ is set to zero, making the driver response the reference, the phase of the kth diffraction line source $\tilde{\theta}_k$ will be given by

$$\tilde{\theta}_k = \frac{-\omega[r_k + (\sqrt{OBD^2 + r_k^2} - OBD)]}{c}$$

where c is the velocity of sound.

The negative sign shows that the diffracted rays are delayed relative to the driver signal.

Using Eq. (2), the magnitude of the driver response at the observation point is given by

$$\tilde{M}_p = \frac{2A}{OBD} .$$

3.2 Step Response

Let the unit step be described by $S_p(t)$. The on-axis step response is then calculated by adding the driver response to appropriately delayed diffraction line source responses. Thus the on-axis step response is given by

$$T_{on}(t) = \tilde{M}_p S_p(t) + \sum_{k=1}^{N} \tilde{M}_k S_p(t - t_k) \quad (7)$$

where

$$S_p(\tau) = 1 \text{ if } \tau > 0, \qquad \text{otherwise } S_p(\tau) = 0$$

and t_k is the time delay between the kth diffraction line source signal and the driver signal. So

$$t_k = \frac{\tilde{\theta}_k}{\omega} = \frac{r_k + (\sqrt{OBD^2 + r_k^2} - OBD)}{c} .$$

$\tilde{\theta}_k$, $\tilde{M}_k$, and $\tilde{M}_p$ are defined in Sec. 3.1.

3.3 Steady-State and Step-Response Simulations

As an illustration the amplitude, phase, and step responses of a point source driver in the center of a 300-mm-radius circular baffle and a 600- by 600-mm square baffle have been calculated using Eqs. (5)–(7). The observer-to-driver distance OBD is 1.0 m, since most measurements are taken at this distance. Figs. 4–11 show the simulations using a DEC10 computer. All the responses are normalized to the free-air driver response, $A = 1$.

It can be seen that the amplitude and phase responses are much flatter using a square baffle instead of a circular one. This occurs because at the observer position the diffracted rays are coherent using the circular baffle,

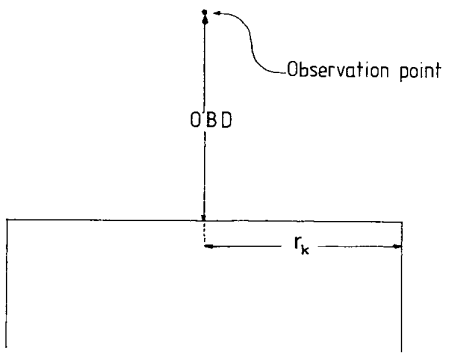


Fig. 3. On-axis geometry.

due to all the diffraction path lengths being equal, whereas for the square baffle, there is a spread of diffraction path lengths causing the diffracted rays to be less coherent. Consequently the circular baffle step response has a sharp downward transition after approximately 1 ms, when all the diffracted rays interfere with the driver signal at the same time. The downward transition in the step response using the square baffle is distributed more in time, because the diffracted rays interfere at slightly different times.

Figs. 4 and 8 can be compared with their corresponding measured responses in [1, p. 23], where there exists good experimental agreement. This suggests that the model is accurate and any slight deviations are probably due to measurement inaccuracies. The other baffle shapes described in [1] were not analyzed because of the computational complexity involved.

4 OFF-AXIS RESPONSE OF LOUDSPEAKER

When moving off axis with regard to the center of the drive unit, the path lengths of the diffracted rays

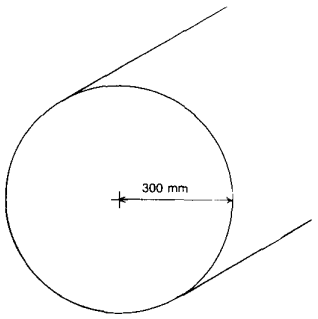


Fig. 7. Baffle shape for computer simulations of Figs. 4–6.

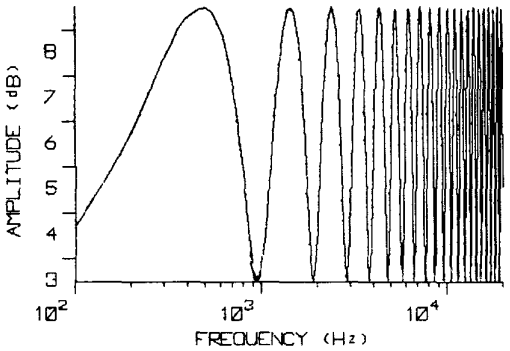


Fig. 4. On-axis amplitude response. Baffle radius 300 mm.

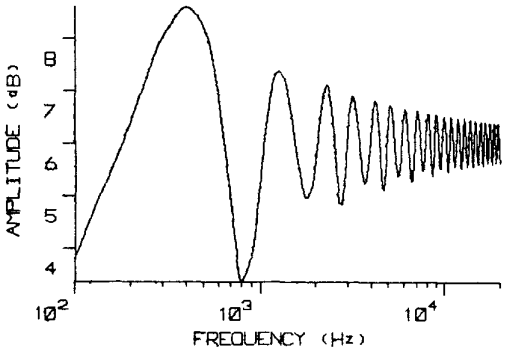


Fig. 8. On-axis amplitude response. Baffle size 600 by 600 mm.

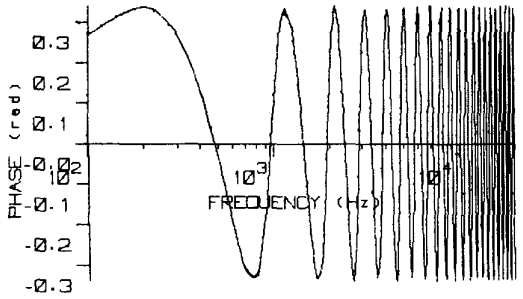


Fig. 5. On-axis phase response. Baffle radius 300 mm.

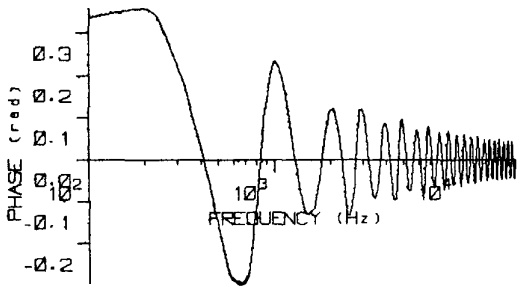


Fig. 9. On-axis phase response. Baffle size 600 by 600 mm.

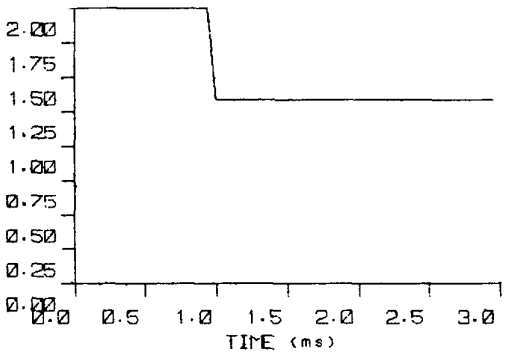


Fig. 6. On-axis step response. Baffle radius 300 mm.

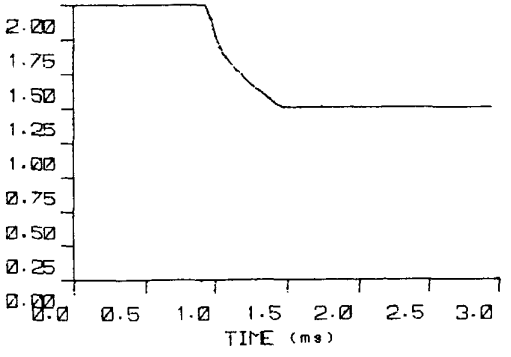


Fig. 10. On-axis step response. Baffle size 600 by 600 mm.

change. Consider the following situation in front of the baffle, as shown in Fig. 12.

Let the observer-to-driver distance be OBD, as before, the distance from the driver to the kth diffraction line source r_k, and the distance off axis OFD.

The phase difference θ_k between the driver response and the response of the kth diffraction line source becomes

$$\theta_k = \frac{-\omega[r_k + r_{OD} - \sqrt{OBD^2 + OFD^2}]}{c} \quad (8)$$

where the distance r_{OD} is shown in Fig. 12.

The corresponding time delay t_k is

$$t_k = \frac{\theta_k}{\omega} = \frac{r_k + r_{OD} - \sqrt{OBD^2 + OFD^2}}{c} . \quad (9)$$

The magnitude of the response produced by the kth diffraction line source M_k is given by

$$M_k = \frac{1}{r_{OD}} \left(\frac{4\pi}{4\pi - \gamma} - 2 \right) \frac{dx}{2\pi r_k} A \quad (10)$$

and the magnitude of the driver response M_p is

$$M_p = \frac{2A}{\sqrt{OBD^2 + OFD^2}} . \quad (11)$$

4.1 Steady-State and Step-Response Simulations

The steady-state and step responses are calculated using Eqs. (5)–(7). All these equations now use the values of θ_k, t_k, M_k, and M_p, as shown in Sec. 4.

The amplitude, phase, and step responses off axis have been computer simulated for a point source driver in the center of a 300-mm-radius circular baffle and a 600- by 6000-mm-square baffle (see Figs. 13–20). All the responses are normalized to the driver's free-air response, $A = 1$. The observer-to-driver distance OBD is set to 1.0 m and the off-axis angle [= arctan(OFD/OBD)] extends from −10° to 10°.

When off axis the diffracted rays are less coherent due to the broader distribution of path lengths. This produces flatter amplitude and phase responses and a less sharp transition around 1 ms in the step response. These changes are most noticeable using the circular baffle.

5 FINITE-SIZED DRIVERS

Since a driver has a finite diameter, which is often significant compared to the smallest dimension of the baffle, the assumption that the output from the driver

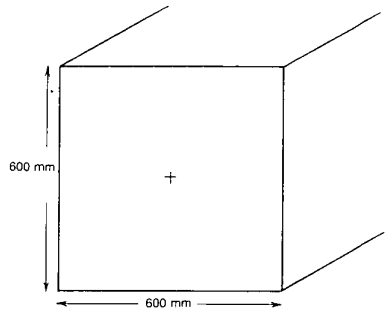


Fig. 11. Baffle shape for computer simulations of Figs. 8–10.

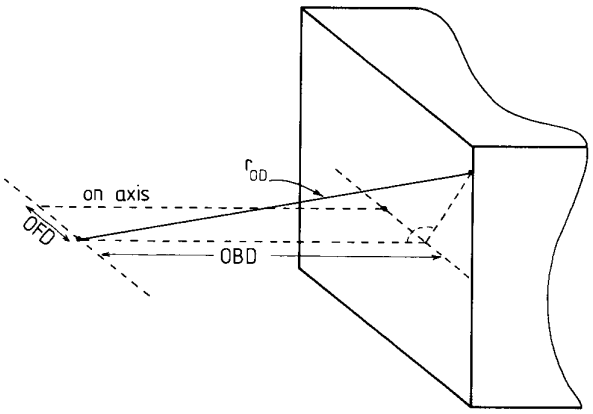


Fig. 12. Off-axis geometry.

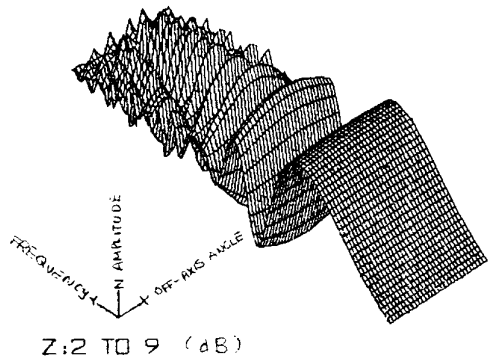


Fig. 13. Off-axis amplitude response. Baffle radius 300 mm.

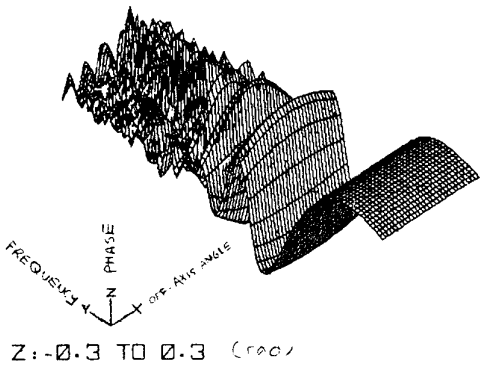


Fig. 14. Off-axis phase response. Baffle radius 300 mm.

is like that of a point source is inaccurate.

So as to approximate to the output from a flat piston drive unit, the surface of the cone is portioned into M equal area elements, each assumed to behave as a point source (see Fig. 21 for $M = 19$).

M is chosen to have the smallest allowable value, provided the amplitude and phase responses have converged in the audio band (up to 20 kHz).

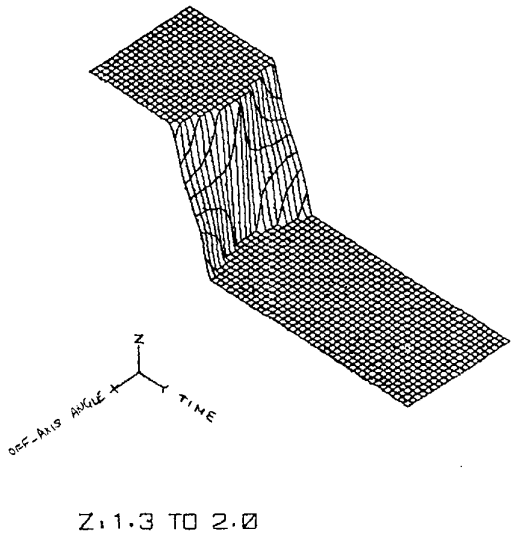


Fig. 15. Off-axis step response. Time range 0–3 ms; baffle radius 300 mm.

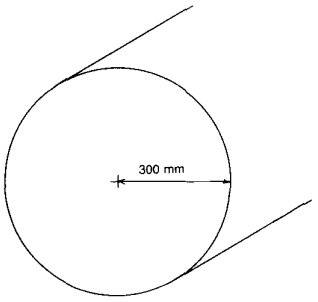


Fig. 16. Baffle shape for computer simulations of Figs. 13–15.

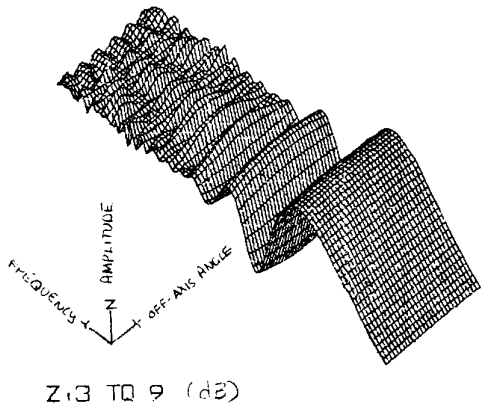


Fig. 17. On-axis amplitude response. Baffle size 600 by 600 mm.

The response of the driver on the baffle is derived by considering the interference of all the point source rays that imitate the driver and their associated diffracted rays.

5.1 On-Axis Steady-State Response

Let the Lth point source on the baffle have a response given by $AR_L(\omega) + j\,AI_L(\omega)$ at a point on axis with the center of the driver. These responses are calculated by adopting a procedure similar to the on-axis steady-state response calculations discussed in Sec. 3.1. However, it must be appreciated that all but one of the point sources which make up the driver are not on axis with the observer. Therefore a phase shift must be added to each point source response and each of the associated diffracted ray responses.

The response of the driver on a baffle $Dr(\omega)$ is then given by

$$Dr(\omega) = \frac{\sum_{L=1}^{M} [AR_L(\omega) + j\,AI_L(\omega)]}{M}.$$

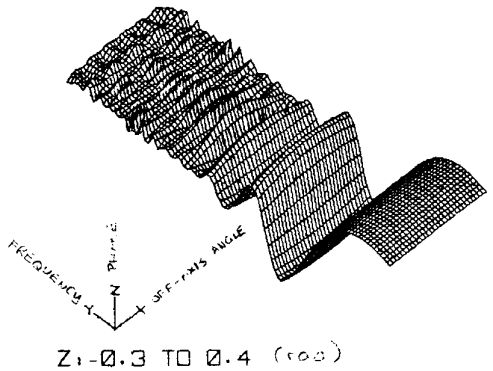


Fig. 18. Off-axis phase response. Baffle size 600 by 600 mm.

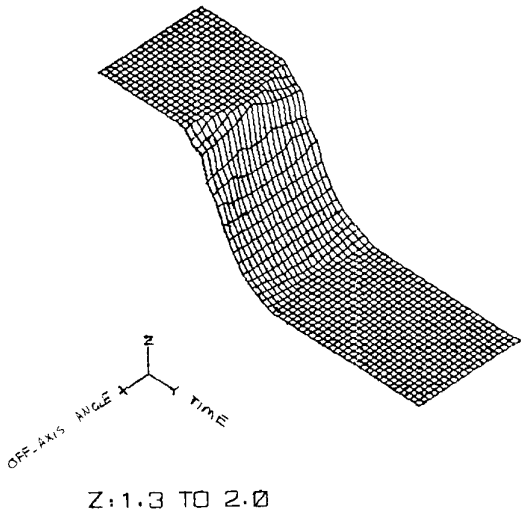


Fig. 19. Off-axis step response. Time range 0–3 ms; baffle size 600 by 600 mm.

This is then normalized to the free-air driver response, that is, Dr(ω)/free-air driver response.

Defining the response of a flat piston driver as PR(ω) + j PI(ω), (see [4] for the derivation), the on-axis normalized amplitude response $M_{con}(\omega)$ is thus

$$M_{con}(\omega) = \left| \frac{\sum_{L=1}^{M} [AR_L(\omega) + j\, AI_L(\omega)]}{M(PR(\omega) + j\, PI(\omega))} \right|$$

$$= \frac{\sqrt{\left[\sum_{L=1}^{M} AR_L(\omega)\, PR(\omega) + \sum_{L=1}^{M} AI_L(\omega)\, PI(\omega)\right]^2 + \left[\sum_{L=1}^{M} AI_L(\omega)\, PR(\omega) - \sum_{L=1}^{M} AR_L(\omega)\, PI(\omega)\right]^2}}{M\sqrt{PR(\omega)^2 + PI(\omega)^2}} . \quad (12)$$

The normalized phase response $P_{con}(\omega)$ is

$$P_{con}(\omega) = \arctan \left[\frac{\sum_{L=1}^{M} AI_L(\omega)\, PR(\omega) - \sum_{L=1}^{M} AR_L(\omega)\, PI(\omega)}{\sum_{L=1}^{M} AR_L(\omega)\, PR(\omega) + \sum_{L=1}^{M} AI_L(\omega)\, PI(\omega)} \right] . \quad (13)$$

5.2 On-Axis Step Response

Let the Lth point source on the baffle have a step response given by $AS_L(t)$ at a point on axis with the center of the driver. These step responses are calculated by adopting a procedure similar to that for the on-axis step response calculation shown in Sec. 3.2. Since all but one of the point sources which imitate the driver are not on axis, a time shift must be added to their response and to the associated diffracted ray responses.

The step response of the driver on the baffle $Dr(t)$ is then

$$Dr(t) = \frac{\sum_{L=1}^{M} AS_L(t)}{M} .$$

The normalized step response $T_{con}(t)$ is thus given by

$$T_{con}(t) = \frac{Dr(t)}{P(t)}$$

$$= \frac{\sum_{L=1}^{M} AS_L(t)}{M\, P(t)} \quad (14)$$

where $P(t)$ is the step response of a flat piston driver, which can be derived by Fourier analysis of the frequency response PR(ω) + j PI(ω).

5.3 Steady-State and Step-Response Simulations

The normalized on-axis amplitude, phase, and step response simulations using Eqs. (12)–(14) for a 300-mm-radius circular baffle and a 600- by 600-mm-square baffle are shown in Figs. 22–29. The observer-to-driver distance OBD is again set to 1 m. 37 point sources are used to imitate the driver, since no discernible change occurs in any of the responses if more point sources are used.

Figs. 22, 23, 26, and 27 show the normalized am-

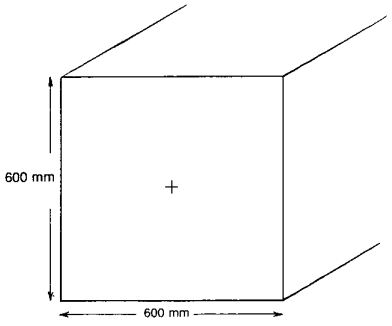


Fig. 20. Baffle shape for computer simulations of Figs. 17–19.

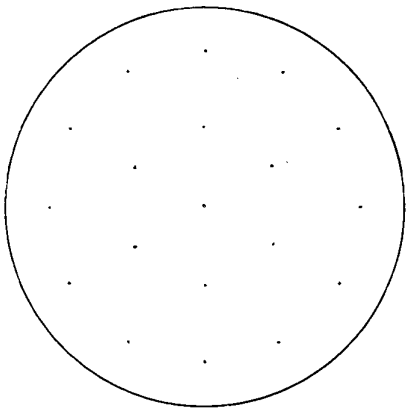


Fig. 21. Driver approximated by point sources.

plitude and phase responses for a 100-mm-radius driver, which can be compared with their corresponding responses using a point source driver as shown in Figs. 4, 5, 8, and 9. Both amplitude and phase have lower peak deviations at higher frequencies using the larger driver. This is caused by an increase in the incoherence of the diffracted rays, due to the differing locations of the point sources which imitate the driver. Consequently the step responses of Figs. 24 and 28 have a less sharp transition around 1 ms, this being most noticeable for the circular baffle. It is therefore advantageous to employ baffles that are as small as possible, ideally fractionally larger than the driver.

6 CONCLUSION

For smoother amplitude, phase, and step responses, the driver should be placed on the baffle such that the diffracted rays are as incoherent as possible. This implies that small irregularly shaped baffles should be used, or the driver located at unequal distances from the sides of small regularly shaped baffles. Rounding of the baffle edges to spatially distribute the edge should also help in this respect.

The application of sound-absorbing materials onto the baffle as first described in [5] is advantageous, since the diffracted rays will be attenuated, which subjectively improves stereo location. However, it must be appreciated that the attenuation is only significant over a limited frequency range dependent on the acoustic properties of the materials used. This will cause a drop in the amplitude response in the regions of high attenuation, leading to tonal coloration unless equalization is used.

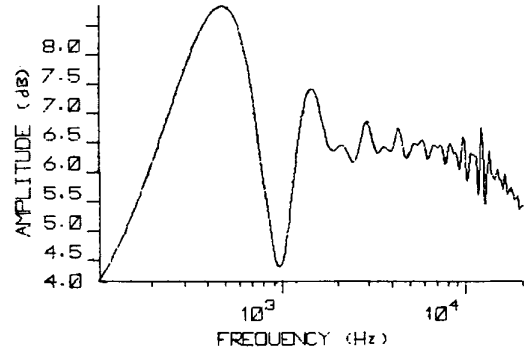


Fig. 22. On-axis amplitude response. Baffle radius 300 mm; driver radius 100 mm; 37 point sources are used.

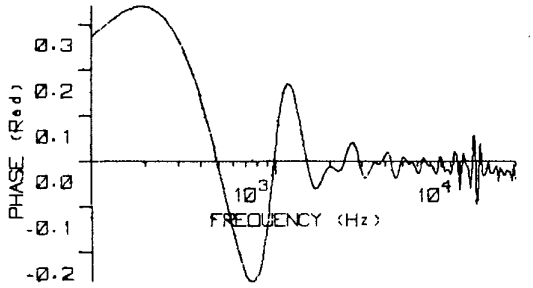


Fig. 23. On-axis phase response. Baffle radius 300 mm; driver radius 100 mm; 37 point sources are used.

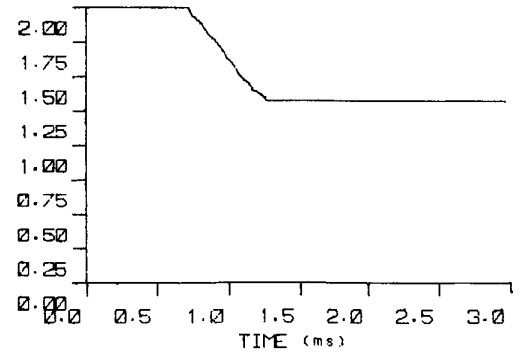


Fig. 24. On-axis step response. Baffle radius 300 mm; driver radius 100 mm; 37 point sources are used.

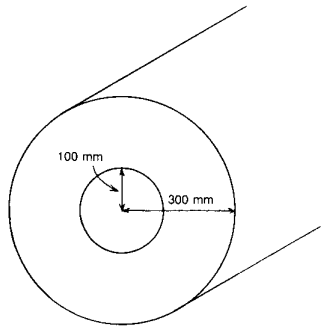


Fig. 25. Baffle shape for computer simulations of Figs. 22–24.

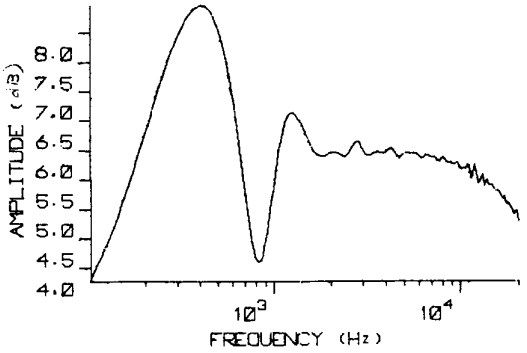


Fig. 26. On-axis amplitude response. Baffle size 600 by 600 mm; driver radius 100 mm; 37 point sources are used.

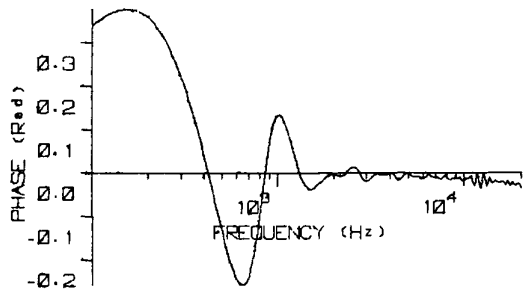


Fig. 27. On-axis phase response. Baffle size 600 by 600 mm; driver radius 100 mm; 37 point sources are used.

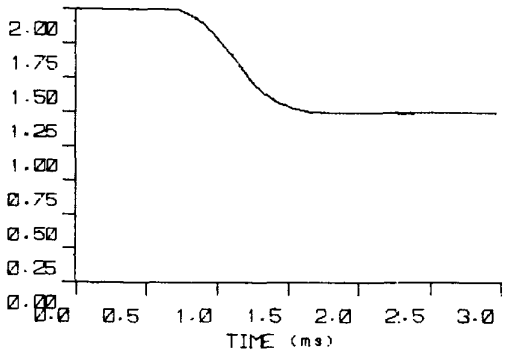


Fig. 28. On-axis step response. Baffle size 600 by 600 mm; driver radius 100 mm; 37 point sources are used.

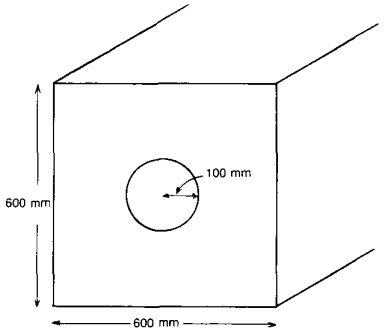


Fig. 29. Baffle shape for computer simulations of Figs. 26–28.

Diffraction is essentially a problem in the time domain, which will lend itself most easily to digital compensation. This area of research is now being actively pursued by the authors.

7 ACKNOWLEDGMENT

This work was supported by the Science and Engineering Research Council.

8 REFERENCES

[1] H. F. Olson, *Acoustical Engineering* (Van Nostrand, Princeton, NJ, 1957), chap. 1, pp. 20–24.

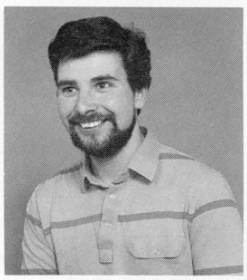
[2] J. B. Keller, "The Geometric Theory of Diffraction," *Symposium on Microwave Optics*, Eaton Electronics Research Laboratory, McGill University, Montreal, Canada (1953 June).

[3] Lord Rayleigh, *The Theory of Sound*, vol. 2 (Macmillan, London, 1878), sec. 280, pp. 100–102.

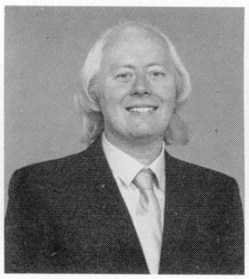
[4] E. Meyer and E. Neumann, *Physical and Applied Acoustics* (Academic Press, New York, 1972), chap. 5, pp. 170–177.

[5] A. Schaumberger, "Impulse Measurement Techniques for Quality Determination in Hi-Fi Equipment, with Special Emphasis on Loudspeakers," *J. Audio Eng. Soc.*, vol. 19, pp. 101–107 (1971 Feb.).

THE AUTHORS



R. Bews



M. Hawksford

Richard Bews is currently a Ph.D. research student working within the Audio Research Group at Essex University, where his studies have included work on loudspeaker diffraction at the baffle edge and the analysis and design of crossover filters in both the analog and digital domains. Prior to his SERC-supported studentship, Bews read for a physics degree, also at Essex University, for which he gained First Class Honours in 1983. He is currently a student member of the AES.

Mr. Bews's lesiure activities also encompass audio engineering and the development of high-performance analog electronics as well as listening to music.

•

Malcolm Hawksford is presently a senior lecturer in the Department of Electronic Systems Engineering at the University of Essex, U.K., where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honours B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship where work on the application of delta modulation to color television was undertaken.

Since his appointment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing, and loudspeaker systems has been undertaken. Dr. Hawksford has written several AES publications that include topics on error correction in amplifiers and oversampling techniques for ADC and DAC systems. His supplementary activities include designing commercial audio equipment and writing articles for *Hi-Fi News*—activities that integrate well with visits to Morocco and France. His leisure activities include listening to music, motorcycling, and motor mechanics. Dr. Hawksford is a member of the IEE, a chartered engineer, and a member of the AES.

Reduction of loudspeaker polar response aberrations through the application of psychoacoustic error concealment

A. Rimell
M.O. Hawksford

Indexing terms: Loudspeaker systems, Psychoacoustics

Abstract: In a loudspeaker system it is important to have a well controlled polar response; however, with conventional multidriver enclosures off-axis phase cancellation will occur at and around the crossover frequency. To generate a uniform polar radiation pattern from any given imperfect loudspeaker cabinet, it is necessary to reduce the perceived off-axis phase cancellation due to the crossover filters. The paper proposes a correction strategy, which is then evaluated in terms of the perceived improvement. By making use of the psychoacoustic theory of masking, and developing a new strategy for time-varying filters, a creative solution was found whereby the crossover filters cause the phase cancellation to be masked from the listener, rendering it inaudible. The proposed strategy is tested using both auditory models and actual listeners as arbiters of performance. Both test methods confirm that with the proposed strategy an improvement in sound quality at an off-axis position is obtained. The on-axis sound, unaffected, remains the in-phase sum of the loudspeaker driver outputs.

1 Introduction

Directionality of sound reproduction systems is an important issue for designers of high quality audio systems, not only for multichannel systems and sound reinforcement systems but also for conventional two-speaker stereo reproduction systems. A major source of polar response irregularity in the midband region (where the ear is most sensitive) is cancellation at and around the crossover frequency due to the difference in the individual driver-to-listener path lengths.

This work proposes a time-varying filter system with filters designed according to a psychoacoustic criterion. The off-axis errors are concealed through the utilisation of a human auditory process known as 'frequency domain masking'. The proposed method positions the crossover frequency such that any off-axis cancellation is masked by the surrounding signal content. Such a system therefore contains time-varying filters because it uses the signal content to determine where to position the crossover frequency. The flexibility of using digital filters to perform the crossover filter function means that it is also possible to include driver equalisation in the crossover filter design. We also show that simply using very high order filters is not a sufficient solution as the perceived error is a function of programme content, crossover frequency and filter order. Fig. 1 is a block diagram of the system which is described in this paper. Section 2 discusses the background theory of loudspeaker crossover filters and psychoacoustics. The main theory and implementation of the time-varying crossover filter is discussed in Section 3. Practical results obtained from listening tests and computer simulation are given in Section 4. Finally, Section 5 draws some conclusions.

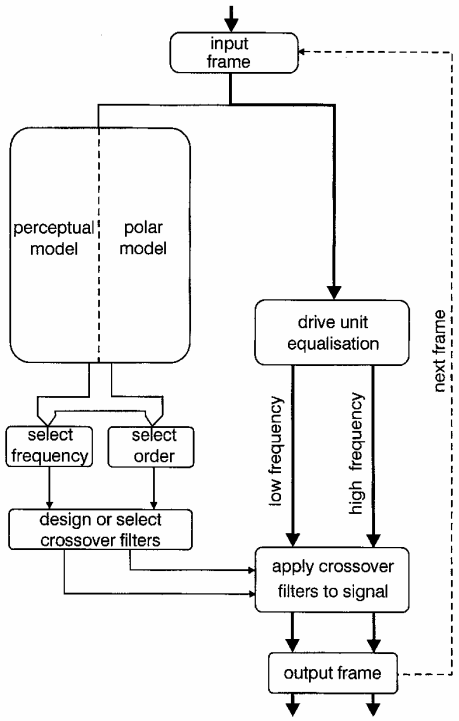


Fig. 1 *Block diagram of system*

IEE Proceedings online no. 19981722

Paper first received 3rd December 1996 and in final revised form 14th October 1997

A. Rimell was with the University of Essex and is now with BT Laboratories, Martlesham Heath, Ipswich IP5 7RE, UK

M.O. Hawksford is with the Centre for Audio Research and Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

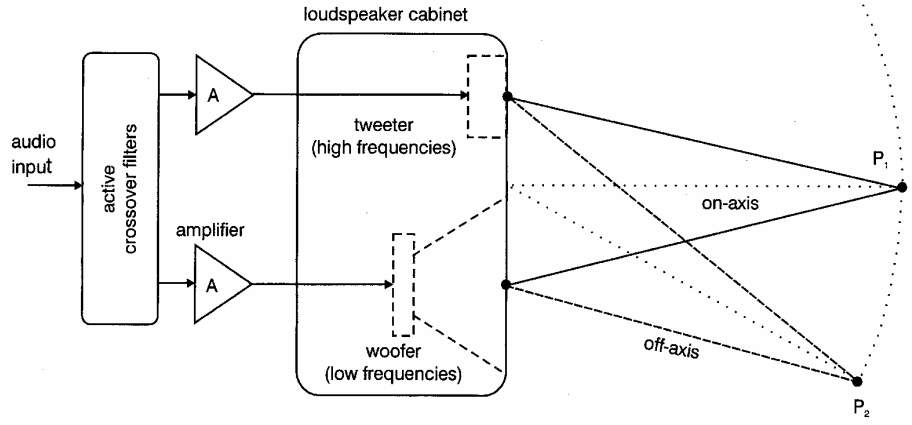


Fig.2 *Active-loudspeaker system*

2 Time-varying crossover filters: Background theory

2.1 Loudspeaker crossover filters

To produce the wide frequency response required for the human hearing system (20Hz to 20kHz), it is necessary to use two or more loudspeaker drive units. This is because standard driver transducers can only produce a portion of the frequency spectrum accurately, due to their bandpass frequency domain characteristics. A typical loudspeaker cabinet will contain two driver units. A 'woofer' recreates the lower half of the audio spectrum and a 'tweeter' recreates the upper half of the spectrum. Three-way systems, employing a midrange unit, are also used; however, for the purpose of this paper only two-way systems are considered. Because each of the two drivers operates within a limited frequency band, it is necessary to prevent a driver being fed frequencies outside its calibrated operating region. If a tweeter were to be fed with the whole frequency spectra it would be damaged and rendered inoperable. The filter for correctly directing frequency components to individual drivers is known as the 'loudspeaker crossover'.

Crossover filters can be implemented by using passive or active filters. Passive filters consist of resistors, inductors and capacitors and are usually placed inside the loudspeaker cabinet. Active crossover filters are usually implemented with op-amps or digital filters and are contained in a separate enclosure. Implementing the crossover filters digitally (such a system is known as a 'digital active system') gives the designer greater flexibility in selecting the most suitable filter frequency response. Fig. 2 shows the connection of one channel of an active system, using two amplifier channels, lowpass and highpass, and hence a two-channel stereo system requires four amplifier channels.

2.2 Crossover filter induced error

Consider the two-way loudspeaker system shown in Fig. 2, consisting of a woofer for the low frequencies and a tweeter for the high frequencies. Because the two drivers are physically separated, the only position where the listener is equidistant from both drivers is on the cabinet central axis. In Fig. 2 it is assumed that both drivers have their acoustic centre on the cabinet baffle; however, in practice the acoustic centres will be behind the baffle. At any off-axis position there will be a differential time delay introduced in the driver paths which causes a phase difference at and around the crossover frequency.

Fig. 3 shows the polar response of a digital 4th-order Butterworth complementary crossover filter pair. In this plot only the crossover response is shown; the speakers are assumed to have uniform directivity and a flat magnitude response. Note that the plot has a flat on-axis frequency response and that there are dips in the off-axis response at and around the crossover frequency of 3kHz. The maximum value of cancellation occurs at around 24°; this is where there is the greatest phase difference between the two drivers, causing maximum cancellation.

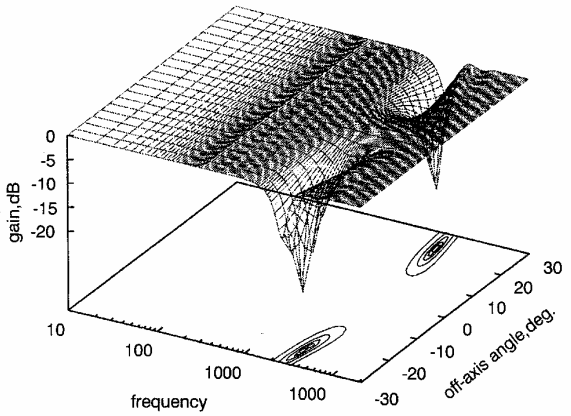


Fig.3 *Loudspeaker-crossover polar response*

2.3 Psychoacoustics

The coding scheme presented in this paper is based on the study of how human listeners perceive sound (psychoacoustics). A detailed description of psychoacoustic theory is beyond the scope of this paper; however, due to its importance a brief description of frequency domain masking is given. The reader is directed towards [1–7] for further information on psychoacoustics.

Frequency domain masking is the name given to the effect where a low intensity sound that can be heard in a quiet environment may cease to be audible in noisy surroundings, and the relative audibility of a particular sound might be diminished by the presence of another sound. By using psychoacoustic models [1, 7] it is possible to determine by how much one tone masks another.

3 Error concealment using time-varying crossover filters

3.1 An overview of the proposed solution

The polar response errors considered in this paper occur in a narrow band of frequencies based around the crossover frequency. Consider the example where an input signal consists of a large set of sine tones, then some of the signal spectrum will contain the polar response errors, some the desired audio signal, and some will be inaudible due to masking. If the part of the spectrum that was masked coincided with that where the off-axis errors occur, then the errors themselves would be masked. To achieve such masking of the error it is necessary to take a short sample of the audio waveform, transform it to the frequency domain, examine the signal spectrum and design the crossover filters so that the frequency of the off-axis errors coincided with that of a portion of the signal which is inaudible due to masking. The filters need to be continuously changing in sympathy with the audio waveform to be coded. The use of time-varying filters requires careful design to ensure that the changing filters do not introduce a modulating noise on to the original audio signal [8].

The crossover filter pair are generated for implementation in an active digital filter system with a linear or minimum-phase response. The filters do not necessarily need to be simple Butterworth type responses (as commonly used in active digital filter systems) and can include equalisation as necessary. The two crossover filters (lowpass and highpass) can be designed such that their on-axis responses always add up to unity, thus giving zero phase cancellation. By ensuring that the on-axis response remains unchanged only the off-axis response is modified, thus reducing the total perceived error in the loudspeaker system.

3.2 Filter selection criteria

In the following subsections we discuss the process of selecting the lowpass crossover FIR filter frequency response. Three basic parameters that can be adjusted are:

1. cutoff frequency (–6dB point)
2. filter order
3. filter shape.

The first two parameters assume a standard filter shape such as a Butterworth or Chebyshev, and the third parameter gives the system complete control of the filter shape. Sections 3.2.1, 3.2.2 and 3.2.3 discuss each of the parameters in greater detail.

3.2.1 Crossover frequency: The simplest method of crossover error concealment involves moving the crossover *frequency* into a position where it is masked by the signal. The model introduced by Johnston [7] was implemented (using a 1024-point FFT and a sampling frequency of 44.1kHz) to determine the masking threshold for the current frame of input signal, an example of which is shown in Fig. 4.

It is desirable to set a region within which the crossover frequency may be selected, in order that the loudspeaker drivers are not subjected to audio information which is outside their rated frequency range. The range selected depends on the physical attributes of the drivers, and in this paper a crossover range of 1.6–4.8kHz is taken. The frequency spectrum is divided up into bands, the positions of which are directly related to the 'auditory filters' used in the human ear [9, 10]. The centre of each band is defined as a possible crossover frequency, as the position of the auditory filter frequencies is closely related to the theory of frequency domain masking [9, 2].

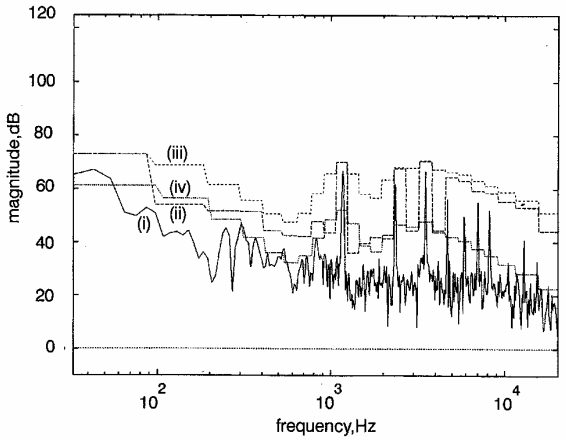


Fig.4 *Power spectrum, P(n), and masking threshold, T(n), for a frame of audio data*
(i) *P(n)*; (ii) *B(i)*; (iii) *C(i)*; (iv) *T(i)*

From Fig. 4 it can be seen that in the band from 2.8–3.2kHz there is no audible signal (i.e. within that band all of the power spectrum is below the masking threshold) and hence if the crossover frequency is set at, say, 3.0kHz then the errors introduced into the signal band will be inaudible. The most suitable band is found by determining which band has the lowest value of signal-to-mask ratio. If the crossover frequency was set at 3.5kHz, which can be seen from Fig. 4 is the frequency of an audible tone, then part of the audible spectrum is distorted and the error is detected when observed from an off-axis position.

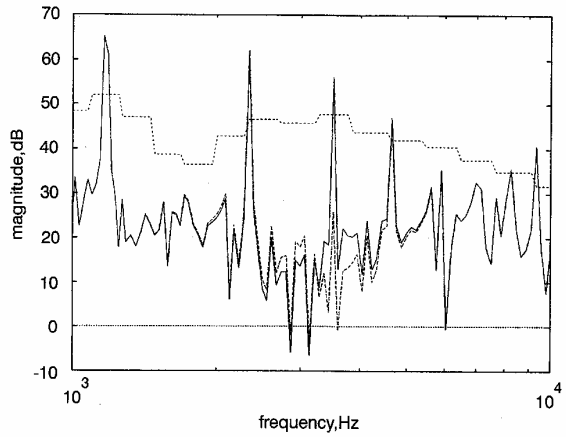


Fig.5 *Perceived off-axis output*
——— f_c = 3kHz; – – – f_c = 3.5kHz; ····· original masking

Fig. 5 shows the power output for two different crossover frequencies, calculated at 20 degrees off-axis (the worst case for the example used). The original sample sequence is the same as that shown in Fig. 4. It can be seen that there is little perceived signal degradation with a crossover frequency of 3.0kHz, but when the crossover frequency is set to 3.5kHz (which could happen if f_c is fixed) it can be seen that the signal tone becomes inaudible. Thus by careful selection of the position of the crossover frequency it is possible to conceal the errors introduced by the off-axis cancellation.

The example used here has a tone-like quality (it is in fact a piano) and thus there is an obvious position in which to place the crossover. When the signal sample is noise-like (i.e. relatively spectrally flat) there may not be such a clear position for the crossover frequency, and thus a strategy for such a situation was developed. For tone-like signal frames the crossover is set to be adaptive and when the signal frame is noise-like the crossover filter remains static, removing the otherwise audible modulation of the noise-like elements. The spectral flatness measure [7] (SFM) given in eqn. 1 is used to determine how tone-like or noise-like the current signal frame is. Theoretical white noise would give an SFM of 0dB, and a theoretical sine wave would give an SFM of $-\infty$dB. A practical 1024-sample frame of white noise gives an SFM of $\simeq$–1dB; a sine wave of the same frame length gives an SFM of $\simeq$–34dB. The non-ideal SFM value of the sine wave is due to the fact that the sine wave is windowed and then an FFT is performed, resulting a some spectral spreading:

$$SFM_{dB} = 10\log_{10}\left(\frac{GM}{AM}\right) \quad (1)$$

where

$$GM = \sqrt[N]{\prod_{n=1}^{N} P(n)} \qquad AM = \frac{1}{N}\sum_{n=1}^{N} P(n)$$

and N is the total number of frequency domain samples in a frame of data.

3.2.2 Filter order: The example given in Figs. 4 and 5 uses standard 4th-order digital Butterworth filters in the crossover. It can be argued that the higher the filter order the less the overlap between the drivers and the less the off-axis cancellation. In this Section we show that the assumption that higher order filters produce less error is not necessarily valid. The perceived error is in fact a function of both crossover frequency and filter order.

Consider the ideal lowpass brick wall filter. This filter's impulse response consists of a sinc function starting at time = $-\infty$ and ending at time = $+\infty$. The sinc time domain response introduces ringing into the system impulse response and thus colours the system output [11]. Ideal brick wall filters are obviously unrealisable and therefore high-order approximations are used, and due to restrictions on filter length the width of the crossover region is finite; thus an area of frequency cancellation remains. If, in the example shown in Fig. 5, a high order filter pair with a crossover frequency of 3500Hz were used, then cancellation of an audible tone would still occur.

Fig. 6 shows the total off-axis perceived error, as calculated by Johnston's psychoacoustic masking model, for different crossover frequencies between 2 and 5kHz with various orders of Butterworth filter pairs. The data used to generate the plot are the same as those used for Figs. 4 and 5 and are thus only valid for the frame of data shown in Fig. 4. It can be seen from Fig. 6 that an 8th-order filter at 3kHz has a lower total error than a 1024th-order filter at 3.5kHz. Assuming that a higher order filter always introduces less system error without also considering the crossover frequency is invalid.

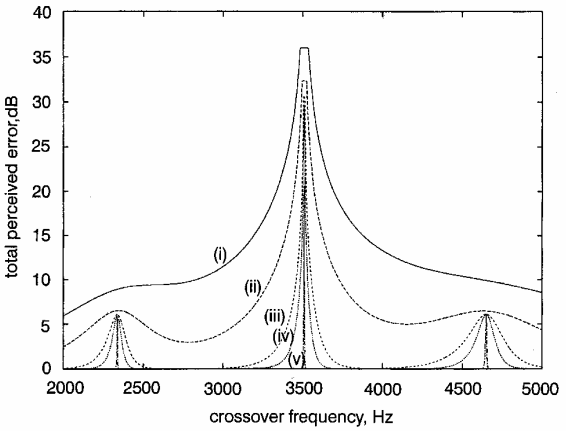


Fig. 6 *Effect of changing filter order*
(i) 4th order; (ii) 8th order; (iii) 32nd order; (iv) 64th order; (v) 1024th order

Consequently, an adaptive crossover system could be set to find the minimum order of the filter that keeps the error within a specified range, and thus reduce time domain dispersions in the crossover filters.

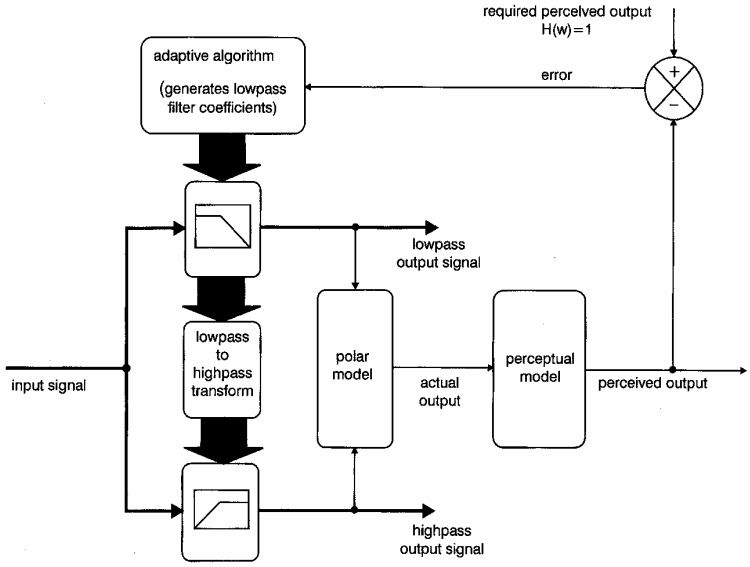


Fig. 7 *Block diagram of adaptive optimisation system*

3.2.3 Filter shape: Sections 3.2.1 and 3.2.2 discussed changing the crossover filter specification whilst using standard filter shapes (such as a Butterworth or Chebyshev); however, with digital filters it is possible to generate arbitrary filter frequency responses. The most advanced implementation of the adaptive crossover system would be to make the whole system adaptive; thus by using an optimisation algorithm it should be possible to generate recursively the lowpass filter coefficients such that the perceived error is minimised. Fig. 7 shows the block diagram of the proposed system.

Filters need to be designed to meet the error concealment criteria, and thus it is necessary to develop an optimal digital filter. As the masking changes continuously with the signal content and as the psychoacoustic model is nonlinear many of the standard filter design techniques are inappropriate. The main constraint is that the filters must retain their lowpass and highpass characteristics, as the main function of loudspeaker crossover filters is to prevent the loudspeaker drivers being fed frequencies outside their normal operating range. This constraint limits the usefulness of traditional adaptive filters, such as LMS and Kalman filters [12, 13], because whilst reducing the error they also destroy the lowpass and highpass characteristics, thus making the filters unusable.

By using a genetic algorithm (GA) [14] it may be possible to develop a filter pair which will produce a minimal amount of audible error [15, 16], whilst retaining the required low/highpass characteristics.

3.3 Equalisation

In practice the loudspeaker drivers used will not have an ideal bandpass frequency response. By taking the drivers' frequency responses it is possible to design a pair of crossover filters such that they include equalisation, thus correcting for anomalies in the driver responses. Using digital filters the equalisation filter design consists of taking the individual driver's frequency response and inverting it. Thus, where there is a dip in the driver's response there will be a peak in the correcting filter's frequency response, and vice versa, thus giving an overall flat passband region.

Because of its physical construction any drive unit will be bandlimited, i.e. there will be a minimum and a maximum frequency that the driver will be able to deliver without distortion. When designing an equalising filter it is important to consider the working range of the driver, and not to try to boost the signal too much at the extreme ends of its operating range. The FIR equalising filter has a frequency response which is the inverse of that of the loudspeaker driver, with a slight modification to prevent out-of-band signals having excessive gain [17–19].

The design of the equalising filters is only carried out once, for a particular set of drive units, unlike the time-varying crossover filter design presented here in which the filters are designed once for each frame of input signal. In a static crossover system (i.e. one where the crossover filter design does not change according to programme content) it is usual to combine the crossover filters with the equalising filters; however, if the filter shapes are being designed within the time-varying loudspeaker crossover filter system (see Section 3.2.3), it is suggested that the equalisation and crossover filtering operations are carried out separately to simplify the time-varying algorithm. Where a look-up table of crossover filter pairs is used (for example a set of 4th-order Butterworth crossover filters with different crossover frequencies), then the equalisation could simply be incorporated into the filters themselves, although the filters would then only be valid for a particular loudspeaker cabinet.

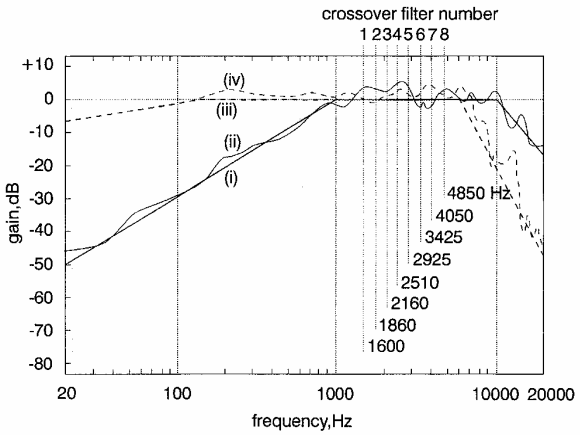


Fig. 8 *The need for driver equalisation*
(i) Tweeter bandpass function (ideal response); (ii) measured tweeter response; (iii) woofer bandpass function (ideal response) (iv) measured woofer response

From an examination of Fig. 8 it can be seen that it is highly desirable to incorporate equalisation into the crossover filters. The eight frequency positions correspond to the centre frequencies of eight pairs of crossover filters (for a scheme such as that described in Section 3.2.1): the frequency values are related to the auditory filter bandwidths [9] as shown in Table 1 [20]. It can be seen that the frequency response of the two drivers at 1600 and 3400 Hz are somewhat different; a change in crossover frequency from one to the other will produce a tonal change due to the change in frequency response. It is important to keep a uniformity in the frequency response over the range of possible crossover frequencies. There will also be a slight change due to the different driver polar responses; however, as long as the low/midband driver is not driven at frequencies which make it highly directional, then the effects are negligible.

Table 1: Frequency bands used in the across program

Lower frequency of band, Hz	Upper frequency of band, Hz	Crossover frequency: centre of band, Hz
1480	1720	1600
1720	2000	1850
2000	2320	2150
2320	2700	2500
2700	3150	2900
3150	3700	3400
3700	4400	4000
4400	5300	4800

3.4 Filter implementation

The program, called *across*, was written in C++ using an object oriented programming style with a proprietary matrix library. Sound files were played on a Silicon Graphics Indy workstation which has an AES/EBU-SPDIF 16 bit digital audio interface [21].

In sections 3.2.1, 3.2.2 and 3.2.3 three methods of crossover filter design were described. In the *across*

program the first of the three methods was implemented, where the crossover frequency is varied. This is achieved by producing a look-up table of filters with crossover frequencies in the 1.6–4.8kHz region. Filter design by use of a genetic algorithm is currently impracticable due to the time taken to generate each filter pair. Using a look-up table of filters the program takes 5 min to process a 20s stereo sound file – this could be further reduced by using assembly language programming on a DSP chip. The crossover filters used included equalisation filters designed specifically for the monitor loudspeakers used to audition the processed audio.

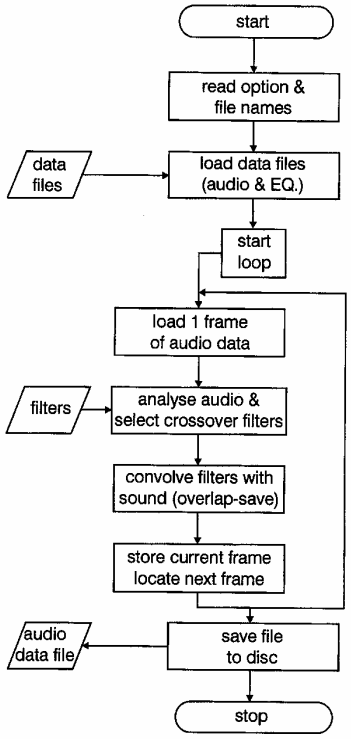


Fig. 9 *System flowchart*

Fig. 9 is a block diagram showing the function of the program: a 16-bit stereo audio sample is read in and analysed, and crossover filters are selected according to the program content and then applied to the input audio. The low/highpass data are then saved to disk for playback.

As the filters are continuously changing (switching between sets stored in a look-up table) it is necessary to interpolate smoothly between the sets of filter coefficients to avoid any processing noise [8]. The use of a perceptually smooth filter interpolation scheme is necessary in a wide range of audio equipment – for example, the equalisation control on a digital mixing desk. This application is different from a digital mixing desk in that the filters need to change very quickly from one setting to another. With equipment that requires human interaction the speed of change is a function of the speed of the user (a user may take a second to change the equalisation filter), whereas the system proposed here may change within a data frame of approximately 20ms, thus increasing the likelihood of introducing processing noise.

Consider the case, for example, where the system alternately switches between two crossover filter pairs. If the filters were instantly switched in and out rather than smoothly interpolated, then we can describe the output signal as a Fourier series representing the time-domain waveform modulated by a square wave (amplitude modulation). It can be shown, by calculation of the Fourier series, that square-wave modulation produces harmonics that decay with $1/n$, where n is the harmonic index. If a linear interpolation were employed to change between filters a triangular-wave modulation would occur, resulting in harmonics which decay at a rate of $1/n^2$. It therefore follows from Fourier series analysis that a sinusiodal interpolation scheme would produce only two harmonics (sum and difference) which are spectrally close to the fundamental frequency.

Pieces of audio were filtered with a pair of time-varying filters using step, linear and sinusoidal interpolation schemes. The resultant signals were played to a panel of listeners (without them knowing which signal was which) and they all noticed modulation tones with the step and linear interpolation. No modulating tones were noticed with the pieces of audio processed with a sinusiodal interpolation scheme; this result was confirmed by passing all of the audio pieces through a psychoacoustic masking model [7]. The model showed that the two harmonics generated with a sinusoidal interpolation scheme were inaudible due to masking effects, whereas with the other interpolation schemes the masking model showed that the harmonic components would indeed by audible.

The system used a frequency domain interpolation scheme, where the filter frequency responses changed smoothly in the frequency domain, thus ensuring that each intermediate filter had a similarly shaped frequency-magnitude response to that of the original crossover filter. As described above, a sinusoidal interpolation envelope was used to interpolate between filters.

4 Results

To test the adaptive crossover algorithm, a selection of speech and musical samples, with a wide variety of styles, were recorded on to a hard disk and then coded with both a fixed and an adaptive crossover algorithm. Two methods of evaluation were used: an off-axis model combined with a psychoacoustic model [1] and listening tests with a group of test subjects.

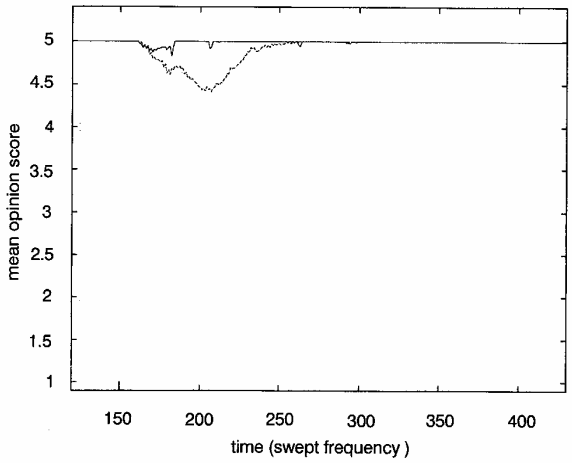


Fig. 10 *Output of perceptual model with MOS scale: Swept sine wave*
——— adaptive; – – – fixed
Scores: 1 Very annoying; 2 annoying; 3 slightly annoying; 4 perceptible, but not annoying; 5 imperceptible

The first test sample consisted of a swept sine wave. As the swept sine wave passes through the crossover region, off-axis cancellation occurs, causing a null at and around the crossover region. Fig. 10 shows the perceived off-axis sound (a comparison of the original on-axis signal with the coded off-axis signal) as determined by a psychoacoustic model. It can be seen that with a fixed (conventional) filter the null around the crossover frequency is clearly audible whereas with the adaptive filter there is little perceived degradation to the signal. The output of the psychoacoustic model was also confirmed by listening to an actual active-crossover loudspeaker system at on-axis and off-axis positions.

Listening tests were carried out with the active-crossover loudspeaker system, examining the other coded samples by comparing the on-axis and off-axis responses. The improvement in sound quality was greatest for tonal signals such as solo classical instruments. For the worst case example (white noise) the fixed and adaptive coding could not be distinguished – in other words for the worst case scenario the sound does not change but in other cases the sound is improved. Owing to the complementary nature of the high- and lowpass crossover filters the on-axis response is always the in-phase sum, correctly recreating the original sound. The tests were carried out in a simulated living room containing carpets, soft chairs and curtains, which was necessary to reduce room effects to a typical level. For a detailed description of the effects a room can have on an audio signal (see [22, 2]).

Table 2: Listening test results of 14 subjects

Listener number	Adaptive better than fixed, %	Fixed better than adaptive, %	Don't know, %	Trained listener?
1	100	0	0	Y
2	95	0	5	Y
3	100	0	0	Y
4	100	0	0	Y
5	100	0	0	Y
6	100	0	0	Y
7	50	0	50	N
8	60	0	40	N
9	100	0	0	N
10	100	0	0	N
11	100	0	0	N
12	50	0	50	N
13	70	0	30	N
14	95	0	5	N

20 random double-blind presentations of a piece of classical guitar music with ten having fixed filters and ten having adaptive filters. For each piece the listener was free to move around the loudspeaker to compare on- and off-axis responses

As an example we shall consider the results of listening tests performed on a piece of classical guitar music coded with fixed and adaptive filters, and listened to at an on- and off-axis position. Table 2 contains the results from the listening tests. Test subjects consisted of a mixture of trained and untrained listeners (a trained listener is one who is experienced in carrying out listening tests and is familiar with the sound of errors in audio systems; an untrained listener is a member of the general public). The results show that almost all of the test subjects perceived the adaptive crossover encoded sample as being more like the original sample than that with the fixed crossover. Two of the 14 test subjects could not distinguish between the two samples and none preferred the fixed filter to the adaptive one. All subjects agreed that there was no audible processing noise caused by the filter coefficients changing with every frame of audio signal.

5 Conclusions

In this paper a method of dealing with errors in two-way loudspeaker systems has been proposed. The methodology employed is based on the principle of error concealment utilising psychoacoustic criteria and is summarised by Fig. 1.

Any multidriver loudspeaker system will have a non-uniform polar frequency response due to the physical separation of the drive units. The amount of nonuniformity is related to the crossover filter order, but high order crossovers will still exhibit off-axis cancellation. Whilst it is not possible to generate a flat polar frequency response with a multidriver system, it is possible to position the error such that it is not detectable by a human listener. To achieve this it is necessary to base the algorithm on an understanding of the human auditory perception system.

It was found that for simple sine wave based test sequences the adaptive system produced a significant improvement in sound quality (i.e. the off-axis errors were successfully masked by the surrounding material). For a wide range of typical audio excerpts an improvement was obtained which a panel of trained and untrained listeners noticed. The system tested comprised a look-up table of 4th-order Butterworth crossover filters. Using a genetic algorithm may produce an even greater perceived improvement but is not practicable at present due to the immense amount of computing power required.

6 References

1 BEERENDS, J., and STEMERDINK, J.: 'A perceptual audio-quality measure based on a psychoacoustic sound representation', *J. Audio Eng. Soc.*, 1992, **40**, (12), pp. 963–978

2 EVEREST, F.A.: 'The master handbook of acoustics'. (McGraw–Hill, 1994), 3rd edn.

3 HOUTSMA, A.: 'Psychophysics and modern digital audio technology', *Philips J. Res.*, 1992, **47**, (1), pp. 3–14

4 MOORE, C.: 'Psychoacoustic considerations in choosing your hi-fi', *in* 'An introduction to the psychology of hearing' (Academic Press, 1989), 3rd edn.

5 RIMELL, A.: 'Perceptual coding in digital audio', *Electronics, the Maplin magazine*, 1995, **14**, (86), pp. 47–50

6 ZWICKER, E., and ZWICKER, U.: 'Audio engineering and psychoacoustics:– matching signals to the final receiver, the human auditory system', *J. Audio Eng. Soc.*, 1991, **39**, (3), pp. 115–126

7 JOHNSTON, J.: 'Estimation of perceptual entropy using noise masking criteria'. Proceedings of IEEE ICASSP, 1988, pp. 2524–2527

8 RIMELL, A., and HAWKSFORD, M.: 'Audibility analysis of processing noise in digital filter morphing schema'. Presented at the 101st convention of the Audio Engineering Soc., 1996

9 MOORE, C.: 'An introduction to the psychology of hearing' (Academic press, 1989), 3rd edn.

10 MOORE, C., and GLASBERG, B.: 'Suggested formulae for calculating auditory-filter bandwidths and excitation patterns', *J. Acoust. Soc. Am.*, 1983, **74**, p. 750

11 RIMELL, A., and HAWKSFORD, M.: 'Digital-crossover design strategy for drive units with impaired and noncoincident polar characteristics', *J. Audio Eng. Soc. (Abstracts)*, 1993, **41**, (12), p. 1065 (Presented at the 95th convention of the Audio Engineering Soc., preprint 3750)

12 FRIEDLANDER, B., and MORF, M.: 'Least squares algorithms for adaptive linear-phase filtering', *IEEE Trans.*, 1982, **ASSP—30**, (3), pp. 381–390
13 HAYKIN, S.: 'Adaptive filter theory' (Prentice–Hall, 1986)
14 HOLLAND, J.: 'Genetic algorithms', *Sci. Am.*, July 1992, **267**, pp. 44–50
15 ETTER, D., HICKS, M., and CHO, K.: 'Recursive adaptive filter design using an adaptive genetic algorithm'. IEEE ICASSP, 1982, pp. 635–638
16 RIMELL, A., and HAWKSFORD, M.: 'The application of genetic algorithms to digital audio filters', *J. Audio Eng. Soc. (Abstracts)*, 1995, **43**, (5), p. 398 (Presented at the 98th convention of the Audio Engineering Soc., preprint 3988)
17 HAWKSFORD, M., and GREENFIELD, R.: 'Efficient filter design for loudspeaker equalisation', *J. Audio Eng. Soc.*, 1991, **39**, (10), pp. 739–751
18 HAWKSFORD, M., and HEYLEN, R.: 'Optimal multi-rate filters for minimum and linear-phase equalisation'. Presented at the 97th convention of the Audio Engineering Soc., 1994, (preprint 3900)
19 WILSON, R.: 'Equalisation of loudspeaker drive units considering both on- and off-axis responses', *J. Audio Eng. Soc.*, 1991, **39**, (3), p. 127
20 ZWICKER, E.: 'Sub-division of the audible frequency range into critical bands', *J. Acoust. Soc. Am.*, 1961, **33**, p. 248
21 AES, : 'AES Recommended practice for digital audio engineering- serial transmission format for linearly represented digital audio data', *J. Audio Eng. Soc.*, 1985, **33**, (12), p. 979
22 RIMELL, A., and HAWKSFORD, M.: 'From the cone to the cochlea: Modelling the complete acoustical path', *J. Audio Eng. Soc. (Abstracts)*, 1996, **44**, (7/8), p. 646 (Presented at the 100th convention of the Audio Engineering Soc., preprint 4240)

Introduction to distributed mode loudspeakers (DML) with first-order behavioural modelling

N.J.Harris and M.O.J.Hawksford

Abstract: A simple equivalent circuit of a distributed mode loudspeaker (DML) is described, which is accessible to engineers not specialised in acoustics. The DML is an acoustic radiator, the electrical, mechanical and acoustical properties of which differ radically from conventional moving coil transducers. DML radiation results from uniformly distributed, free vibration in a stiff, light panel and not pistonic motion. To enable acoustic engineers to use existing software programs to model their application of DML technology, an efficient equivalent circuit is developed. Within the constraints of the model, velocities and displacements of the various elements can be calculated, and the radiated acoustic power and pressure predicted.

List of symbols

Panel parameters:

E = Young's modulus, Pa
ρ = mass density, kg/m^3
ν = Poisson's ratio
η_{mech} = mechanical loss factor
η_{rad} = radiation loss factor each side of panel
γ = power ratio (radiation loss/total loss)
h = thickness of panel, m
B = bending rigidity of panel, Nm
μ = mass per unit area of panel, kg/m^2
$v(\omega)$ = velocity of travelling wave in panel (tangential to panel)
$k_p(\omega)$ = wave number of travelling wave in panel (tangential to panel)
Z_p = mechanical impedance of panel, kg/s

Note that, for an isotropic panel $B = h^3/12\ E/(1 - \nu^2)$ and $\mu = h\rho$

Acoustic parameters:

P = acoustic power, W
U = velocity, m/s
R_r = acoustic radiation resistance, kg/s
X_r = acoustic radiation reactance, kg/s
Z_r = acoustic radiation impedance, kg/s
ρ_0 = density of air, kg/m^3
c = speed of sound in air, m/s
k = acoustic wave number = ω/c, m^{-1}
a = radius of circular piston, m

Mechanical parameters (normal to panel):

x, x_p, x_m = displacements, m
u, u_p, u_m = velocities, m/s
F = force, N
Y_p = specific velocity (mobility) = u_p/F, m/Ns

General:

ω = angular frequency, rad/s
bar over variable = RMS value (e.g. $\bar{u}$)
single underline = vector (e.g. $\underline{x}$)
double underline = matrix (e.g. $\underline{\underline{M}}$)

IEE Proceedings online no. 20000390
DOI: 10.1049/ip-20000390
Paper received 9th February 2000

N.J. Harris is with New Transducers Ltd., Huntingdon PE18 6AY, UK
M.O.J. Hawksford is with the Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

1 Introduction

The aim of this paper is to introduce a simple yet efficient equivalent circuit of a distributed mode loudspeaker (DML), which is accessible to engineers not specialised in acoustics. (DML is a term coined by New Transducers Limited (NXT) to describe a loudspeaker working according to their teaching. NXT is a registered trademark of New Transducers Ltd, a subsidiary of NXT plc.) Using this equivalent circuit, the velocities and displacements of the various elements can be modelled and the radiated acoustic power predicted.

In the 1920s the ideal loudspeaker was conceived to operate as a rigid piston with all points on the radiating surface moving in phase. This design aim applied irrespective of whether the loudspeaker diaphragm was driven from a moving armature, or later, from a moving coil. Operating a diaphragm in this manner imposes two fundamental requirements to maintain an even frequency response. First, the diaphragm has to be sufficiently small, compared with the wavelength of sound in air, to approximate to a point source and secondly, the whole diaphragm must move with the same acceleration so that, because of pistonic behaviour, it can be considered as a lumped moving mass. There have been a few notable exceptions to this design philosophy. For example, Bertagni (Sound Advance Systems Inc., California, USA) envisaged a 'timpanic' diaphragm with controlled break-up and Manger (Manger-Schallwandler, Mellrichstadt, Germany) exploited damped bending waves.

Contrary to normal pistonic diaphragm behaviour, a DML is an acoustic radiator, the electrical, mechanical and acoustical properties of which differ radically from conventional moving-coil transducers such that a new loudspeaker class is defined. A DML is distinguished by acoustic radiation that emanates from uniformly distributed, free bending wave vibration induced in a stiff, light panel and not from pistonic motion. Optimisation techniques have been developed and are the subject of many international patents by New Transducers Ltd (e.g. PCT/GB96/02145, September 1996). These solve complicated differential equations defining motion in such structures and enable stiff, light panels to be designed that exhibit an extremely uniform modal density, the prerequisite for distributed mode behaviour. The two fundamental requirements for a pistonic radiator do not apply to a DML, the latter being both arbitrarily scalable in size and the antithesis of a rigid body.

To understand and describe DML operation, a simple mechanical model is first derived, from which an equivalent circuit is developed. Using this equivalent circuit, the velocities and displacements of the various elements can be determined. Also, within the constraints imposed by the model parameters, the radiated acoustic power can be estimated from the panel mean velocity.

Currently, there is no technical terminology that adequately describes the acoustic radiation of a DML. In some respects a DML is a better approximation to a point source than a piston, but it also exhibits both temporal and spatial decorrelation, which are unusual yet distinctive DML traits. An open-baffled DML could be described as a 'diffuse dipole', although it is neither completely diffuse nor completely dipolar. Likewise, a DML within an infinite baffle could be described as a 'diffuse monopole'. A brief description of the DML is presented as background, although it is not proposed to discuss in depth the acoustics of this class of radiator. More complete descriptions of the acoustical properties of the DML are to be found elsewhere, for example [1–4] for physical acoustics and [5, 6] for psychoacoustics.

2 Piston loudspeakers

Initially, we review briefly the core theory that describes the conditions for a pistonic acoustic radiator to yield a flat power response, from which the corresponding results for a DML can be extrapolated.

For a pistonic loudspeaker to achieve frequency independent power transfer, the diaphragm velocity must be inversely proportional to frequency, or in other words, the acceleration must be constant. This is achieved when the loudspeaker operates as a mass-controlled device, such that a constant force produces a constant acceleration, a condition normally achieved above the fundamental resonant frequency. This requirement for a constant acceleration is a direct result of the frequency dependent real part of the radiation impedance of the piston [7]. The complex radiation impedance Z_r of a vibrating piston is given as

$$Z_r = \pi a^2 \rho_0 c \{R_1(2ka) + jX_1(2ka)\} \quad (1)$$

where

$$R_1(x) = 1 - 2J_1(x)/x \quad (2)$$

$$X_1(x) = (4/\pi)[x/3 - \text{higher terms}] \quad (3)$$

a = radius of piston (m) ρ_0 = density of air (kg/m^3)

c = velocity of sound in air (m/s) k = wave number

Defining U as the diaphragm velocity, the excitation force F and forward acoustic radiation power P are related by

$$F = Z_r U \quad \text{and} \quad P = \frac{1}{2}U^2 R_r \quad (4)$$

Consequently, observing the expression for power, a flat power response is achieved where the diaphragm velocity is inversely proportional to frequency and where the real part of the radiation resistance is proportional to frequency squared, a condition met when $ka < 1$ and when the loudspeaker operates above its fundamental resonance. For $ka > 1$, the radiation impedance becomes approximately constant and equal to $\pi a^2 \rho_0 c$, as can be seen from eqn. 2. Since the impedance characteristic changes from a 12dB/octave slope to a constant slope around $ka = 1$, while the mass controlled operation is retained assuming an ideal piston, a second-order high-frequency roll-off in the power response occurs.

A small acoustic radiator approximates a point source, which produces spherically symmetric acoustic radiation. At higher frequencies, when the source size becomes large with respect to the wavelength of sound in air, the radiation beams in the forward direction. This beaming maintains an on-axis pressure proportional to the acceleration and accounts for the loss of acoustic power. Consequently, the pressure response of a typical cone loudspeaker is limited at high frequency as illustrated in Fig. 1.

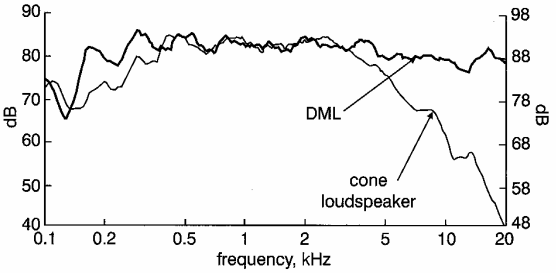


Fig. 1 *Measured on-axis pressure responses of DML and 140mm cone in baffle of identical size*
Levels adjusted for equality in pass-band of piston

3 Distributed mode loudspeakers

The radiation impedance of a typical pistonic loudspeaker is normally small compared with the mass of the cone and therefore has little effect on its motion. However, it dominates the loudspeaker's performance at the mechanical-to-acoustical interface, where the frequency dependence of the radiation resistance defines the required mechanical performance (i.e. a frequency independent acceleration).

The radiation impedance is also small compared with the impedance of a typical DML. However, in contrast to the pistonic loudspeaker, its effect on the acoustical performance is also small. The mechanical losses in a DML constructed from stiff but light panel material are typically very low, where the principal damping mechanism is attributable to acoustic radiation. The power ratio γ defined as

$$\gamma = \frac{2\eta_{rad}}{\eta_{mech} + 2\eta_{rad}}$$

can be as high as 98%, and is often as high as 90%. Therefore, a good approximation when modelling the mechanical behaviour of the DML is to assume that all the mechanical input power is radiated. This approximation holds for frequencies at which the loudspeaker is truly modal and where the panel is sufficiently large to be self-baffling. If the panel is mounted in a baffle, then this additional constraint applies to the total baffle size. At lower

frequencies, a simple diffraction model can provide a fairly accurate extension to the high-frequency case.

Acoustic radiation from each element of a DML results from surface motion normal to the plane of the panel that is induced by bending waves that propagate across the surface. Because bending waves are dispersive (i.e. the wave velocity is a function of frequency, as shown in eqn. 5) [8], a good approximation is to consider the panel as a randomly vibrating area. The radiation intensity from such an area is shown in [9] to depend on the square of the mean velocity, hence the requirement is for constant velocity excitation. To achieve constant velocity with constant force, the mechanical impedance must be resistive. An infinite panel operating in a bending wave mode meets this criterion [10], where expressions for bending wave velocity $v(\omega)$, wave number $kp(\omega)$ and mechanical impedance Z_p are quoted below. Note that $v(\omega)$ is the in-plane velocity, and should not be confused with panel velocity u_p, normal to the panel surface.

$$v(\omega) = \omega^{0.5}\left(\frac{B}{\mu}\right)^{0.25} \tag{5}$$

$$k_p(\omega) = \omega^{0.5}\left(\frac{\mu}{B}\right)^{0.25} \tag{6}$$

$$Z_p = 8\sqrt{B\mu} \tag{7}$$

4 Mechanical model

To model a physical system, assumptions are required. Because the DML is considered to be in a state of random vibration, any existing panel motion is uncorrelated to any new applied input, and therefore appears as an infinite plate (see [11] for a definitive justification of a statistical approach to mechanical impedance). Additionally, because the panel has low mechanical loss, all the energy supplied to the panel is assumed dissipated by acoustic radiation. These assumptions have been shown to give useful results and measurements confirm that to calculate the radiated acoustic power, only the mechanical power delivered to the panel need be calculated [12]. If the impedance analogue is used, where voltage is analogous to force and velocity is analogous to current, the radiated pressure is proportional to the mean velocity in the panel.

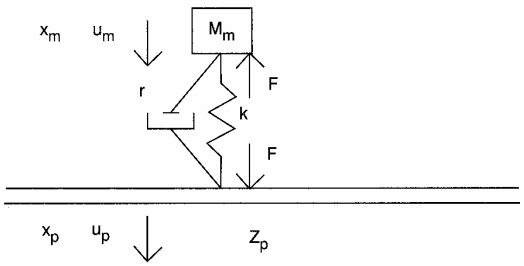


Fig. 2 *Mechanical elements and forces for panel driven by damped mass-spring oscillator*

Given that the DML is a resistance-controlled device and that the acoustic radiation need not be considered in detail, the equivalent circuit can be developed from Fig. 2. This structure represents a simplified version of the 'inertial magnet driver' application [13]. The coupled equations of motion are given in eqns. 8 and 9:

$$M_m\frac{d^2x_m}{dt^2} + r\left(\frac{dx_m}{dt} - \frac{dx_p}{dt}\right) + k\,(x_m - x_p) - F = 0 \tag{8}$$

$$Z_p\frac{dx_m}{dt} + r\left(\frac{dx_p}{dt} - \frac{dx_m}{dt}\right) + k(x_p - x_m) + F = 0 \tag{9}$$

If the driving force is assumed to be sinusoidal with angular frequency, ω, and using the same symbols to refer to the peak values of variables:

$$F(t) \equiv Fe^{j\omega t} \quad \text{(and similarly for } x_m \text{ and } x_p\text{)}$$

$$\omega^2 M_m x_m - j\omega r(x_m - x_p) - k(x_m - x_p) - F = 0 \tag{10}$$

$$j\omega Z_m x_p - j\omega r(x_m - x_p) - k(x_m - x_p) - F = 0 \tag{11}$$

or in matrix form, separating the stiffness, mass and resistance matrices:

$$(\underline{\underline{K}} - \omega^2\underline{\underline{M}} + j\omega\underline{\underline{R}})\,\underline{x} - \underline{F} = 0$$

or

$$\underline{x} = (\underline{\underline{K}} - \omega^2\underline{\underline{M}} + j\omega\underline{\underline{R}})^{-1}\,\underline{F} \tag{12}$$

where

$$\underline{\underline{M}} = \begin{pmatrix} M_m & 0 \\ 0 & 0 \end{pmatrix} \qquad \underline{\underline{K}} = \begin{pmatrix} k & -k \\ -k & k \end{pmatrix}$$

$$\underline{\underline{R}} = \begin{pmatrix} r & -r \\ -r & r \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & Z_p \end{pmatrix}$$

$$\underline{x} = \begin{pmatrix} x_m \\ x_p \end{pmatrix} \qquad \underline{F} = F\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

So the specific velocity, or mobility Y_p in the panel is given by:

$$Y_p = \frac{u_p}{F} = \frac{j\omega x_p}{F} \tag{13}$$

$$Y_p = j\omega(0 \quad 1) \times \begin{pmatrix} k - \omega^2 M_m + j\omega r & -k - j\omega r \\ -k - j\omega r & k + j\omega r + j\omega Z_p \end{pmatrix}^{-1}\begin{pmatrix} -1 \\ 1 \end{pmatrix} \tag{14}$$

$$Y_p = \frac{\omega^2 M_m}{(\omega^2 M_m(Z_p + r) - Z_p k) - j\omega(kM_m + rZ_p)} \tag{15}$$

By inspection, and noting that the velocity in the spring and damper is the difference between the velocities in the mass and panel, the equivalent circuit using the impedance analogue can be drawn as in Fig. 3. It is then a relatively straightforward task to verify that the ratio of panel velocity u_p to force F matches that given by the reciprocal of eqn. 15, i.e.

$$Zm_{eff} = Z_p\left(1 - \frac{k}{\omega^2 M_m}\right) + r + \frac{1}{j\omega}\left(k + \frac{rZ_p}{M_m}\right) \tag{16}$$

5 Practical implementation of the equivalent circuit for a moving-coil motor

Figs. 2 and 3 represent a DML panel driven by an idealised point source. If the motor system is considered to be a moving coil, M_m represents the mass of the magnet, cup and pole piece. The spring and damper assembly represent a means of attachment of the motor to the panel. To account for the effect of the coil, a mechanical mass M_c is added in series with Z_p that is equal to the mass of the

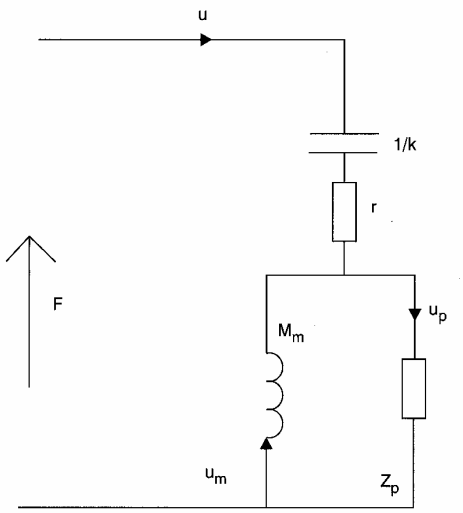


Fig. 3 *Impedance analogue model of DML panel*
$Zm_{eff} = F/u_p$
$u = u_p - u_m$

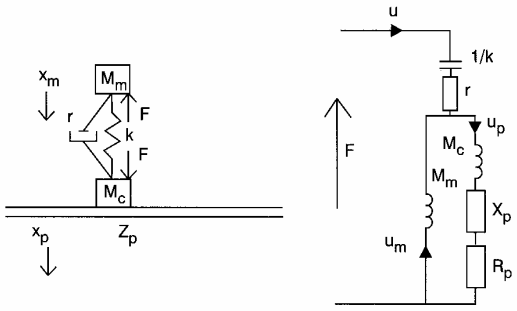


Fig. 4 *Impedance analogue model of DML panel with moving-coil motor system*
$Z_p = R_p + jX_p$

voice coil. Additionally, the impedance Z_p is only real for a point source and generally becomes complex for a finite diameter voice coil. The reactive component X_p is small except at high frequencies, where it should be recalled that pseudorandom fluctuations in X_p due to modal behaviour are not considered in this simple model. The high-frequency component of X_p is systematic and therefore of importance. Fig. 4 shows a model of such a system together with its equivalent circuit.

The effective mechanical impedance relating u_p to F for the model in Fig. 4 is

$$Zm_{eff} = Z'_p\left(1 - \frac{k}{\omega^2 M_m}\right) + r + \frac{1}{j\omega}\left(k + \frac{rZ'_p}{M_m}\right) \quad (17)$$

where

$$Z'_p = R_p + jX_p + j\omega M_c$$

However, at high frequencies eqn. 17 can be simplified, where if all terms involving negative powers of ω are ignored, $Zm_{eff} = Z'_p + r$, i.e.

$$Zm_{eff} \approx (R_p + r) + j(X_p + \omega M_c) \approx R_p + j\omega M_c \quad (18)$$

which gives the high-frequency limit f_{max} for the DML as

$$f_{max} \approx \frac{R_p}{2\pi M_c} \quad (19)$$

A similar simplification yields the low-frequency limit f_{min}, where, ignoring k,

$$\frac{1}{Zm_{eff}} \approx \frac{1}{R_p} + \frac{1}{j\omega M_m} \quad \text{so} \quad f_{min} \approx \frac{R_p}{2\pi M_m} \quad (20)$$

Alternatively, if the influence of M_m is small and stiffness dominates:

$$f_{min} \approx \frac{k}{2\pi R_p} \quad (21)$$

6 Modelling results

To complete the model, a gyrator and coil impedance must be included as in Fig. 5. To evaluate the model variables in the analysis, the complete electromechanical circuit is coded into a commercially available electroacoustic simulator (e.g. AkAbak™, formerly Panzer & Partner, now supported by New Transducers Limited, Huntingdon, UK). The electrical and mechanical domains are constructed from the mechanical equivalent circuit shown earlier, with the addition of the transfer characteristics of the moving-coil exciter. These parameters include the magnet moving mass, compliance, shove factor (usually written as '*BL*', where *B* is the magnet strength, and *L* is the conductor length. The authors have not used this in order to avoid confusion with the panel parameter, B) and voice-coil DC resistance, also the component values R_s and C_s are r and $1/k$, respectively.

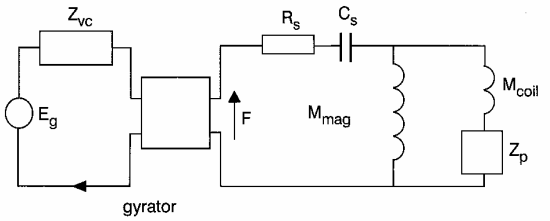


Fig. 5 *Complete electromechanical schematic for DML panel and exciter*

Solving the above circuit enables the mean driving-point velocity, and hence the sound pressure to be evaluated. The following results were obtained from such a model. A panel of relatively low mechanical impedance (~10kg/s) was used.

6.1 Terminal impedance

The reactive nature of the traditional moving mass loudspeaker is reflected in the terminal impedance, giving a classical low-frequency electrical resonance. Since the DML panel is approximately resistive, its terminal impedance is substantially flat. The two classes of loudspeaker are compared in Fig. 6. Notice the evidence of modal activity apparent in the measurement of the DML, which indicates the degree of approximation made in the model. Practical experience shows that these modes are less evident on higher impedance panels, or at higher frequencies. Interestingly, the cone loudspeaker example also shows evidence of modal activity at about 500Hz, which is not untypical, so is not entirely pistonic in its behaviour.

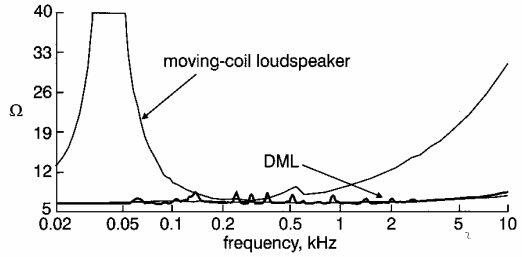


Fig. 6 *Terminal impedance of DML panel with exciter from model compared with measurement and cone drive unit of same nominal impedance*

6.2 Velocities

Fig. 7 shows plots of panel velocity overlaid with measured pressure, taken at 1m. The panel is truly modal from about

140Hz (the lowest mode is at about 45Hz), and self-baffling from about 250Hz. The high-frequency extension of this particular panel is actually better than predicted by the model and results from additional compliance between the voice coil and the panel. The model can readily be extended to include this effect. Ripples in the pressure response and the progressive attenuation below 250Hz are primarily due to diffraction. Some improvement in low-frequency accuracy can be obtained by modelling the lowest resonance of the panel, but the resulting increase in model complexity makes this unattractive.

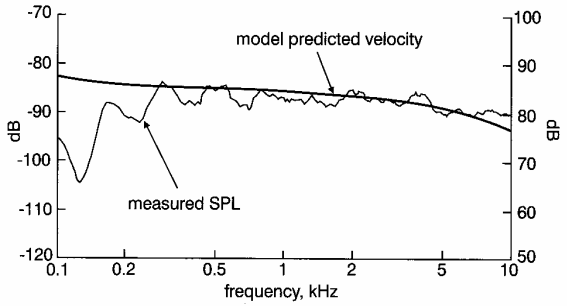


Fig. 7 *Velocity of DML panel and exciter from model, with measured SPL at 1m*

7 Conclusions

An electromechanical model has been presented which enables engineers to use existing software programs to investigate the application of DML technology to their acoustic problems. Although more advanced models have been developed [3, 14], this simple model offers an efficient alternative while retaining good accuracy. Given that a stiff, light panel can be designed to have an optimal modal distribution, together with low loss, it has been shown that the simplified model can accurately predict acoustic pressure and acoustic power, where it is necessary to calculate only the mean velocity in the panel.

The bandwidth of the DML is seen from eqns. 18–20 to depend only on the ratio of magnet mass, coil mass, and suspension stiffness to the panel mechanical impedance. The panel properties affect the sensitivity and frequency limits only via the mechanical impedance. It is possible to design a single DML to be substantially flat in pressure and power response over a wide audio bandwidth (exceeding 100Hz to 20kHz) without any electrical filters, a task impossible to achieve with conventional loudspeaker technology.

8 References

1 PANZER, J., and HARRIS, N.: 'Distributed-mode loudspeaker radiation simulation'. Proceedings of 105th AES convention, September 1998, (preprint 4783)
2 BANK, G.: 'The intrinsic scalability of the distributed-mode loudspeaker'. Proceedings of 104th AES convention, May 1998, (preprint 4742)
3 AZIMA, H., and HARRIS, N.: 'Boundary interaction of diffuse field distributed mode radiators'. Proceedings of 103rd AES convention, September 1997, (preprint 4635)
4 HARRIS, N., and HAWKSFORD, M.: 'Measurement and simulation results comparing the binaural acoustics of various direct radiators'. Proceedings of 107th AES convention, September 1999, (preprint 5015)
5 HARRIS, N., FLANAGAN, S., and HAWKSFORD, M.: 'Stereophonic localization in rooms, comparing conventional and distributed-mode loudspeakers'. Proceedings of 105th AES convention, September 1999, (preprint 4794)
6 HARRIS, N., and FLANAGAN, S.: 'Stereophonic localisation in rooms, comparing the distributed-mode loudspeaker with conventional two-way cone-based loudspeakers'. Institute of Acoustics, Proceedings of reproduced sound 14, Windermere, October 1998
7 KINSLER, L.E., FREY, A.R., COPPENS, A.B., and SANDERS, J.V.: 'Fundamentals of acoustics' (Wiley, 1982, 3rd edn.) Section 8.12
8 MORSE, P.M., and INGARD, K.U.: 'Theoretical acoustics' (McGraw Hill, 1968, first Princeton University Press edn.), Section 5.3.1 and 5.1.6
9 MORSE, P.M., and INGARD, K.U.: 'Theoretical acoustics' (McGraw Hill, 1968, first Princeton University Press edn.), Section 7.4.28
10 MORSE, P.M., and INGARD, K.U.: 'Theoretical acoustics' (McGraw Hill, 1968, first Princeton University Press edn.), Section 5.3.19
11 CREMER, L., HECKL, B., and UNGAR, B.B.: 'Structure-borne sound' (Springer-Verlag, 1988, 2nd English edn.), Chap. IV.4, especially pp 327–333
12 GONTCHAROV, V., HILL, N., and TAYLOR, V.: 'Measurement aspects of distributed mode loudspeakers'. Proceedings of 106th AES convention, May 1999
13 'NXT White Paper'. (C) New Transducers Ltd., 1996
14 PANZER, J., and HARRIS, N.: 'Distributed mode loudspeaker simulation model'. Proceedings of 104th AES convention, May 1998, (preprint 4739)

PAPERS

Distortion Reduction in Moving-Coil Loudspeaker Systems Using Current-Drive Technology*

P. G. L. MILLS** AND M. O. J. HAWKSFORD

University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK

The performance advantages of current-driving moving-coil loudspeakers is considered, thus avoiding thermal errors caused by voice-coil heating, nonlinear electromagnetic damping due to $(Bl)^2$ variations, and high-frequency distortion from coil inductive effects, together with reduced interconnect errors. In exploring methods for maintaining system damping, motional feedback is seen as optimal for low-frequency applications, while other methods are considered. The case for current drive is backed by nonlinear computer simulations, measurements, and theoretical discussion. In addition, novel power amplifier topologies for current drive are discussed, along with methods of drive-unit thermal protection.

0 INTRODUCTION

The moving-coil drive unit is by far the most widely used electroacoustic transducer in both high-performance studio and domestic audio installations, as well as in general-purpose sound reinforcement. Consequently it has attracted numerous studies to investigate its inherent distortion mechanisms (see, for example, [1]–[11]), which as a consequence are well understood. Much work has also been carried out on improving drive-unit linearity by the application of motional feedback techniques, which provide a useful enhancement in performance at low frequencies. Improvements to the basic regime of motional feedback have been made by including an additional current feedback loop [12], [13], which is reported to reduce high-frequency distortion. This method is a specific implementation of what we will term *current drive*, a subject that, it is felt, has not received the attention it deserves.

This paper therefore aims to explore in detail the benefits of current drive in reducing the dependence of drive-unit performance on motor system nonlinearities, in particular the voice-coil resistance which undergoes significant thermal modulation.

In a conventional voltage-driven system (one where the power amplifier output voltage is regarded as the information-representing quantity), the current is initially limited by the series elements of voice-coil resistance and inductance, together with the interconnect and amplifier output impedance. A force related to the current in the system then acts on the drive unit moving elements as a result of the motor principle, and once motion occurs, an electromotive force is induced in the coil to oppose the applied signal voltage, thus constraining the magnitude of current flow. The accuracy to which the drive-unit velocity responds to the applied signal is, therefore, dependent on the series elements in the circuit, and any signal-related changes in their value will result in distortion.

The voice-coil resistance is of specific concern, as it is usually a dominant element. As a result of self-heating in excess of 200°C, a significant increase in coil resistance occurs of typically 0.4%/°C for copper, leading to sensitivity loss, lack of damping, and crossover misalignment. In their paper, Hsu et al. [6] concluded that a satisfactory method of compensating for this effect had yet to be found.

At higher audio frequencies, the coil inductance also becomes significant, resulting in a loss of sensitivity. In addition, the inductance suffers dynamic changes with displacement, providing a distortion mechanism which is further complicated by eddy current coupling to the pole pieces in the magnetic circuit [14, pt. 1]. A further problem is distortion mechanisms at the amplifier–loudspaker interface, such as interconnect errors [14, pt. 4] and interface intermodulation distortion

* Manuscript received 1988 February 17.

** Now at Tannoy Ltd., Rosehall Industrial Estate, Coatbridge, Strathclyde, ML5 4TF, UK.

[15]–[17].

To overcome these limitations, the drive unit should be current rather than voltage controlled and interfaced directly to a power amplifier configured as a current source, thus offering a high output impedance. The performance advantages of this technique are discussed in detail, supported by computer simulation of the nonlinear system together with objective measurement on a prototype two-way active loudspeaker system.

To complement this study, the application of both motional feedback and noninteractive frequency response shaping as a means of aligning the drive unit Q to the required value is discussed. Finally, the topic of current source power amplifier design is considered along with the presentation of some novel types of circuit topology, while the subject of drive-unit protection under current drive is also examined.

The technique of current drive in active loudspeaker systems is seen as being of particular importance in view of the performance advantages demonstrated over conventional systems in terms of both reduced linear and nonlinear distortion. For high-quality system design, current drive is seen as the more logical methodology, with voltage drive appearing as the result of established practice and convenience.

1 LOUDSPEAKER PERFORMANCE UNDER VOLTAGE DRIVE

In order to establish a performance reference, this section considers motor system linearity for a moving-coil drive unit under conventional voltage drive. For the tests, a Celestion SL600 135-mm-diameter bass–midrange driver was used, mounted in its enclosure. It was chosen partly due to the excellent cone and surround behavior, meaning that the distortion contribution of these elements is small.

To enable performance predictions under general signal excitation to be made, the variation of parameters with coil displacement was measured. This is shown in graphic form, in Fig. 1 for Bl product, compliance, and coil inductance. The linear parameters for the model are given in Table 1, which explains the terminology and also the equivalence between the electrical model and the mechanical model used. The approach broadly

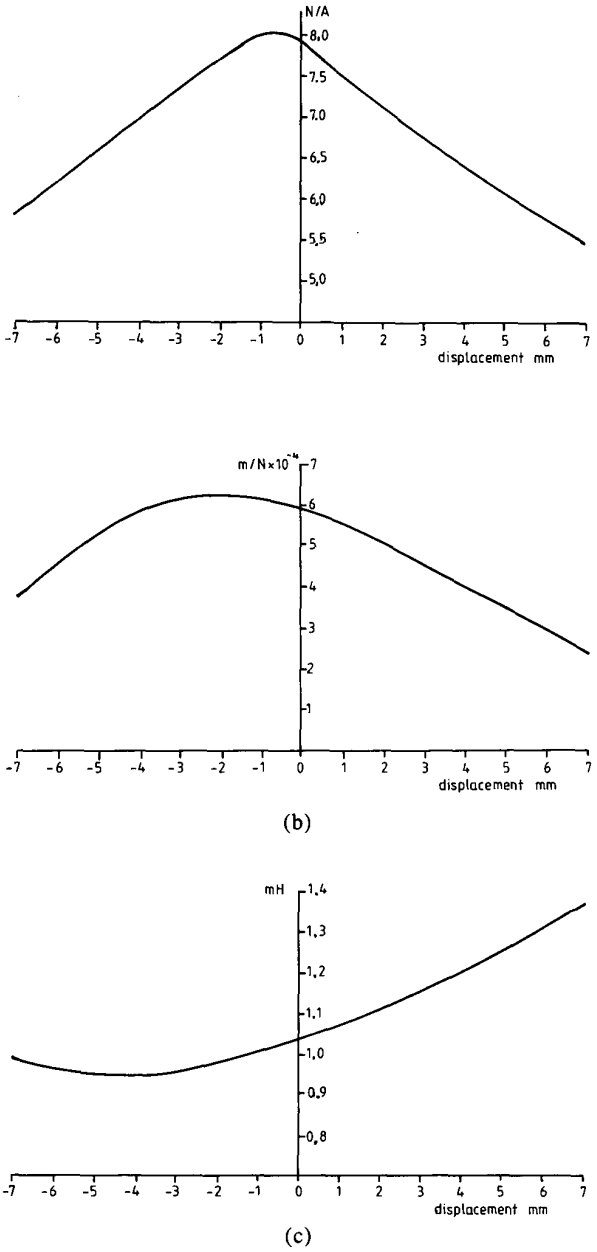


Fig. 1. Variation of model parameters with displacement. Negative displacement indicates motion toward magnet. (a) Bl product. (b) Mechanical compliance. (c) Electric coil inductance.

Table 1. Model parameters for example drive unit.

Parameter	Electrical model	Mechanical model
Voice-coil resistance	$R_e = 7.0\ \Omega$	$R_{me} = (Bl)^2/R_e$ kg/s
Voice-coil inductance	L_e*	$C_{me} = L_e/(Bl)^2$ m/N
Enclosure compliance	$L_{cmb} = C_{mb}(Bl)^2$	$C_{mb} = 750 \times 10^{-6}$ m/N
Suspension compliance	$L_{cms} = C_{ms}(Bl)^2$	C_{ms}* m/N
Moving mass	$C_{mes} = M_{ms}/(Bl)^2$	$M_{ms} = 0.0183$ kg
Mechanical resistive losses	$R_{es} = (Bl)^2/R_{ms}$	$R_{ms} = 2.4336$ kg/s
Source impedance	Z_g (assume zero)	$Z_{mg} = (Bl)^2/Z_g$ kg/s

Bl = force factor (N/A)*

* Indicates nonlinear elements.

follows that of Small [18], except that a mechanical model is used in preference to the acoustic one. Fig. 2(a) shows the equivalent electrical model for the drive unit, connected to an amplifier and interconnect of series source impedance Z_g, showing the mechanical impedance as a lumped quantity Z_m. Analysis of this model gives the transfer function between amplifier output voltage and cone velocity,

$$u = \frac{V_0 Bl}{Z_m [Z_s + (Bl)^2/Z_m]} \quad (1)$$

where

u = cone velocity, meters per second
V_0 = amplifier source voltage, volts
B = flux density for motor system, tesla
l = coil length in field B, meters
Z_m = lumped mechanical impedance, kilograms per second
Z_s = lumped electric impedance (Z_g, R_e, and sL_e), ohms.

Referring the mechanical impedance to the "primary" of the Bl transformer to show its constituents gives the electrical model of Fig. 2(b), while referring the electrical parameters to the "secondary" results in the mechanical model of Fig. 2(c). Both these models are useful in the forthcoming discussion, although emphasis is placed on the mechanical system.

The mechanical model forms the basis of a transient analysis procedure, which can readily incorporate nonlinear parametric variations. The details of this approach

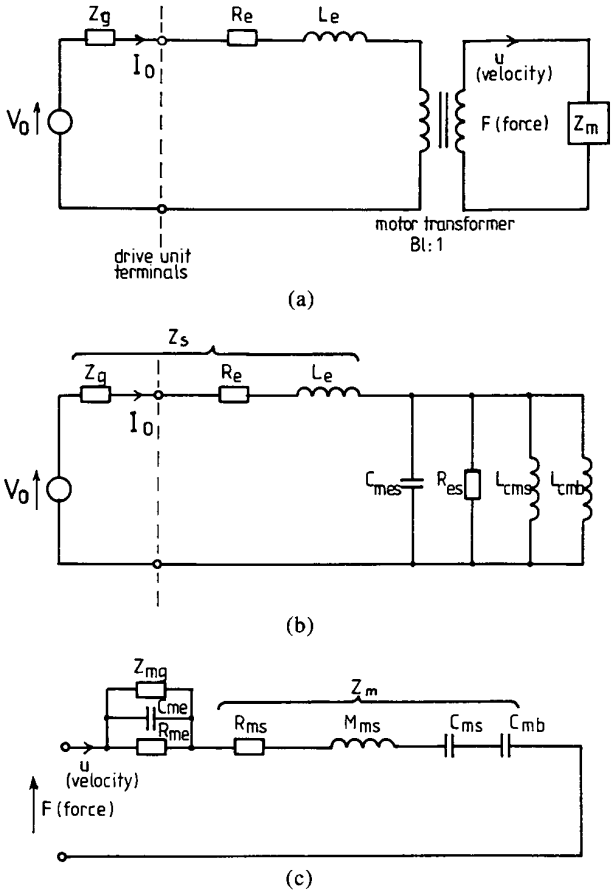


Fig. 2. Modeling of drive unit in sealed enclosure, under voltage drive. (a) Basic electromechanical model. (b) Electrical model. (c) Mechanical model.

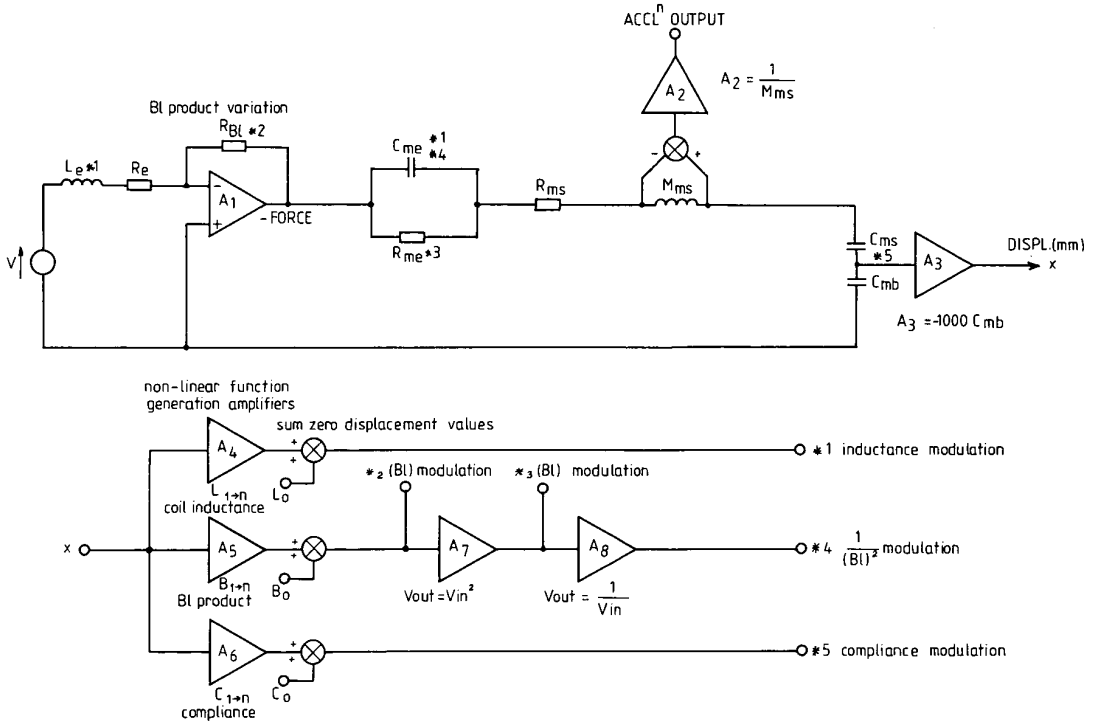


Fig. 3. Simplified nonlinear model for voltage-driven simulation.

were given in an earlier paper [19], where it was seen to avoid the approximations forced by an analytical solution. It may be contrasted to that of Kaizer [3], where drive-unit nonlinearity was modeled by Volterra series expansion. Fig. 3 shows the simplified circuit for the computer model, where it is assumed that the driving amplifier has zero source impedance. The output quantity acceleration is derived from the voltage across the moving mass element M_{ms}, while a signal proportional to displacement is obtained from the voltage across the enclosure compliance C_{mb}. This displacement voltage is used to drive three nonlinear amplifiers A_4, A_5, and A_6, whose transfer characteristics are expressed as polynomials, representing the measured variation in coil inductance, Bl product, and compliance, respectively. A technique was adopted whereby the three nonlinear functions were first represented by a Fourier series from which the corresponding polynomials were generated. A 30th-order approximation to each function was deemed necessary to avoid undue error. The zero displacement values for these parameters were then summed by constant factors L_0, B_0, and C_0. Each of the modulation outputs in the diagram is coded by an asterisk and number to indicate which circuit parameters it modulates.

To produce distortion predictions from this model, a sine wave input is used and the system allowed to reach steady state. A single cycle is then sampled as input data to a fast Fourier transform, which indicates the relative amplitude of the distortion harmonics.

At 100 Hz, with the source voltage chosen to give a current of 1 A peak, a reasonable relation between theoretical and measured distortion spectra can be seen in Fig. 4. The model slightly overestimates most distortion components, probably due to errors in measuring the nonlinear parameters. However, at high frequency the model does not prove usable, due to factors such as the complicated nature of eddy current losses and hysteresis effects in the magnetic circuit. The measured 3-kHz at 1A peak distortion spectra are shown in Fig. 5, while intermodulation products between 50-Hz and 1-kHz sine wave inputs of equal amplitude are shown in Fig. 6.

The effect of voice-coil heating is a major problem under voltage drive, and it is interesting to note the severe difficulty in obtaining these measurements due to the sensitivity loss and frequency response errors which occur as the coil heats up. A further problem caused by heating is that of crossover misalignment in the case of passive systems. To illustrate this effect, Fig. 7 shows an idealized two-way second-order crossover aligned to 3.4 kHz. The drive units are represented by resistive elements, and the overall system transfer function is evaluated by effectively subtracting the high-pass and low-pass outputs. Fig. 8 then compares the system transfer function arising from coil heating to 200°C with the intended response at 20°C. Although oversimplistic, this model does show that large errors can result.

Errors due to interconnect effects are also seen to be of importance. Measurements of the error across a selection of 5-m interconnects have revealed errors up to 15 dB below the main signal. It is worth noting that the error is a function of drive-unit–crossover impedance and, while mainly linear, also contains a nonlinear component due to the nonlinear nature of the load.

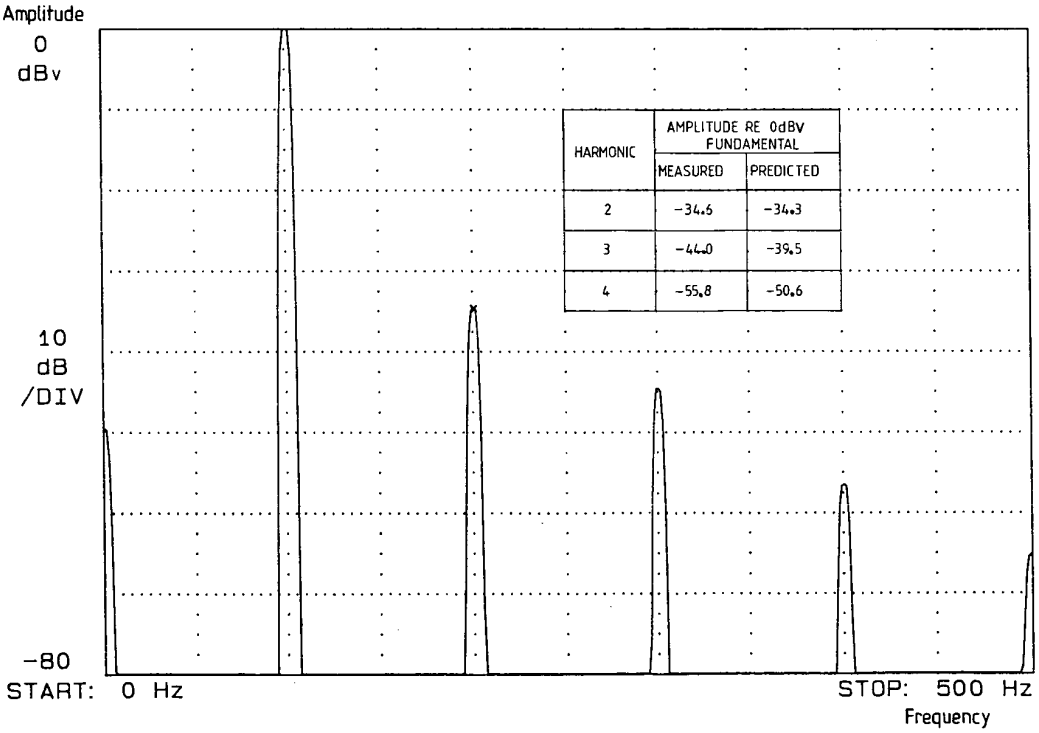


Fig. 4. Measured 100-Hz harmonic distortion, voltage driven. Table compares with predicted result.

2 THE CASE FOR CURRENT DRIVE

To assess the performance advantages of a current-driven moving-coil drive unit, a similar procedure is adopted to that of Sec. 1, though a current source is substituted for the voltage source, with output impedance assumed infinite. An immediate consequence of this strategy is that the series elements of coil resistance, coil inductance (with attendant eddy current losses), and interconnect lumped series elements, together with *Bl* and the lumped mechanical impedance no longer influence the instantaneous driving current. The significance of this observation is best illustrated by examining the current-driven velocity transfer function,

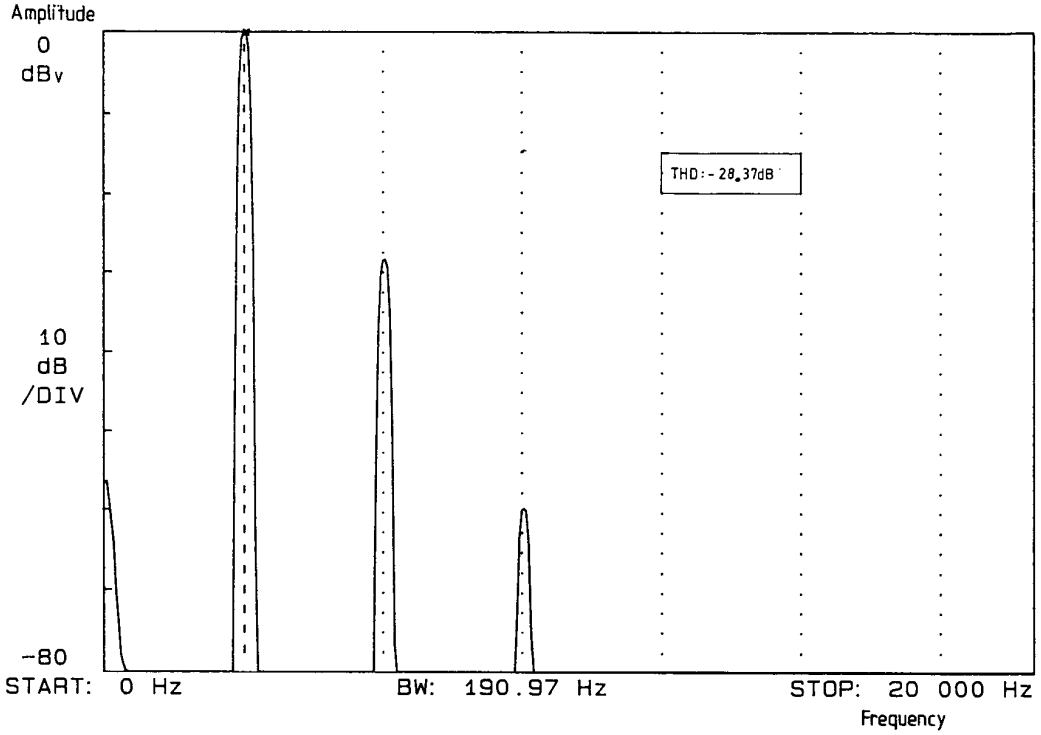


Fig. 5. Measured 3-kHz harmonic distortion, voltage driven.

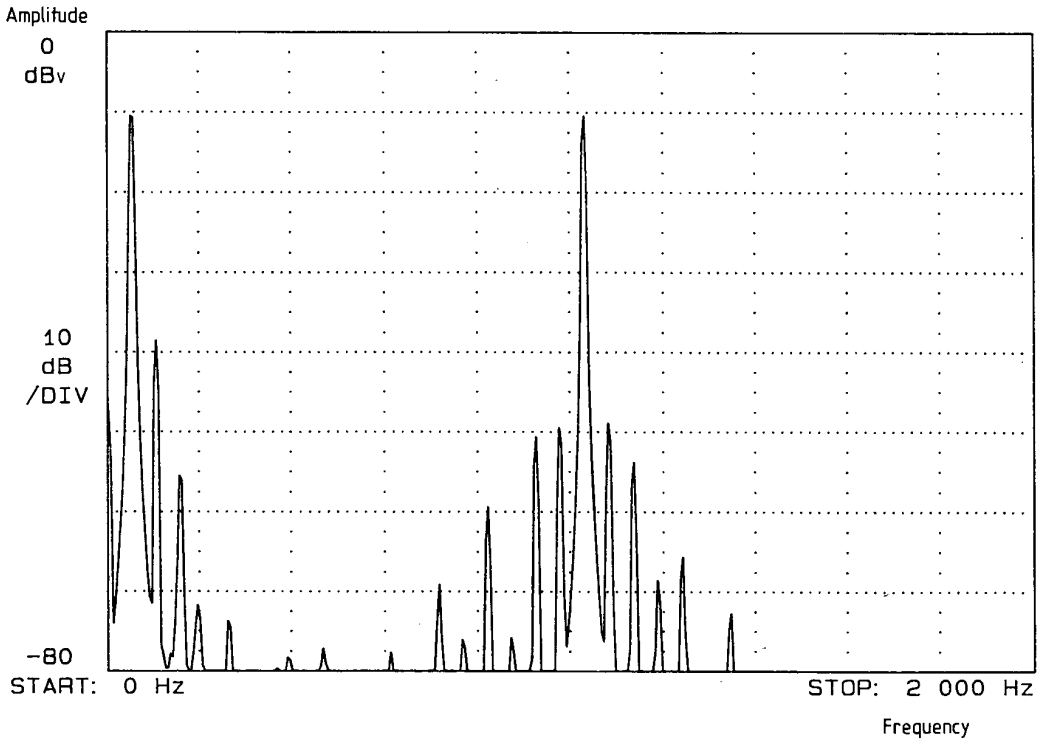


Fig. 6. Measured 50-Hz and 1-kHz intermodulation distortion, voltage driven.

$$u = \frac{I_0 Bl}{Z_m} \quad (2)$$

where I_0 is the amplifier output current in amperes.

Comparison with the voltage-driven case, Eq. (1), shows that for current drive, the transfer function is of a simpler form, independent of the terms Z_s and $(Bl)^2$. It is therefore anticipated that lower distortion will result from elimination of the term $[Z_s + (Bl)^2/Z_m]$. Performance is therefore free of any linear and nonlinear contributions from Z_s, the $(Bl)^2$ term, and shows a reduced dependence on compliance nonlinearity within Z_m, together with any frequency-dependent nonlinear interactions. The mechanical model for the drive unit and enclosure is then reduced to that of Fig. 9.

To demonstrate this claim, a transient analysis at 100 Hz with 1-A peak drive current was performed, using the nonlinear model of Fig. 10. A reasonable match between measured and predicted distortion is again obtained, as shown by Fig. 11.

Comparing this result with the voltage-driven case (Fig. 4) shows a measured and predicted distortion reduction of around 9 dB for the second harmonic, with third- and fourth-order products being reduced by between 3 and 7 dB, depending on whether the measured or the predicted values are taken (the predicted results yielding the better distortion reduction).

Regarding the 3-kHz at 1A peak measurement given in Fig. 12, this shows a substantial reduction of over 26 dB to the voltage-driven result in Fig. 5. Likewise, the 50-Hz–1-kHz intermodulation distortion is improved, as indicated by Fig. 13.

These results show the importance of eliminating the distortion contributions of the $(Bl)^2$ and Z_s terms in a relative comparison between current drive and voltage drive. Thus at the high-frequency end of the drive unit's operating range, the elimination of performance dependence on coil inductance modulation and eddy current losses is seen to be a valuable asset. Further, the current-driven system is completely free from any voice-coil thermal effects. Although the argument is based on a bass–midrange drive unit, with significant cone displacement, tweeters were found to benefit also, with a more modest 3–7-dB measured distortion reduction across the band, along with the elimination of coil-heating effects.

Finally, the performance independence on linear interconnect errors is also welcome when a low shunt capacitance cable is chosen—a high resultant series inductance being of no significant consequence.

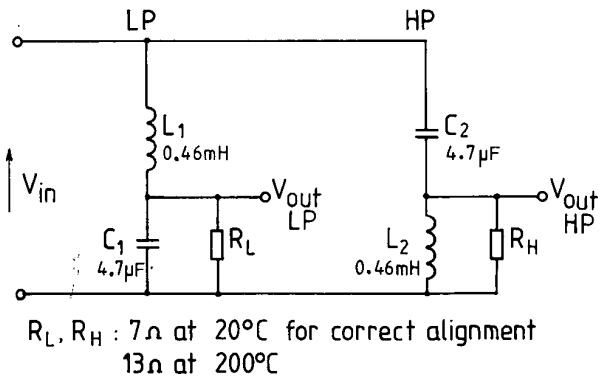


Fig. 7. Idealized two-way system with second-order crossover.

3 DRIVE-UNIT TRANSFER FUNCTION ALIGNMENT UNDER CURRENT DRIVE

Small signal analysis reveals that under current drive, there is a change in frequency response compared with the voltage-driven case, the principal cause being the loss in electromagnetic damping from the low-impedance voice-coil circuit. Consequently, the drive unit Q at fundamental resonance rises to that determined by the mechanical parameters—generally too high for optimal system alignment. To illustrate this, Fig. 14 compares the measured frequency responses of our example drive unit under both current drive and voltage drive. The rise in output around the fundamental resonance under current drive should be noted, along with a reduction in high-frequency rolloff due to the voice-coil inductance no longer appearing in the system transfer function.

In order to realign the acoustic transfer function, three methods have been investigated.

3.1 Electronic Equalization Using Open-Loop Compensation

The addition of a low-level equalizer to redefine the low-frequency alignment of a drive unit under voltage drive is a well-documented technique [20], [21]. The approach is equally applicable to current drive. If the drive-unit transfer function is of the form

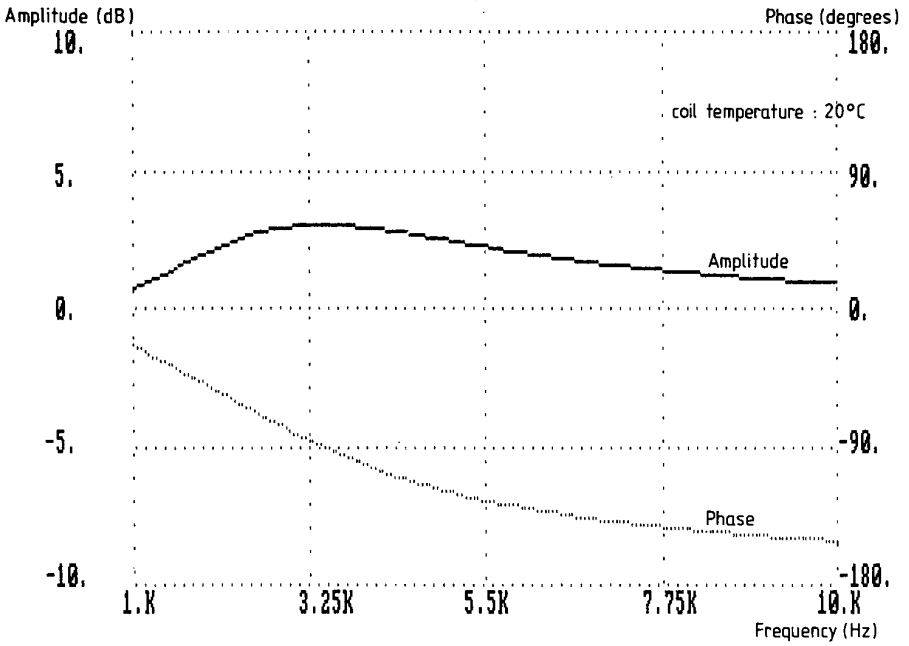
$$G(s) = \frac{s^2 T_s^2}{s^2 T_s^2 + sT_s/Q_m + 1}$$

where T_s is the time constant of fundamental resonance, in seconds, and Q_m is the mechanical drive-unit Q, and the desired low-frequency target alignment is written

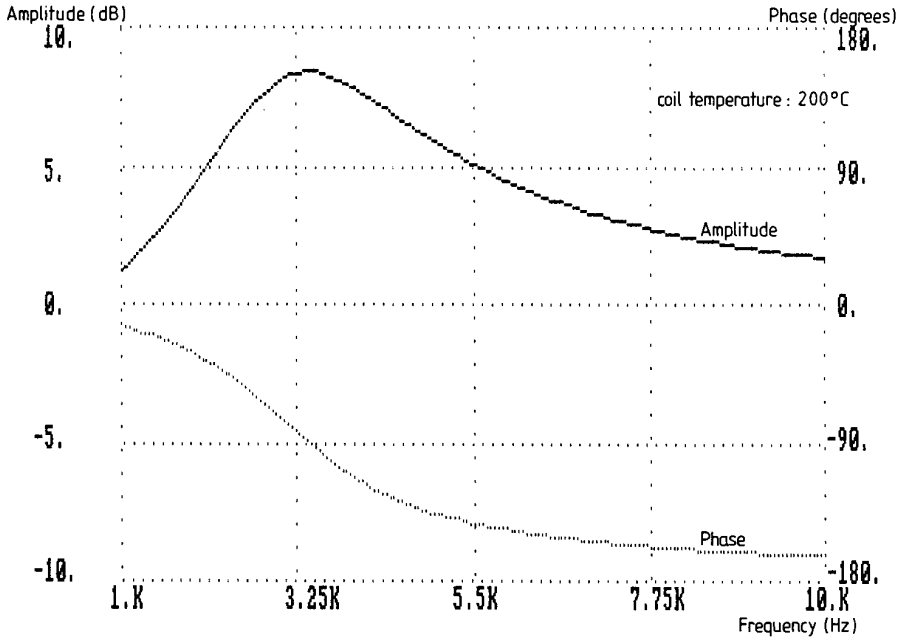
$$G_t = \frac{s^2 T_c^2}{s^2 T_c^2 + sT_c/Q_c + 1}$$

where T_c is the redefined system time constant, in seconds, and Q_c is the compensated Q value, then, assuming a second-order low-frequency alignment is retained, the equalizer transfer function is defined:

$$X(s) = \left(\frac{s^2 T_c^2}{s^2 T_c^2 + sT_c/Q_c + 1}\right) \times \left(\frac{s^2 T_s^2 + sT_s/Q_m + 1}{s^2 T_s^2}\right)$$



(a)



(b)

Fig. 8. Summed high-pass and low-pass outputs for idealized two-way system. (a) at 20°C. (b) at 200°C.

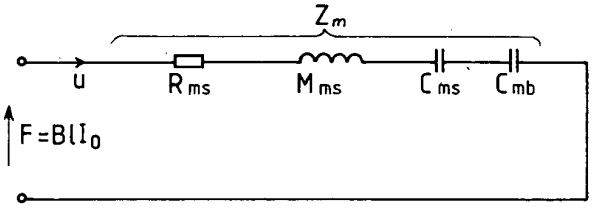


Fig. 9. Mechanical model of drive unit in sealed enclosure under current drive.

that is,

$$X(s) = \frac{s^2T_c^2 + sT_c^2/T_sQ_m + T_c^2/T_s^2}{s^2T_c^2 + sT_c/Q_c + 1} . \quad (3)$$

The equalizer used for experimental purposes is represented by the cascaded integrator structure of Fig. 15. Comparing the transfer function of this system with $X(s)$ in Eq. (3), the time constants, are

$$T_1 = Q_cT_c$$

$$T_2 = \frac{T_c}{Q_c}$$

and the summing constants,

$$C_1 = 1$$

$$C_2 = \frac{Q_cT_c}{Q_mT_s}$$

$$C_3 = \frac{T_c^2}{T_s^2} .$$

This gives the ability to redefine the system Q and, if required, provide low-frequency extension.

Computer simulation of the equalizer and the example drive-unit–enclosure combination shows in Fig. 16(a) the overall system response for a Q realignment to 0.7071 with no resonant frequency shift, while Fig. 16(b) shows the effect of a resonance realignment to 40 Hz, with $Q = 0.7071$.

The disadvantage of this approach is the sensitivity to drive-unit mechanical parameter changes. To investigate this effect, the drive-unit mechanical parameters were subjected to ±20% tolerance and a Monte Carlo analysis based on 25 trials carried out to show the effect of random parametric variations within this range. The results reveal a 2-dB response error standard deviation around the area of fundamental resonance in both cases. However, in practice, mechanical parameter variations are likely to be better controlled with a well-engineered drive unit.

A novel technique of altering low-frequency realignment, which has been described in [22], is the "ace bass" system after Stahl. This method relies on

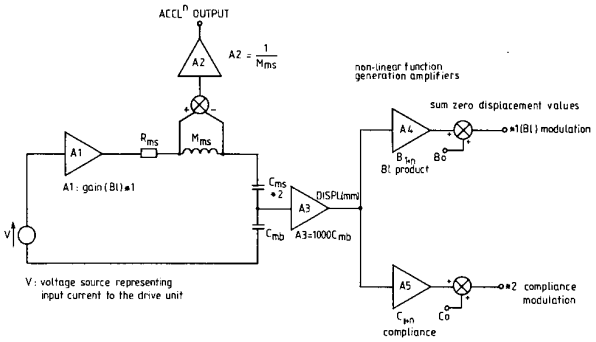


Fig. 10. Simplified nonlinear model for current-driven simulation.

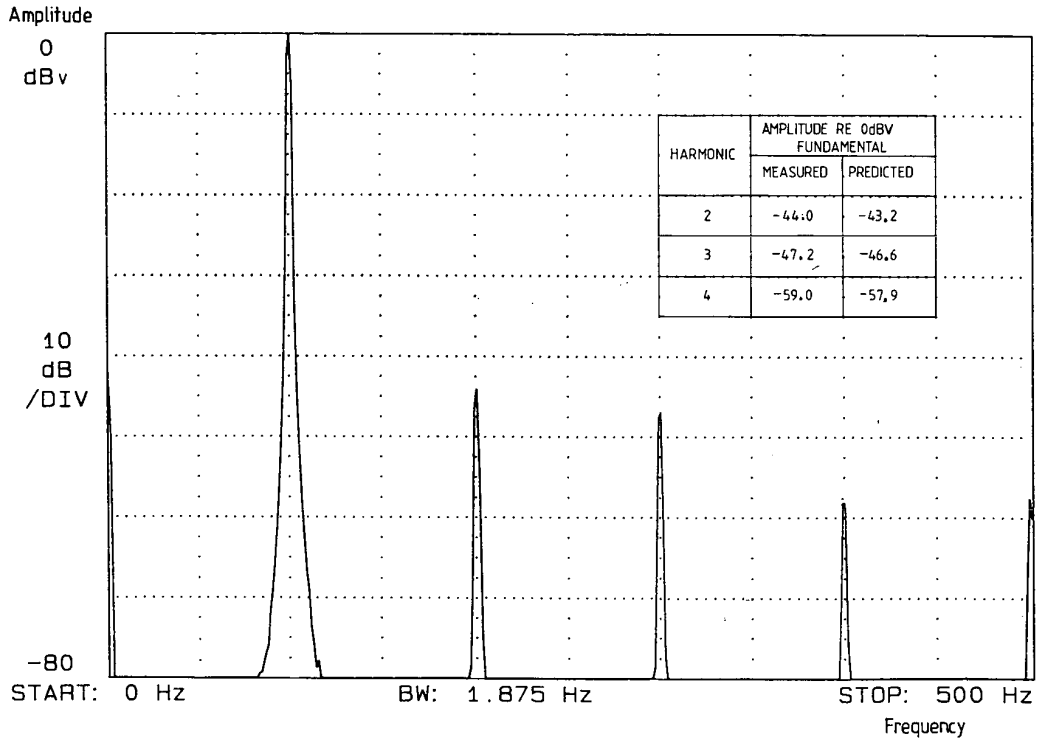


Fig. 11. Measured 100-Hz harmonic distortion, current driven. Table compares with predicted result.

providing the power amplifier with both a negative output resistance to cancel the drive-unit resistance and a synthesized parallel reactance in effect to modify the drive-unit mechanical parameters. While this technique does exhibit insensitivity to mechanical parameter variations (less than 1-dB standard deviation error on the same basis as the open-loop compensator for 40-Hz realignment), it does, as the author admits, incur problems due to voice-coil heating. The error for our example drive unit is shown in Fig. 17, by computer simulation with the voice-coil temperature at 20°C (reference) and increased to 200°C, where the low-

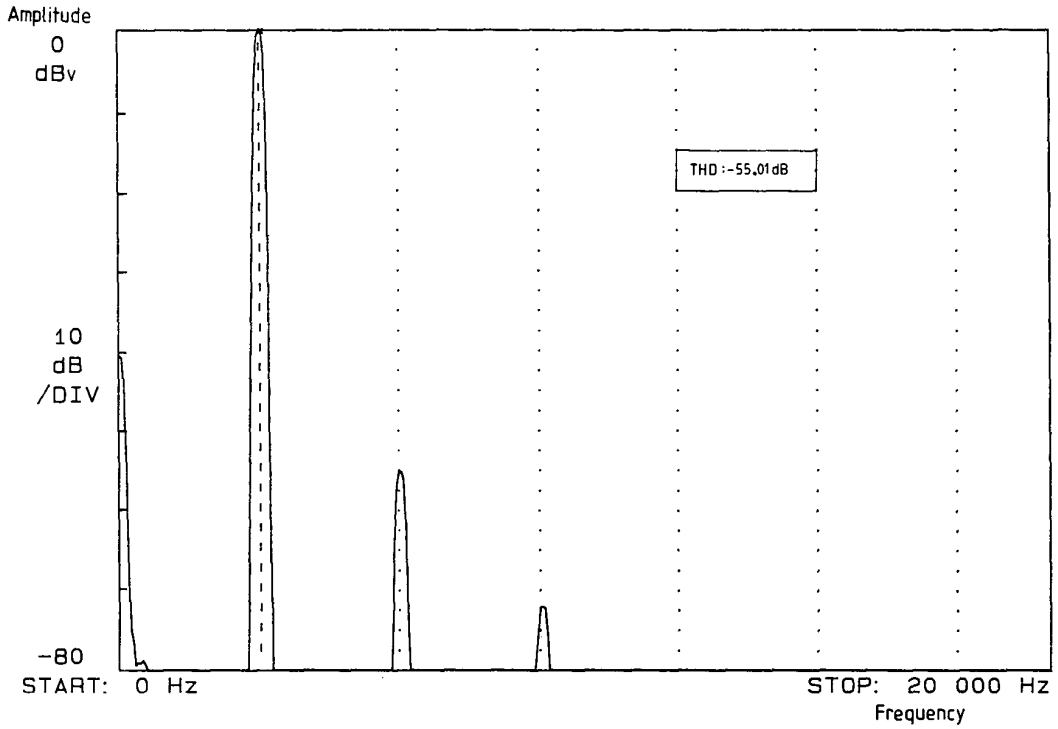


Fig. 12. Measured 3-kHz harmonic distortion, current drive.

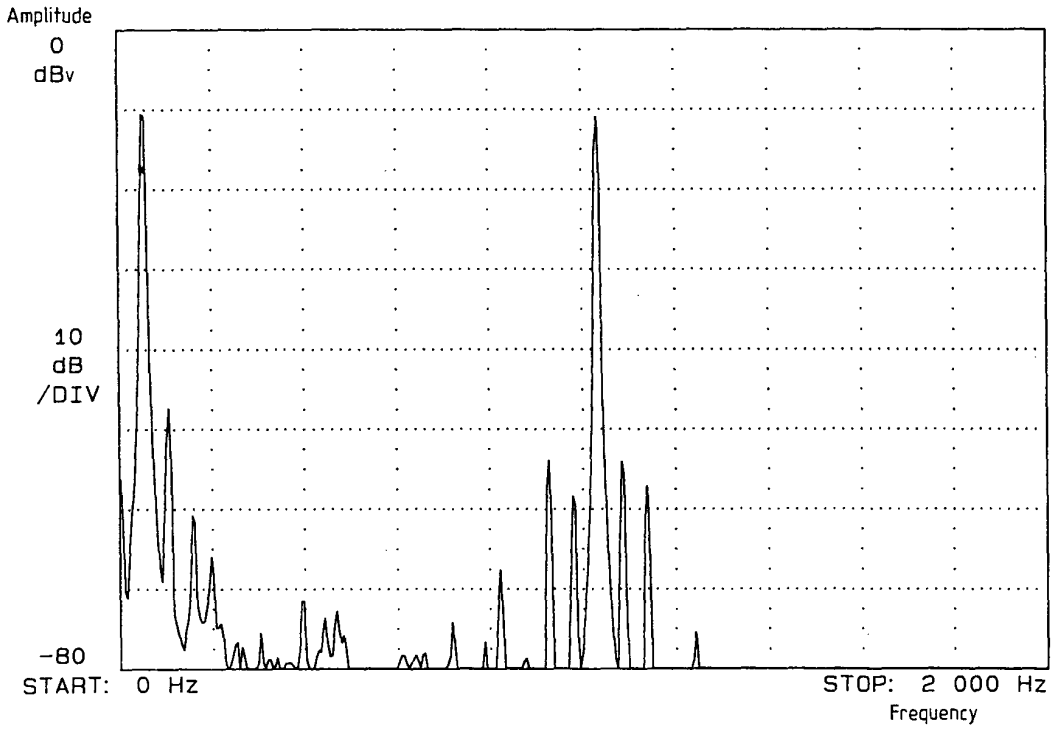
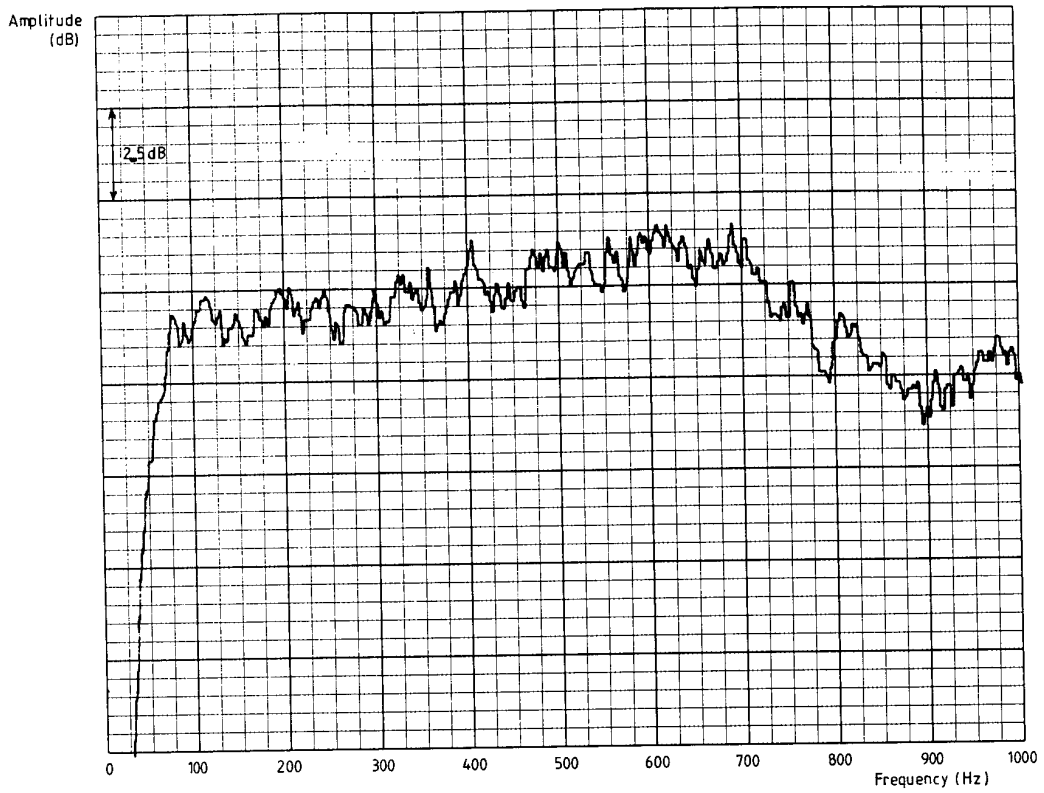
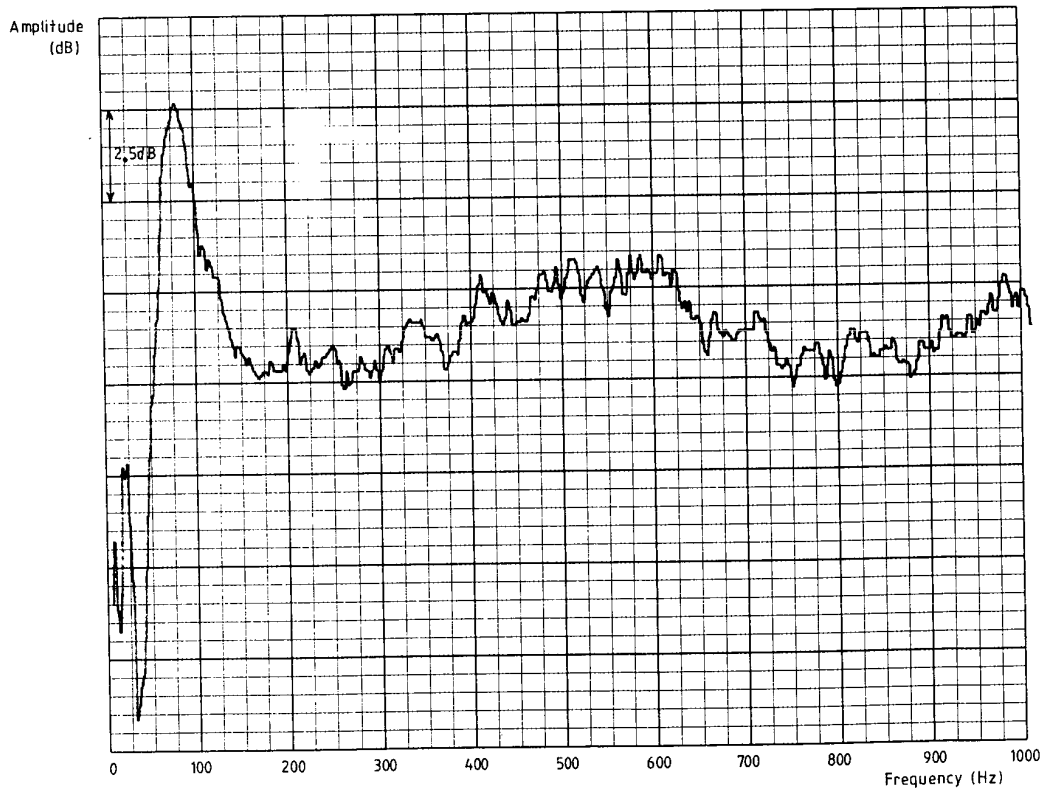


Fig. 13. Measured 50-Hz and 1-kHz intermodulation distortion, current driven.



(a)



(b)

Fig. 14. Measured frequency response of example drive unit. (a) Voltage drive. (b) Current drive.

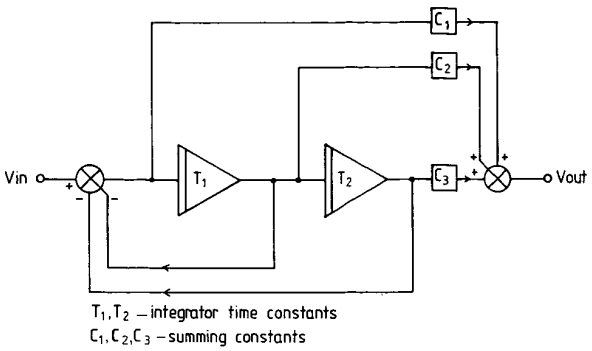
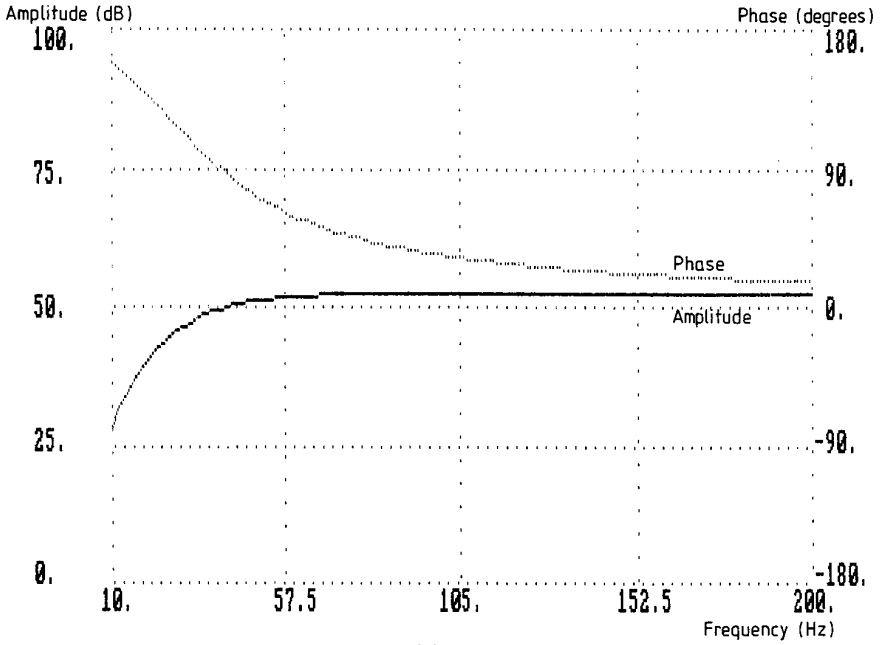


Fig. 15. Representation of open-loop equalizer.

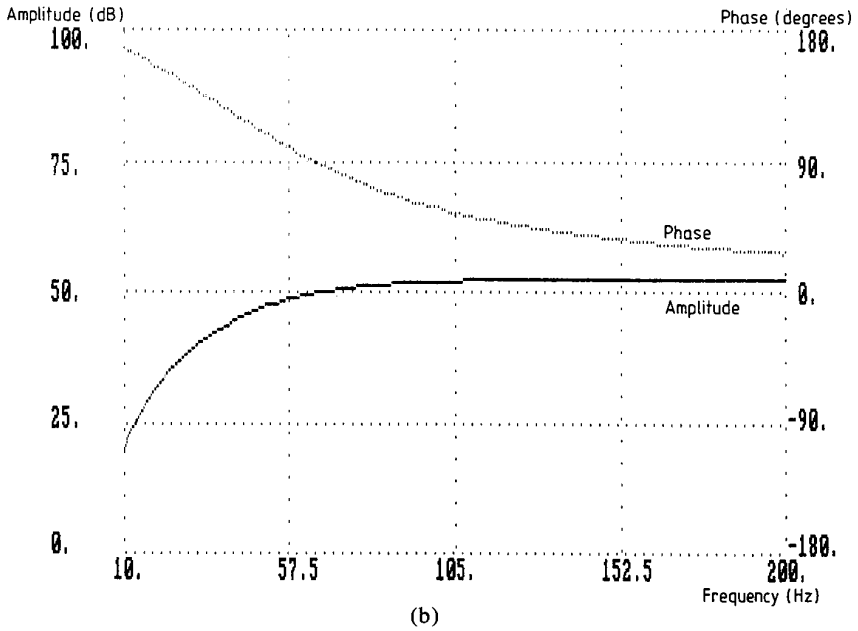
frequency alignment was set to 40-Hz resonance with Q = 0.7071. Where low-frequency extension is required, the extra power needed to combat the drive unit's falling response means that coil heating effects are particularly troublesome, hence reinforcing the need for current drive in this type of application.

3.2 Motional Feedback

Motional feedback is considered the optimal method for Q alignment of low-frequency drive units under current control and was consequently incorporated into the prototype development system. Our earliest reference to the technique is due to Voight in 1924 [23],



(a)



(b)

Fig. 16. Simulated current-driven system response with equalizer. (a) No resonance change (65 Hz), Q = 0.7071. (b) Resonance lowered to 40 Hz, Q = 0.7071.

where the lack in damping from an open-loop tube output stage led to similar problems as faced under pure current drive. The approach was later abandoned for general use when Black [24] formalized negative feedback techniques and amplifier output impedance could be reduced. Since then there has been much interest in motional feedback in high-performance applications, using a variety of sensing methods to obtain velocity, displacement, or acceleration feedback [25]–[28].

The method selected in this study was to wind a sensing coil over the primary drive coil for reasons of cost effectiveness, with little additional complexity over a standard drive unit. The penalty of this mechanical simplicity is that as well as generating a signal proportional to cone velocity, there is also transformer coupling that induces an error from the driving coil into the sensing coil. Methods used to deal with this effect have included the use of additional neutralizing coils [29], [30]. In this case, the rather different method of electronic compensation has been adopted. Fig. 18 shows how this has been achieved, together with a

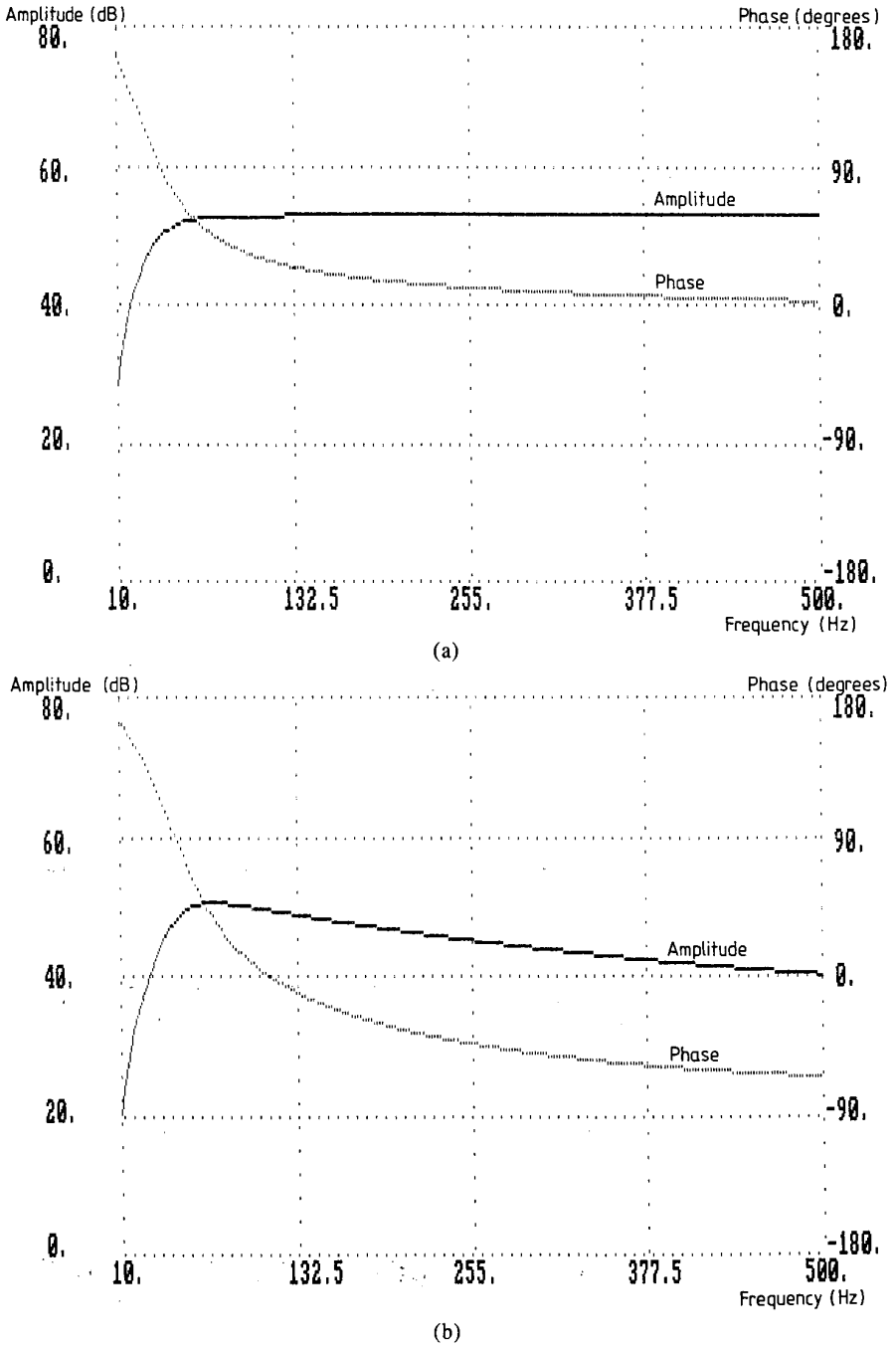


Fig. 17. "Ace bass" system frequency response simulations. Resonance set to 40 Hz, $Q = 0.7071$. (a) Voice-coil temperature 20°C. (b) Voice-coil temperature 200°C.

block diagram of the prototype system. So that the sensing coil does not reintroduce thermal errors, it must be interfaced to a high-input impedance buffer amplifier. The coupling error compensator consists of a filter matched to the transfer function of the coupling error characteristic, which increases by around 15 dB per decade up to 1 kHz at zero coil displacement. Some change in both the magnitude and the slope of this error is apparent with coil displacement, together with an unpredictable response above 1 kHz. To reduce this high-frequency residual error, the second-order low-pass filter at 1 kHz in the velocity feedback loop proves effective and in any event is required to maintain loop stability. Measurements have shown no increase in high-frequency distortion (at 3 kHz) over the open-loop case, indicating the effectiveness of this method.

To analyze the system, consider the simplified representation of Fig. 19, consisting of transconductance power amplifier, drive unit with sensing coil, and feedback path. The output voltage from the sensing coil V_s is written

$$V_s = (Bl)_s u$$

where $(Bl)_s$ is the sensing coil Bl product, in newtons per ampere, and u is the cone velocity, in meters per second. Also,

$$I_0 = [V_{in} - k(Bl)_s u]\, g_m$$

where

I_0 = amplifier output current, amperes
V_{in} = input voltage, volts
k = feedback constant
g_m = amplifier transconductance, siemens.

Using Eq. (2),

$$u = (Bl)\,[V_{in} - k(Bl)_s u]\, Y_m\, g_m$$

where Y_m is the admittance of the mechanical drive-unit model of Fig. 9. Thus

$$u = \frac{V_{in}\,(Bl)\,g_m}{1/Y_m + k(Bl)_s\,(Bl)\,g_m} . \tag{4}$$

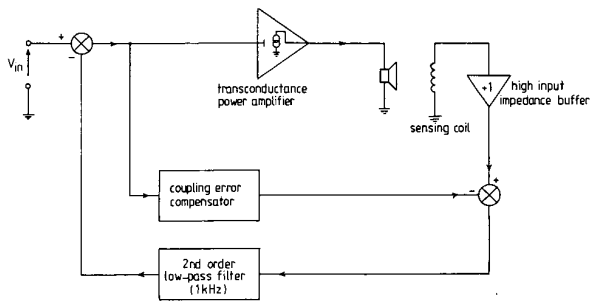


Fig. 18. Block diagram of prototype motional feedback system.

Examining the mechanical drive-unit model reveals

$$Y_m = \frac{sC_{mt}}{s^2 M_{ms}\, C_{mt} + sC_{mt}\, R_{ms} + 1} \tag{5}$$

where C_{mt} is the total mechanical compliance of the system, that is,

$$C_{mt} = \frac{C_{ms}\, C_{mb}}{C_{ms} + C_{mb}} .$$

Substituting this result into Eq. (4) gives

$$u = [V_{in}\,(Bl)sC_{mt}\, g_m][M_{ms}C_{mt}\,\{s^2 + s(1/M_{ms})\,[R_{ms} + (Bl)_s\,(Bl)kg_m] + 1/M_{ms}\, C_{mt}]^{-1} \tag{6}$$

(ω_0/Q and ω_0^2)

Thus, for a second-order system,

$$Q = \frac{1}{(C_{mt}/M_{ms})^{0.5}\,[R_{ms} + g_m\, k\,(Bl)\,(Bl)_s]} \tag{7}$$

which may be rearranged to give

$$k = \frac{1/Q\,(M_{ms}/C_{mt})^{0.5} - R_{ms}}{(Bl)\,(Bl)_s\, g_m} . \tag{8}$$

Investigation of system performance was carried out as for the open-loop case, with nonlinear transient analysis followed by measurement. It should be noted that in the prototype, the sensing coil followed the same Bl profile as the main coil, although to achieve a further low-frequency distortion reduction over the open-loop case, a more elaborate linear sensing mechanism is required. The transient analysis model will not be detailed, as it follows the earlier methodology—the velocity feedback signal was derived from the voltage across the mechanical resistance R_{ms}, and the feedback path loop stability filter was set to 1 kHz, second order. No transformer coupling effects were included in the model. For a 100-Hz at 1A peak sine-wave excitation, the measured distortion spectra are shown in

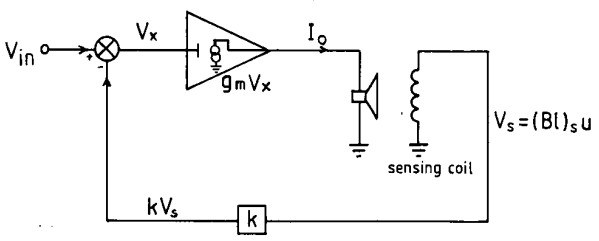


Fig. 19. Simplified motional feedback model for analysis purposes.

Fig. 20, along with comparative data from the simulation for a system Q of 0.7071. The results are seen to be broadly similar to the open-loop case for nonlinear Bl sensing, with a distortion reduction of around 4 dB on the second and third harmonics resulting from predicted linear velocity sensing.

The measured frequency response (Fig. 21) is seen to be flatter than both the voltage-driven and the open-

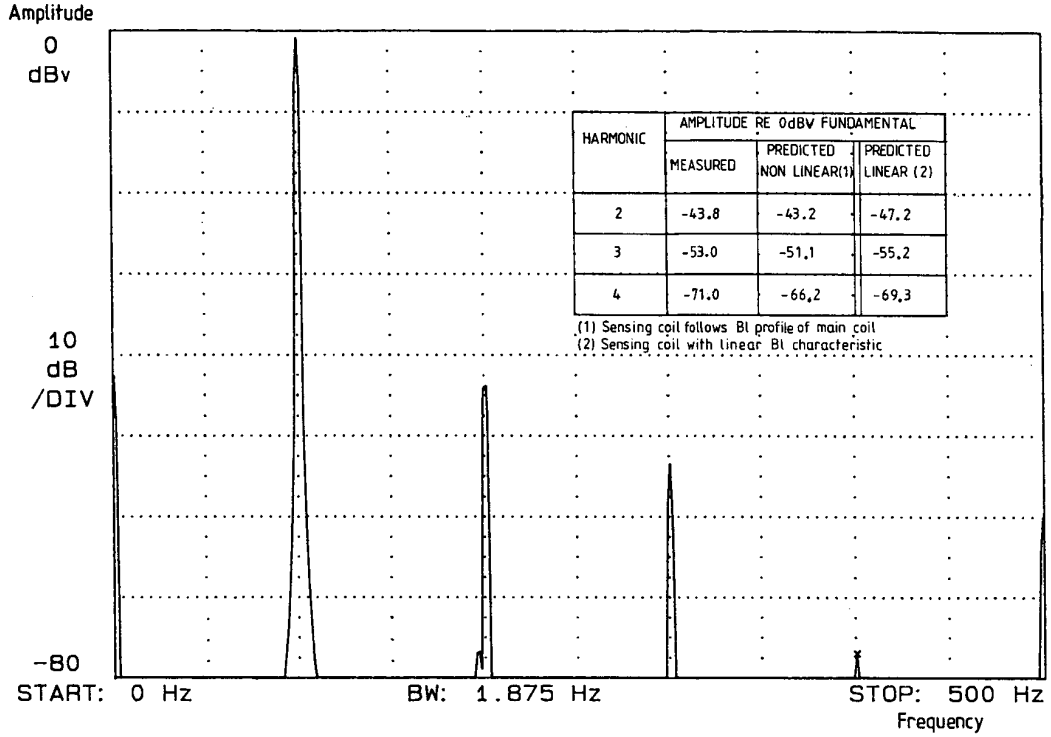


Fig. 20. Measured 100-Hz harmonic distortion, current driven with velocity feedback. Table compares with predicted results for both linear and nonlinear sensing-coil Bl profiles.

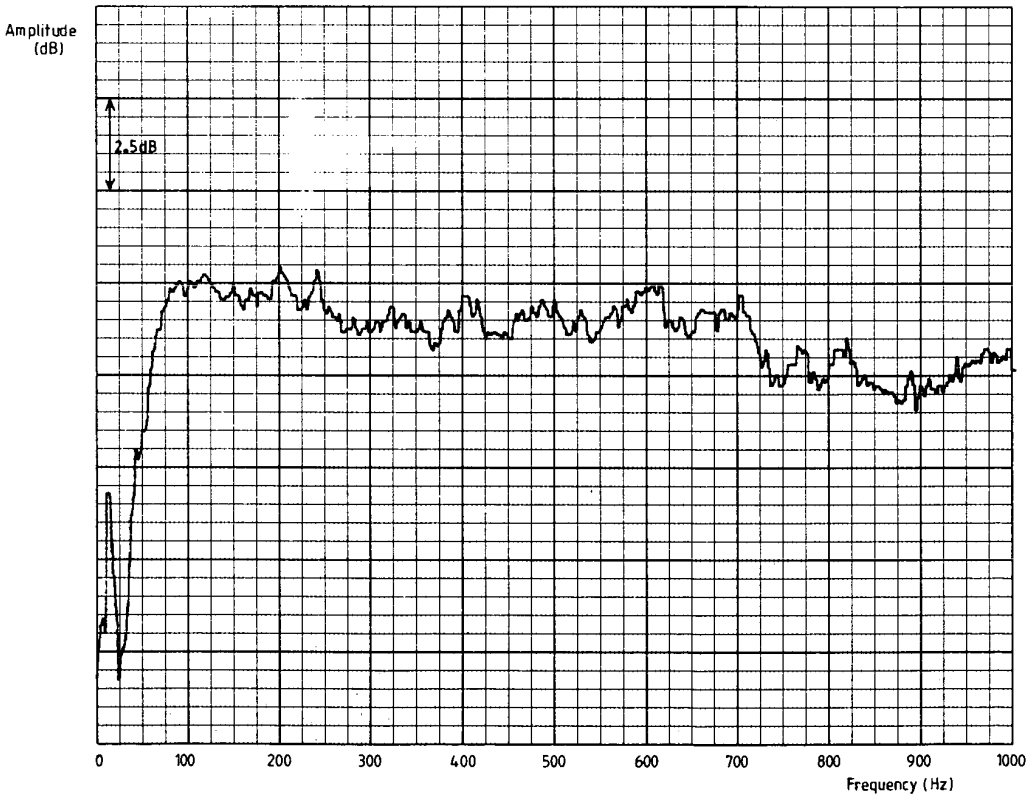


Fig. 21. Measured closed-loop frequency response for velocity feedback current-driven case, Q = 0.7071.

loop current-driven cases (Fig. 14). The effect of altering the feedback factor k is shown by the measured step responses in Fig. 22 for $Q = 3.0$ (open loop), $Q = 1.0$, and $Q = 0.7$.

3.3 High-Frequency Drive Unit with Electrically Conductive Former

At high frequencies, the preferred damping method is to use a drive unit with inherent electromagnetic damping through a conductive coil former. With such a device, experiment has shown a negligible contribution from voice-coil damping under voltage drive. Due to the improved linearity of high-frequency drive units, resulting from low coil displacement, the distortion reduction of current drive is less marked (around 3–7-dB reduction for drive units tested). However, the advantages in terms of freedom from thermally induced response errors and power compression are still valid.

4 POWER AMPLIFIER TOPOLOGIES FOR CURRENT DRIVE

A voltage-driven system requires a power amplifier with adequate bandwidth, low distortion, and a low output impedance which is linear and frequency independent. With current drive, the latter requirement translates to a high output impedance, which again should be linear and frequency independent. Also, the current demand under voltage drive [19], [31]–[34] becomes a problem of voltage demand under current drive. Consequently the maximum current delivery is known, which aids amplifier protection, as the system is inherently self-limiting.

The most basic strategy for generating a high output impedance is by the use of negative current feedback from a sensing resistor in the loudspeaker ground return [12], [13]. A typical configuration is shown in Fig. 23. Analysis of this system reveals that the transconductance g_m is given by

$$g_m = \frac{I_0}{V_{in}} = \frac{A}{(Z_0 + Z_L) + R_f(1 + A)} \qquad (9)$$

where

I_0 = load current, amperes
V_{in} = input voltage, volts
Z_o = open-loop output impedance, ohms
Z_L = drive-unit impedance, ohms
R_f = current-sensing resistor, ohms
A = forward gain of amplifier.

Consequently the output impedance Z_e may be written

$$Z_e = (1 + A)R_f + Z_o \; . \qquad (10)$$

This configuration, although a feasible solution, has two main limitations. First, the forward gain of the amplifier is frequency dependent, falling with increasing frequency as a result of its dominant pole. As a consequence, the output impedance falls with a similar characteristic. Second, the loudspeaker impedance is both frequency dependent and nonlinear, leading to a modulation of the system transconductance [Eq. (9)] and being analogous to interface distortion in a conventional voltage amplifier.

To investigate this effect in more detail, consider an error function E_1 which is defined as

$$E_1 = \frac{g_m}{g_t} - 1$$

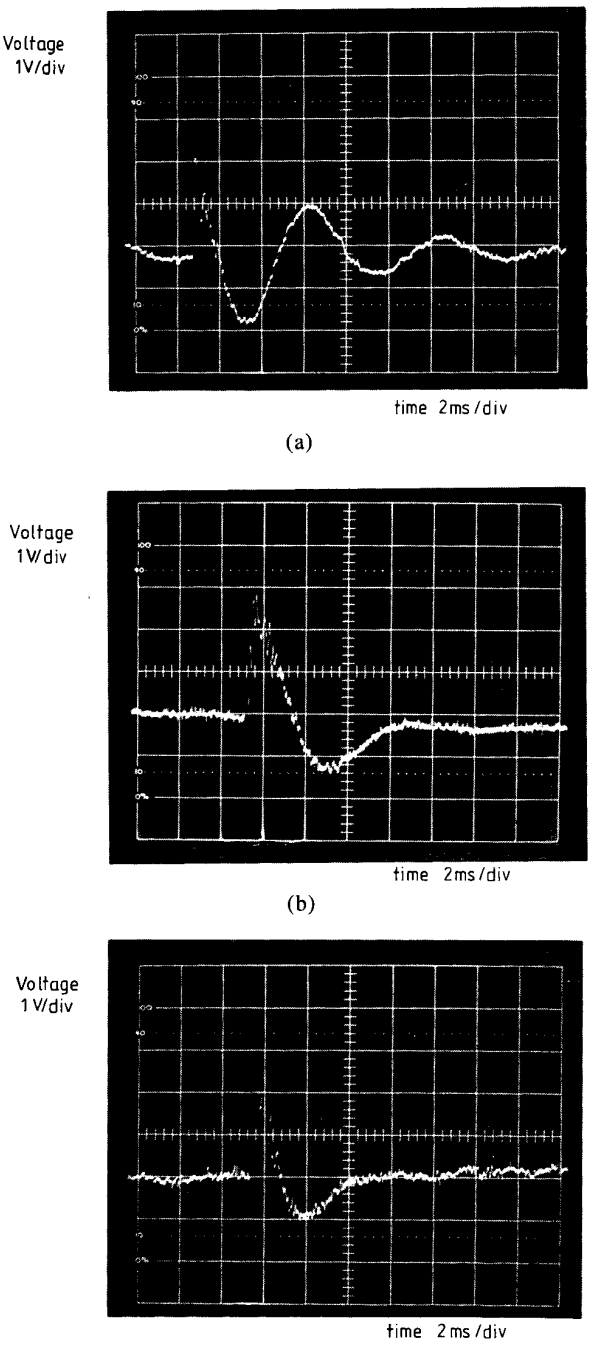


Fig. 22. Measured step responses for velocity feedback current-driven system. (a) $Q = 3.0$ (no feedback). (b) $Q = 1.0$. (c) $Q = 0.7$.

where g_t is the target transconductance, that is, $g_t = 1/R_f$. Thus, using Eq. (9),

$$E_1 = \frac{AR_f}{(Z_o + Z_L) + R_f(1 + A)} - 1$$

that is,

$$E_1 = -\frac{Z_o + R_f + Z_L}{(Z_o + Z_L) + R_f(1 + A)} .$$

Assuming $(1 + A)R_f >> (Z_o + Z_L)$,

$$E_1 \approx -\frac{Z_o + R_f + Z_L}{AR_f} \approx -\frac{Z_L}{AR_f} . \quad (11)$$

If we suppose as a numerical example that E_1 should be less than 0.1%, then from Eq. (11),

$$A > \frac{1000Z_L}{R_f} .$$

Hence, if R_f is set to 0.5 Ω and Z_L assumes a maximum value of 20 Ω, then $A_{dB} > 92$ dB. This is seen to be a high open-loop gain to maintain and illustrates well the limitations of the current feedback technique, particularly as Z_L is nonlinear.

A more optimal solution to the problem is to provide a cascaded open-loop grounded-base isolation stage in the amplifier structure to isolate the transconductance amplifier from the load, as shown in Fig. 24. Several advantages result from this enhanced technique.

1) Output impedance is essentially independent of the transconductance amplifier A_t, being a function of the grounded-base isolation stage.

2) Performance of amplifier A_t is isolated from the nonlinear load Z_L, thus eliminating interface distortion through loop gain modulation [see Eq. (11)].

3) Amplifier A_t can, if desired, operate in class A with its own supply $\pm V_{s1}$, which may be of low value to minimize power dissipation.

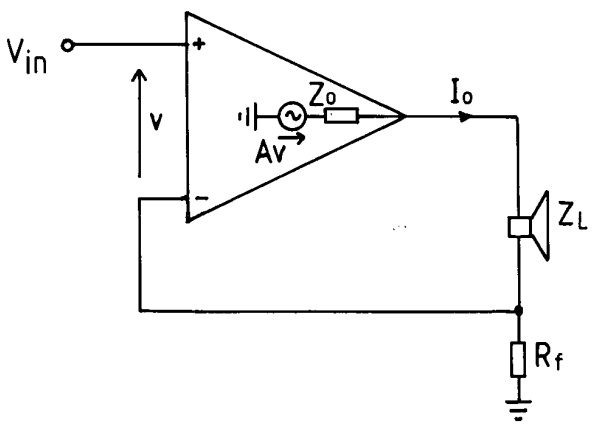


Fig. 23. Basic current feedback derived transconductance amplifier.

4) The grounded-base stage can operate in class AB, with a small standing current giving minimal distortion penalty, as current $I_L = I_0$, except for any base current leakage to ground.

5) Unlike the topology of Fig. 23, the loudspeaker load is referenced to ground, which simplifies installation and reduces the effect of interconnect capacitance at high frequencies.

6) If points P and Q (Fig. 24) are coincident, grounding-related errors are reduced due to signal currents forming well-defined closed paths.

7) The circuit topology is effectively a complementary cascode, therefore offering performance advantages in bandwidth and linearity.

8) As the grounded-base stage operates open loop, it does not degrade the loop gain and bandwidth characteristic of amplifier A_t.

9) The supply voltages $\pm V_{s2}$ can, in principle, be made adaptive to increase efficiency, when used in conjunction with a predictive digital processor.

Fig. 25 shows an alternative power amplifier topology, this time taking the form of a current gain stage. It must therefore be fed from a transconductance preamplifier. It has the advantage of having a ground-referenced power supply $\pm V_{s1}$ for the current amplifier A_i, meaning that in practice, several amplifiers in an active system may share a common supply, reducing complexity and cost.

Several prototype amplifiers have been built using these techniques. The first was based on the Fig. 24 complementary cascode configuration and operated with the transconductance amplifier A_t in class A with error feedback correction [35], [36], while the second was based on the Fig. 25 topology and used class AB operation for the current gain amplifier A_i with more extensive error correction and also moderate overall feedback. Both amplifiers were evaluated in terms of

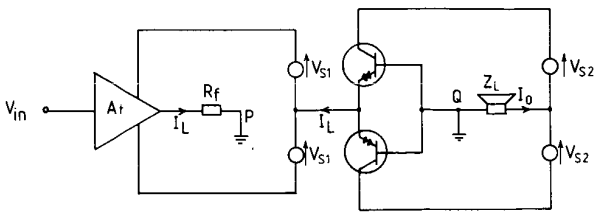


Fig. 24. Transconductance power amplifier using grounded-base output stage in complementary cascode configuration.

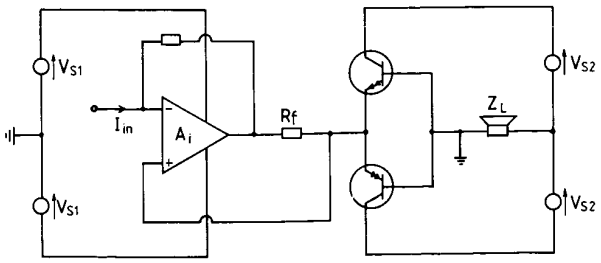


Fig. 25. Alternative power amplifier topology.

conventional measurements (Table 2) and were found comparable to typical high-performance voltage power amplifiers. On a practical note, it is judged important to provide adequate high-frequency current gain in the common-base stage to avoid distortion due to base leakage current to ground.

5 POWER AMPLIFIER AND DRIVE-UNIT PROTECTION UNDER CURRENT DRIVE

As the maximum available current from the transconductance power amplifier is an inherent design parameter, it is therefore self-limiting, so the system is simpler to protect. Indeed, this self-limiting characteristic implies that the designer need not be so concerned about protection circuitry, which has been cited as a source of degradation [37], [38].

Unlike the voltage power amplifier, which requires a series switching element for loudspeaker protection against offsets and other fault conditions, the current power amplifier requires a shunting element across the loudspeaker, thus avoiding the problems of contact degradation with time. Also a series fuse may be added without signal impairment, whereas with a voltage power amplifier, thermal modulation of the fuse wire resistance offers a source of distortion. Whereas a voltage amplifier requires short-circuit protection, a current amplifier is sensitive to open-circuit conditions. However, tests on the experimental amplifiers constructed have not given rise to a failure mode under open circuit.

A major factor concerning system reliability is drive-unit thermal failure. With conventionally powered loudspeakers, the coil current (and hence power dissipation) falls as temperature increases due to the thermal coefficient of the coil, giving a degree of protection. Elaborate protection systems have, however, been described for loudspeakers under voltage drive [39], [40]. Under current drive, no such self-limiting occurs. Indeed, it is an effect we are seeking to avoid. Thus, particularly for high-power and high-reliability installations, a method of sensing voice-coil temperature is required. This is best explained with reference to the block diagram system of Fig. 26. In addition to the main transconductance power amplifier and drive unit, a second low-power transconductance amplifier is provided to drive an impedance scaled model of the drive unit. In this model, R_e' represents the voice-coil resistance at room temperature and Z_m' the drive-unit motional impedance. After taking account of the scaling factors in the current level and impedance of the reference network, the difference in voltage across the drive unit and reference network is obtained by a differential amplifier. The rms values of the input voltage V_{in}, which is proportional to the drive-unit current and of the differential amplifier output, are then fed to a divider network to produce a voltage V_{out} representing any increase in coil resistance due to heating. Presuming the temperature coefficient of the coil material is known, a measure of temperature is then determined.

The output voltage V_{out} may be used to drive a comparator to shut down power to the loudspeaker drive unit at a predetermined temperature. Alternatively, it can be used to progressively attenuate the drive-unit current to a safe level, or to provide curtailment of low-frequency extension to reduce power dissipation. The latter techniques are of particular interest to studio monitors, where high reliability and continuity of operation are paramount.

6 PROTOTYPE TWO-WAY ACTIVE LOUDSPEAKER SYSTEM

The ideas presented in this paper have been incorporated into a working prototype two-way active loud-

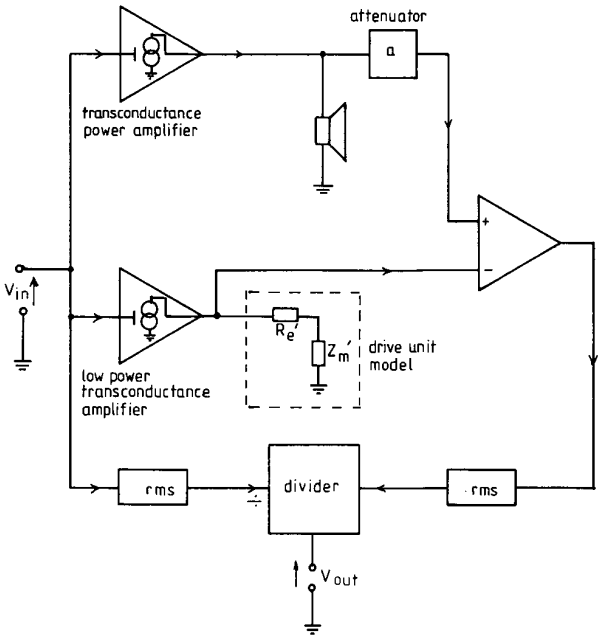


Fig. 26. Drive-unit thermal protection system for current drive.

Table 2. Performance comparison of prototype power amplifiers.

Measurement parameter	Test condition/notes	Class A design	Class AB design
Rated output power	8 Ω resistive load	75 W average	75 W average
Total harmonic distortion re rated power	20 Hz	−88 dB	−79 dB
	1 kHz	−84 dB	−86 dB
	20 kHz	−79 dB	−68 dB
Intermodulation distortion at rated power	19 kHz and 20 kHz at equal levels	<−90 dB	−86 dB
Hum and noise	Unweighted re full power	−91 dB	−90 dB
Small-signal bandwidth	−3 dB	dc–50 kHz	0.1 Hz–50 kHz

speaker system, based on the Celestion SL600 loudspeaker, because its high drive-unit quality and low level of enclosure coloration would theoretically make the benefits of current drive apparent on audition.

The system employs a discrete low-level electronic crossover, feeding individual current power amplifiers. One of the power amplifiers, based on the topology of Fig. 25, running in class AB with error correction is shown in detail in Fig. 27. Both power amplifiers are mounted on the loudspeaker stand, which aids thermal dissipation, with transformers mounted on the base to give mechanical stability. Fig. 28 shows the complete assembly. The crossover, motional feedback control circuits, coupling error compensator, and transconductance line amplifiers are housed in a separate enclosure (Fig. 29), which also incorporates level controls for the input signal. The crossover time constants are each independently adjustable for trimming, to enable comparison with the original voltage-driven loudspeaker to be made on a fair basis.

7 CONCLUSIONS

This paper has presented an alternative approach to the amplifier–loudspeaker interface, where numerous advantages have been cited through technical discussion, nonlinear computer modeling, and measurement. The principal advantages of current drive are seen to be an elimination of performance dependence on voice-coil resistance (which is thermally modulated) and also coil-inductive effects, which give rise to high-frequency distortion, along with nonlinear electromagnetic damping due to Bl variations. The technique is similarly insensitive to the lumped series elements of the amplifier–loudspeaker interconnect. However, it is often necessary to lower the system Q caused by the loss of amplifier-generated damping, either by open-loop compensation, by special drive-unit design, or by motional feedback, where the latter is regarded as the optimal method at low frequencies.

Having attempted a broad coverage of the principles of current drive, it is hoped that a greater interest in and awareness of the technique will result. The authors perceive digital signal processing (DSP) as being integral to further developments in terms of crossovers, motion feedback signal processing, and drive-unit protection against thermal and excursion damage. The subject of digital crossover design has already been researched in depth within the Group. This resulted in a two-way working system, operating on the data stream from a CD player [41]. A further area of DSP to be investigated is compensation for the drive-unit Bl profile with coil displacement sensing, which under voltage drive would not be so amenable to correction, due to the more complicated nature of nonlinearities present in the system transfer function. In addition, DSP techniques have the potential to improve transconductance power amplifier efficiency by using modulated switched-mode power supplies.

While the research has been directed at moving-coil drive units, there is no reason why current drive should not be applied to ribbon transducers, which are often mechanically well damped and would benefit from removal of the matching transformer needed under voltage drive.

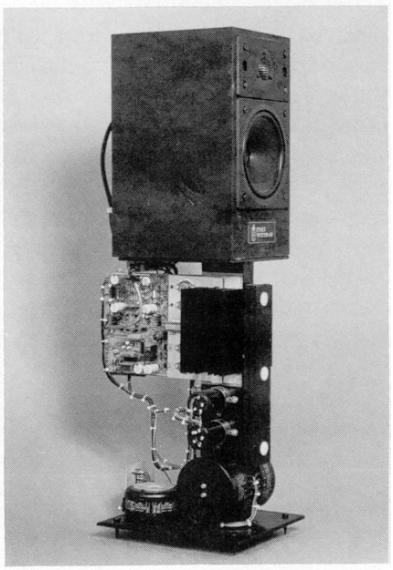


Fig. 28. Prototype active loudspeaker system.

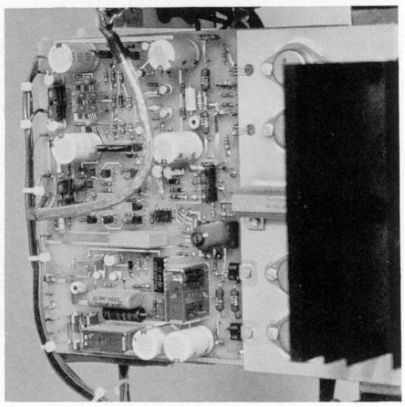


Fig. 27. View of prototype power amplifier based on Fig. 25.

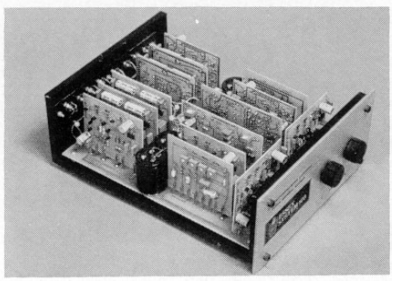


Fig. 29. View of control unit. System includes low-level crossovers, velocity feedback circuitry, and transconductance line amplifiers.

The technique of current drive should find wide applications in both high-performance domestic and studio applications, where it is felt that useful performance gains will be made over conventional systems.

8 ACKNOWLEDGMENT

This research was initially supported by the Science and Engineering Research Council of Great Britain under a research studentship program. Our thanks are due to Ed Form and more recently Graham Bank of Celestion International for providing the loudspeaker system used as the basis for the experimental work, along with the provision of modified drive units.

9 PATENT PROTECTION

The authors would like to point out that aspects of the work documented in this paper are the subject of a U.K. patent application.

10 REFERENCES

[1] H. D. Harwood, "Loudspeaker Distortion Associated with Low-Frequency Signals," *J. Audio Eng. Soc.*, vol. 20, pp. 718–728 (1972 Nov.).

[2] A. Dobrucki and C. Szmal, "Nonlinear Distortions of Woofers in the Fundamental Resonance Region," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 389 (1986 May), preprint 2344.

[3] A. J. M. Kaizer, "Modeling of the Nonlinear Response of an Electrodynamic Loudspeaker by a Volterra Series Expansion," *J. Audio Eng. Soc.*, vol. 35, pp. 421–433 (1987 June).

[4] R. J. Newman, "Do You Have a Sufficient Quantity of Acoustical Benzin? Aspects Related to the Significance of Diaphragm Excursion," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 388 (1986 May), preprint 2342.

[5] M. R. Gander, "Dynamic Linearity and Power Compression in Moving-Coil Loudspeakers," presented at the 76th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 32, pp. 1008–1009 (1984 Dec.), preprint 2128.

[6] T. S. Hsu, S. H. Tang, and P. S. Hsu, "Electromagnetic Damping of High-Power Loudspeakers," presented at the 79th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 33, p. 1011 (1985 Dec.), preprint 2297.

[7] C. A. Henricksen, "Heat-Transfer Mechanisms in Loudspeakers: Analysis, Measurement, and Design," *J. Audio Eng. Soc.*, vol. 35, pp. 778–791 (1987 Oct.).

[8] W. J. Cunningham, "Nonlinear Distortion in Dynamic Loudspeakers due to Magnetic Effects," *J. Acoust. Soc. Am.*, vol. 21, pp. 202–207 (1949 May).

[9] J. R. Gilliom, P. L. Boliver, and L. C. Boliver, "Design Problems of High-Level Cone Loudspeakers," *J. Audio Eng. Soc. (Project Notes/Engineering Briefs)*, vol. 25, pp. 294–299 (1977 May).

[10] M. R. Gander, "Moving-Coil Loudspeaker Topology as an Indication of Linear Excursion Capability," *J. Audio Eng. Soc.*, vol. 29, pp. 10–26 (1981 Jan./Feb.).

[11] H. F. Olson, "Analysis of the Effects of Nonlinear Elements upon the Performance of a Back-Enclosed, Direct Radiator Loudspeaker Mechanism," *J Audio Eng. Soc.*, vol. 10, pp. 156–163 (1962 Apr.).

[12] J. A. M. Catrysse, "On the Design of Some Feedback Circuits for Loudspeakers," presented at the 73rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 364 (1983 May), preprint 1964.

[13] R. A. Greiner and T. M. Sims, Jr., "Loudspeaker Distortion Reduction," *J. Audio Eng. Soc.*, vol. 32, pp. 956–963 (1984 Dec.).

[14] M. J. Hawksford, "The Essex Echo—Unification," *Hi-Fi News Rec. Rev.*, pt. 1 (1986 May), pt. 2 (1986 Aug.), pt. 3 (1986 Oct.), pt. 4 (1987 Feb.).

[15] M. N. T. Otala and J. Lammasniemi, "Intermodulation Distortion at the Amplifier–Loudspeaker Interface," presented at the 59th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 26, p. 382 (1978 May), preprint 1336.

[16] J. Lammasniemi and M. Otala, "Power Amplifier Design Parameters and Intermodulation Distortion at the Amplifier–Loudspeaker Interface," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 380 (1980 May), preprint 1608.

[17] R. R. Cordell, "Open-Loop Output Impedance and Interface Intermodulation Distortion in Audio Power Amplifiers," presented at the 64th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 27, p. 1022 (1979 Dec.), preprint 1537.

[18] R. H. Small, "Direct-Radiator Loudspeaker System Analysis," *J. Audio Eng. Soc.*, vol. 20, pp. 383–395 (1972 June).

[19] P. G. L. Mills and M. J. Hawksford, "Transient Analysis: A Design Tool in Loudspeaker Systems Engineering," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, p. 386 (1986 May), preprint 2338.

[20] S. H. Linkwitz, "Loudspeaker System Design," *Wireless World*, pp. 79–83 (1978 Dec.).

[21] W. M. Leach, Jr., "Active Equalization of Closed-Box Loudspeaker Systems," *J. Audio Eng. Soc.*, vol. 29, pp. 405–407 (1981 June).

[22] K. E. Stahl, "Synthesis of Loudspeaker Mechanical Parameters by Electrical Means: A New Method for Controlling Low-Frequency Loudspeaker Behavior," *J. Audio Eng. Soc.*, pp. 587–596 (1981 Sept.).

[23] P. G. A. H. Voight, "Improvements in or Relating to Thermionic Amplifying Circuits for Telephony," UK Patent 231972 (1924 Jan.).

[24] H. Black, "Inventing the Negative Feedback Amplifier," *IEEE Spectrum*, pp. 55–60 (1977 Dec.).

[25] J. A. Klaassen and S. H. de Koning, "Motional Feedback with Loudspeakers," *Philips Tech. Rev.*, vol. 29, no. 5, pp. 148–157 (1968).

[26] D. de Greef and J. Vandewege, "Acceleration Feedback Loudspeaker," *Wireless World*, pp. 32–36 (1981 Sept.).

[27] G. J. Adams, "Adaptive Control of Loudspeaker Frequency Response at Low Frequencies," presented at the 73rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 361 (1983 May), preprint 1983.

[28] E. de Boer, "Theory of Motional Feedback," *IRE Trans. Audio*, pp. 15–21 (1961 Jan./Feb.).

[29] A. F. Sykes, "Damping Electrically Operated Vibration Devices," UK Patent 272622 (1926 Mar.).

[30] R. L. Tanner, "Improving Loudspeaker Response with Motional Feedback," *Electronics*, pp. 142 ff. (1951 Mar.).

[31] I. Martikainen, A. Varla, and M. Otala, "Input Current Requirements of High-Quality Loudspeaker Systems," presented at the 73rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 364 (1983 May), preprint 1987.

[32] D. Preis and J. Schroeter, "Peak Transient Current and Power into a Complex Impedance," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 386 (1986 May), preprint 2337.

[33] M. Otala and P. Huttunen, "Peak Current Requirement of Commercial Loudspeaker Systems," *J. Audio Eng. Soc.*, vol. 35, pp. 455–462 (1987 June).

[34] J. Vanderkooy and S. P. Lipshitz, "Computing Peak Currents into Loudspeakers," presented at the 81st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, pp. 1036–1037 (1986 Dec.), preprint 2411.

[35] M. J. Hawksford, "Distortion Correction in Audio Power Amplifiers," *J. Audio Eng. Soc.*, vol. 29 (*Engineering Reports*), pp. 27–30 (1981 Jan./Feb.).

[36] M. J. Hawksford, "Power Amplifier Output-Stage Design Incorporating Error-Feedback Correction with Current-Dumping Enhancement," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 960 (1983 Dec.), preprint 1993.

[37] T. Holman, "New Factors in Power Amplifier Design," *J. Audio Eng. Soc. (Engineering Reports)*, pp. 517–522 (1981 July/Aug.).

[38] M. Huber, "Important Aspects of Power Amplifiers," *Studio Sound*, pp. 66–74 (1985 Nov.).

[39] H. D. Harwood, "Improvements Relating to Loudspeakers," UK Patent 1520156 (1976 Mar.).

[40] D. R. von Recklinghausen, "Dynamic Equalizer System for Loudspeakers," UK Patent 2050754A (1980 Jan.).

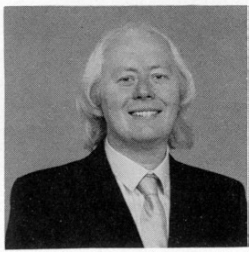
[41] R. M. Bews, "Digital Crossover Networks for Active Loudspeaker Systems," Ph.D. dissertation, University of Essex, UK (1987 Sept.).

THE AUTHORS

P. Mills

Paul Mills graduated with a B.Eng. degree in engineering science and industrial management from Liverpool University in 1980. After working for GEC in power and control systems engineering, he began a 3-year period of postgraduate study at the University of Essex in 1983, during which time he formulated the ideas presented in this paper. He then taught at Essex for two years in the Electronics Systems Engineering Department, while writing his Ph.D. thesis on active loudspeaker systems with transconductance amplification. In 1988 July he was appointed senior design engineer with Tannoy Limited. He is a member of the Audio Engineering Society.

•



M. O. Hawksford

Malcolm Omar Hawksford is a senior lecturer in the Department of Electronic Systems Engineering at the University of Essex, where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honors B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship, where the field of study was the application of delta modulation to color television and the development of a time compression/time multiplex system for combining luminance and chrominance signals.

Since his employment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing, and loudspeaker systems has been undertaken. Since 1982 research into digital crossover systems has begun within the group and, more recently, oversampling and noise shaping investigated as a means of analog-to-digital/digital-to-analog conversion. Dr. Hawksford has had several AES publications that include topics on error correction in amplifiers and oversampling techniques. His supplementary activities include writing articles for *Hi-Fi News* and designing commercial audio equipment. He is a member of the IEE, a chartered engineer, a fellow of the AES, and a member of the review board of the *Journal*. He is also a technical adviser for *HFN* and *RR*.

PAPERS

Efficient Filter Design for Loudspeaker Equalization*

RICHARD GREENFIELD AND MALCOLM OMAR HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, UK

The advent of digital storage of audio signals and the availability of high-speed digital signal processing devices facilitate the implementation of high-order filter functions for loudspeaker equalization. A method is presented for generating a digital model of a loudspeaker system from which an efficient compensation filter using an IIR structure is derived. Application of this technique allows the relative attributes of FIR and IIR structures to be distinguished and an equalization filter to be realized which provides simultaneous magnitude and phase equalization.

0 INTRODUCTION

From the conception of the Compact Disc (CD) player in 1982, digital signal processing (DSP) has had an ever-increasing role to play in audio systems. The digital serial output port that is now commonplace on CD players greatly facilitates the application of DSP to the many processes that take place in audio systems and, indeed, permits additional processing, which would otherwise be costly or impractical with analog techniques.

Loudspeaker systems, which are the most nonideal components of an audio system, potentially stand to benefit the most from the powerful processing capabilities available with modern digital signal processors. Crossover networks, traditionally, have been the source of both practical and theoretical consternation. For some of the reasons now discussed, these would appear to be a good starting point on the road to loudspeaker system improvement. A problem encountered with analog crossover networks arises from the sensitivity of the loudspeaker response to component tolerances. Digitizing the crossover, therefore, reduces these effects, as the component values are now simply coefficients in a digital signal processor. (Tolerances are now in the form of quantized coefficients.) Another undesirable feature inherent in analog crossovers, discussed by Lipshitz and Vanderkooy [1], is the tradeoff between overall phase linearity and polar response. Bews [2], in a Ph.D. thesis, proposes the use of FIR filters for crossover networks. The linear-phase property obtainable with FIR filters enables the high- and low-pass filters to remain in phase, with the combined output also being linear phase. Hence improved lobing error and transient response are achieved. A further advantage of digital crossovers, discussed in [2], is the ability to incorporate drive unit compensation which, similarly, improves the polar and transient responses of the system.

These advantages may well result in the use of digital crossovers in the high end of the audio market or studio applications. Digital filters, however, have to be part of an active loudspeaker system which requires separate power amplifiers for each of the drivers. This aspect will probably price digital crossover systems outside much of the domestic market. An alternative strategy, therefore, is to use a passive loudspeaker system in conjunction with an outboard digital equalizer. The availability of digital outputs from CD players and outboard digital-to-analog converters (DACs) makes the concept of a standard digital equalizer structure, with coefficients matched to commercially available loudspeaker systems, a most versatile and attractive proposition. An outboard equalizer further offers the ability to compensate, at least in part, for cabinet errors such as diffraction at the cabinet edges and internal reflections from within the cabinet. Thus a complete compensation system, taking into account crossover, driver, and cabinet errors, is feasible. Being flexible in nature, the specific equalization task can be tailored either to the loudspeaker systems or to their environment. For ex-

* Presented at the 86th Convention of the Audio Engineering Society, Hamburg, Germany, 1989 March 7–10; revised 1991 April 8.

ample, one may choose to ignore the more directional aberrations, giving a greater "depth of field" to the equalization, though a less focused one.

Owing to the potential of digital equalizers in acoustic applications, the subject has recently received much attention, a sample of which is given in [3]–[5]. This paper proposes an efficient method of designing the equalizer. Sec. 1 briefly considers some of the issues surrounding loudspeaker equalization, such as what form of equalization provides the best perceived performance. Sec. 2 is a brief discourse on current equalizer designs and surmises that the total reliance on FIR filters places a heavy demand on existing digital signal processors. This could be alleviated with the use of IIR filters. In Sec. 3, an equalizer system derived from an IIR model of a loudspeaker system is described and offered as a more efficient solution to the equalizer problem. Sec. 4 gives applications of this design strategy, including a discussion on its use in the evaluation of the subjective importance of phase distortion.

1 EQUALIZER DESIGN CONSIDERATIONS

Two important considerations need to be taken into account before design is commenced, 1) what is the desired loudspeaker system response? and 2) what needs to be equalized?

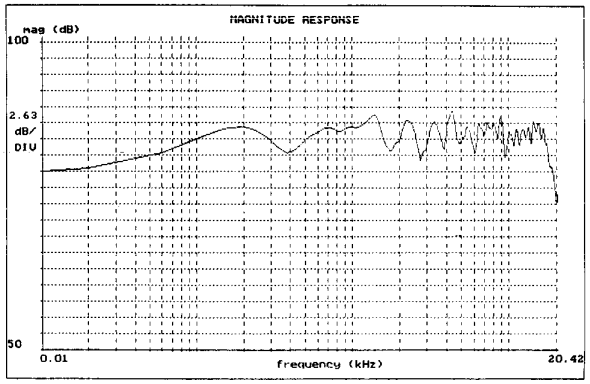
As well as being of fundamental concern to both loudspeaker and equalizer design, these two points touch on precarious psychoacoustic issues. The subject of this paper is aimed purely at equalizing an individual loudspeaker system's response; no attempt will be made to equalize for room effects. The concept of a perfect loudspeaker system is a matter for debate, but for the purpose of this paper, an ideal loudspeaker system will be considered as having flat magnitude and phase responses, which remain constant over all forward space (that is, a flat polar response). This corresponds to a system that instantly produces an impulse over the entire listening space (when driven by an impulse), which is an unrealistic aiming point for an equalization scheme for two reasons.

1) There must be some delay, as physical delays are incurred in the signal transmission path which cannot be equalized. (It is not possible to generate negative time.)

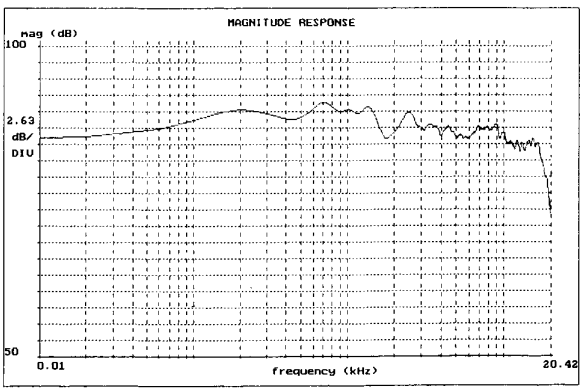
2) The polar response is a function of multiway crossover networks and the physical construction of the loudspeaker system. Hence no amount of equalization can achieve this requirement in a simple stereo system.

A better approach is to aim for a flat-magnitude response with a pure delay (linear phase) over a finite listening space (or a function thereof). With this concept of a realistic loudspeaker system's response, the practical target requirements of the equalizer need to be clarified.

First, consideration must be given to the measurements on which the equalizer should be based. The on-axis measurement appears to be an obvious choice. But bearing in mind the requirement for equalization over a listening space, the value of correcting small aberrations that occur on the on-axis response and do not occur on the off-axis responses seems suspect. Indeed, this form of equalization may prove detrimental to off-axis responses. An alternative strategy is to use an averaged measurement formed from a number of responses taken over the finite listening space. Fig. 1 shows the responses of a three-way loudspeaker system measured on axis and averaged over a a number of points in space. The averaged response is formed from nine responses measured on a grid spanning $\pm 15°$ vertically and $\pm 30°$ horizontally. (All measurements are taken at 1 m from the assumed acoustic center of the loudspeaker.) Note that both of the responses shown are, effectively, the anechoic response of the loudspeaker, as they are derived from an impulse measurement which has been truncated prior to the first room reflection. A number of interesting features are revealed by a comparison of the two plots. First, the average measurement is significantly smoother than the axial response. This indicates that a number of the ripples in the magnitude response are directional (probably caused by diffraction effects), and therefore do not want specific equalization. If the loudspeaker system response is assumed to be of finite order (that is, it can be modeled by a polynomial of finite order),



(a)



(b)

Fig. 1. Measured loudspeaker magnitude responses. (a) On-axis response. (b) Spatially averaged response.

a smoothing function in the frequency domain can be represented as an effective reduction in the order of the loudspeaker system response. Thus a practical advantage is gained with the averaged response, as the equalizer based on the inverse response will be correspondingly of lower order. Referring back to Fig. 1, the averaged response exhibits a downward tilting response relative to the axial measurement. The downward tilt is due to the increasingly directional higher frequency response, a trait common to most loudspeaker systems. An equalizer based on the averaged measurement is therefore likely to produce a perceived high-frequency emphasis effect that will be audibly objectionable.

It is the authors' belief that a better solution involves some form of compromise between the axial- and averaged-response-based equalization schemes. For example, the average is formed from a set of weighted measurements, that is, the on-axis measurement is normalized to unity, with a reduction in weight as the angle off axis is increased. Alternatively, an averaged measurement can be used with compensation for the high-frequency rolloff. These ideas are largely conjecture at the moment, and before any definitive conclusions can be made on this subject, more work is required on the subjective effects of the various equalization schemes. It may be that different loudspeaker systems require differing equalization strategies. For example, the response exhibited by asymmetrical constructions would not necessarily benefit from the same equalization strategy as that required by symmetrical constructions.

A point regarding low-frequency equalization is worth mentioning at this stage. Unless the loudspeaker system in question is specifically designed to accept bass extension, its practice is not recommended from either digital-implementation or high-fidelity contexts. From the digital perspective, a boost in magnitude of any frequency band demands an increase in signal dynamic range at the output of the digital filter. As all current digital systems work on a fixed-word-length architecture (even if the digital processor has floating-point capability, the signal ultimately has to be presented to a finite-precision DAC), a gain boost in one frequency band actually translates to attenuation in all other bands. Consequently the digital filter will exhibit a reduced dynamic range and possibly suffer from other undesirable distortion artifacts caused by the signal requantization. From a mechanical perspective, bass extension will extend the diaphragm displacement, overdriving the loudspeaker and causing excessive nonlinear distortion. Because of these considerations, low-frequency equalization will not generally be considered a function of the digital equalizer. For the equalizer design process used in this paper, in order to prevent the optimization scheme from attempting low-frequency equalization, the signal presented to the algorithm must be preequalized. This is achieved by deriving an equalizer based on the low-frequency electroacoustic model of a loudspeaker system. Such equalization techniques are discussed elsewhere (see, for example, Greiner [6]), and therefore will not be considered further here. Note that this equalization is performed in software at a "postprocessing" stage, and not in hardware, prior to the loudspeaker measurement. Similar consideration must be given to the high-frequency rolloff of the loudspeaker. Fortunately in nonoversampled systems, there is usually no need to pay specific attention to this point, as the response of the digital filter is bounded by the Nyquist frequency. This is generally below the frequency at which tweeters begin to fall off.

2 CURRENT DIGITAL EQUALIZATION SCHEMES

The use of FIR filters in digital equalization schemes seems a logical choice. Briefly, functions with arbitrary magnitude and phase are readily generated. This is particularly useful for equalization schemes where the equalizer is determined by an impulse response derived from the complementary frequency response of the loudspeaker system (taking into account the low-frequency equalization). Further, FIR filters are well suited to optimization techniques, giving simplified design tools and also the possibility of adaptive equalizers. For a more enlightening description of FIR filter design and implementation, the reader is referred to one of the numerous books on the subject, such as [7], [8].

Jensen [3] adopts a deterministic approach where the system is made up of three filters, each covering a fraction of the audio band. The lower frequency band filters operate on subsampled data which, after filtering, are then oversampled back to the original sampling frequency and combined with the higher frequency bands to form the broad-band equalizer. The decimation and interpolation process is necessary to provide sufficient resolution of equalization over the entire frequency range, which cannot be achieved with a simple system using existing digital signal processors, as the number of coefficients is prohibitive.

A different approach, adopted by Mourjopoulos [4], for example, uses optimization procedures such as the least mean squares (LMS) algorithm. In [4], Mourjopoulos describes the equalization of a system response as a deconvolution process. Here the error function, used to optimally generate the equalizer coefficients, is formed from the difference between the desired output and the convolution of an FIR filter (the equalizer) with the system's response. Typically, in acoustic applications, direct implementation of this method is confined to low-sampling-rate systems or otherwise restricted bandwidth systems. This is because the number of coefficients required to equalize low-frequency anomalies or high-Q resonances is too great for existing digital signal processors.

The limiting factor in all FIR-based equalizers is that the amount of processing involved is directly related to the impulse duration required by the equalization task. The application of IIR filters is therefore attractive where the extended impulse response possible is better

suited to the equalization requirements. The following section describes an IIR approach to the equalization problem.

3 EQUALIZER DESIGN

Although it may be possible to design an IIR equalizer analytically, the complicated nature of both the filter type and the functions to be equalized makes this approach impractical. An optimization approach, therefore, appears preferable and is the method used here. The mixed-phase nature exhibited by most loudspeaker systems creates severe problems for IIR equalization schemes. The presence of zeros in the right-hand side of the s plane, in mixed-phase signals, leads to an acausal equalizer response. This implies that any "true" IIR equalizer would be unstable. (It would have poles in the right-hand side of the s plane.) It is therefore not possible to use the direct deconvolution approach adopted in [4], [5]. Thus an alternative strategy is needed.

As the response of all loudspeaker systems (one would hope) is both causal and stable, another strategy is to form a model of the loudspeaker response. Using the coefficients obtained from the model, the minimum- and excess-phase components can be isolated and dealt with separately. (The excess-phase component is the objectionable factor, demanding the acausal equalizer response.) The design of an equalizer using this approach requires four steps:

1) Loudspeaker system modeling using an IIR filter
2) Separation of minimum- and excess-phase components
3) Formation of minimum-phase equalizer (magnitude equalizer)
4) Formation of excess-phase equalizer (all-pass equalizer).

3.1 Loudspeaker System Modeling

This section describes a technique which generates coefficients of an IIR model using the LMS algorithm. The LMS algorithm is particularly well suited to FIR structures (tapped delay lines) where the correlation cancellation loop (CCL), described by Morgan and Craig [9], is readily employed. The stability complications encountered in IIR filters generally require more advanced algorithms, such as SHARF by Larimore et al. [10]. However, if, as in the case of system modeling, the system, and therefore the model also, is driven only by an impulse, the modeling process has a number of simplifications. The process can now be decomposed into two parts: the time period when the feedforward coefficients have no effect on the model output, and the period when both feedback and feedforward coefficients contribute to the output. Thus simpler optimization, combined with deterministic approaches, can be used to find the feedback and feedforward coefficients, respectively. This aspect is illustrated by the following.

Consider the direct from 1 IIR structure of Fig. 2.

The output $y(n)$, when the filter is driven by $x(n)$, is

$$y(n) = \sum_{i=0}^{N} x(n - i)b_i + \sum_{j=1}^{M} y(n - j)a_j \ . \quad (1)$$

If the filter is driven by $\delta(n)$,

$$y(n) = \sum_{i=0}^{N} \delta(n - i)b_i + \sum_{j=1}^{M} y(n - j)a_j \ . \quad (2)$$

For $n > N$,

$$\sum_{i=0}^{N} \delta(n - i)b_i = 0 \ . \quad (3)$$

Hence the output of the filter will be

$$y(n) = \sum_{j=1}^{M} y(n - j)a_j \ . \quad (4)$$

If $a_0 = 0$ is assumed, then

$$y(n) = \sum_{j=0}^{M} y(n - j)a_j \quad (5)$$

which is the response of an FIR filter excited by its own previous outputs. Thus it would appear that the feedback coefficients can be found using standard Wiener estimation techniques. The function to be minimized, giving the optimum coefficient set, in the LMS sense, is

$$I = \sum_{n=N+1}^{L-1} e^2(n) = \sum_{n=N+1}^{L-1} [y(n) - \hat{y}(n)]^2 \quad (6)$$

$$I = \sum_{n=N+1}^{L-1} \left[y(n) - \sum_{j=1}^{M} \hat{y}(n - j)a_j \right]^2 \quad (7)$$

where $y(n)$ is the system response, $\hat{y}$ is the estimated

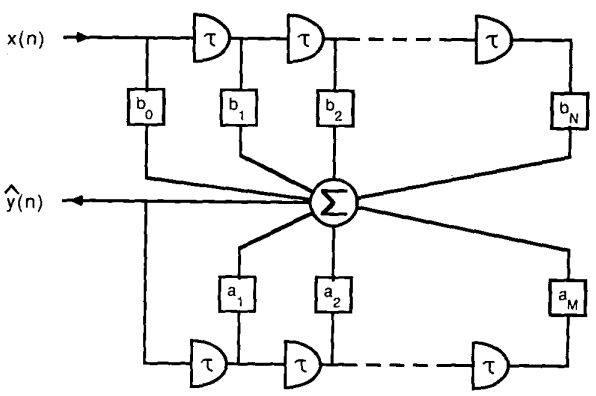


Fig. 2. Direct form 1 IIR filter structure.

response, and $e(n)$ is the error between the desired and the estimated responses. L is the length of the impulse response over which the optimization is being performed. Note that the function given in Eq. (7) differs from the standard LMS problem, as the signal driving the filter is not known a priori. Therefore a direct solution to the normal equations[1] is not possible. Instead, the approach used to minimize the function in Eq. (7) uses a gradient search (or steepest descent) algorithm. The gradient search method attempts to approximate the gradient of the error function by iteratively searching along the line of steepest descent (given by an estimate of the current error gradient function). The filter coefficient values are found by iteratively evaluating the function

$$\boldsymbol{A}(n + 1) = \boldsymbol{A}(n) - \mu \frac{\partial e^2(n)}{\partial \boldsymbol{A}} \tag{8}$$

where $\boldsymbol{A}$ denotes the coefficient vector and μ is a damping factor. Substituting the error function $e^2(n)$ from Eq. (7) into Eq. (8) leads to the simple coefficient update formula

$$a_j(n + 1) = a_j(n) + 2\mu e(n)\hat{y}(n - j) \tag{9}$$

from which we see that no a priori knowledge is required of the filter input, as only the current error value and previous filter outputs are required to form an estimate of the error gradient.

At this stage it is worth discussing a couple of points about the algorithm. It is possible to find a direct solution (as opposed to the iterative gradient search method) by making a slight modification to Eq. (7). If the modeling filter is driven by the known system response, instead of the estimated response, the optimization becomes the standard solution of the LMS problem. This will indeed, give an optimum solution over the period for which it has been optimized. However, there is no control of the response of the filter beyond this point. This procedure has been attempted, and has been found to work satisfactorily, in cases where the system response is well behaved and of finite order. When applied to acoustic measurements, the modeling algorithm nearly always produced an unstable filter. We return now to the gradient search algorithm given in this paper. The coefficient updates are made based on an estimate of the error gradient. If, at one point, an estimate is generated which produces an unstable filter, the error term will begin to diverge. This divergence effectively indicates that the last update was not made in the direction of steepest descent and, therefore, will force a reversal in the search direction, bringing the model back into a stable condition. Thus provided the initial coefficients give a stable filter (generally the feedback coefficients are set to zero), the model should remain in a stable condition. If the modeling process does continue to diverge, a smaller damping factor μ should be tried. To the authors' knowledge, there is no way to optimally determine the damping factor, as the input to the modeling filter is not known a priori. To date trial-and-error methods (and some educated guesses gained with experience) have been used to determine a suitable damping factor. It should be noted that the smaller the damping factor, the more accurate the final model will be. As this is not a real-time application, the reduction in the algorithm's convergence rate is not a crucial factor, while its stability is. Note also that the model should be optimized over a time period where the training signal is clearly converging. This will pull the model into a similarly convergent pattern.

Let us now determine the feedforward coefficients. Having found the feedback coefficients, a simple method for obtaining the feedforward coefficients is as follows:

Consider time $n = 0$

$$y(0) = b_0\delta(0)$$

$$>> b_0 = y(0)$$

then at time $n = 1$

$$y(1) = b_1\delta(n) + a_1y(0)$$

$$>> b_1 = y(1) - a_1y(0)$$

and so on, until time $= N$,

$$b_N = y(N) - a_1y(N - 1) - a_2y(N - 2) - \cdots - a_Ny(0) \; . \tag{10}$$

Eq. (10) is a general equation, which is used to find all of the feedforward coefficients from a knowledge of the impulse response and the previously found feedback coefficients.

The procedures described for finding the feedforward and feedback coefficients are used in a program which calculates a specified-order model from a given system impulse response. Fig. 3(a) shows the frequency and time domain responses of a commercial loudspeaker system and Fig. 3(b) shows the corresponding responses of a 40th-order model. The order of the model is arbitrary. However, the accuracy of the model is directly related to its order, which is seen in both the residual mean squared error and the response plots. Fig. 3(c) shows the responses of a 55th-order model, where the improved accuracy is readily observed. An interesting feature of this modeling procedure is that, although it is optimized in the time domain, the model tends to lock onto the most dominant resonances that are apparent in the frequency domain. This is observed in Fig. 3(a) and (b), where the lower order model follows a smoothed version of the loudspeaker response. The higher order model bears the same general characteristic, but has

[1] The details of LMS techniques will not be covered here. For the interested reader there are a number of texts available on the subject. See, for example, [11].

also managed to pick up some finer detail in the frequency response. An explanation for this phenomenon comes from noting that the response of an IIR filter can be formed from the summation of geometric series. Consider the frequency response of an IIR filter, given by the z transform of Eq. (2),

$$H(z) = \frac{\sum_{i=0}^{N} b_i z^{-i}}{1 - \sum_{j=1}^{M} a_j z^{-j}} . \quad (11)$$

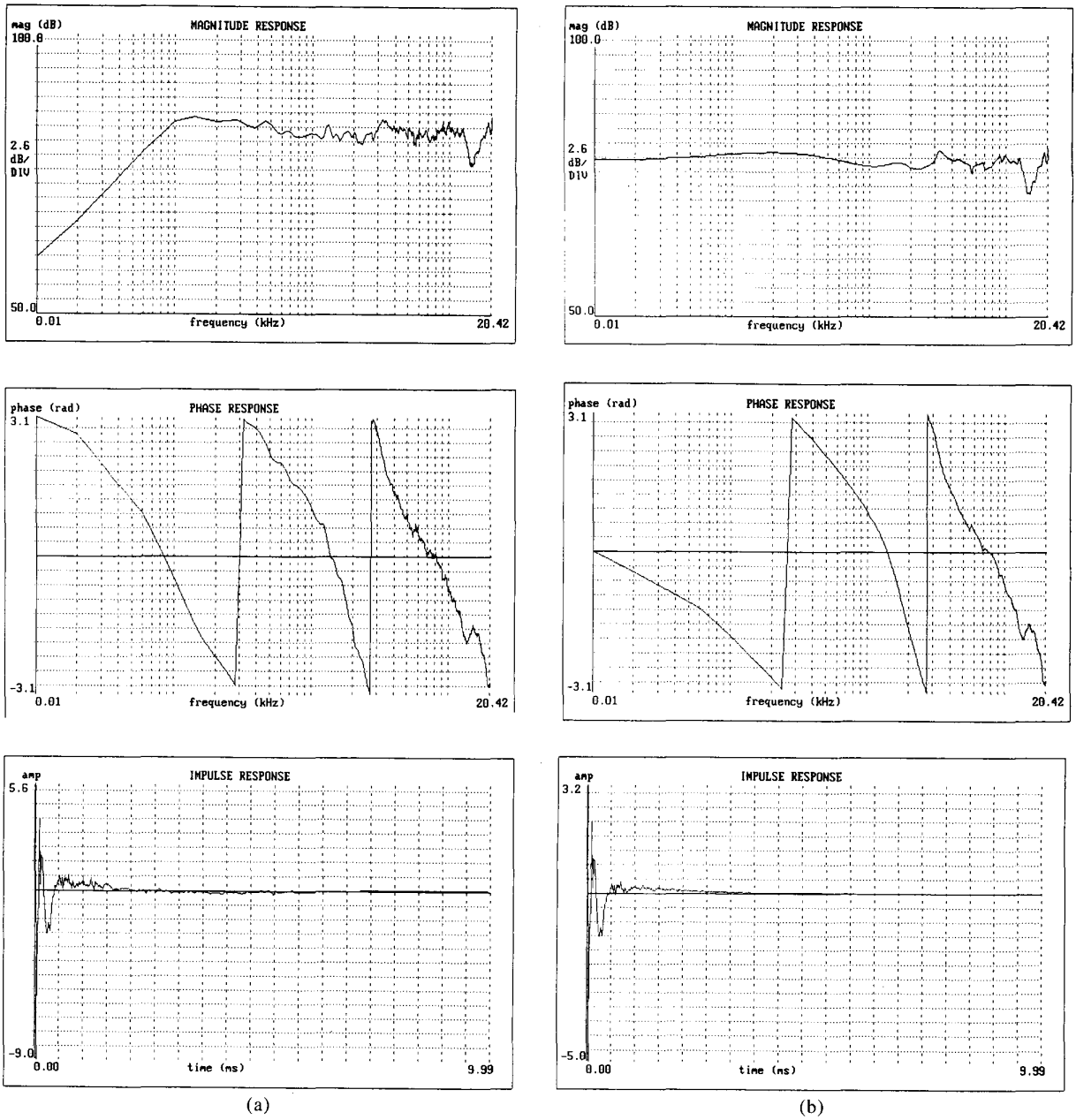
This can be split into its partial fraction expanson (for the simple case of no repetitive poles), giving

$$H(z) = z \sum_{i=1}^{M} \frac{R_i}{z - \alpha_i} + R_{M+1} \quad (12)$$

where α_i are the filter poles and R_i are the residues found from the filter zeros. Eq. (12) transforms into the summation of geometric series in the time domain, giving

$$h(n) = \sum_{i=1}^{M} R_i \alpha_i^n + R_{M+1}\delta(n) . \quad (13)$$

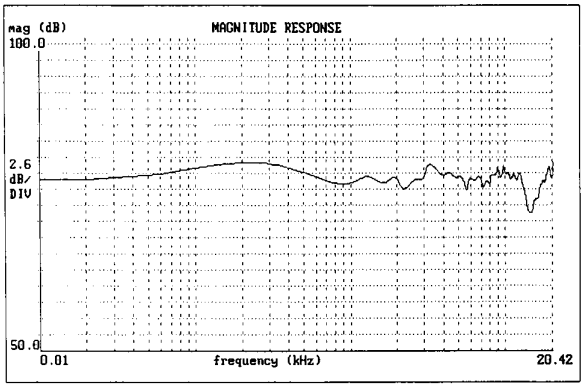
Eq. (13) shows how the pole locations in the frequency

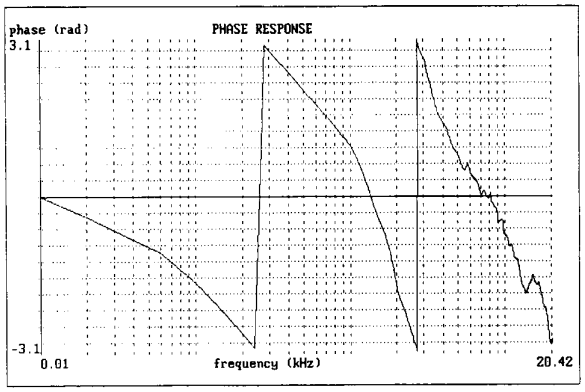


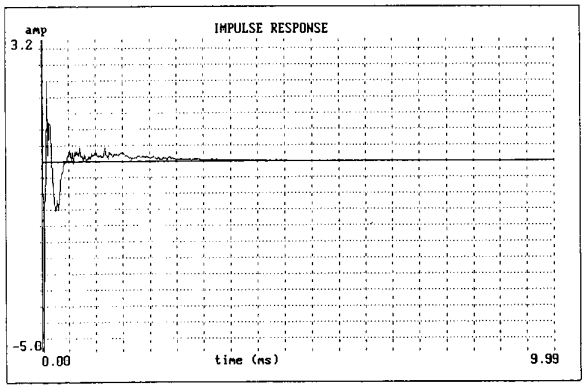
(a) (b)

Fig. 3. Loudspeaker system responses. (a) Measured. (b) 40th-order model. (c) 55th-order model.

domain are directly related to the time domain response of the IIR filter. The correlation cancellation loop [Eq. (9)] works by locking onto correlations within the signal. It is therefore likely that the most dominant (or high-Q) resonances will be the first to be picked up by the modeling process. This accounts for the lower order model's ability to extract the main features of the loudspeaker response and discard the finer details first. While this aspect of the modeling process is in most cases desirable, from a subjective perspective, one must be aware that all-pass responses are also capable of exhibiting very high Q resonances. All-pass responses, if not inaudible, are certainly less audibly significant than most amplitude anomalies. Yet, the model will lock onto these resonances in preference to weaker amplitude resonances.







(c)

Fig. 3. continued.

3.2 Separation of Minimum- and Excess-Phase Components

Separation of the minimum and excess-phase components requires the factorization of a z-domain polynomial into its poles and zeros. In the z domain a mixed-phase function is one in which zeros are present outside the unit circle (excess-phase zeros). For example, Fig. 4(a) shows a pole–zero plot of a typical mixed-phase function in the z domain. A mixed-phase function can be represented by all-pass and minimum-phase functions in cascade, as follows:

$$H(z) = \frac{(z - n_m)(z - n_e)}{(z - d_1)(z - d_2)}$$

$$H(z) = \frac{(z - n_m)(zn_e^* - 1)}{(z - d_1)(z - d_2)} \frac{z - n_e}{zn_e^* - 1} \qquad (14)$$

$$H(z) = H_{min}(z)H_{ap}(z)$$

where n_m and n_e denote roots inside and outside the unit circle, respectively, and * denotes the complex conjugate.

In essence, the minimum-phase function is formed simply by replacing the excess-phase zeros by their reflection about the unit circle. Similarly, the all-pass function is formed from the excess-phase zeros and poles, which are the excess-phase zeros reflected about the unit circle. To demonstrate this graphically, the pole–zero plots of the minimum-phase and all-pass functions are given in Fig. 4(b) and (c), respectively.

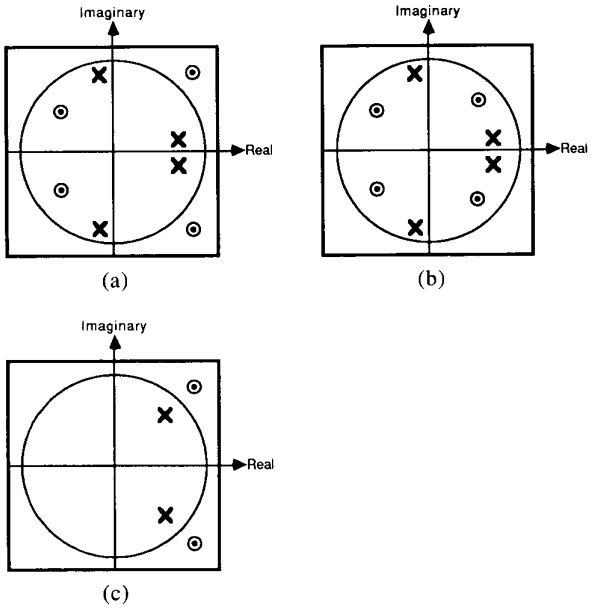


Fig. 4. Pole and zero plots. (a) Fourth-order mixed-phase function. (b) Corresponding fourth-order minimum-phase function. (c) Corresponding second-order all-pass function.

From Eq. (14) the minimum-phase function contains all its zeros inside the unit circle and is therefore invertible (the criterion for stability in the z domain being that all poles must be contained within the unit circle), while the all-pass function contains all its zeros outside the unit circle and is therefore noninvertible. Now an all-pass function, as its name suggests, passes all frequencies with unity gain. Therefore the minimum-phase function contains all the magnitude characteristics of the loudspeaker system response. Magnitude and minimum-phase equalization can, therefore, be performed by the inverse minimum-phase function, while equalization of the resulting all-pass response requires an additional equalizer based on the all-pass function.

3.3 Formation of Minimum- and Excess-Phase Equalizers

The minimum-phase equalizer is the reciprocal of the minimum-phase function found in the Sec. 3.2. The polynomial in the z domain is converted into the filter coefficients using the relationship observed between Eqs. (1) and (11). This produces an IIR filter capable of equalizing all magnitude effects over the entire frequency range. Application of this type of equalizer is discussed in the following section.

The excess-phase equalization is considerably more problematical. As was shown in preceding sections, an IIR equalizer based directly on the inverse function is not possible. Looking at the problem from a different perspective, the all-pass characteristic is a result of pure phase or delay effects, which can be equalized with the time-reversed version of the phase or delay effects. Thus,

$$e^{j\omega\sigma}e^{-j\omega\sigma} = 1 \qquad (15)$$

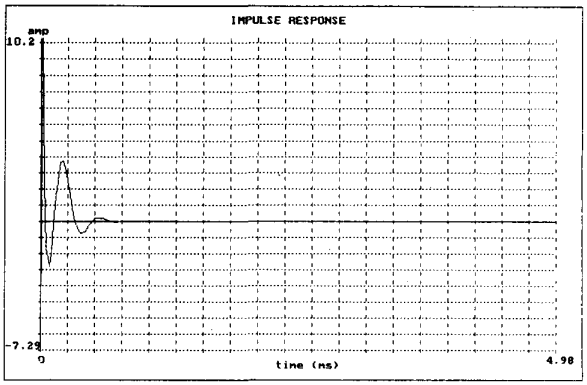
where ω is the frequency variable and σ is the frequency-dependent phase function. Although negative time is unrealizable, it is possible to introduce an overall delay. Hence the phase equalization is affected as follows:

$$e^{j\omega\sigma}e^{j\omega(\tau-\sigma)} = e^{j\omega\tau} .$$

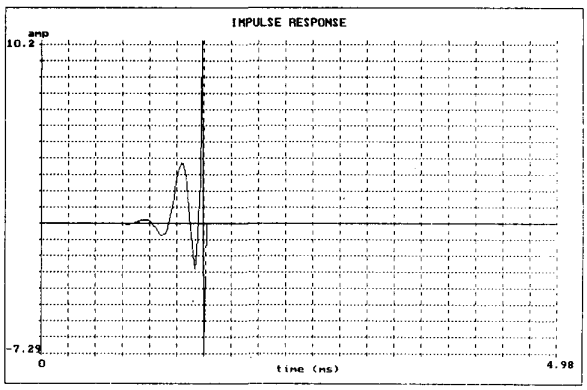
which is a pure delay or linear phase shift. Note that the time delay τ must be chosen such that $\tau - \sigma > 0$ for all ω.

Although approximations to a pure delay can be made with IIR filters, it seems more expedient to make use of FIR techniques where delays are inherent of the structure. Thus the phase equalizer is formed from the time-reversed, time-shifted, and time-windowed impulse response. The all-pass impulse response is derived from the all-pass function found in Sec. 3.2. Fig. 5(a) gives an example of an all-pass impulse response and Fig. 5(b) gives the corresponding impulse used to form the equalizer.

As an aside, a useful feature of the filter derivation process is that it allows a limited amount of control of the impulse duration requirements that are demanded by the excess-phase equalizer. A major concern with all FIR filters is the duration of the impulse response. Typically, some form of time window is applied to the impulse response in order to constrain its duration. Now, applying a time window to an all-pass response will result in the response no longer being all pass. Thus unless the filter is of sufficient length to completely accommodate the mixed-phase equalizer, only minimum-phase equalization should be attempted. This results in an all-pass response from the equalized system, which certainly is preferable to the amplitude distortions that would have been incurred had the excess-phase equalizer impulse response been prematurely truncated. Referring back to Eq. (14), we note that the all-pass function is formed from a set of matched pole–zero pairs. Elimination of any pole–zero pairs, from the set of pairs, will not affect the amplitude response (that is, an all-pass filter remains all pass). Using this property, one can discard any slowly decaying all-pass resonances from the all-pass function and use the remaining pairs to equalize the remaining phase distortion. In loudspeaker systems, the all-pass resonances that are likely to cause a problem will be with low-frequency crossovers. If the pole–zero pairs associated with this crossover region are discarded, in most cases an FIR filter of sufficient length can be derived to deal with the not so demanding, but more significant midband crossover distortion. To demonstrate the possibilities of this procedure, Fig. 6(a) shows a fifth-order all-pass



(a)



(b)

Fig. 5. Impulse responses. (a) All-pass function. (b) Corresponding equalizer function.

response obtained from a measured loudspeaker. Along with the all-pass response caused by a crossover network centered at 2.5 kHz, the all-pass function contains two slowly decaying resonances (one at approximately 100 Hz and the other at 20 kHz). The pole–zero pairs associated with these two troublesome resonances are removed, giving the responses shown in Fig. 6(b). The duration of the impulse is now reduced considerably. Using the reduced-order function to derive a 70th-order FIR equalization filter gives the quasi-excess-phase equalized response shown in Fig. 6(c). The midband phase distortion has been removed almost entirely

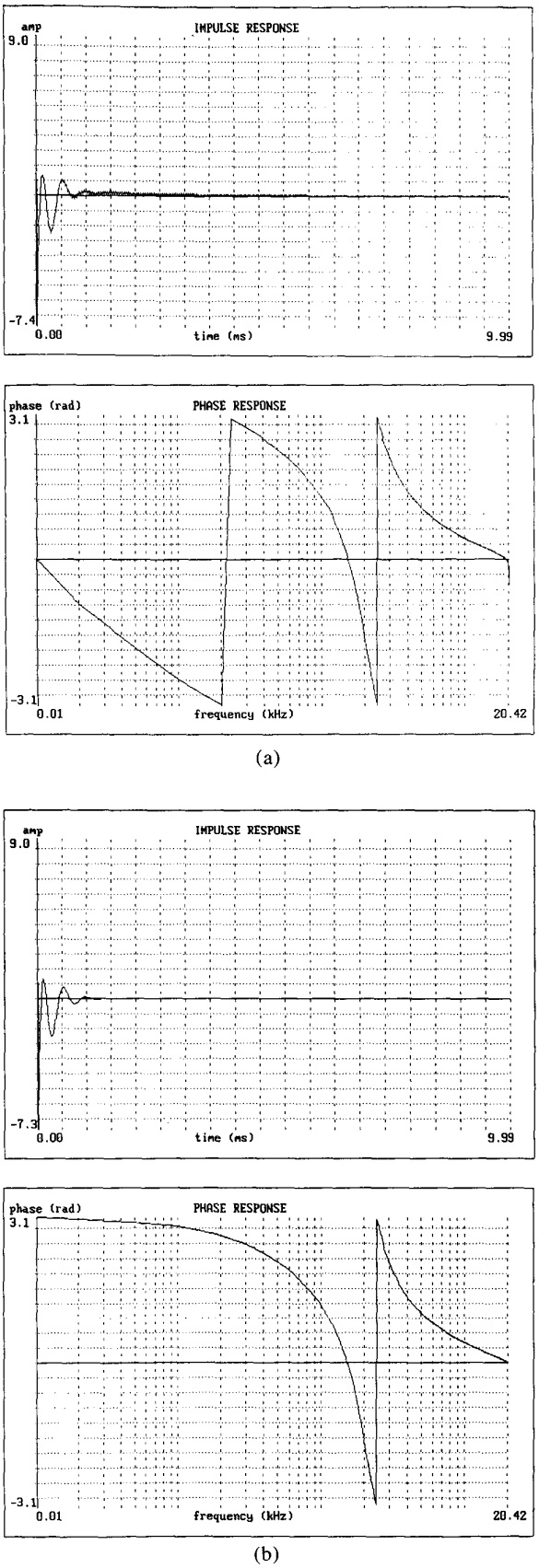


Fig. 6. Impulse and phase responses. (a) Fifth-order all-pass filter. (b) Third-order all-pass filter derived from (a). (c) Fifth-order all-pass network after equalization by reduced-order all-pass filter. (d) Magnitude response of all-pass equalizer derived from original fifth-order all-pass function truncated at 70 samples.

without compromising the amplitude response. For reflection, Fig. 6(d) shows the magnitude response of a 70th-order excess-phase equalizer derived from the original fifth-order all-pass function using a rectangular window. Note the severe amplitude distortion now incurred.

This completes the description of loudspeaker equalizer design. The remainder of this paper discusses various applications where this form of equalization may be of use.

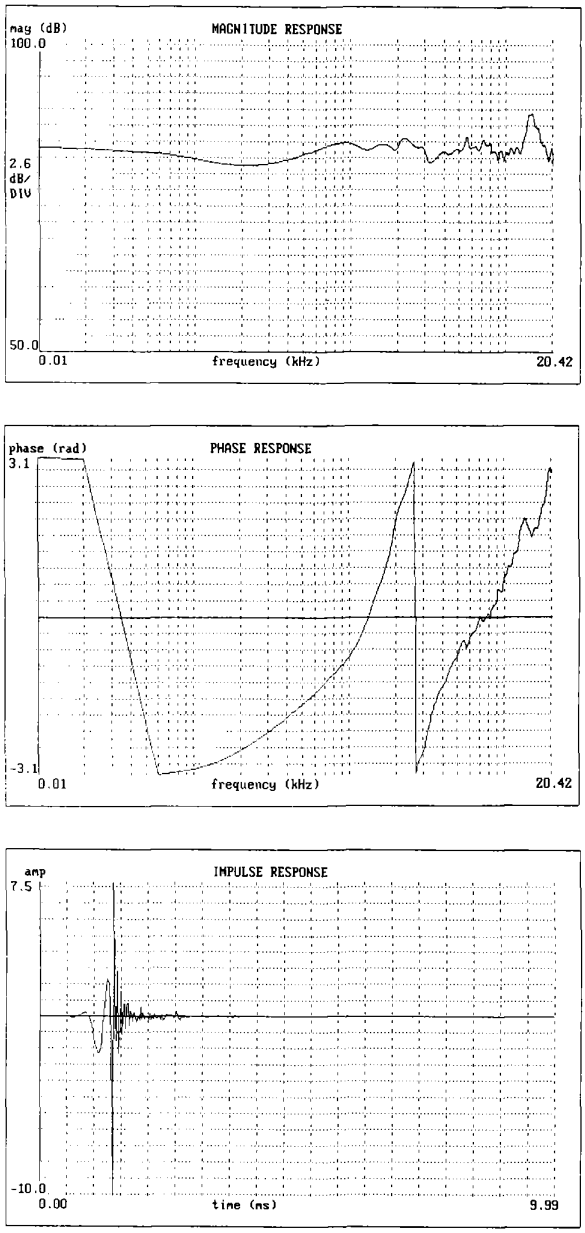
4 APPLICATIONS OF EQUALIZATION TECHNIQUE

The principal application is the equalization of a loudspeaker system response. Secondary issues, which will also be discussed, are its application to the subjective effects of phase distortion and to loudspeaker system measurement.

For the purpose of this paper, the equalizers demonstrated are derived from the on-axis response measurement. This is done here because of the ease of observing the effect of the equalizer on the loudspeaker system's response. This form of equalization also has application to loudspeaker system measurement and quality assessment, which will be discussed later. The data on which the equalizer is based are obtained from an impulse measurement technique similar to that of Berman and Fincham [12], which provides an impulse response in a digital form suitable for direct input into a computer. The frequency response measurements of a commercial loudspeaker system (obtained from the fast Fourier transform of the impulse response) are shown in Fig. 3(a) and the corresponding responses of a 55th-order model in Fig. 3(c). The equalizer responses, formed from the cascade of a 70th-order FIR filter (excess-phase equalizer) and the 55th-order IIR filter (minimum-phase and magnitude equalizer) are shown in Fig. 7(a). The measured equalized responses, given in Fig. 7(b), show significant improvement in the impulse response and, consequently, the frequency responses as well. The equalized responses presented in this paper were obtained using a real-time digital filtering system built around the TMS320C25 digital signal processor. The loudspeaker equalizer uses the AES/EBU digital interface and is therefore compatible with most two-box CD systems (CD transport and outboard DAC). The prototype digital equalizer is shown in Fig. 8.

The method of deriving an equalizer described in this paper enables some subjective evaluation of the loudspeaker's inherent phase distortion to be performed (as opposed to additionally imposed phase distortion). Sec. 3 discussed the implementation of separate equalizers, a magnitude with minimum-phase equalizer, and an excess-phase equalizer. If just the former is used, the resultant response is all pass, that is, only phase distortion is incurred. The minimum-phase (magnitude) equalized responses are given in Fig. 7(d). The magnitude responses of Fig. 7(d) and (b) are similar, while the phase responses and consequently the impulse responses are quite different. Thus the application of these equalizers will permit the subjective effects of the excess-phase distortion exhibited by the loudspeaker to be readily observed and compared to measured performance. Such experiments have been carried out and were reported elsewhere [13].

Equalization also has application in loudspeaker system measurement and quality assessment. One deficiency of the impulse measurement technique is the truncation of the impulse before the first reflection [12].



(a)

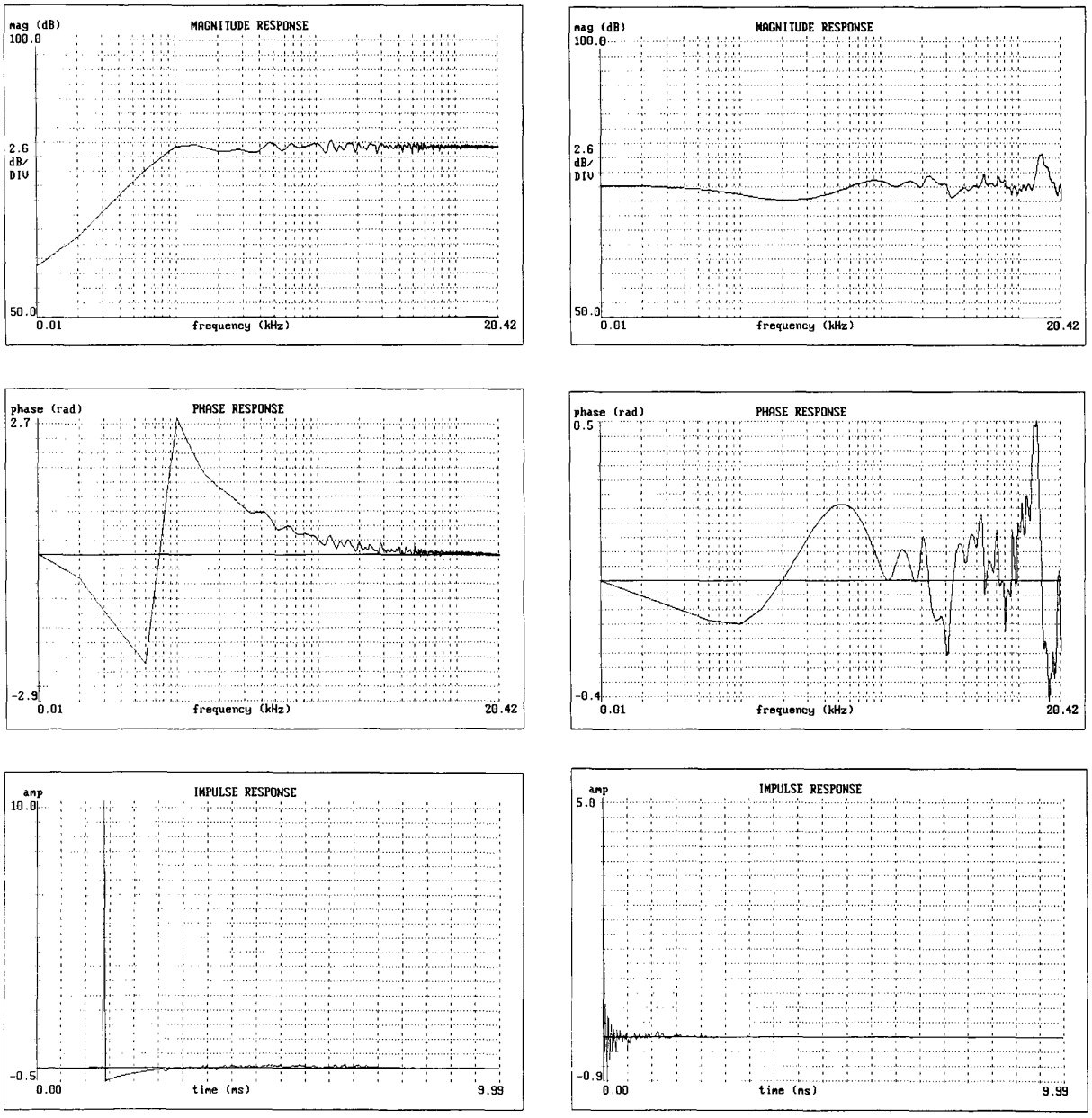
Fig. 7. (a) Magnitude and phase equalizer responses. (b) Measured loudspeaker responses after magnitude and phase equalization. (c) Magnitude and minimum-phase equalizer responses. (d) Measured loudspeaker responses after magnitude and minimum-phase equalization.

This leads to a loss of information, particularly in the low frequencies. In a follow-up paper to [12], Fincham [14] proposes the use of an equalizer to reduce the duration of the impulse tail. Hence a reduced amount of information is lost by the truncation. The techniques described in this paper are similarly applicable to the enhancement of acoustic measurements. Indeed, there may be some additional gain to be had by removing higher frequency artifacts from the measurement. These, too, may lead to an extended impulse response.

Another application of equalization is in loudspeaker system quality assessment where, briefly, if the loudspeaker system is equalized on axis, deterioration of the off-axis responses would show deficiencies of the system. Three-dimensional plots of the time or frequency responses versus measurement angle of the equalized loudspeaker would illustrate valuable information about the performance of a loudspeaker system.

5 CONCLUSION

An equalizer design technique which uses both FIR and IIR filters has been presented. The function of the IIR filter is to equalize the magnitude response of the loudspeaker. It is believed that, in most instances, an IIR structure will perform this task far more efficiently



(b) (c)

Fig. 7. continued.

than what can be achieved using FIR structures. The FIR filter is used to deal with any acausal requirements of the equalizer. Measured responses, taken from a real-time digital equalizer, demonstrate significant improvements in the loudspeaker's linear transfer function. A method was introduced for constraining the impulse response of the FIR filter. This allows practical implementation of the excess-phase equalizer without compromising the amplitude response, although the amount of excess-phase equalization is now reduced.

A feature of the design process is the ability to separate the minimum- and excess-phase components of the system. Thus the audible effects of loudspeaker excess-phase distortion can be assessed and compared to measured responses. The subjectivity of phase distortion is a controversial topic, and this technique may provide valuable contributions toward it.

Finally, the application of the scheme to loudspeaker system measurement and quality assessment is suggested, which could prove useful in loudspeaker system design and manufacture.

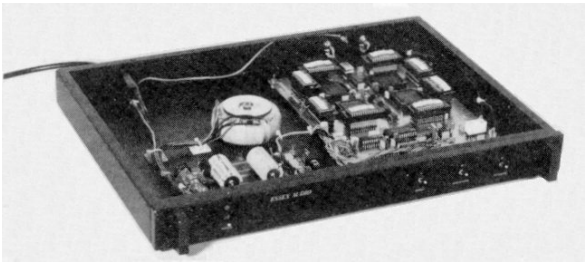
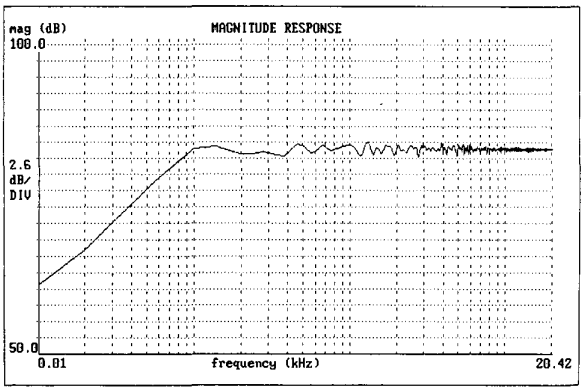
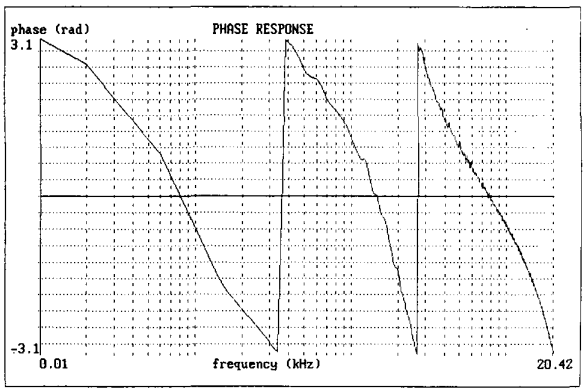
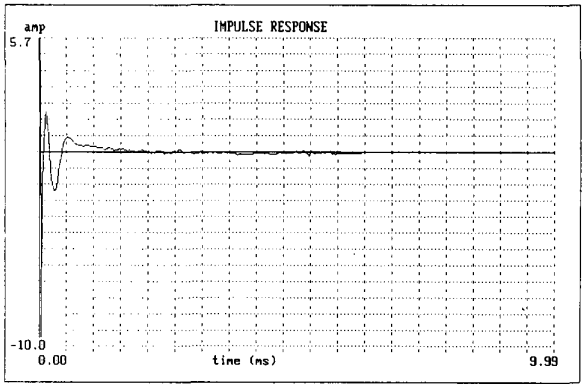


Fig. 8. Digital loudspeaker equalization system.







(d)

Fig. 7. continued.

6 REFERENCES

[1] S. P. Lipshitz and J. Vanderkooy, "In-Phase Crossover Network Design," *J. Audio Eng. Soc.*, vol. 34, pp. 889–894 (1986 Nov.).

[2] R. Bews, "Digital Crossover Networks for Active Loudspeaker Systems," Ph.D. thesis, University of Essex, UK, 1988.

[3] J. A. Jensen, "A New Principle for an All-Digital Preamplifer and Equalizer," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 35, pp. 994–1003 (1987 Dec.).

[4] J. Mourjopoulos, "Digital Equalization Methods for Audio Systems," presented at the 84th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 36, p. 384 (1988 May), preprint 2598.

[5] P. A. Nelson and S. J. Elliot, "Least Square Approximations to Exact Multiple Point Sound Reproductions," *Proc. Inst. Acoust.*, vol. 10, pt. 7, pp. 151–186 (1988).

[6] R. A. Greiner and M. Schoessow, "Electronic Equalization of Closed-Box Loudspeakers," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 31, pp. 125–134 (1983 Mar.).

[7] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1972).

[8] A. Antoniou, *Digital Filter Analysis and Design* (McGraw-Hill, New York, 1979).

[9] D. R. Morgan and S. C. Craig, "Real-Time Adaptive Linear Prediction Using the Least Mean Squares Gradient Algorithm," *IEEE Trans. Acoust., Speech, Signal-Process.*, vol. ASSP-24, pp. 494–507 (1976 Dec.).

[10] M. G. Larimore, J. R. Treichler, and C. R. Johnson, "SHARF: An Algorithm for Adapting IIR Digital Filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 428–440 (1980 Aug.).

[11] W. H. Press et al., *Numerical Recipes in C: The Art of Scientific Programming* (Cambridge University Press, Cambridge, UK, 1988).

[12] J. M. Berman and L. R. Fincham, "The Application of Digital Techniques to the Measurement of Loudspeakers," *J. Audio Eng. Soc.*, vol. 25, pp. 370–384 (1977 June).

[13] R. G. Greenfield and M. O. J. Hawksford, "The Audibility of Loudspeaker Phase Distortion," presented at the 88th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 38, p. 384 (1990 May), preprint 2927.

[14] L. R. Fincham, "Refinements in the Impulse Testing of Loudspeakers," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 966 (1983 Dec.), preprint 2055.

THE AUTHORS

R. Greenfield

M. O. J. Hawksford

Richard Greenfield received the B.Sc. degree in electronic engineering (telecommunications) from the University of Essex. He continued at the University of Essex, working with the audio research group under the supervision of Dr. M. O. J. Hawksford. His interests lie specifically with digital loudspeaker equalization, which forms part of a study which has now been accepted for the conferment of the Ph.D. degree.

Mr. Greenfield is now with Essex Electronic Consultants at the University of Essex. His main area of consultancy includes digital audio and related subjects.

•

Malcolm Omar Hawksford is a reader in the Department of Electronic Systems Engineering at the University of Essex, where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honors B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship, where the field of study was the application of delta modulation to color television and the development of a time compression/time multiplex system for combining luminance and chrominance signals.

Since his employment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing, and loudspeaker systems has been undertaken. Since 1982 research into digital crossover systems has begun within the group and, more recently, oversampling and noise shaping investigated as a means of analog-to-digital/digital-to-analog conversion.

Dr. Hawksford has had several AES publications that include topics on error correction in amplifiers and oversampling techniques. His supplementary activities include writing articles for *Hi-Fi News* and designing commercial audio equipment. He is a member of the IEE, a chartered engineer, a fellow of the AES and of the Institute of Acoustics, and a member of the review board of the *AES Journal*. He is also a technical adviser for *HFN* and *RR*.

PAPERS

Asymmetric All-Pass Crossover Alignments*

M. O. J. HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester, Essex CO4 35Q, UK

A family of asymmetric crossover filters is investigated that expands the established all-pass set. A general method of determining complementary crossover filters with an all-pass composite response is presented together with a method of reflecting the crossover asymmetry about the crossover region. Several example filters are included. CAD techniques are used to verify the basic approach, and a simple application to a satellite–subwoofer loudspeaker is described where asymmetry in the crossover frequency response and the resulting overlapping of the high-pass and low-pass responses are shown to improve the effectiveness of filter attenuation. However, because of asymmetry in the high-pass and low-pass phase responses and the resulting polar response irregularity in association with noncoincident drive units, the application regime for these alignments is restricted to low frequency.

0 INTRODUCTION

A set of low-frequency crossover alignments is described that exhibit composite all-pass characteristics, but have individual high-pass and low-pass transfer functions of different order. Of specific interest is the alignment using a first-order high-pass filter, although more general higher order asymmetric alignments are also discussed. This work is presented for two reasons. First the existence of exact asymmetric filters as opposed to approximate filters (Hillerich [1], although the time-delay compensation is acknowledged) generalizes the already well documented crossovers belonging to the all-pass family. Second the application of this class of filter to low-frequency crossover design can represent a useful simplification in network topology. However, the main application regime is restricted to low frequency, primarily because of asymmetric phase response and the association of lobing errors [2] when there are significant time delays between drive-unit acoustic centers.

As an example, the simplest alignment is described in Section 4, which combines a broad-bandwidth (satellite) loudspeaker with a subwoofer. The system uses a first-order network in the satellite loudspeaker channel, yet accommodates a second-order network of lower cutoff frequency in the subwoofer feed to enhance suppression of high-frequency signal components. The performance of two sets of higher order filters is presented in Section 5 using a minimum of reactive elements with state-variable active synthesis. CAD analysis confirms the performance of these systems and demonstrates both the all-pass characterization of the composite response and the interchange of low-pass and high-pass responses using an s-domain transformation.

1 ASYMMETRIC ALL-PASS ALIGNMENTS

The general expression for an all-pass alignment is

$$kA_H(s) + A_L(s) = \frac{P_n(-s)}{P_n(s)} \quad (1)$$

where $A_H(s)$ and $A_L(s)$ are the high-pass and low-pass transfer functions of a two-way crossover and $k = 1$

* Presented at the Conference of the Institute of Acoustics, Reproduced Sound 6, 1990 November; an expanded version was presented at the 93rd Convention of the Audio Engineering Society, San Francisco, CA, 1992 October 1–4; revised 1992 November 23.

for even n or $k = -1$ for odd n, n being the order of the all-pass polynomial $P_n(s)$.

To develop a method for identifying asymmetric crossovers we express the transfer functions $A_L(s)$ and $A_H(s)$ as

$$A_L(s) = \frac{LN(s)}{LD(s)}, \qquad A_H(s) = \frac{HN(s)}{HD(s)}$$

where, from Eq. (1),

$$k\frac{HN(s)}{HD(s)} + \frac{LN(s)}{LD(s)} = \frac{P_n(-s)}{P_n(s)} .$$

Rearranging in terms of the low-pass filter numerator polynomial LN(s),

$$LN(s) = \frac{LD(s)}{P_n(s)\ HD(s)} [P_n(-s)\ HD(s) - kP_n(s)\ HN(s)] .$$

Since LN(s) is a polynomial of finite order, then

$$LD(s) = P_n(s)\ HD(s) \tag{2}$$

and,

$$LN(s) = P_n(-s)\ HD(s) - kP_n(s)\ HN(s) . \tag{3}$$

Hence by specifying both the high-pass filter denominator polynomial and the polynomial $P_n(s)$, the transfer function of the corresponding low-pass filter can be derived. The following subsections describe a range of crossover examples where the orders of LD(s) and HD(s) differ.

1.1 First-Order High-Pass Filter with Polynomial $P_n(s)$ of Order n

The high-pass filter is defined as first order, where

$$A_H(s) = \frac{b_1 s}{1 + b_1 s} \tag{4}$$

that is,

$$HN(s) = b_1 s$$

$$HD(s) = 1 + b_1 s .$$

The polynomial $P_n(s)$ of order n is expressed as

$$P_n(s) = \sum_{r=0}^{n} a_r s^r . \tag{5}$$

Hence the denominator of the low-pass filter LD(s) follows from Eq. (2),

$$LD(s) = (1 + b_1 s) \sum_{r=0}^{n} a_r s^r . \tag{6}$$

In evaluating the numerator LN(s) from Eq. (3) there are two conditions: for n odd, $k = -1$ while for n even, $k = 1$. These conditions ensure that terms in s^{n+1} cancel. They also represent the relative inverted and noninverted connections between high-pass and low-pass channels. Hence for odd n,

$$LN(s)\Big|_{n\ \text{odd}} = \sum_{r=0}^{(n-1)/2} s^{2r} [a_{2r} + (2b_1 a_{2r} - a_{2r+1})\ s] \tag{7}$$

and for even n,

$$LN(s)\Big|_{n\ \text{even}} = a_0 - \sum_{r=1}^{n/2} [a_{2r-1}\ s^{2r-1} + (2b_1 a_{2r-1} - a_{2r}) s^{2r}] . \tag{8}$$

Using Eqs. (6)–(8) the low-pass filter transfer function $A_L(s)$ can be evaluated for any order of polynomial $P_n(s)$ when matched to the first-order high-pass function of Eq. (4), where the all-pass function follows directly from Eq. (1).

1.2 Polynomial $P_n(s)$, Order $n = 4$

An $n = 4$ example demonstrates the evaluation procedure and allows simple modification to $n = 2$ by setting appropriate coefficients to zero. From Eqs. (6) and (8) the transfer function for $A_L(s)$ is given as

$$A_L(s) = \frac{LN(s)}{LD(s)} = \frac{a_0 - a_1 s - a_3 s^3 - (2b_1 a_1 - a_2)s^2 - (2b_1 a_3 - a_4)s^4}{(1 + b_1 s)(a_0 + a_1 s + a_2 s^2 + a_3 s^3 + a_4 s^4)} \tag{9}$$

and the all-pass function as

$$\frac{P_4(-s)}{P_4(s)} = \frac{a_0 - a_1 s + a_2 s^2 - a_3 s^3 + a_4 s^4}{a_0 + a_1 s + a_2 s^2 + a_3 s^3 + a_4 s^4} . \tag{10}$$

The order of the polynomial LN(s) can be minimized [thus maximizing the high-frequency rate of attenuation

of $A_L(s)$] by setting

$$a_4 = 2b_1a_3$$

$$a_2 = 2b_1a_1$$

and normalizing the dc gain to 1, that is, $a_0 = 1$. Hence

$$A_L(s) = \frac{1 - a_1s - a_3s^3}{(1 + b_1s)(1 + a_1s + 2b_1a_1s^2 + a_3s^3 + 2b_1a_3s^4)} \quad (11)$$

$$\frac{P_4(-s)}{P_4(s)} = \frac{1 - a_1s + 2b_1a_1s^2 - a_3s^3 + 2b_1a_3s^4}{1 + a_1s + 2b_1a_1s^2 + a_3s^3 + 2b_1a_3s^4} . \quad (12)$$

1.3 Polynomial $P_n(s)$, Order $n = 2$

Eqs. (11) and (12) are modified to order $n = 2$ by setting coefficient $a_3 = 0$,

$$A_L(s) = \left(\frac{1 - a_1s}{1 + b_1s}\right) \frac{1}{1 + a_1s + 2b_1a_1s^2} \quad (13)$$

$$\frac{P_2(-s)}{P_2(s)} = \frac{1 - a_1s + 2b_1a_1s^2}{1 + a_1s + 2b_1a_1s^2} . \quad (14)$$

In Eq. (13) we note the special case where, if $a_1 = b_1$, $A_L(s)$ becomes second order in cascade with a first-order all-pass transfer function.

However, a characteristic revealed by Eq. (11) is that as the order of n is increased, the high-frequency rate of attenuation approaches only 12 dB per octave. Hence there is little advantage in seeking larger values of n, particularly as the phase distortion described by the all-pass transfer function [Eqs. (10) and (12)] becomes more severe. This observation is generalized in Section 2.

1.4 Polynomial $P_n(s)$, Order $n = 3$

A similar procedure is followed, but this time $A_L(s)$ is derived from Eqs. (6) and (7), where, for $n = 3$,

$$A_L(s) = \frac{LN(s)}{LD(s)} = \frac{a_0 + a_2s^2 + (2b_1a_0 - a_1)\,s + (2b_1a_2 - a_3)\,s^3}{(1 + b_1s)(1 + a_1s + a_2s^2 + a_3s^3)} \quad (15)$$

$$\frac{P_3(-s)}{P_3(s)} = \frac{1 - a_1s + a_2s^2 - a_3s^3}{1 + a_1s + a_2s^2 + a_3s^3} . \quad (16)$$

Again $LN(s)$ can be reduced in order by setting

$$a_3 = 2b_1a_2$$

$$a_1 = 2b_1a_0$$

$$a_0 = 1$$

and hence

$$A_L(s) = \frac{1 + a_2s^2}{(1 + b_1s)(1 + 2b_1s + a_2s^2 + 2b_1a_2s^3)} . \quad (17)$$

The denominator $D(s)$ of Eq. (17) can be factorized as

$$D(s) = (1 + b_1)(1 + 2b_1s + a_2s^2 + 2b_1a_2s^3)$$

$$= (1 + b_1s)(1 + ps)(1 + qs + rs^2) .$$

Assigning values to both b_1 and p and then comparing coefficients,

$$D(s) = (1 + b_1s)(1 + ps)(1 - ps)(1 + 2b_1s)$$

and

$$a_2 = -p^2 .$$

Then Eq. (17) reduces to

$$A_L(s) = \frac{1}{(1 + b_1s)(1 + 2b_1s)} \quad (18)$$

$$A_H(s) = \frac{b_1s}{1 + b_1s} \quad (19)$$

$$\frac{P_3(-s)}{P_3(s)} = \frac{P_1(-s)}{P_1(s)} = \frac{1 - 2b_1s}{1 + 2b_1s} . \quad (20)$$

This alignment is possibly the most useful as $A_L(s)$ remains second order and is all pole, while the all-pass function is first order, thus offering reduced phase distortion. Also Eqs. (18) and (19) show that the second low-pass filter pole is located at one-half the frequency of the high-pass filter pole. Consequently $A_L(s)$ has a greater attenuation at and above the high-pass filter break frequency, yet retains only gradual curvature in the amplitude response. In fact at the high-pass filter 3-dB break frequency, $|A_L(s)|$ exhibits 10-dB attenuation and the frequency of maximum group-delay distortion is one-half of the 3-dB break frequency of $A_H(s)$.

2 LIMITS TO DISPARITY IN ORDER OF ASYMMETRIC ALIGNMENTS

Higher order alignments can be identified using the procedures of Section 1; examples are tabulated in Section 3. However, we describe a second-order/fifth-order example to demonstrate that there is a limit in crossover usefulness as the disparity in filter order is increased.

For $n = 3$ let $P_3(s) = a_0 + a_1 s + a_2 s^2 + a_3 s^3$ and

$$A_H(s) = \frac{b_2 s^2}{1 + b_1 s + b_2 s^2} . \quad (21)$$

A fifth-order low-pass filter $A_L(s)$ follows as

$$A_L(s) = \frac{LN(s)}{LD(s)} = \frac{a_0 - (a_1 - a_0 b_1)s + (a_2 + 2a_0 b_2 - a_1 b_1)s^2 - (a_3 - a_2 b_1)s^3 + (2a_2 b_2 - a_3 b_1)s^4}{(a_0 + a_1 s + a_2 s^2 + a_3 s^3)(1 + b_1 s + b_2 s^2)} .$$

Setting $a_0 = 1$ and reducing the order of the numerator $LN(s)$ by equating coefficients in s^2, s^3, and s^4 to zero,

$$A_L(s) = \frac{1 - (a_1 - \sqrt{2b_2})s}{\{1 + a_1 s + (a_1\sqrt{2b_2} - 2b_2)s^2 + [2a_1 b_2 - (2b_2)^{3/2}]s^3\}(1 + \sqrt{2b_2}\, s + b_2 s^2)} \quad (22)$$

$$\frac{P_3(-s)}{P_3(s)} = \frac{1 - a_1 s + (a_1\sqrt{2b_2} - 2b_2)s^2 - [2a_1 b_2 - (2b_2)^{3/2}]s^3}{1 + a_1 s + (a_1\sqrt{2b_2} - 2b_2)s^2 + [2a_1 b_2 - (2b_2)^{3/2}]s^3} . \quad (23)$$

These equations show that with a second-order high-pass filter there is little advantage in seeking a polynomial $P_n(s)$ of order greater than 2 as there is no improvement in the rate of attenuation of $A_L(s)$ at high frequency, a result that mirrors the first-order case of Section 2.

Hence we may generalize by saying that for a high-pass filter of order r, the order of polynomial $P_n(s)$ should not exceed r if zeros in $A_L(s)$ are to be avoided, whereon it follows that the maximum order of $A_L(s)$ is $2r$.

A further reduction in order is possible in Eq. (22) by setting $a_1 = \sqrt{2b_2}$, where the all-pass and low-pass responses become first order while the high-pass filter remains second order. However, this interchange in transfer function order is approached more efficiently using the transformation described in Section 3.

3 ALTERNATIVE ASYMMETRY ALIGNMENTS

Following the procedure described in the previous sections further all-pass alignments can be identified where the low-pass filters have an order greater than the high-pass filters. However, by changing the order of the asymmetry using the s-domain transformation $\tau s \rightarrow 1/\tau s$ the range of theoretical alignments can be further extended.

This transformation can be applied to all the identified high-pass and low-pass filter pairs, as indicated in Table 1, where the notation $A_{LX}{}^T$ represents a low-pass filter of order x that is transformed (^{T}) from a high-pass prototype of order x. A similar notation is used also for the reverse transformation of a high-pass filter derived from a low-pass prototype.

The prototype alignments identified using the procedures in this paper together with their transformed filter pairs are tabulated in the following, where it is shown that the composite phase response is invariant of transformation in all cases except Section 3.5.

3.1 First-Order/Second-Order Filters

$$A_{L2} = \frac{2}{(2 + s\tau)(1 + s\tau)}$$

$$A_{H2}^T = \frac{2(s\tau)^2}{(1 + 2s\tau)(1 + s\tau)}$$

$$A_{H1} = \frac{s\tau}{2 + s\tau}$$

$$A_{L1}^T = \frac{1}{1 + 2s\tau}$$

$$A_{L2} + A_{H1} + A_{H2}^T + A_{L1}^T = 1$$

$$A_{L2} - A_{H1} = A_{L1}^T - A_{H2}^T = \frac{1 - s\tau}{1 + s\tau} .$$

3.2 Second-Order/Third-Order Filters

$$A_{L3} = \frac{2}{(1 + s\tau)[2 + 2s\tau + (s\tau)^2]}$$

$$A_{H3}^T = \frac{2(s\tau)^3}{(1 + s\tau)[1 + 2s\tau + 2(s\tau)^2]}$$

$$A_{H2} = \frac{(s\tau)^2}{2 + 2s\tau + (s\tau)^2}$$

$$A_{L2}^T = \frac{1}{1 + 2s\tau + 2(s\tau)^2}$$

$$A_{L3} - A_{H2} = A_{L2}^{T} - A_{H3}^{T} = \frac{1 - s\tau}{1 + s\tau}.$$

3.3 Second-Order/Fourth-Order Filters

$$A_{L4} = \frac{2}{[1 + s\tau + (s\tau)^2][2 + 2s\tau + (s\tau)^2]}$$

$$A_{H4}^{T} = \frac{2(s\tau)^4}{[1 + s\tau + (s\tau)^2][1 + 2s\tau + 2(s\tau)^2]}$$

$$A_{H2} = \frac{(s\tau)^2}{2 + 2s\tau + (s\tau)^2}$$

$$A_{L2}^{T} = \frac{1}{1 + 2s\tau + 2(s\tau)^2}$$

$$A_{L4} + A_{H2} = A_{L2}^{T} + A_{H4}^{T} = \frac{1 - s\tau + (s\tau)^2}{1 + s\tau + (s\tau)^2}.$$

3.4 Third-Order/Fifth-Order Filters

$$A_{L5} = \frac{2\sqrt{2}}{[1 + \sqrt{2}\,s\tau + (s\tau)^2][2\sqrt{2} + 4s\tau + 2\sqrt{2}(s\tau)^2 + (s\tau)^3]}$$

$$A_{H5}^{T} = \frac{2\sqrt{2}(s\tau)^5}{[1 + \sqrt{2}s\tau + (s\tau)^2][1 + 2\sqrt{2}s\tau + 4(s\tau)^2 + 2\sqrt{2}(s\tau)^3]}$$

$$A_{H3} = \frac{(s\tau)^3}{2\sqrt{2} + 4s\tau + 2\sqrt{2}(s\tau)^2 + (s\tau)^3}$$

$$A_{L3}^{T} = \frac{1}{1 + 2\sqrt{2}s\tau + 4(s\tau)^2 + 2\sqrt{2}(s\tau)^3}$$

$$A_{L5} + A_{H3} = A_{L3}^{T} + A_{H5}^{T} = \frac{1 - \sqrt{2}s\tau + (s\tau)^2}{1 + \sqrt{2}s\tau + (s\tau)^2}.$$

3.5 Third-Order/Sixth-Order Filters

$$A_{L6} = \frac{1}{(1 + s\tau)[1 + s\tau + (s\tau)^2][1 + 2s\tau + 2(s\tau)^2 + 2(s\tau)^3]}$$

$$A_{H6}^{T} = \frac{(s\tau)^6}{(1 + s\tau)[1 + s\tau + (s\tau)^2][2 + 2s\tau + 2(s\tau)^2 + (s\tau)^3]}$$

$$A_{H3} = \frac{(s\tau)^3}{(1 + s\tau)[1 + s\tau + (s\tau)^2]}$$

$$A_{L3}^{T} = \frac{1}{(1 + s\tau)[1 + s\tau + (s\tau)^2]}$$

$$A_{L6} - A_{H3} = \frac{1 - 2s\tau + 2(s\tau)^2 - 2(s\tau)^3}{1 + 2s\tau + 2(s\tau)^2 + 2(s\tau)^3}$$

$$A_{L3}^{T} - A_{H6}^{T} = \frac{2 - 2s\tau + 2(s\tau)^2 - (s\tau)^3}{2 + 2s\tau + 2(s\tau)^2 + (s\tau)^3}.$$

In this final example the phase response is not invariant to the transformation.

4 TWO-WAY LOUDSPEAKER SYSTEM USING AN ASYMMETRIC CROSSOVER WITH A FIRST-ORDER HIGH-PASS FILTER IN THE SATELLITE CHANNEL

As an example, consider a satellite loudspeaker (lower midrange to high-frequency band of operation) with a transfer function $A_S(s)$ which exhibits a second-order high-pass response with a 3-dB break frequency of approximately 70–80 Hz. This system is to be interfaced with a subwoofer system which has a low-frequency transfer function $A_W(s)$. In the frequency range of 100–500 Hz both satellite and subwoofer show well-behaved responses. The crossover alignment selected for this example is given in Section 1, where $A_L(s)$, $A_H(s)$, and $P_1(-s)/P_1(s)$ are described by Eqs. (18), (19), and (20), respectively.

In Fig. 1 the two-way system is shown, where the asymmetric high-pass and low-pass filters are implemented using passive RC circuit elements. The crossover frequency (−3 dB) is set at 200 Hz and the filters are designed using the equations presented in Fig. 1. The

Table 1.

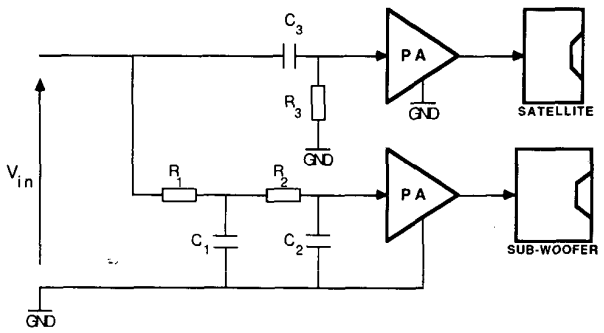
Filter characteristic	Symmetric alignments				Asymmetric alignments 3.1	3.2	3.3	3.4	3.5	Transformed asymmetric alignments 3.1	3.2	3.3	3.4	3.5
Low pass	A_{L1}	A_{L2}	A_{L3}	A_{L4}	A_{L2}	A_{L3}	A_{L4}	A_{L5}	A_{L6}	A_{L1}^{T}	A_{L2}^{T}	A_{L2}^{T}	A_{L3}^{T}	A_{L3}^{T}
High pass	A_{H1}	A_{H2}	A_{H3}	A_{H4}	A_{H1}	A_{H2}	A_{H2}	A_{H3}	A_{H3}	A_{H2}^{T}	A_{H3}^{T}	A_{H4}^{T}	A_{H5}^{T}	A_{H6}^{T}

subwoofer in the example has an extended response to 20 Hz and $Q = 0.5$, while the satellite has an undamped natural resonance of 70 Hz and $Q = 0.7$. Computed results are then presented in Fig. 2 (a) and (b), which shows individual satellite and subwoofer amplitude and phase responses both with and without the associated crossover filters, while Fig. 3 describes the overall response that includes both the asymmetric crossover alignment and the drive unit responses.

The composite amplitude response reveals a smooth characteristic with low error, validating the effectiveness of the alignment. Also by setting the crossover to 200 Hz, the output of the satellite loudspeaker is curtailed adequately at low frequency, thus reducing distortion through excessive cone excursion, while the subwoofer commences its attenuation region at 100 Hz rather than 200 Hz and, being second order, achieves a respectable attenuation at midrange and high frequencies. The attraction of this system is the low circuit complexity and the achievement of adequate signal attenuation for each drive unit by using the overlapping crossover response offered by an asymmetric alignment.

5 ASYMMETRIC CROSSOVER ACTIVE SYNTHESIS

The asymmetric crossover alignments of order greater than 2 can be synthesized using a state-variable topology similar to those proposed earlier [3] for the Linkwitz–Riley LR-4 crossover. The advantage of this approach is both the reduced number of reactive circuit elements and the more direct active-circuit signal path for the critical high-frequency channel. Two examples of asymmetric crossovers are presented in this section



Design: Let satellite 3 dB break frequency = f_0 Hz

Set capacitors C_2, C_3, then $R_1 = 1/(12\pi f_o C_2)$

$R_2 = 9R_1$

$C_1 = 8C_2$

$R_3 = 1/(2\pi f_o C_3)$

Fig. 1. Two-way active loudspeaker using asymmetric crossover with passive low-level circuity.

that include the transformed alignments discussed in Section 3.

The two sets of topology are shown in Figs. 4 and 5, where integrator time constants are aligned to match the corresponding transfer functions defined in Sections 3.3 and 3.4, respectively. To validate the performances of these four crossover networks, Figs. 6 and 7 show the frequency responses, including both the transformed and the composite responses, where the phase responses in all four cases are normalized to be symmetric about the defined crossover frequency $f_c = 1$ Hz. It should be noted that each filter section in Figs. 6 and 7 introduces a gain of 6 dB. However, these have been accounted for in generating the frequency response results. Also the curve for each composite amplitude response is reduced by 6 dB for greater clarity.

To demonstrate the means of calculating integrator time constants, the third-order section of the third-order high-pass/fifth-order low-pass state-variable filter of Fig. 5(a) is analyzed; the low-pass, high-pass, and composite phase transfer functions are described in Section 3.4.

The denominator of the low-pass transfer function $A_{L5}(s)$ is presented as the product of a second-order and a third-order polynomial. Since the high-pass section contains the same third-order term, we choose to split the filter into two cascaded stages, namely, a third-order stage from which the high-pass function emerges followed by a cascaded second-order low-pass section from which the final low-pass output is derived.

In Fig. 5(a) the topology uses a unity-gain difference amplifier in the feedback path of the third-order section. Examinations of the cascade of inverting integrators reveals the output of this amplifier to be

$$V_d = V_{high}\left(\frac{1}{sR_1C_1} + \frac{1}{s^2R_1C_1R_2C_2} + \frac{1}{s^3R_1C_1R_2C_2R_3C_3}\right).$$

Also from the input amplifier of the selected topology,

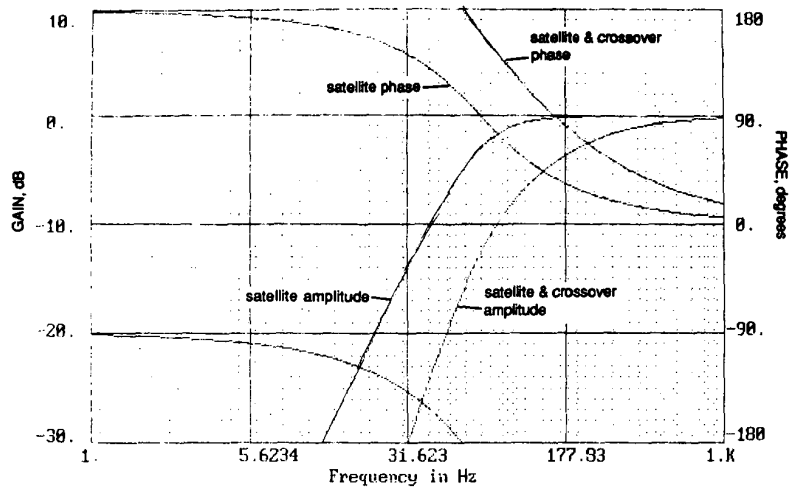
$$V_{high} = 2V_{in} - V_d .$$

Hence substituting for V_d and rearranging,

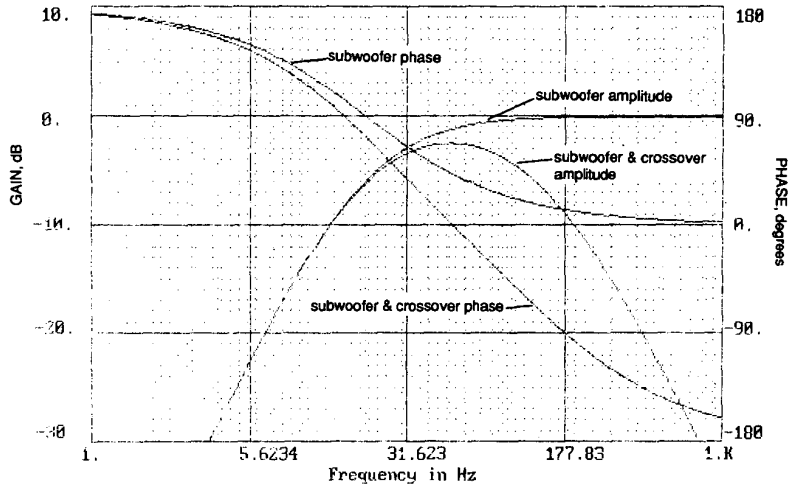
$$V_{high}\left(1 + \frac{1}{sR_1C_1} + \frac{1}{s^2R_1C_1R_2C_2} + \frac{1}{s^3R_1C_1R_2C_2R_3C_3}\right) = 2V_{in} .$$

This expression can be compared with the third-order high-pass transfer function, which after rearrangement yields

$$V_{high}\left[1 + \frac{2\sqrt{2}}{s\tau} + \frac{4}{(s\tau)^2} + \frac{2\sqrt{2}}{(s\tau)^3}\right] = V_{in} .$$



(a)



(b)

Fig. 2. Amplitude and phase responses with and without crossover filter. (a) Satellite. (b) Subwoofer.

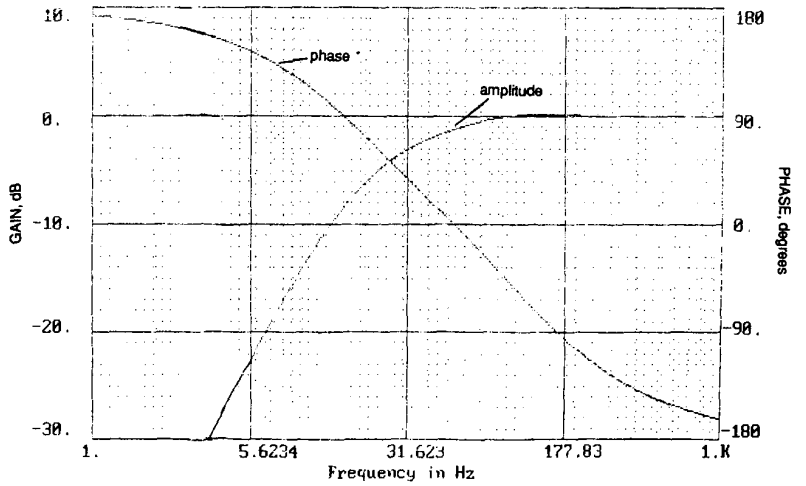


Fig. 3. Composite amplitude response of satellite and subwoofer with asymmetric crossover filters.

Hence, comparing coefficients,

$$R_1C_1 = \frac{\tau}{2\sqrt{2}}$$

$$R_1C_1R_2C_2 = \frac{\tau^2}{4}$$

that is,

$$R_2C_2 = \frac{\tau}{\sqrt{2}}$$

$$R_1C_1R_2C_2R_3C_3 = \frac{\tau^3}{2\sqrt{2}}$$

that is,

$$R_3C_3 = \sqrt{2}\tau \ .$$

Observing the expression for the all-pass phase response, the crossover frequency occurs at $f_c = 1/2\pi\tau$. Hence the respective time constants follow, as shown in Fig. 5(a).

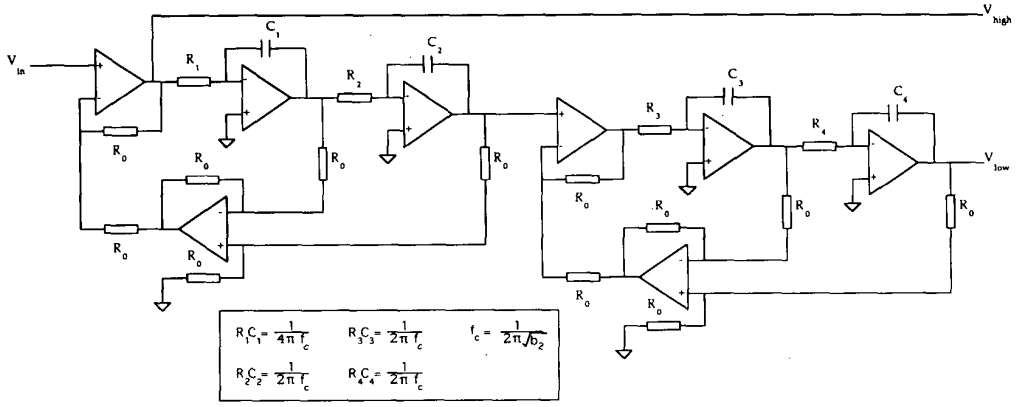
The corresponding low-pass output V_{L3} of the third-order filter section is related to V_{high} by

$$V_{L3} = \frac{-V_{high}}{s^3R_1C_1R_2C_2R_3C_3}$$

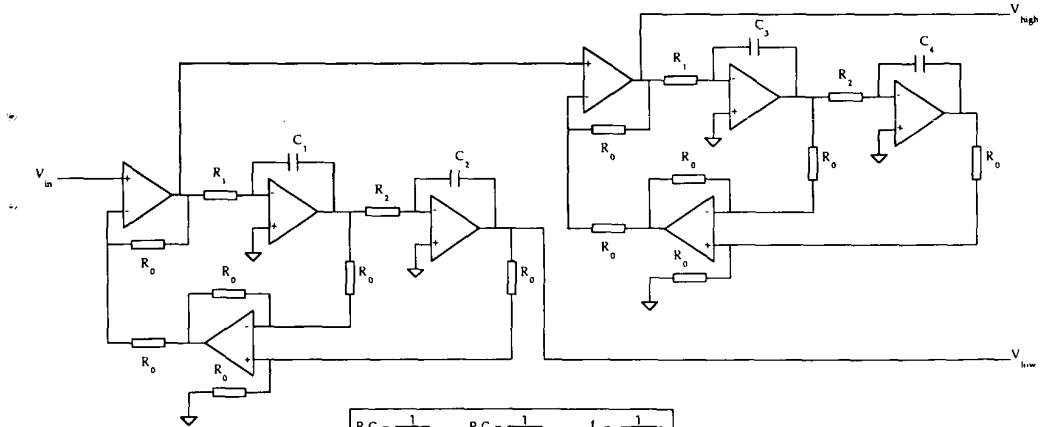
which in turn drives the second-order low-pass section. A similar analysis procedure then applies to this second stage to realize the time constants R_4C_4 and R_5C_5, again presented in Fig. 5(a).

6 CONCLUSION

A set of asymmetric all-pass crossovers up to the combination of third-order high-pass/sixth-order low-pass alignments has been described. The defining equations and their corresponding reverse-order transformations are summarized in Section 3. The pairing of high-pass/low-pass filter alignments against the order



(a)



(b)

Fig. 4. Asymmetric crossover second-order/fourth-order transform pair circuit synthesis. (See Section 3.3). (a) A_{L4}, A_{H2}. (b) A_{L2}^T, A_{H5}^T.

of all-pass polynomials is also summarized in Fig. 8 for those filters analyzed, including both symmetric and asymmetric cases. However, the analysis can be extended to any filter of order r, where, as discussed in Section 2, the maximum useful filter order for the complementary amplitude response is $2r$, although some filters in the range $r + 1$ to $2r$ may be possible within an asymmetric alignment with a corresponding reduction in order of the all-pass polynomial. For example, A_{L3}/A_{H6} has a third-order all-pass polynomial, while A_{L3}/A_{H5} has a second-order polynomial. Indeed, Fig. 8 suggests that for the set of realizable asymmetric filters the order of the all-pass polynomial is the difference in order between the high-pass/low-pass filter pairs. However, if filters of order greater than $2r$ are sought, then the associated zeros in the numerator of $A_L(s)$ must be accepted, which limit the ultimate rate of attenuation, even though the zeros may cause a higher initial rate of attenuation. Fig. 8 also classifies the well-known symmetric all-pass filters, although these are not discussed here.

The first-order high-pass example described in Section 1 is particularly useful because the low circuit overhead bodes well for minimizing signal impairment, and the overlapping of the transfer functions also increases their effective attenuation. However, for the higher order alignments Section 5 showed that efficient feedback systems can both realize the desired transfer functions while using a minimum of reactive circuit elements and aid time-domain synchronization of the low-pass and high-pass impulse responses.

7 REFERENCES

[1] B. Hillerich, "Acoustic Alignment of Loudspeaker Drivers by Nonsymmetrical Crossovers of Different Orders," *J. Audio Eng. Soc.*, vol. 37, pp. 691–699 (1989 Sept.).

[2] S. P. Lipshitz and J. Vanderkooy, "A Family of Linear-Phase Crossover Networks of High Slope Derived by Time Delay," *J. Audio Eng. Soc.*, vol. 31, pp. 2–20 (1983 Jan./Feb.).

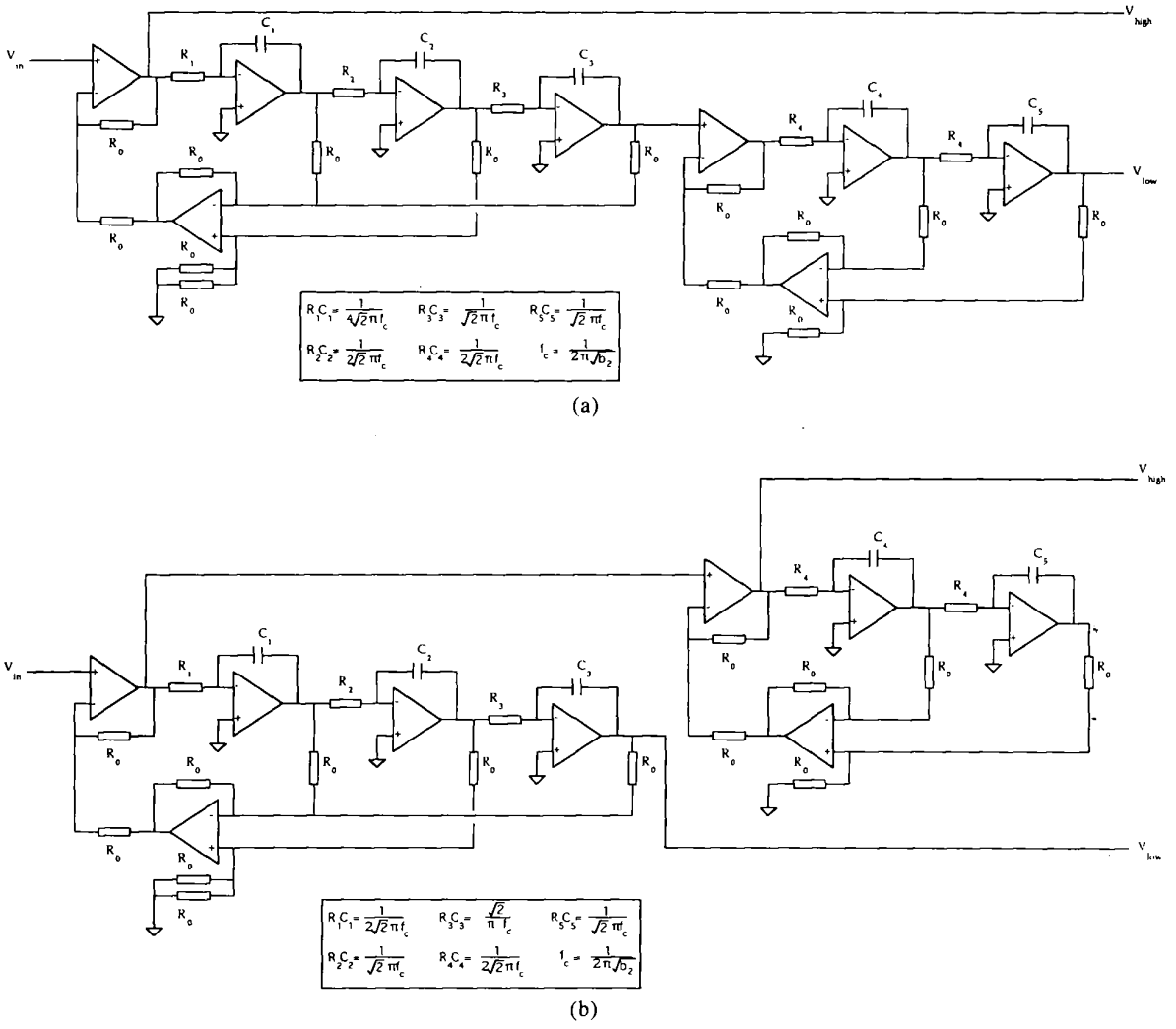


Fig. 5. Asymmetric crossover third-order/fifth-order transform pair circuit synthesis. (See Section 3.4.) (a) $A_{L5}\,A_{H3}$. (b) A^T_{L3}, A^T_{H5}.

[3] D. A. Bohn, "A Family of Circuit Topologies for the Linkwitz–Riley (LR-4) Crossover Alignment," presented at the 85th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 36, p. 1024 (1988 Dec.), preprint 2697.

8 BIBLIOGRAPHY

R. M. Bews, "Digital Crossover Networks for Active Loudspeaker Design," Ph.D. dissertation, University of Essex, Colchester, UK (1987).

D. A. Bohn, "A Fourth-Order State-Variable Filter for Linkwitz–Riley Active-Crossover Designs," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 960 (1983 Dec.), preprint 2011.

D. A. Bohn, "An Eight-Order State-Variable Filter for Linkwitz–Riley Active-Crossover Design," presented at the 85th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 36, p. 1024 (1986 Dec.), preprint 2697.

J. D'Appolito, "Active Realisation of Multiway All-Pass Crossover Systems," presented at the 76th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 32, p. 1008 (1984 Dec.), preprint 2125.

P. Garde, "All-Pass Crossover Systems," *J. Audio Eng. Soc.*, vol. 28, pp. 575–584 (1980 Sept.).

S. H. Linkwitz, "Active Crossover Networks for Non-coincident Drivers," *J. Audio Eng. Soc.*, vol. 24, pp. 2–8 (1976 Jan./Feb.).

S. P. Lipshitz and J. Vanderkooy, "Use of Frequency Overlap and Equalization to Produce High-Slope Linear-Phase Loudspeaker Crossover Networks," *J. Audio Eng. Soc.*, vol. 33, pp. 114–126 (1985 Mar.).

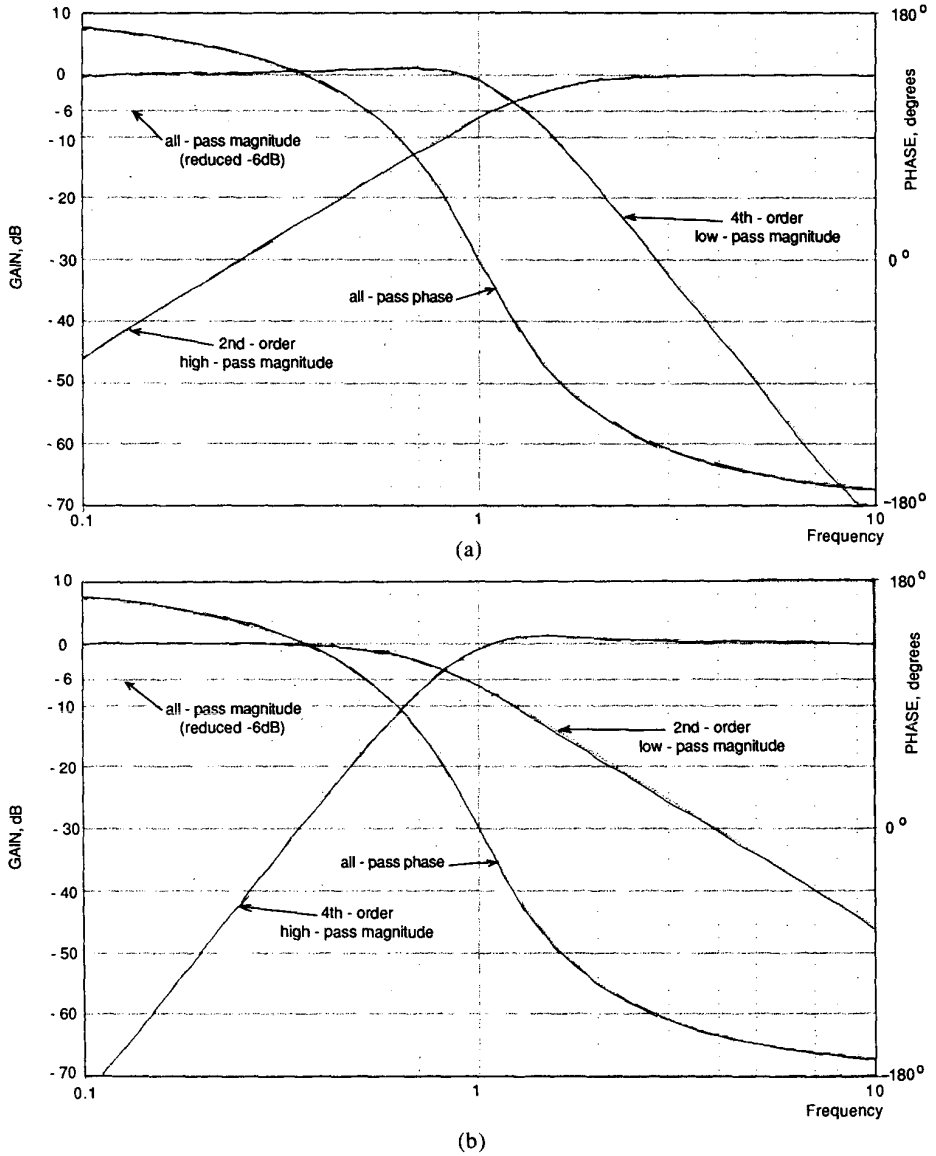


Fig. 6. Low-pass, high-pass, and all-pass responses. (a) Corresponding to Fig. 4(a). (b) Corresponding to Fig. 4(b).

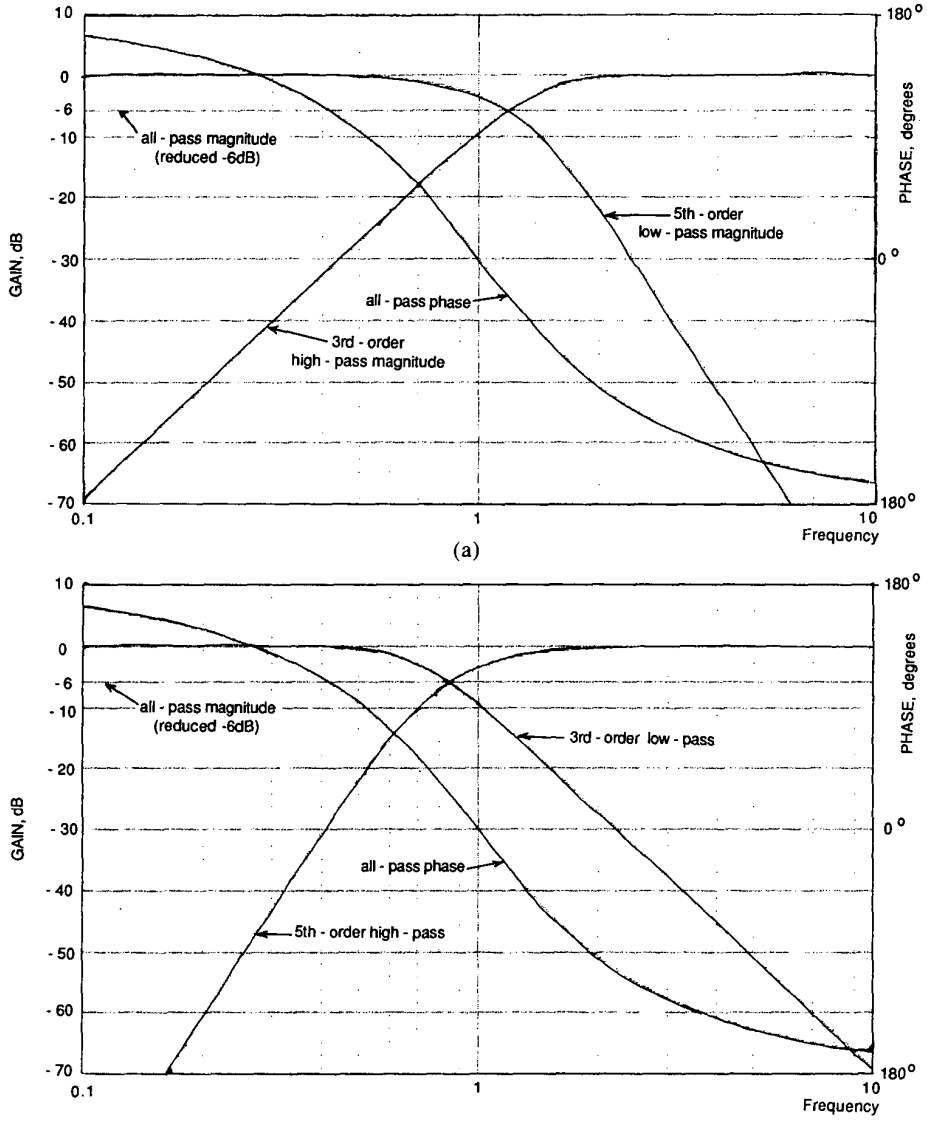


Fig. 7. Low-pass, high-pass, and all-pass responses. (a) Corresponding to Fig. 5(a). (b) Corresponding to Fig. 5(b).

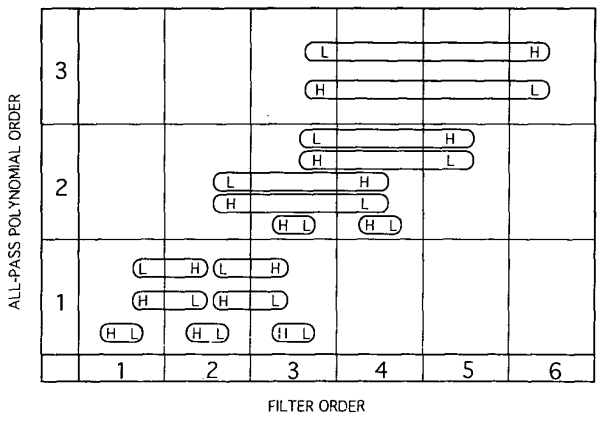


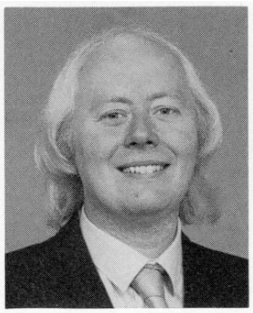
Fig. 8. Map of crossover filters with analzyed all-pass composite responses.

J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters, *IEEE Trans. Audio Electroacoust.* (1973 Dec.); also in *Programs for Digital Signal Processing*, IEEE Press, New York, 1979, pp. 5.1.1–13.

M. Rutz, "Digital Implementation of Loudspeaker Crossover Systems," Internal Research Rep., Dept. of Electronic Systems Engineering, University of Essex, Colchester, UK (1982).

R. H. Small, "Constant-Voltage Crossover Network Design," *J. Audio Eng. Soc.*, vol. 19, pp. 12–19 (1971 Jan.).

THE AUTHOR



Malcolm Omar Hawksford is a reader in the Department of Electronic Systems Engineering at the University of Essex, where his principal interests are in the fields of electronic circuit design and audio engineering. Dr. Hawksford studied at the University of Aston in Birmingham and gained both a First Class Honors B.Sc. and Ph.D. The Ph.D. program was supported by a BBC Research Scholarship, where the field of study was the application of delta modulation to color television and the development of a time compression/time multiplex system for combining luminance and chrominance signals. Since his employment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing, and loudspeaker systems has been undertaken. Since 1982 research into digital crossover systems has begun within the group and, more recently, oversampling and noise shaping investigated as a means of analog-to-digital/digital-to-analog conversion.

Dr. Hawksford has had several AES publications that include topics on error correction in amplifiers and oversampling techniques. His supplementary activities include writing articles for *Hi-Fi News* and designing commercial audio equipment. He is a member of the IEE, a chartered engineer, a fellow of the AES and of the Institute of Acoustics, and a member of the review board of the *AES Journal*. He is also a technical adviser for *HFN* and *RR*.

On the Dither Performance of High-Order Digital Equalization for Loudspeaker Systems*

R. G. GREENFIELD AND M. O. J. HAWKSFORD, *AES Fellow*

Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

Experimental results derived using real-time hardware together with supporting software simulations reveal that for infinite impulse response digital filters low-level chaos acts as self-dither, which both decorrelates truncation distortion and aids linearization of low-level digital-to-analog conversion nonlinearity. The process is examined for a range of filter responses, in particular those that can match the requirements for loudspeaker equalization.

LIST OF SYMBOLS

a_j = feedback coefficients
b_i = feedforward coefficients
$d(n)$ = filter-shaped quantization noise sequence
$E\{\}$ = expectation operator
$H(z)$ = system transfer function
M = length of feedback tapped delay line
N = length of feedforward tapped delay line
$q(n)$ = quantization noise sequence
$Q(z)$ = quantization noise frequency spectrum
$x(n)$ = input data sequence
$X(z)$ = input data frequency spectrum
$y(n)$ = noiseless output data sequence
$Y(z)$ = noiseless output data frequency spectrum
$\hat{y}$ = noisy output data sequence
$\hat{Y}$ = noisy output data frequency spectrum
z = z-transform operator

0 INTRODUCTION

The application of digital techniques to audio has been increasing at an almost exponential rate over the past few years, with digital systems starting to overtake the role previously played by analog functions. In the digital domain these functions map to mathematical operations performed on the data, often generating fractional quantities or increased signal levels, both of which increase the word length needed to represent the data. However, owing to practical limitations governed ultimately by the digital-to-analog converter (DAC), requantization of the signal back to the working number of bits will be required at some point. To give an example, this scenario is commonplace in the equalization of a loudspeaker's response operating to the 16-bit CD format. If, say, the loudspeaker exhibits a 3-dB dip at the crossover region, a matched equalizer dictates a corresponding peak at this frequency, thereby increasing the dynamic range of the outgoing signal by 3 dB. To be compatible with the rest of the system the signal must be scaled by 1 bit either at the input or at the output of the digital equalizer. Scaling has serious consequences as it not only introduces nonlinear distortion but will further strip any dither (below 1 bit in this case) that previously may have been added to the signal. In general any processes that involve the multiplication of two or more numbers can result in word lengths greater than that of the input signal. Thus even filters with unity gain in the passband are not devoid of quantization effects.

The problem of quantization distortion is not a new one, and its first noted manifestation was in the realm of video transmission. The effects of finite quantization in audio are similar to those in video, and the accepted solution, based on the human averaging process on rapidly changing signals, is also the same. The solution, generically known as dithering, is dealt with in depth by a series of papers by Vanderkooy et al. [1]–[3]. In essence dither is a low-level additive signal (such as a white noise) that causes the signal to jump randomly between quantization levels which, when averaged by the ear, forms a smooth transition between the levels. A side effect of dithering is an increase in total noise. However, the benign nature of this noise is preferable to the harmonic distortion artifacts caused by the highly

* Presented at the 88th Convention of the Audio Engineering Society, Montreux, Switzerland, 1990 March 13–16; revised 1995 August 8.

nonlinear quantization process, and as such dithering is now considered mandatory in high-quality audio.

The subject of this paper arose from the authors' concern about the high levels of quantization distortion introduced by a loudspeaker equalizer. The equalizer was implemented partially with high-order infinite impulse response (IIR) filters, which are themselves notoriously noisy, and there was no desire to further degrade the noise performance by adding digital dither. With this in mind harmonic distortion measurements were made on a system consisting of the Sony CP-553 ESD CD player and DAS-703 ES outboard DAC with the equalizer inserted within the AES/EBU interface. The test signals were low-level dithered sinusoids taken from the *Hi Fi News* test disc CD II HFN 015. To the authors' surprise the effect of the equalizer was to actually reduce the harmonic distortion generated by the Sony DAC. Fig. 1 shows the frequency response of the −70-dB sine wave with and without loudspeaker equalization. There is an amplitude reduction in the equalized fundamental frequency which is in sympathy with the response of the equalizer. However, the improvement in harmonic distortion performance relative to the noise floor is apparent.

The reduction in distortion is attributed to the noise generation mechanism inherent in the structure of IIR filters, which decorrelates the quantization distortion from the primary signal at both the equalizer filter and the DAC outputs. The roundoff or truncation noise response of IIR filters is a subject that has received much attention, with an emphasis on minimizing either its magnitude or its audibility. The paradox here is that the presence of this noise actually fulfills the function performed by digital dither that would otherwise be needed to restore system linearity. This paper proposes the analogy between additive digital dither and the fed back quantization error coincident with IIR structures. By way of computer simulations and real-time measurements the significance of the filter frequency response and the corresponding noise probability distribution function (PDF) is examined. A parallel between noise shaping and IIR filter noise is drawn with some subjective implications considered, particularly in relation to loudspeaker equalization.

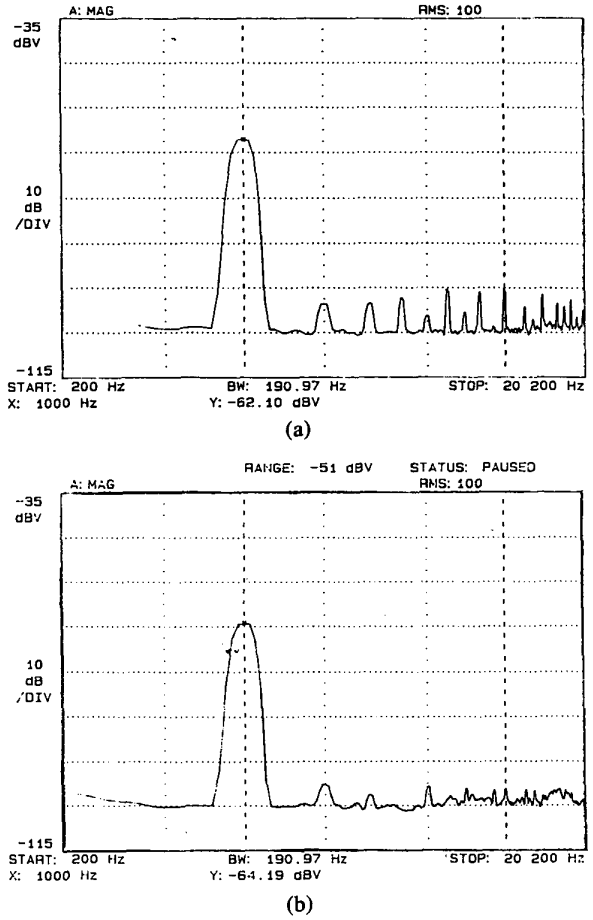


Fig. 1. Harmonic distortion of Sony DAC. (a) Before loudspeaker equalization. (b) After loudspeaker equalization.

1 QUANTIZATION NOISE IN IIR FILTERS

With all finite word length IIR structures there will be some node at which data must be quantized (the term quantized is used generally to mean either truncated or rounded) and fed back to a prior node. This process modifies the noise spectrum at the output, depending on both the filter's frequency response and its structure. For reasons beyond the scope of this paper (see Dattorro [4]) the filter structures most widely used are the direct form 1 (DF1) or the transposed direct form 2 (TDF2). We shall concentrate on the DF1 structure as, while exhibiting similar noise response, the DF1 structure gives a clearer insight into the processes involved.

A typical nonsubtractive dither scheme is shown in Fig. 2. Here a digitized noise source of finite word length and with peak magnitude somewhere in the region of the least significant bit (LSB) is added to the data stream before the quantizer.[1] Compare this now to the DF1 structure of Fig. 3 with a 16/32 architecture (16-bit word length, 32-bit accumulator). The input to the feedback path is $y(n) + q(n)$, where $y(n)$ is the true[2] output and $q(n)$ is a 16-bit error term representing a fractional quantity below the LSB. $q(n)$ is not strictly a random quantity

[1] For more details on digital dithering the reader is referred to Lipshitz and Vanderkooy [3].

[2] We shall consider that after processing, a 32-bit number represents the true response whereas a 16-bit number is the quantized version.

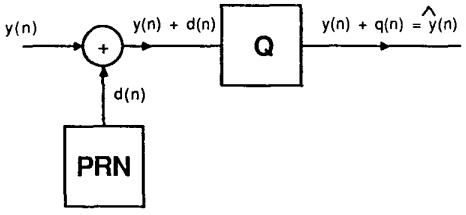


Fig. 2. Additive digital dither using pseudorandom number generator. Q—quantizer.

as it is both input-signal dependent and deterministic. However, for most signals present distinct values of $q(n)$ can be assumed to be statistically independent. While this latter statement is not mathematically correct, for most signals encountered in audio applications there will be some wide-band signal content in the signal presented to the quantizer. This will be true not only for recorded performances which may be considered dithered by analog noise prior to analog-to-digital conversion, but also for electronically generated signals which must be dithered prior to 16-bit (or higher) quantization in order for themselves to be free of quantization distortion. The effect of feedback in the IIR structure increases both the quantization noise power output from the filter and the moment of the noise sequence. If we assume that the initial quantization error term is independent from the output signal $y(n)$, then this term will be recirculated back to the input of the filter in an infinite cycle which, following the central limit theorem, will result in a Gaussian PDF noise sequence at the output of the filter. It is this broad assumption of independence between distinct quantization noise samples that determines whether IIR filters are self-dithering.[3] This will be the subject of investigation in the following sections. However, in order to analyze the effect of the IIR filter on the output quantization noise, we will assume that the quantization noise is statistically independent. Thus the autocorrelation of $q(n)$ can be given by

$$E\{q(n) \cdot q(n-k)\} = 0|_{k\neq 0} \tag{1}$$

where $E\{\ \}$ is the expectation operator.

The signal presented to the quantizer is then the summation of $y(n)$ and a delayed modified version of $q(n)$, designated $d(n)$, and is given by

$$y(n) + d(n) = \sum_{i=0}^{N} b_i x(n-i) + \sum_{j=1}^{M} a_j y(n-j) + \sum_{j=1}^{M} a_j q(n-j) . \tag{2}$$

If Eq. (1) holds, $d(n)$ and $y(n)$ are uncorrelated, thus decorrelating the quantization error $q(n)$ from the output sequence $y(n)$. The sequence $d(n)$, discussed later in more detail, is therefore analogous to the dither source of Fig. 2. Fig. 4(a) shows a computer simulation of the harmonic distortion generated by a 16-bit quantizer on a low-level (−78-dB) sinusoid with pseudo analog dither.[4] Even though the signal is dithered using rectangular PDF dither, there is still some evidence of correlation between the quantization distortion and the output signal $y(n)$. This is probably due to deficiencies in the random-number generating algorithm used or an insufficient integration period. Fig. 4(b) shows the resultant spectrum of the same sinusoid after digital filtering by a model of an IIR equalizer with no additional dither at the output quantizer. The spectral plots demonstrate that the quantization at the output of the equalizer is linear, more so than the directly quantized signal. This characteristic supports the hypothesis that the feedback mechanism further randomizes the dither component input to the quantizer. Thus in this instance, the assumption that the quantization noise sequence $q(n)$ is independent of $y(n)$ appears to hold.

2 NOISE-SHAPING PROPERTY OF IIR FILTERS

The motivation for the study of this topic was inspired by DAC linearity measurements which proved to be equalizer dependent. To understand this phenomenon it is necessary to examine the effect the filter has on the noise response. With reference to Fig. 3, the output of the filter before final quantization is given by

$$y(n) + d(n) = \sum_{i=0}^{N} b_i x(n-i) + \sum_{j=1}^{M} a_j \hat{y}(n-j) . \tag{3}$$

[3] This assumption may fail when the filter transfer function exhibits broad-band attenuation, thus stripping low-level activity in the signal prior to the quantization node. While this is a subject that requires further investigation, it is not generally the case with most filter requirements, particularly those of loudspeaker equalizers. It should be noted that the noise generated by IIR structures is certainly not sufficient for the dither requirements of digital attenuators where triangular or higher moment PDF dither should be used prior to quantization.

[4] The dither is simulated by a floating-point pseudorandom number generator with uniform PDF and peak-to-peak amplitude of ±0.5 LSB.

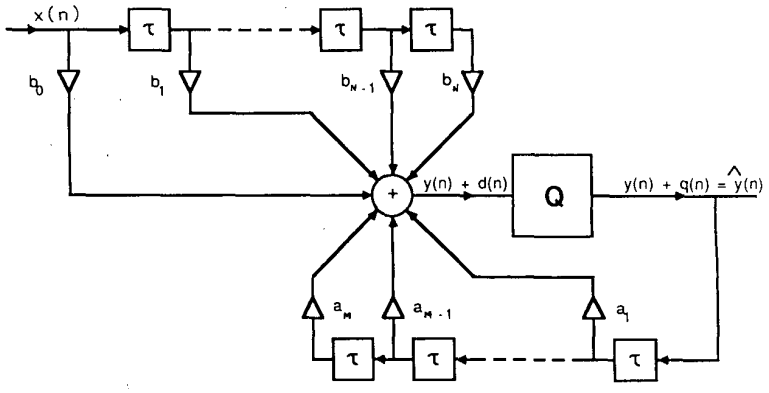


Fig. 3. IIR filter structure with quantizer at output.

Now $\hat{y}(n) = y(n) + q(n)$, where $\hat{y}(n)$ incorporates $d(n)$, allowing us to write

$$\hat{y}(n) - q(n) = \sum_{i=0}^{N} b_i x(n-i) + \sum_{j=1}^{M} a_j \hat{y}(n-j) . \quad (4)$$

If we represent $q(n)$ in the frequency domain by $Q(z)$, the frequency response at the output of the filter is given by

$$\hat{Y}(z) = X(z) \frac{\Sigma_{i=0}^{N} b_i z^{-i}}{1 - \Sigma_{j=1}^{M} a_j z^{-j}} + \frac{Q(z)}{1 - \Sigma_{j=1}^{M} a_j z^{-j}} . \quad (5)$$

Thus we can observe that, for the DF1 structure, the quantization noise is spectrally shaped by the poles of the filter. This is important as it gives insight into the effect the filter has on system linearity.

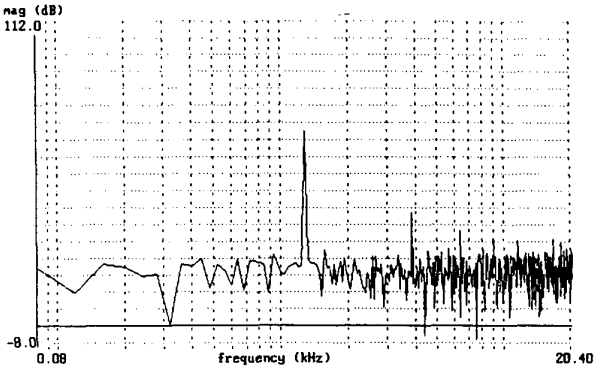
Using Eq. (5), three second-order all-pass filters were designed and implemented in a real-time system using the TMS320C25 digital signal processor. All-pass functions were chosen so that the fundamental amplitude of the test signal should not be altered while allowing different pole responses, thereby modifying only the noise performance of the filter. The three pole responses chosen were low-frequency emphasis, high-frequency emphasis, and approximately flat. To complement the real-time measurements, computer simulations evaluated three other filter characteristics—frequency response of the filter poles, noise spectrum, and noise PDF. The pole spectral response is simply found by evaluating the function

$$H(z) = \frac{1}{1 - \Sigma_{j=1}^{M} a_j z^{-j}} . \quad (6)$$

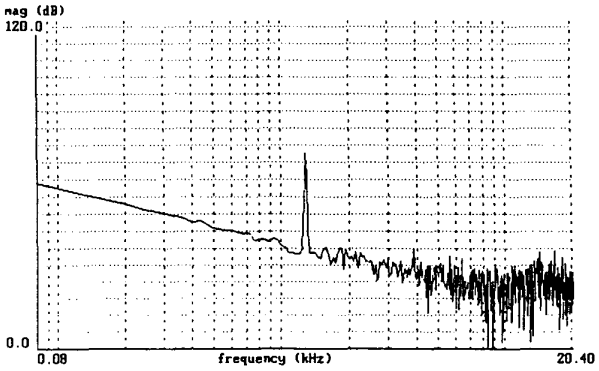
The noise spectral density and PDF are calculated using the simulation scheme shown in Fig. 5. The actual filter is the real-time system program run on the TMS320C25 IBM PC evaluation module. The ideal filter is a computer simulation operating with 64-bit floating-point arithmetic. All filters are input the same computer-generated sinusoidal signal. At the output the simulated signal is subtracted from the actual signal, leaving only the noise generated by the actual filter from which its spectrum and PDF are calculated.

Figs. 6, 7, and 8 show the following plots for the three all-pass filters: (a) the harmonic distortion at the output of the Sony DAC, (b) the pole frequency response, (c) the noise spectrum, and (d) the noise PDF. These plots reveal some interesting points which support the assumption made in Eq. (1). The pole frequency response and noise frequency response match very closely, indicating that $Q(z)$ is approximately spectrally white. Second, there is no indication of any harmonic relationship between the noise spectrum and the output signal, implying that the noise sequence $q(n)$ is independent of $y(n)$. Finally, observing the spread of noise variances, the noise PDFs of the three filters appear to conform, albeit crudely, to the set of bell-shaped contours characteristic of a Gaussian distribution. These three points indicate that the quantization error sequence $q(n)$ approaches a Gaussian noise sequence.

Consider next the relation between pole frequency response and improvement in DAC linearity. Comparing plots (a) and (b) in Figs. 6–8, it is evident that the filters with greatest noise gain result in the greatest DAC linearity improvement (the high-frequency emphasis, low-frequency emphasis, and flat filters have noise gains of 37 dB at 20 kHz, 35 dB at 10 Hz, and 1 dB at 10 Hz to 20 kHz, respectively). The same conclusions can be drawn by observing the spread of the noise PDF plots. The increase in noise power generated by the filters causes the peak-to-peak noise levels to exceed the boundaries of additional DAC quantization levels, thus further decorrelating differential DAC nonlinearity. We can conclude from this result that the effective improve-



(a)



(b)

Fig. 4. Simulation of harmonic distortion on low-level sinusoid. (a) 16-bit linear quantizer. (b) IIR filter designed for loudspeaker equalization using a 16/32-bit processor architecture.

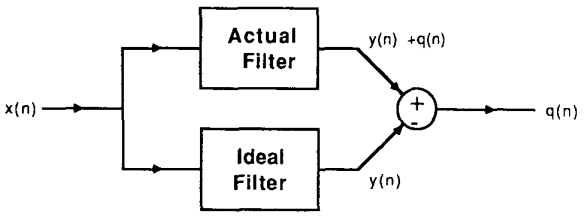
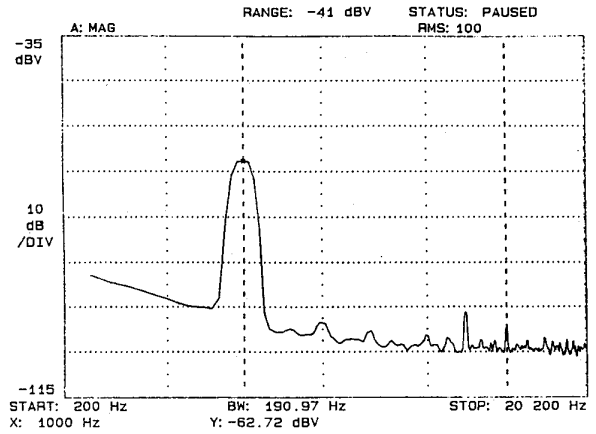
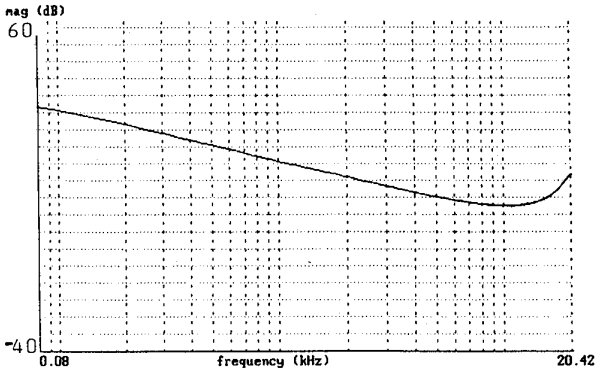


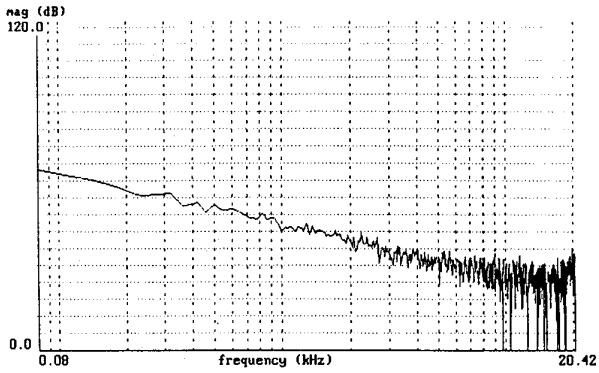
Fig. 5. Simulation of noise generated by filters used in real-time systems.



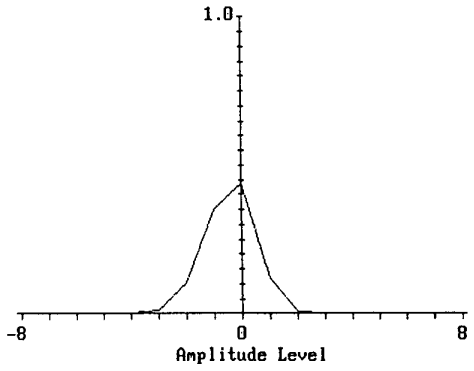
(a)



(b)

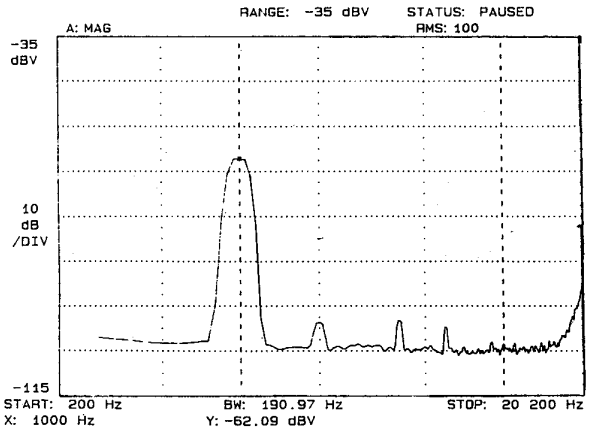


(c)

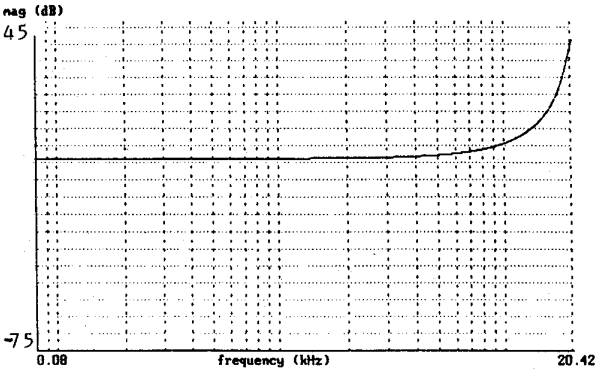


(d)

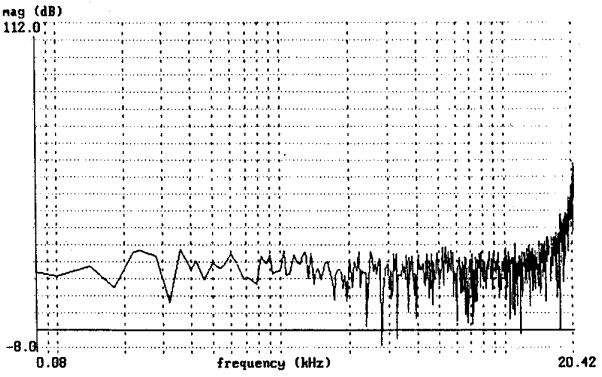
Fig. 6. Responses of low-frequency noise emphasis all-pass filter. (a) Harmonic distortion. (b) Pole response. (c) Noise spectrum. (d) Noise PDF.



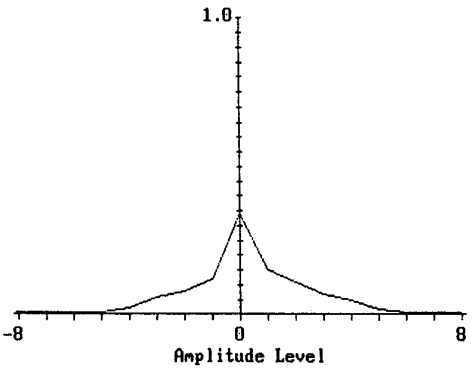
(a)



(b)



(c)

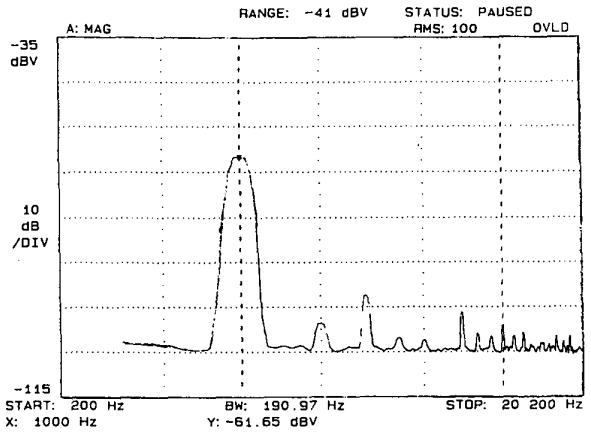


(d)

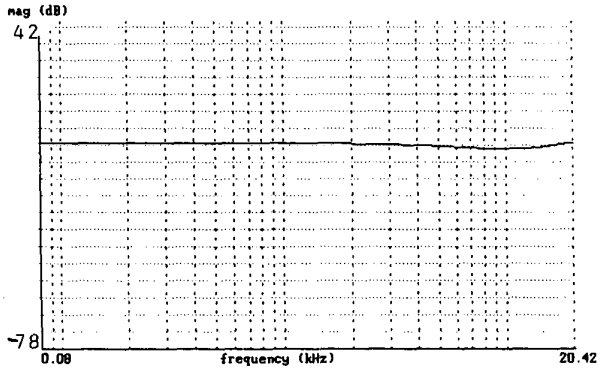
Fig. 7. Responses of high-frequency noise emphasis all-pass filter. (a) Harmonic distortion. (b) Pole response. (c) Noise spectrum. (d) Noise PDF.

ment in overall system linearity is related directly to the pole transfer function of the IIR filters. The foregoing comments relate only to multibit DACs as to date the authors have not performed the same experiment on single-bit DACs.

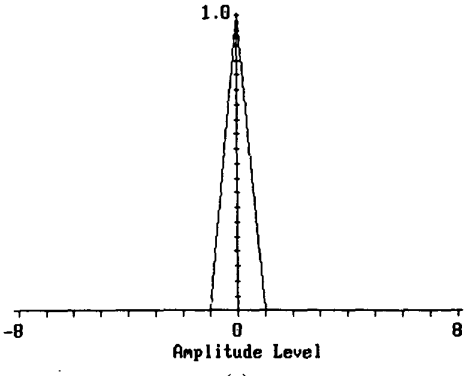
To compare the measured results with a subjective evaluation, an experiment was constructed where the signal from the DAC was amplified to audible levels and monitored through headphones. With all three filters the effect of linearization was audible, with the filtered signal sounding purer and smoother.



(a)



(b)

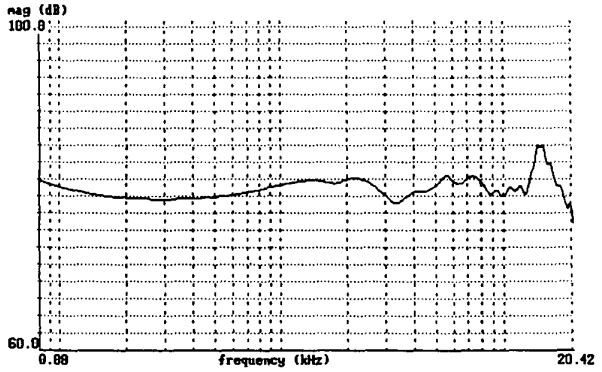


(c)

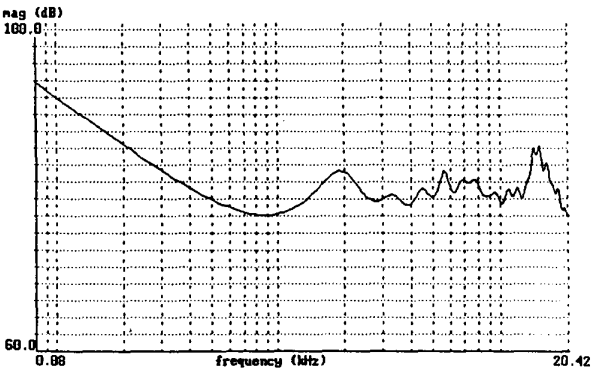
Fig. 8. Responses of flat noise response all-pass filter. (a) Harmonic distortion. (b) Pole response. (c) Noise PDF. *Note*: The noise with this filter had peak magnitude below the LSB. Hence with the 16-bit simulation it was truncated to zero.

3 DISCUSSION ON THE SUBJECTIVE IMPORTANCE OF QUANTIZATION NOISE

In Section 2 we saw how IIR filters modified the spectrum and power of the quantization noise which, at the expense of the signal-to-noise ratio (SNR), improved the system linearity. Using this knowledge it may be possible to design a system with a prescribed amplitude response while positioning the poles in locations such that the greatest noise gain occurs at frequencies insensitive to the ear. Recently much attention has been given to the subject of optimal noise shaping (see Lipshitz et al. [5]) whereby the quantization noise is shifted into regions of the ears' lowest sensitivity. Under certain conditions and with appropriate filter design it may be possible to include noise shaping in the IIR filter without a cost penalty. For example, consider the equalization transfer function for a high-quality loudspeaker system. Here the predominant demands on the equalizer will tend to be at the extremes of the drive units' frequency ranges, as one would expect the drive units to behave well within their designed bands of operation. Thus in such cases the natural pole locations are likely to be situated in the low mid bass and upper treble regions of the audio band. Fig. 9 shows the response of a loudspeaker equalizer and the corresponding pole transfer function. Comparing the pole transfer function with the Fletcher–Munsen contours shown in Fig. 10 (taken from [6]), we see that the noise-shaping function approaches the characteristic



(a)



(b)

Fig. 9. Frequency response. (a) Loudspeaker equalizer. (b) Poles of loudspeaker equalizer.

shape of minimal audible threshold contours, with the greatest boosts in noise energy occurring below 500 Hz and above 15 kHz. While this is not a general result, the authors have found from experience with many loudspeaker equalizers that the rule of thumb is typical. In other digital filtering applications there is no guarantee that the filter poles will be so conveniently placed. However, it may be possible that with appropriate pole–zero matching, a prescribed amplitude response can be met with semioptimal pole positioning. This may lead to some excess phase distortion, which may not be desirable in some circumstances. The subject of filter design for optimal pole placement, with respect to noise shaping, is an area where further investigation is needed.

4 CONCLUSION

In this paper we described the conditions necessary for the requantization of digital data and the detrimental effect that this process can have on system linearity. At the expense of increased noise, additive dither completely linearizes the quantization process. A parallel between fed back quantization noise, in the DF1 IIR filter structure, and additive dither was drawn. Initially, for the purposes of analysis, the quantization error sequence was assumed to be a random process with a uniform PDF. It was hypothesized that the effect of feedback would be an infinite convolution of the noise PDFs, resulting in a Gaussian noise distribution at the output of the filter. In Section 2 software-simulated results and measured results were presented that supported the hypothesis drawn in Section 1, namely, that the quantization noise at the output of the IIR filter does approach a Gaussian PDF and is independent of the output signal. The conclusion given here relies on the condition that the filter be input a typical audio signal. There will of course be cases where the quantization noise will be correlated to the primary output signal, such as low-level square waves or, more dramatically, digital silence. The latter condition raises the interesting point of the filters' noise modulation performance. If there is digital silence at the input to the filter, the output will be either digital silence or limit cycles. Limit cycles can be detected and muted, but the digital silence is a clear example of signal-dependent noise modulation. It is the authors' belief that, given that the quantization noise is approximately Gaussian, the only occurrence of noise modulation will be when there is digital silence, and therefore it is not considered detrimental to the overall performance of the filter. In listening tests similar to the one described in Section 2, there was no obvious pumping of the noise floor when low-level slowly varying music was played through the system. This aspect of the self-dithering property of IIR filters is, however, a subject of ongoing research. In conclusion of this section, the results presented here, from both the objective and the subjective experiments, demonstrate that under normal audio conditions the application of dither to the DF1 IIR filter structure is not necessary and only serves to worsen the noise performance of these structures.

The second part of this paper dealt with the noise-shaping property of the DF1 IIR structure. With the aid of experimental and simulated results we demonstrated the effective improvement of DAC linearity caused by the increase in noise power. The precise characteristics of the noise-shaping property is controlled by the pole transfer function of the IIR filter, and it was suggested that where possible the main boosts in noise power should be concentrated in regions of low hearing sensitivity. This bodes well for loudspeaker equalization as these regions are often close to where the natural pole locations will lie, hence giving the greater noise gain in these regions. Results of both objective and listening tests indicate that it is possible to obtain a subjectively preferable performance from what is generally considered a fundamental flaw of IIR filters.

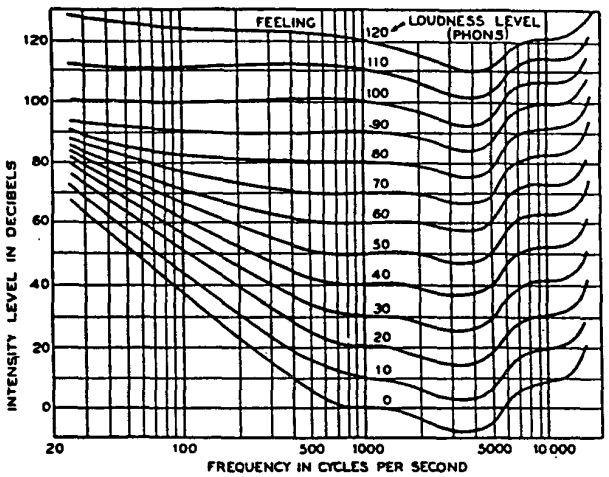


Fig. 10. Fletcher–Munson contours of equal loudness.

5 REFERENCES

[1] J. Vanderkooy and S. P. Lipshitz, "Resolution below the Least Significant Bit in Digital Systems with Dither," *J. Audio Eng. Soc.*, vol. 32, pp. 106 113 (1984 Mar.).

[2] J. Vanderkooy and S. P. Lipshitz, "Dither in Digital Audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975 (1987 Dec).

[3] S. P. Lipshitz and J. Vanderkooy, "Digital Dither," extended version of the same title presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 382 (1986 May).

[4] J. Dattorro, "The Implementation of Recursive Filters for High-Fidelity Audio," *J. Audio Eng. Soc.*, vol. 36, pp. 851–878 (1988 Nov.).

[5] S. P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally Audible Noise Shaping," presented at the 88th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 38, p. 382 (1990 May), preprint 2916.

[6] F. Langford-Smith, *Radio Engineer's Handbook*, 4th ed. (Ilife Books, 1963).

THE AUTHORS

Richard Greenfield received a B.Sc. degree in electronic engineering (telecommunications) from the University of Essex. He continued at the University of Essex, working with the audio research group under the supervision of Malcolm Hawksford, where he gained his Ph.D. His interests lie in the application of digital signal processing to audio systems with emphasis on digital loudspeaker equalization.

After gaining his doctorate Dr. Greenfield was employed by Wivenhoe Enterprises as a consultant engineer. The bias of the work was mainly with digital audio systems where he designed products such as digital loudspeaker equalizers and digital active loudspeaker systems. Other areas in which he has worked are in the design of digital to analog converters and FM tuners.

Dr. Greenfield is currently employed as a lecturer at the University of Essex where he continues to play an active role within the audio research group.

The biography for Malcolm O. J. Hawksford was published in the 1995 October issue of the *Journal*.

PAPERS

Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design*

MALCOLM OMAR HAWKSFORD, *AES Fellow*

Centre for Audio Research and Engineering, Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

Loudspeaker response data can be decomposed into minimum- and excess-phase components. Subjectively less significant elements can then be filtered and a new family of response curves computed. Applications to both cumulative-decay spectra (CDS) and energy–time curves (ETC) are discussed, where noncausal attributes are corrected. A versatile technique for producing digital crossover alignments is presented together with the implications on CDS and ETC. A novel method of masking polar response errors within the crossover transition band is also proposed.

0 INTRODUCTION

In this paper we explore alternative methods of presenting measured data of the linear imperfections arising in loudspeaker systems. There are a number of ways that the linear distortion of a loudspeaker can be described, which include impulse response, frequency and phase responses, cumulative-decay spectra (CDS), and energy–time curves (ETC). In practice all these data formats, although excluding certain performance attributes, are essentially equivalent and can be seen to be generic upon the impulse response. The impulse response is considered here as the basis response of a loudspeaker and is selected because it allows direct time-domain editing to eliminate reflections from objects other than the loudspeaker itself. There are of course numerous ways in which the impulse response can be obtained, which include direct measurement, maximum-length sequence excitation with cross correlation, and time-delay spectrometry. All these techniques have been widely researched.

A multi-drive-unit loudspeaker system requires a crossover filter [1] to spectrally divide the audio signal into appropriate low-pass, band-pass, and high-pass responses so as to match the characteristics of the drive units in terms of frequency response, power handling, and polar response. The accuracy to which these filters are designed and matched to the drive units contributes significantly to the overall performance of the loudspeaker. However, because of practical constraints there are residual imperfections in the loudspeaker transfer function, which in general also shows polar response aberrations. Three main categories of loudspeaker systems can be identified: the passive system using passive crossover networks (possibly with external analog/digital equalization), active loudspeakers using analog crossovers, and active systems employing digital filters. It is well known that for the first two categories, except for the first-order case and the restricted class of zero-phase constant-voltage filters [2], [3] the crossover networks must exhibit an overall all-pass characteristic [2], [4]–[6] and that this in itself will contribute significantly to the time dispersion revealed within the impulse response. However, in the digital implementation it is straightforward (though not essential) to produce linear-phase filter transfer functions [7] that do not exhibit an all-pass characteristic.

Audibility of phase distortion in a linear system has been widely debated and is especially relevant to loudspeaker design. Under certain conditions phase distortion is clearly audible. For example, when a nonsymmetric signal is time reversed, although the long-term amplitude spectrum is unchanged, there are obvious perceptual differences. Nevertheless, in examples of more modest phase distortion, such as those encountered in typical loudspeaker crossover networks, the human ability to detect phase distortion is reduced. For example, it is generally accepted that the phase distortion intro-

*Presented at the 100th Convention of the Audio Engineering Society, Copenhagen, Denmark, 1996 May 11–14, under the title "Minimum-Phase Signal Processing for Loudspeaker Systems," revised 1996 October 29.

duced by a fourth-order Linkwitz–Riley alignment [6] has minimal subjective consequence where results of research on determining the audible threshold all-pass phase distortion have been presented [8].

Our own informal experiments [9] have also been conducted on more complicated forms of phase distortion using digital equalization networks. For example, a two-stage digital equalizer was implemented where the first stage performed minimum-phase equalization (to achieve a flat on-axis spectrum) while a second stage performed excess-phase equalization, the latter filter having a flat amplitude response. This process permitted auditioning the contribution of excess-phase distortion while keeping the amplitude response the same. Measurements showed that the excess-phase distortion introduced significant time-domain error in the impulse response, yet auditioned in mono, the audible change both on an impulsive signal and on music was not perceptible. Hence provided no other characteristics of a loudspeaker are modified, it follows that amplitude response errors affect subjective performance to a very much greater extent than phase response errors. However, it should also be reported that in later experiments, using the Gerhard loudspeaker layout (GLL) [10], where accurately matched, digital and active low-diffration loudspeakers were located in a widely spaced configuration designed to minimize the contributions of early reflections, correcting the excess distortion was perceived to improve certain attributes of the stereophonic sound stage. Nevertheless, under more normal listening conditions using, for example, a conventional IEC floor plan, this subtlety can be lost due to greater loudspeaker–room interaction.

Consequently, because of the lower audible significance of modest phase distortion, it is expedient to process measured data to exclude some attributes of phase distortion in order to declutter the final display and thus make the diagnosis of loudspeaker imperfections more accessible together with an improved subjective—measurement correlation.

It is proposed that the displayed data should still include some attributes of phase distortion, where these are selected using minimum-phase criteria. As such, the display retains causal integrity, yet simultaneously reduces clutter due to the excess-phase distortion of lower subjective significance. However, for accurate stereophonic sound reproduction phase is believed to be of greater significance. Consequently phase distortion should still be displayed, but segregated onto a separate display.

The ETC is also used as an energy–time or envelope function, but has been criticized for exhibiting apparent noncausal attributes. A modified ETC is proposed, which corrects for the precursive response elements using minimum-phase processing. However, to match typical loudspeaker responses, the ETC is further modified by incorporating the excess-phase response of a loudspeaker. As such, an ETC results that closely matches both causality and envelope criteria.

Finally, some additional aspects of minimum-phase and linear-phase crossover filter design are discussed relating to equalization and polar response, and a flexible digital design program based on Butterworth prototype filters is presented. Also, two new classes of crossover filter are proposed, designated a stochastic alignment and a sinusoidal frequency interleaved alignment. These alignments are shown to have application in enhancing the off-axis frequency response of a loudspeaker.

The mathematics in this paper are presented for conciseness using vectors and matrices in a MATLAB[1]-like notation, where the functional operators are defined in Appendix 1. Also, a computer listing is presented in Appendix 2, which permits the evaluation of all the major features and designs presented in the paper together with an opportunity to experiment with alternative design variations.

1 MINIMUM-PHASE SYSTEM THEORY

Most loudspeaker drive units, excluding the crossover networks, exhibit a minimum-phase frequency response, meaning that the log-amplitude response and the phase response are uniquely related by the Hilbert transform [12]. It also implies that the group delay (that is, the time differential of the phase response) is minimum within the bounds of the system amplitude response and requirements of causality. However, analog crossover alignments are generally nonminimum phase, where the composite response can be decomposed into minimum-phase and excess-phase components, the latter characterized by a constant-gain amplitude response. If the amplitude response is corrected by equalization, then the normalized and equalized minimum-phase response becomes unity at all frequencies, but the excess-phase response can only be approximately linearized by performing a convolution of its corresponding impulse response with a truncated but time-reversed version, resulting in an overall time delay. This time delay is a direct consequence of the requirements of causality.

The Fourier transform $G(f)$ of a continuous-time function $g(t)$ (where t is time and f frequency) consists of the summation of both even-order and odd-order functions, which constitute the real and imaginary components of $G(f)$,

$$G(f) = \int_{-\infty}^{\infty} g(t)e^{-j2\pi ft}\,dt$$

$$= \int_{-\infty}^{\infty} g(t)\cos(2\pi ft)\,dt + j\int_{-\infty}^{\infty} g(t)\sin(2\pi ft)\,dt$$

where taking the inverse Fourier transform,

$$g(t) = g_{even}(t) + g_{odd}(t)\,.$$

In practice both even and odd functions $g_{even}(t)$ and

[1] MATLAB is a registered trade name.

$g_{odd}(t)$ exist for all time and have even and odd symmetry, respectively, about $t = 0$.

For a causal system and where the excitation is applied at $t = 0$, it must follow that $g(t) = 0$ for $t < 0$, which implies that for $t < 0$,

$$g_{even}(t) = -g_{odd}(t) .$$

The problem is therefore to calculate the functions $g_{even}(t)$ and $g_{odd}(t)$ such that the amplitude response matches that of the amplitude response of $G(f)$. There is a unique solution to this problem, where the relationship yields a complex Fourier transform, which defines the minimum-phase response that directly implies a causal relationship. In the present work, which involves only discrete-time signal vectors, the following procedure for calculating the minimum-phase response has been adopted using the (zero-phase) magnitude response as input.

Computation of the minimum-phase discrete-time sequence[2]: Let the input magnitude spectrum be |c|, where we assume here a sampled-data system such that the spectrum is discrete. In practice this can be found directly either from the loudspeaker impulse response h(n) by calculating the absolute value of the corresponding Fourier transform, or from the Fourier transform c itself, that is,

$$c = \text{fft}(h(n))$$

where fft(.) is the fast Fourier transform operator of a discrete normalized spectrum and n is a vector 1:N, with N being a power of 2.

Assume that the causal minimum-phase impulse response derived from the magnitude spectrum of c is $h_{min}(1{:}N/2)$, and that it is decomposed into even and odd symmetric functions as follows.

1) First construct an even sequence $h_e(n)$ from $h_{min}(1{:}N/2)$,

$$h_e(N/2+1{:}N) = 0.5\, h_{min}(1{:}N/2)$$

and

$$h_e(N/2{:}-1{:}\,1) = 0.5\, h_{min}(1{:}N/2) .$$

2) Then derive an odd-symmetric sequence $h_o(n)$ from $h_{min}(1{:}N/2)$,

$$h_o(N/2+1{:}N) = 0.5\, h_{min}(1{:}N/2)$$

and

$$h_o(N/2{:}-1{:}\,1) = -0.5\, h_{min}(1{:}N/2) .$$

Here the functions $h_e(n)$, $h_o(n)$, and $h_e(n) + h_o(n)$ have the form shown in Fig. 1.

[2] *Editor's Note*: Since most of the equations in this computation use MATLAB-like notation, we deviate here from *Journal* style by not italicizing variables for the sake of consistency with the program in Appendix 2.

The corresponding spectra of $h_e(n)$ and $h_o(n)$ follow:

$$d_e = \text{fft}(h_e(n))$$

$$d_0 = \text{fft}(h_o(n)) .$$

The minimum-phase function $h_{min}(1{:}N/2)$ is zero for negative time (located here at the center of the display). From the construction described in 1) and 2) it has a spectrum c_{min},

$$c_{min} = d_e + d_0 = \text{real}(d_e) + j\,\text{imag}(d_0)$$

that is,

$$c_{min} = \text{real}(\text{hilbert}(d_e)) + j\,\text{imag}(\text{hilbert}(d_e))$$

$$= \text{hilbert}(d_e) .$$

Here hilbert is a complex operator, which incorporates the odd and even symmetry of the spectrum, where the real part equates to d_e and the imaginary part to d_0. However, the spectra d_e and d_0 must be chosen so that $|c_{min}| = |c|$, which is effectively a statement of the minimum-phase computation task. This can be achieved using a logarithmic substitution, where if $d_e = \log(|c|)$,

$$c_{min} = \exp(\text{hilbert}(\log(|c|)))$$

$$= \exp(\log(|c|) + j\,\text{imag}(\text{hilbert}(\log(|c|))))$$

$$= |c|\,\exp(j\,\text{imag}(\text{hilbert}(\log(|c|)))) .$$

Consequently the amplitude spectrum |c| is preserved and the phase response is related to the logarithm of the amplitude response.

When processing discrete vectors of length N and

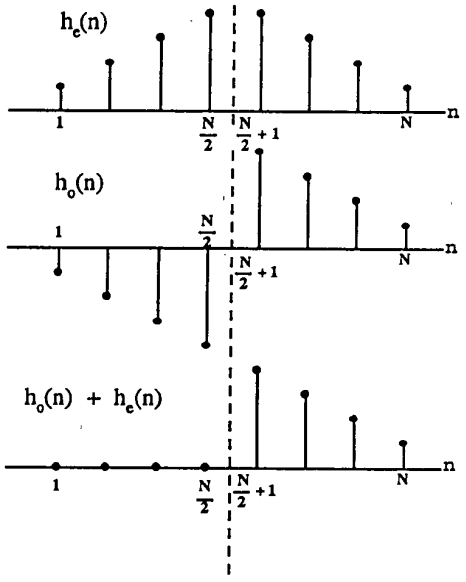


Fig. 1. Causal impulse response decomposed into even and odd symmetric impulse responses.

power 2, the MATLAB operator[3] hilbert(.) can be used to compute the discrete minimum-phase impulse response $h_{min}(n)$ from a magnitude spectrum $|c|$,

$$h_{min}(n) = \text{real(ifft(exp(conj(hilbert(log(abs(c))))))) .}$$

The even and odd functions $h_e(n)$, $h_o(n)$ then follow directly as single-line MATLAB commands,

$$h_e(n) = \text{real(ifft(real(exp(conj(hilbert(log(abs(c))))))))}$$

$$h_o(n) = \text{imag(ifft(imag(exp(conj(hilbert(log(abs(c))))))))}$$

which give the same sequences as those stated in 1) and 2). The excess-phase response $\varphi_{exc}(n)$ can be calculated from the arguments of the complex spectra c and c_{min} as

$$\varphi_{exc}(n) = \text{angle}(c) - \text{angle}(c_{min})$$

where the excess-phase impulse response $h_{exc}(n)$ is

$$h_{exc}(n) = \text{real(ifft(exp}(\varphi_{exc}(n))))$$

Fig. 2 shows an example series of waveforms illustrating a truncated impulse response of a loudspeaker, the corresponding magnitude spectrum, the odd and even time functions required for the response to be causal, the derived minimum-phase impulse response $h_{min}(n)$, and the excess-phase impulse response $h_{exc}(n)$.

Corollary 1: Observations on Minimum-Phase Equalization with FIR Filters

An observation of relevance to loudspeaker equalization concerns the relationship between an inverse finite-impulse response (FIR) equalization filter and the derived minimum-phase impulse response that results after equalization.

Consider a nonminimum-phase measurement of a loudspeaker expressed as an impulse response $h_l(m)$. The response is time edited to a shorter sequence $h_{ledit}(n)$ to eliminate first and subsequent reflections and is zero padded to give power-of-two samples n to enable fast Fourier transform processing. The magnitude response $c_m(n)$ is calculated as

$$c_m(n) = \text{abs(fft}(h_{ledit}(n)))$$

and an inverse frequency response $c_i(n)$ formed using element-by-element division ./ as

$$c_i(n) = \text{ones}(n)./c_m(n) .$$

A minimum-phase equalizer impulse response $h_{meq}(n)$ can be derived as

$$h_{meq}(n) = \text{real(ifft(exp(conj(hilbert(log(abs}(c_i(n))))))))$$

where, using one-dimensional convolution, the equalized sequence $h_{leq}(l)$ is

$$h_{leq}(l) = \text{conv}(h_{ledit}(n), h_{meq}(n)) .$$

Finally, a minimum-phase version $h_{mleq}(l)$ is calculated,

$$h_{mleq}(l) = \text{real(ifft(exp(conj(hilbert(log(abs(fft}(h_{leq}(l))))))))) .$$

Fig. 3(a) shows a truncated loudspeaker impulse response (40-sample duration) and the minimum-phase impulse response $h_{mleq}(n)$ derived using a truncated 59-tap FIR filter, where the equalized response is magnified by a factor of 70 to emphasize the detail in the tail of this response. The highlighted result is that equalization has achieved a total suppression of the minimum-phase response $h_{mleq}(l)$ over 59 samples, equal to the inverse FIR equalizer length. Consequently, the number of taps of an FIR equalizer should ideally exceed the number of samples over which the minimum-phase derived sequence of the time-edited impulse response is significant. It should be noted that although the minimum-phase equalizer response is completely suppressed for 59 samples, there remains a residual tail for samples >59 as the equalizer empties from the excitation of the finite (truncated) loudspeaker impulse response. This is a consequence of the inverse fast Fourier transform filter design. An optimization procedure [7] could produce a different error distribution, which can take greater account of the tail, although there would then be some irregularity within the main response.

Corollary 2: Observations on Minimum-Phase Equalization with IIR Filters

A similar procedure can be followed using an infinite-impulse response (IIR) filter design. In this case a least-mean-square fit IIR polynomial is generated to match the inverse spectrum of the truncated loudspeaker impulse response. Fig. 3(b) shows the truncated (40-sample) loudspeaker impulse response together with the overall minimum-phase equalized response derived using an IIR filter with 59 coefficients. Again, the filter is capable of almost total suppression of the impulse over the first 40 samples, with near complete suppression up to 59 samples. A residual error then occurs for samples >59. However, in this instance the tail in the response is much reduced, where a scale factor of 70 000 has been used to expand the error.

These two examples demonstrate the behavior of minimum-phase digital equalization using both an asymmetric FIR and an IIR filter when the result of equalization is observed devoid of excess-phase distortion. This format is useful as it enables the error (observed in the minimum-phase equalized response) resulting from finite-length equalizers to be placed at a given period in time after the excitation. In this sense the error can be seen as a short echo that does not contaminate the first few milliseconds of the overall impulse response. This detail is generally hidden when uncorrected excess-phase distortion is included.

2 CUMULATIVE-DECAY SPECTRUM

The cumulation delay spectrum (CDS) is a linear transform operation performed upon the impulse response of a loudspeaker that is designed to display en-

ergy storage as a function of time and frequency. As such it is possible to identify which frequency regions of the audio spectrum are exhibiting time dispersion through resonant modes, for example.

The basic CDS can be generated using a rectangular window function that is made to slide along the impulse response h(n) in discrete steps (Fig. 4). At each location of the window function the magnitude of the frequency response is calculated. When the transform is plotted for each window location, a three-dimensional matrix is formed, which defines the CDS.

The Hankel matrix operator can form the two-

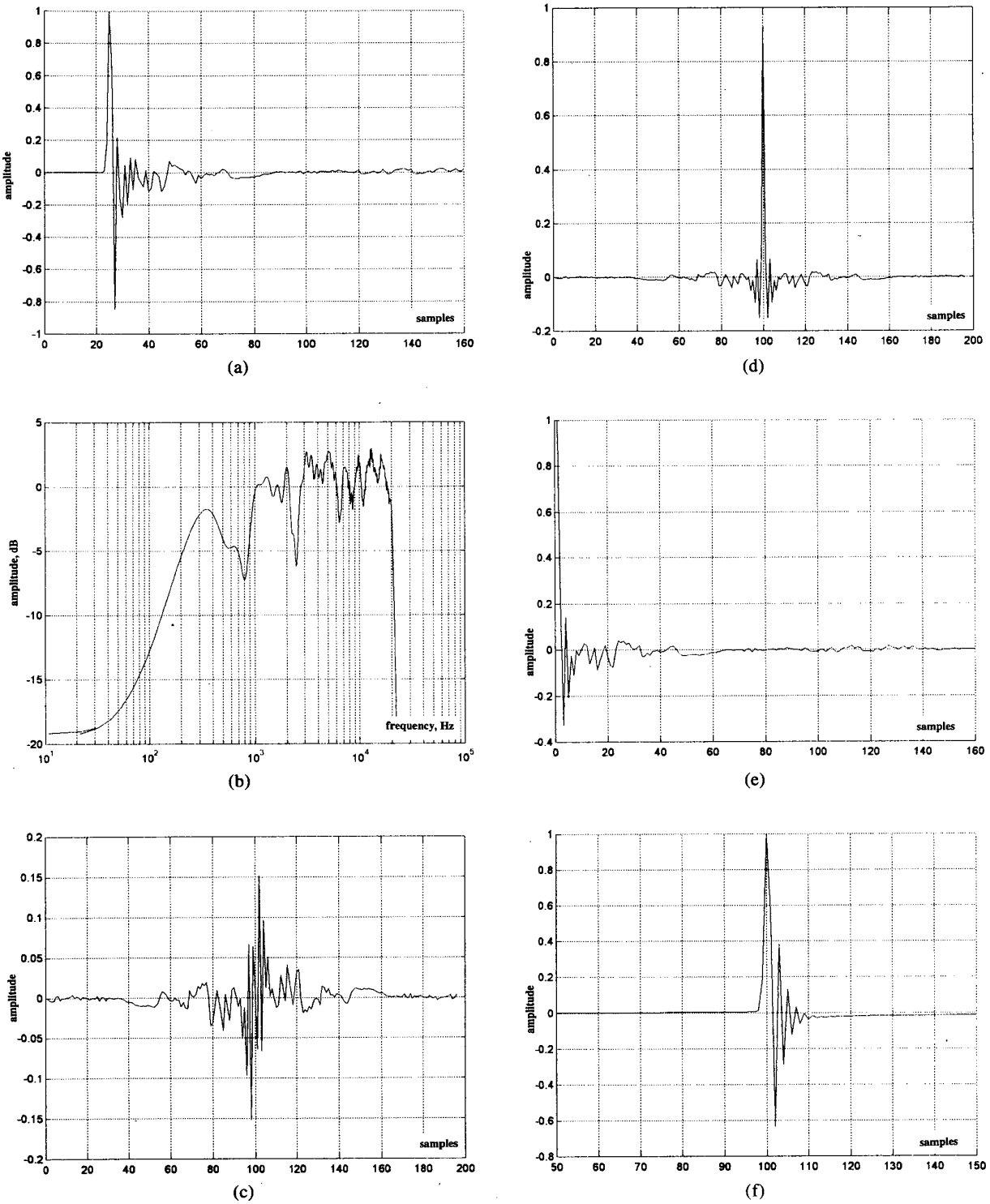


Fig. 2. (a) Truncated loudspeaker impulse response. (b) Magnitude spectrum of loudspeaker impulse response in (a). (c) Odd-symmetry impulse response supporting causality. (d) Even-symmetry impulse response supporting causality. (e) Minimum-phase impulse response derived from (a). (f) Excess-phase impulse response from response in (a).

dimensional rectangular window function that is required for the CDS. Here each row is replicated and windowed but shifted to the left, whereas the right-hand elements are progressively set to zero. Shifting the elements to the left is of no consequence as only the magnitude of the Fourier transform is taken when calculating the CDS. For illustration, a 5 by 5 Hankel matrix of a sequence h(1:5) has the form

$$\text{hankel(h(1:5))} = \begin{bmatrix} h(1) & h(2) & h(3) & h(4) & h(5) \\ h(2) & h(3) & h(4) & h(5) & 0 \\ h(3) & h(4) & h(5) & 0 & 0 \\ h(4) & h(5) & 0 & 0 & 0 \\ h(5) & 0 & 0 & 0 & 0 \end{bmatrix}$$

where the rows of the Hankel matrix reveal the windowed and left-shifted impulse response. The CDS can then be plotted directly using the (MATLAB) "mesh" command, which takes element values of a matrix and plots them vertically as a three-dimensional graph. Thus,

$$\text{CDS(n)} = \text{mesh(abs)fft(hankel(h(n)))))} .$$

In Section 1 a loudspeaker impulse was discussed, where Fig. 2 showed the measured response h(n) and the derived minimum-phase response $e_{min}(n)$. Following the matrix calculation of the CDS, we define a new display $CDS_{min}(n)$ calculated from the minimum-phase impulse response $e_{min}(n)$, where

$$CDS_{min}(n) = \text{mesh(abs(fft(hankel}(e_{min}(n)))))) .$$

Fig. 5(a) and (b) shows the corresponding CDS for both the direct impulse response and the minimum-phase impulse response, where the latter reveals a significant eduction in clutter due to the suppression of excess phase. Fig. 5(c) shows the CDS based on the excess-phase impulse response $e_{exc}(n)$. It is evident that for this loudspeaker much of the time dispersion is exhibited in the excess-phase response, whereas the minimum-phase CDS displays a more constrained dispersion characteristic. If the lower audibility of phase distortion is considered as discussed in the Introduction, then we propose that the modified minimum-phase CDS will give a closer subjective correlation to loudspeaker performance. However, we advocate that the excess-phase CDS should also be presented, but that the separation into excess- and minimum-phase displays is more useful to the loudspeaker designer.

To reduce some of the finer detail in the display, a two-dimensional convolution window or mask can be applied to the CDS matrix. Following image processing

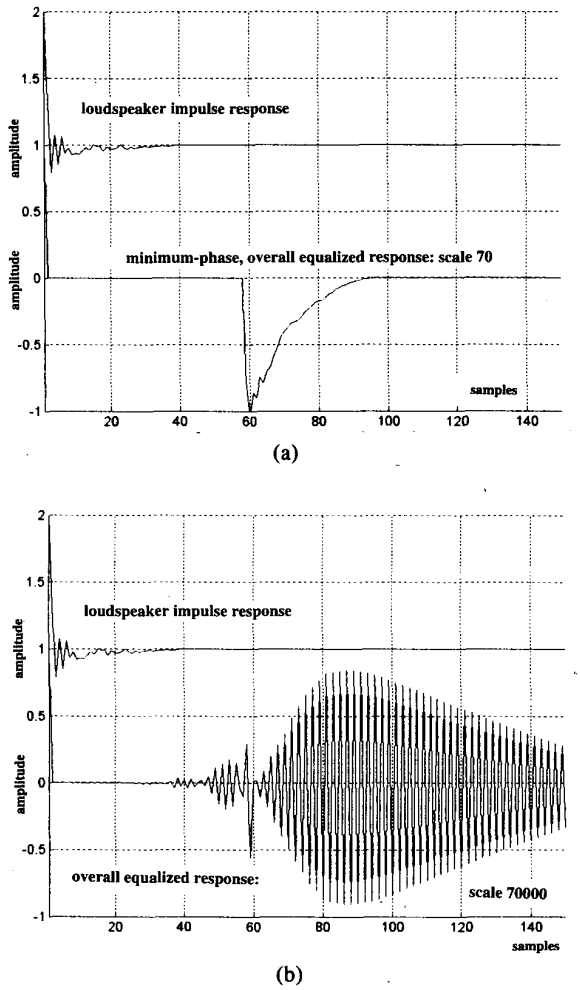


Fig. 3. (a) 59-tap FIR equalization; (a) minimum-phase responses. (b) 59-tap IIR equalization; minimum-phase responses.

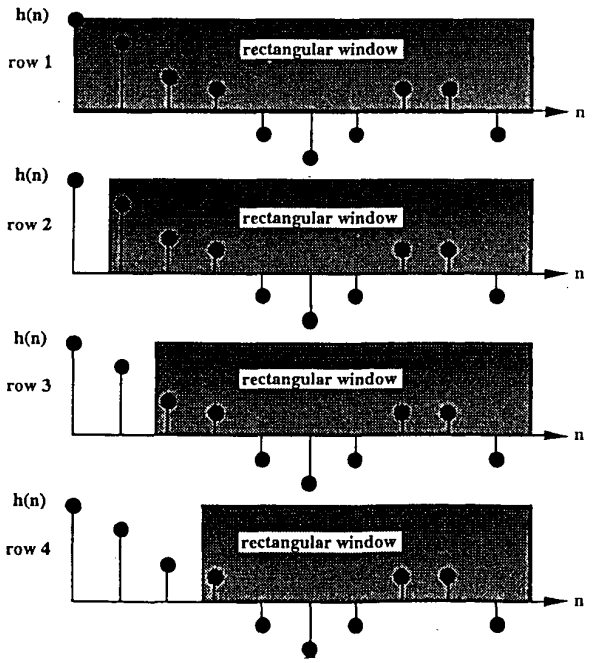


Fig. 4. Construction of CDS using progressive displacement of a rectangular window over the loudspeaker impulse response.

applied to the CDS matrix. Following image processing practice, a Gaussian window shows a combination of good smoothing plus low image artifacts. A Gaussian window $CDS(n)_{win}$ of order n can be computed from multiple two-dimensional convolutions starting from a simple 2 by 2 matrix, where the coefficients are normalized so that their summation is unity,

$$CDS(2)_{win} = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}.$$

This process is similar to one-dimensional convolution where the technique was applied in the generation of a Gaussian crossover filter [12]. A third-order window $CDS(3)_{win}$ then follows by two-dimensional convolution using operator notation conv2,

$$CDS(3)_{win} = |\text{conv2}(CDS(2)_{win}, CDS(2)_{win})|_{norm} .$$

This yields a 3 by 3 mask, where

$$CDS(3)_{win} = \begin{bmatrix} 0.0625 & 0.125 & 0.0625 \\ 0.125 & 0.25 & 0.125 \\ 0.625 & 0.125 & 0.0625 \end{bmatrix}.$$

Note how the mask is square and the order of the mask equals the number of rows (or columns) in the matrix. The convolution can be continued by iteration to generate a family of matrices to order m, where

$$CDS(m)_{win} = |\text{conv2}(CDS(m-1)_{win}, CDS(m-1)_{win}|_{norm} .$$

Fig. 6 shows a seventh-order filter mask. A filtered spectrum CDS_{minw} is then formed by a two-dimensional convolution between actual CDS and the filter mask,

$$CDS_{minw} = \text{conv2}(CDS_{min}(n)\, CDS_{win}(m)) .$$

As examples of this smoothing process, Fig. 7 shows a third-order and a seventh-order mask applied to the CDS in Fig. 5(a).

3 ENERGY–TIME CURVE

The energy–time curve (ETC) can be used to display time dispersion in a loudspeaker as an envelope function of a complex time sequence. Examining the impulse response of a loudspeaker, such as h(n) illustrated in Fig. 2(a), shows a function that varies from negative to positive values with multiple zero crossings. This typical behavior is common to all audio signals with zero mean. However, from a subjective stance when listening, for example, to a sine wave, we perceive a continuous

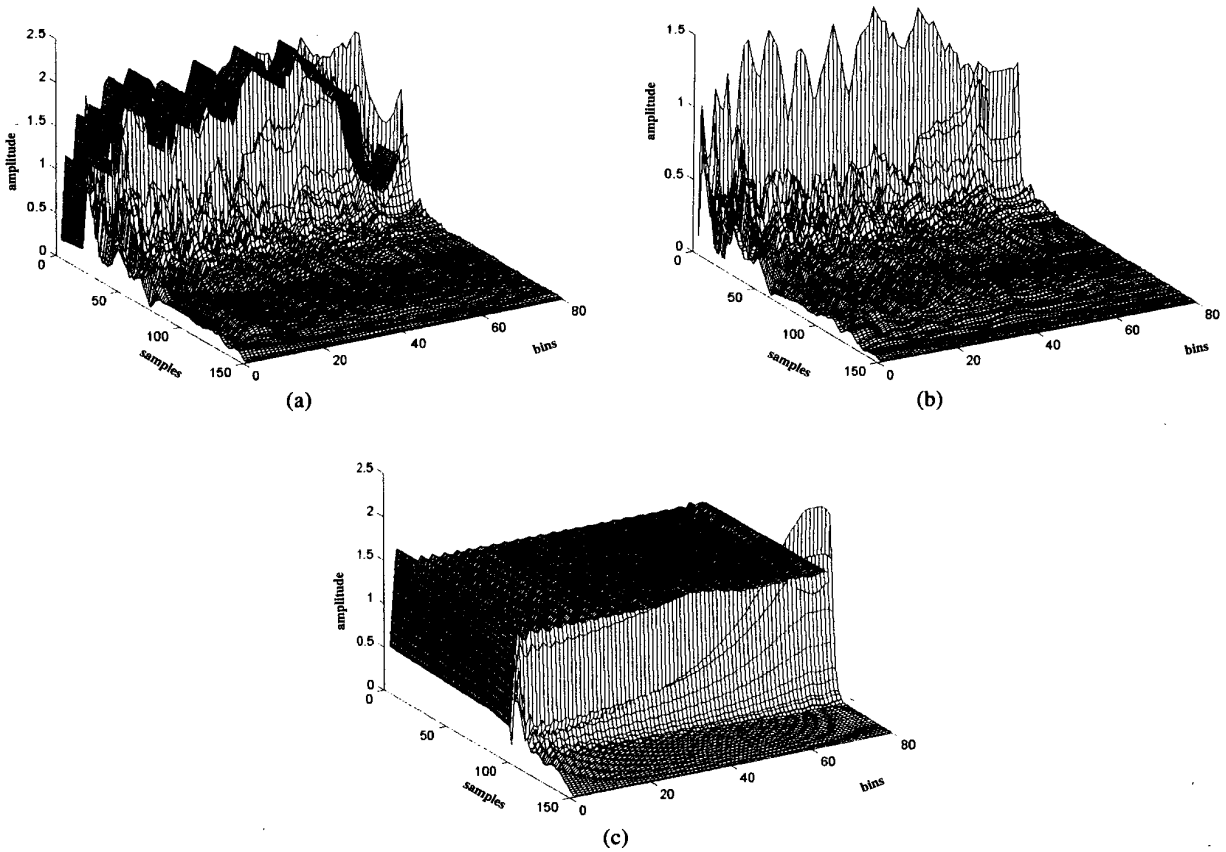


Fig. 5. CDS. (a) Derived from loudspeaker impulse response. (b) Derived from loudspeaker minimum-phase impulse response. (c) Derived from excess-phase loudspeaker impulse response.

acoustic object that does not exhibit periodic moments of silence. A functional presentation that describes intensity variations but ignores subjectively obscure signal zeros is therefore desirable. In effect the zeros in the time-domain waveform do not necessarily correspond to silences. We infer aspects of pitch and loudness as continuous attributes in which derivative information as well as amplitude information contribute to the overall percept. Simplistically, the signal peaks appear to carry similar weight to signal derivatives at the signal zero crossings.

3.1 Determination of ETC

It is therefore expedient to select an algorithm in which both amplitude and derivative information are included explicitly, from which an envelope function can be derived. The ETC uses the Hilbert transform to calculate a complex time function, the amplitude spectrum of which is essentially identical to that of the amplitude spectrum of h(n). The Hilbert transform is an operator that yields a complex time function having real and imaginary parts. This analytic function must therefore be described in two-dimensional space and consequently has real and imaginary time sequences. If we derive an envelope function that corresponds to the magnitude of this complex time function, this is the ETC.

The Hilbert transform ht(n) of the impulse sequence h(n) can be represented by the operation

$$ht(n) = \text{hilbert}\{h(n)\} = ht_r(n) + j\, ht_i(n)$$

where $ht_r(n)$ and $ht_i(n)$ are the real and imaginary time sequences, respectively [with $ht_r(t)$ corresponding to h(n) in this notation]. It is informative to examine the Fourier transform of a typical sequence h(n), where Fig. 8(a)–(c) displays $ht_r(n)$, $ht_i(n)$, and their shared magnitude spectrum. Fig. 8(d) shows the magnitude spectrum of the complex time function ht(n). It should be noted that all the in-band spectra are identical while the reflected spectral components of ht(n) are zero, confirming the complex form of ht(n). The envelope et(n) of ht(n) is calculated as

$$et(n) = [ht_r(n)^2 + ht_i(n)^2]^{0.5} = \text{abs}\{ht(n)\}\,.$$

Fig. 8(e) and (f) shows the resulting ETC, et(*n*), and its corresponding magnitude spectrum.

3.2 Minimum-Phase Corrected ETC

However, the nature of the Hilbert transform is such that it exhibits a precursive response that appears to contravene causality. In practice this does not happen, as the function should be considered in conjunction with an overall delay. Nevertheless the presence of a precursor to the main impulse event is counterintuitive where

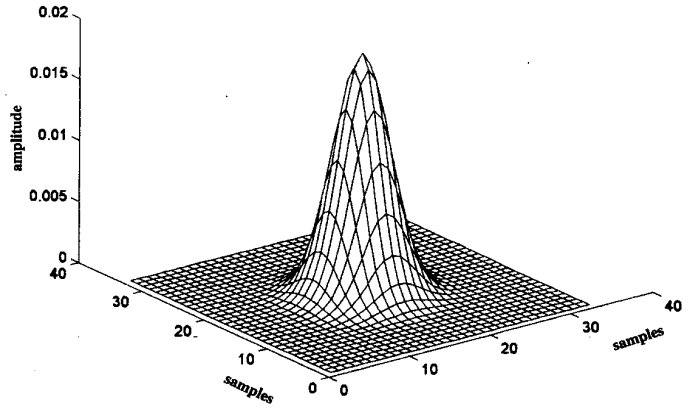


Fig. 6. Seventh-order two-dimensional filter mask.

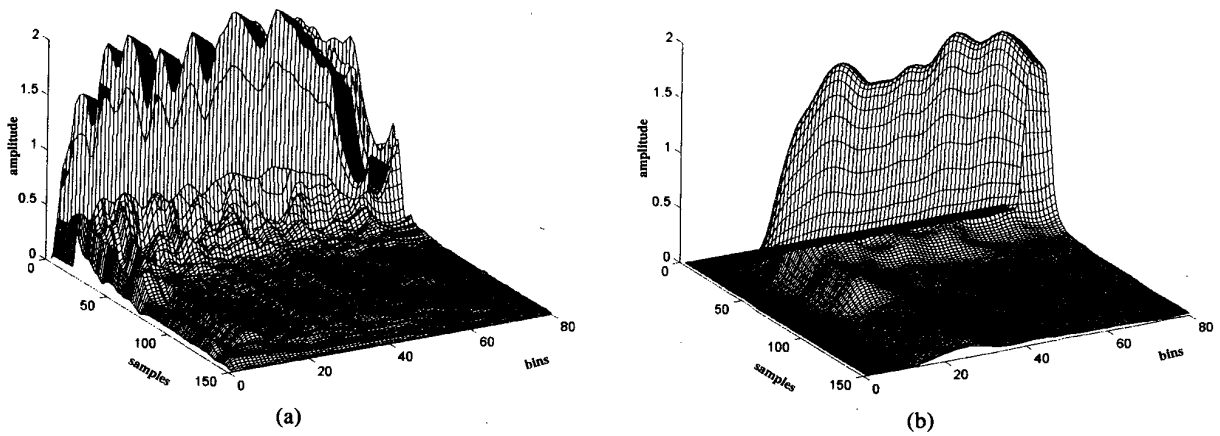


Fig. 7. CDS as Fig. 5(a). (a) Filtered with third-order mask. (b) Filtered with seventh-order mask.

the peaks in the ETC and the loudspeaker impulse response should also remain time aligned. To rectify this anomaly, the envelope function et(n) can be further processed to form a minimum-phase envelope, where the envelope spectrum is preserved but the precursor region is removed in line with minimum-phase theory. The calculation of $et_{min}(n)$, the minimum-phase ETC, is similar to that described in Section 1, where the salient operations are as follows.

1) Calculation of amplitude spectrum fet(n) of the

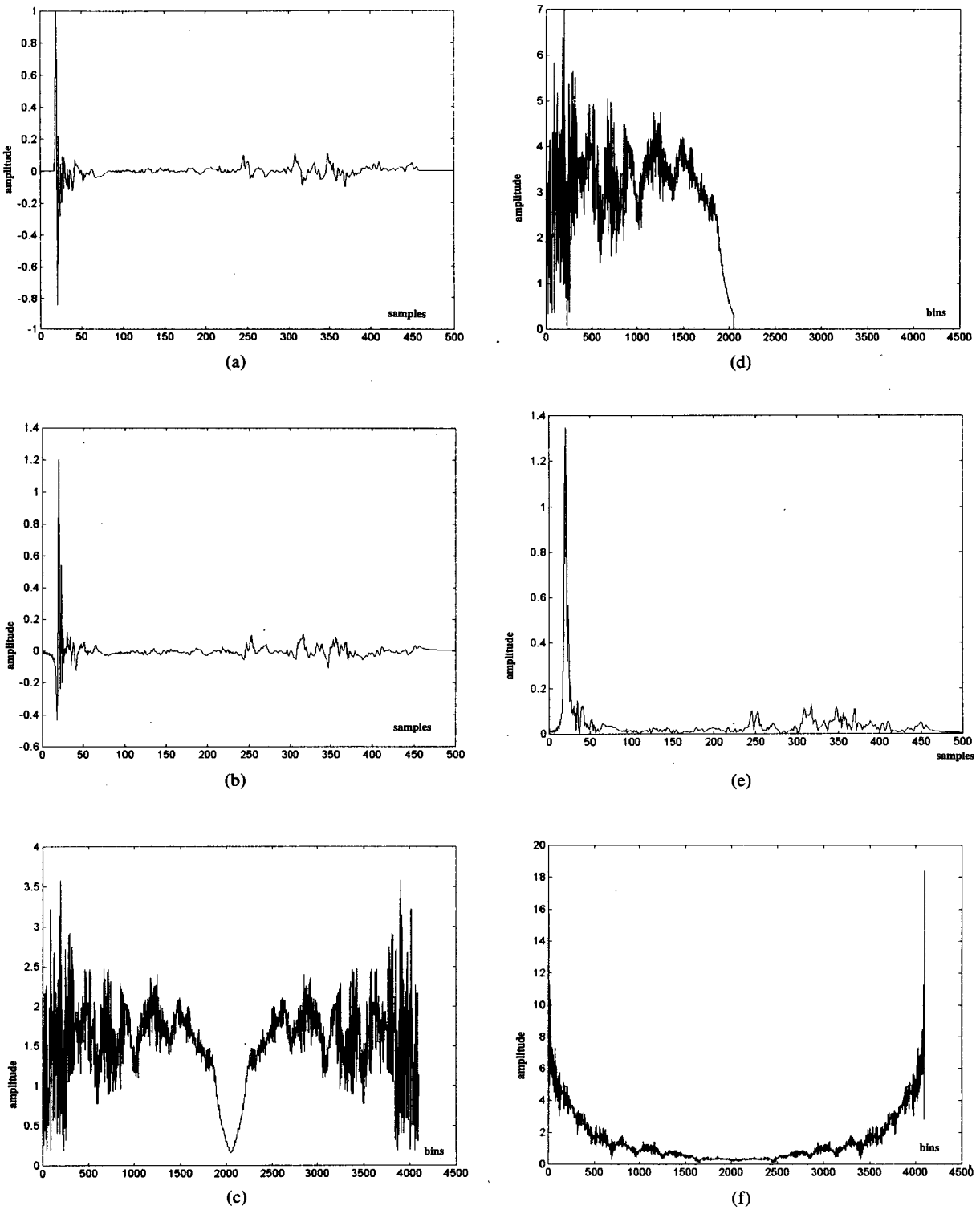


Fig. 8. (a) Real part $ht_r(n)$ of Hilbert transform, impulse response. (b) Imaginary part $ht_i(n)$ of Hilbert transform. (c) Magnitude spectrum of real part of Hilbert transform. (d) Magnitude spectrum of complex-time Hilbert transform. (e) ETC formed from magnitude of Hilbert transform. (f) Magnitude spectrum of ETC shown in (e).

ETC sequence et(n),

$$fet(n) = abs(fft(et)) .$$

2) The minimum-phase impulse response derived from the original ETC envelope then follows,

$$et_{min}(n) = real(ifft(exp(conj(hilbert(log(fet(n))))))$$

whereby the corrected ETC is $abs(et_{min}(n))$. Fig. 9(a) shows a minimum-phase loudspeaker impulse response, the corresponding nonminimum-phase ETC, and the minimum-phase ETC.

3.3 Excess-Phase Compensated Minimum-Phase ETC

The minimum-phase corrected ETC conforms to the derived minimum-phase loudspeaker impulse. However, when compared against the measured loudspeaker response that includes excess phase, then anomalies in the envelope of the ETC are evident. It is proposed to correct the minimum-phase ETC by convolution with the excess-phase impulse response $h_{exc}(n)$, as defined in Section 1, where the compensated response ETC et_{exc} is calculated from the complex ETC $et_{min}(n)$,

$$etc_{exc} = abs(conv(et_{min}(n),h_{exc}(n))) .$$

Fig. 9(b) illustrates the improvement in envelope matching, where the loudspeaker impulse response including excess phase, the nonminimum-phase ETC, and the excess-phase compensated ETC are shown.

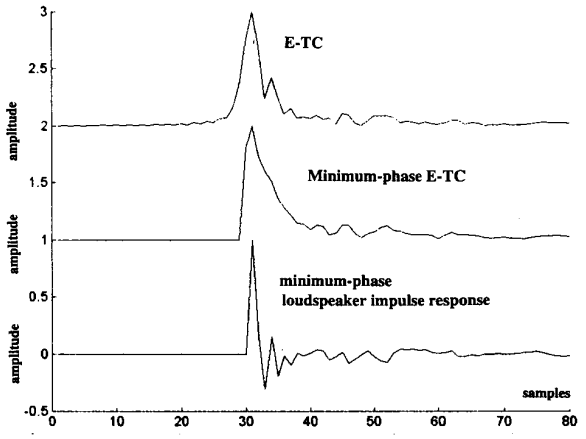


Fig. 9. (a) ETC and minimum-phase ETC of loudspeaker impulse response. (b) Excess-phase corrected ETC.

4 CUMULATIVE-DECAY ROOT SPECTRUM

An alternative display mode related to the roots of the polynomial describing the loudspeaker can be defined and is called cumulative-decay root spectrum (CDRS). This technique is well matched to the minimum-phase impulse response as this response is invertible, the roots being stable. Consequently the minimum-phase polynomial can be expressed as a partial-fraction expansion, where the roots of the denominator form stable (convergent) terms, that is,

$$h_{min}(z) = a_0 + \sum_{r=1}^{n} \frac{N_r(z)}{(z - \alpha_r + j\beta_r)(z + \alpha_r + j\beta_r)}$$

Here $N_r(z)$ represents the numerator associated with each root derived in a partial-fraction expansion. The roots can be computed on a least-mean-squares basis so that the more terms, the better the approximation. However, limiting the number of roots is a straightforward method of filtering the display. In effect, the partial expansion approximation generates a parallel array of second-order filters, which can be defined in terms of their respective damping Q_r and undamped natural resonant frequency ω_r.

The impulse response corresponding to a specific second-order section can be calculated directly as a discrete-time sequence. For the rth term in the partial-fraction expansion the impulse response $h_{min}(r, n)$ is computed and the corresponding minimum-phase ETC, that is, $et_{min}(r, n)$, is derived as described in Section 3. This process is repeated for each root, and the corresponding undamped natural resonances are computed. The ETCs are then assembled onto a three-dimensional display, each being directed along the time axis and either located at the appropriate resonant frequency on the frequency axis or simply presented in rank order of resonant frequency on an integer scale. Fig. 10 shows two CDRS that correspond to the loudspeaker impulse response displayed in Fig. 8. The roots were determined from the 59-tap FIR equalizer filter polynomial discussed in Corollary 1 in Section 1. Fig. 10(a) is a linear CDRS display, whereas Fig. 10(b) plots each ETC (forming the CDRS) on a decibel scale to reveal more low-level detail.

5 LINEAR AND MINIMUM-PHASE FILTER EXAMPLES

In this section a number of examples of signal processing relevant to audio are considered. The examples

demonstrate techniques of frequency response tilting, psychoacoustic response shaping for subjective enhancement, and constant-amplitude crossover filters of arbitrary slope. For each class of filter, the impulse response is presented in both linear-phase and minimum-phase formats, where the results can be interpreted using both CDS and ETC by means of the program in Appendix 2.

5.1 Constant-dB-Gain Logarithmic Frequency Response Tilting

A filter characteristic useful for subjectively tailoring an audio signal uses a tilt function, which if plotted on a graph of gain in decibel against logarithmic frequency gives a line of constant slope. The following routine computes a frequency domain vector a_j describing the tilt function from a unit vector $m_x(n)$: Let {slope} denote filter slope in decibels per octave. Define

$$n_j = \{\text{slope}\}/(20 \log_{10}(2)) \quad \text{and} \quad f_j = (m_l * m_h)^{0.5} .$$

Here m_l and m_h are the lower and upper frequency limits of the Fourier transform and f_j is the geometric mean, whereby

$$a_j(n) = (m_x(n)/f_j)^{.nj}$$

where the operator .nj implies that all elements of the vector $m_x(n)/f_j$ are raised to the power n_j. A linear-phase time-domain sequence ta_j can then be computed using the inverse Fourier transform,

$$ta_j(n) = \text{real}(\text{ifft}(a_j(n)))$$

and a minimum-phase impulse response $ta_{jmin}(n)$ follows from the procedure in Section 1,

$$ta_{jmin}(n) = \text{real}(\text{ifft}(\exp(\text{conj}(\text{hilbert}(\log(\text{abs}(a_j))))))) .$$

Fig. 11(a) shows four examples of linear-phase time-domain responses corresponding to filter slopes of [−4, −2, 0, 2, 4] dB per octave, whereas Fig. 11(b) displays the corresponding minimum-phase time-domain responses.

5.2 Psychoacoustic Weighting Functions

A second weighting function that can usefully modify the audio frequency response was proposed by Blauert [13]. This characteristic attempts to improve subjective performance by frequency-selective attenuation in the frequency range where the ear–brain is most sensitive. The amplitude-frequency response in decibels is de-

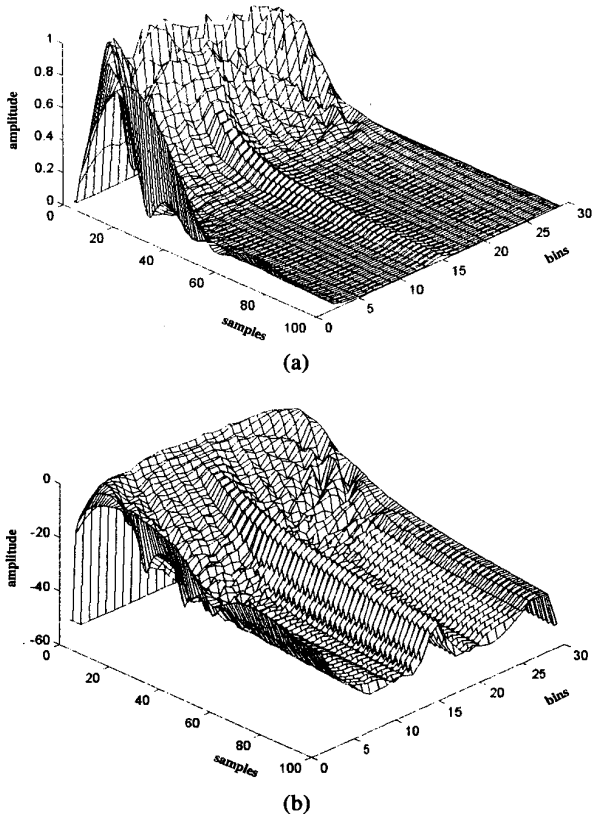


Fig. 10. CDRS. (a) Linear vertical-scale display (59-tap filter). (b) Decibel vertical-scale display (59-tap filter).

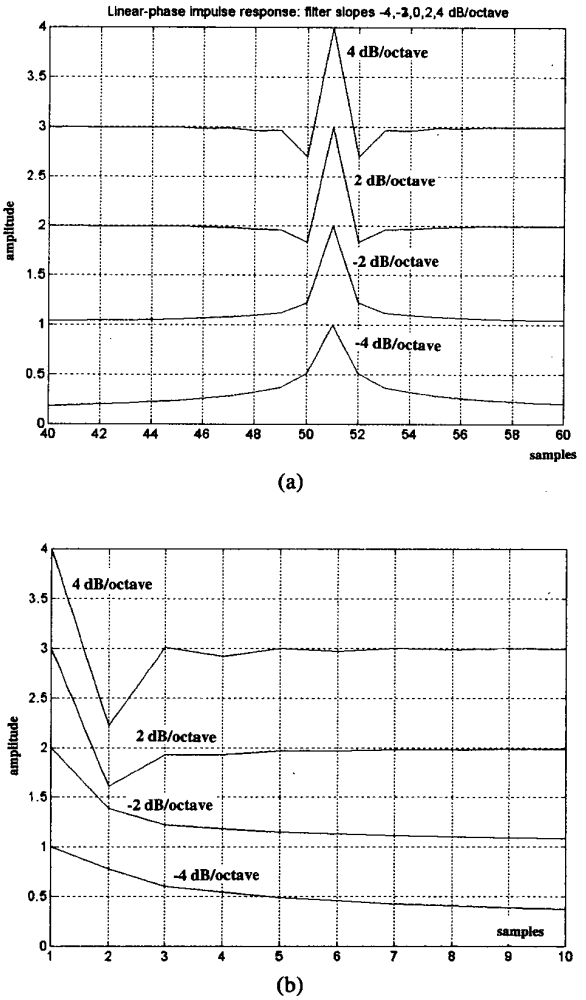


Fig. 11. Impulse responses of frequency tilting. (a) Linear phase. (b) Minimum-phase.

scribed as

$$\text{gain}_{\text{dB}} = \begin{cases} -2.5\log(f/3\text{ kHz}), & \text{for } f < 3\text{ kHz} \\ 6.4\log(f/3\text{ kHz}), & \text{for } f > 3\text{ kHz} \end{cases}$$

where f is the frequency in kilohertz and gain_{dB} is expressed in decibels. The frequency response characteristic is shown in Fig. 12(a), whereas the corresponding linear-phase and minimum-phase impulse responses are shown in Fig. 12(b).

5.3 Generalized Linear- and Minimum-Phase Crossover Filters with Arbitrary Slope

Digital filters can be used in active loudspeaker systems to implement highly accurate crossover functions, where with correctly designed drive-unit frequency response equalization, near-theoretic crossover target responses can be synthesized. The normal constraints imposed by analog processing do not apply where it is possible to separate the phase and amplitude responses with respect to both the composite response c(f) and the individual high-pass and low-pass responses hp(f) and lp(f). (Note that extending the number of bands is a natural extension of this process where the same procedures apply.)

The target transfer functions can be chosen to be either linear phase or minimum phase, yielding either an overall linear-phase composite response or an all-pass composite response. Some relevant observations can be made which are helpful in choosing suitable crossover filter targets.

1) If the linear-phase transfer functions hp(f) and lp(f) have identical phase responses, then the composite response is also linear phase, meaning that for an FIR implementation the composite filter introduces pure time delay and each filter must have the same number of coefficients.

2) If the individual transfer functions hp(f) and lp(f) are each minimum phase, then the composite response is either "constant voltage" [2], [3] or all pass. Examples are the first-order filter and the fourth-order Linkwitz–Riley (LR-4) [1], [2], [4], [6], [14].

3) Asymmetric filters [15] could include a combination of minimum-phase and nonminimum-phase hp(f) and lp(f) transfer functions.

4) An all-pass composite filter response can be phase equalized [7], implying that hp(f) and lp(f) transfer functions are each nonminimum phase when excess-phase correction is incorporated.

5) Desirable qualities for hp(f) and lp(f) are that they sum to form either a linear-phase or an all-pass function and in particular that the individual filters hp(f) and lp(f) can have the same phase responses to minimize lobing errors in the composite polar response [2], [6], [18].

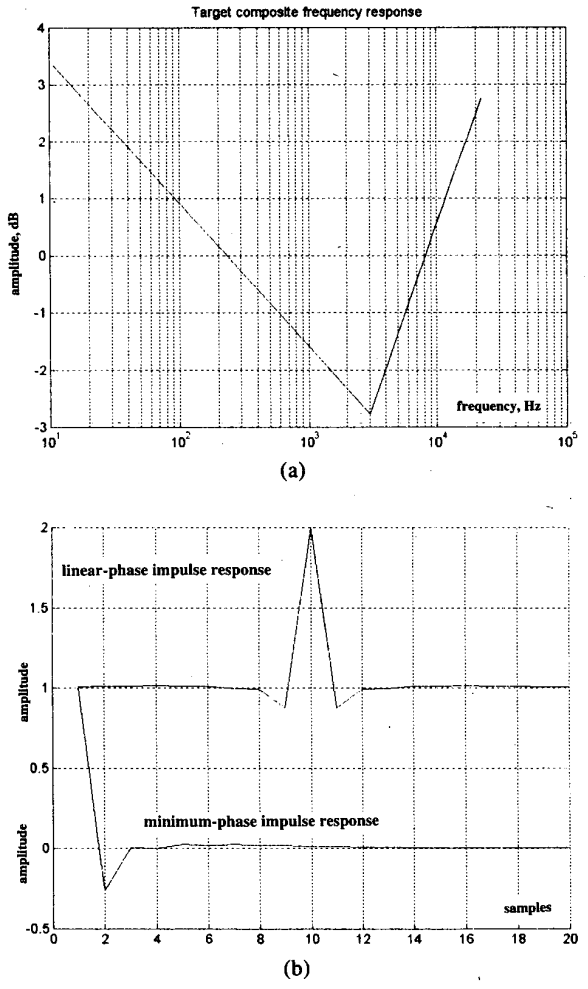


Fig. 12. (a) Psychoacoustic weighting function. (b) Linear- and minimum-phase impulse responses.

5.3.1 Linear-Phase Crossover Filters

To illustrate this process, a filter set is presented that is derived from a Butterworth amplitude response, but with modifications to guarantee a constant, composite amplitude response. The low-pass and high-pass Butterworth amplitude responses $\text{LPF(f)}_{\text{Butt}}$ and $\text{HPF(f)}_{\text{Butt}}$ are defined as

$$\text{LPF(f)}_{\text{Butt}} = \frac{1}{[1 + (f/f_{\text{xl}})^2]^{0.5k}}$$

$$\text{HPF(f)}_{\text{Butt}} = \frac{(f/f_{\text{xh}})^k}{[1 + (f/f_{\text{xh}})^2]^{0.5k}}$$

where

$$f_{\text{xl}} = \frac{f_{\text{ol}}}{(2^{2/k} - 1)^{0.5}} \quad \text{and} \quad f_{\text{xh}} = \frac{f_{\text{oh}}}{(0.5^{-2/k} - 1)^{0.5}}.$$

Here f_{ol} and f_{oh} are the 6-dB crossover frequencies. (Normally $f_{\text{ol}} = f_{\text{oh}}$, but they can be unequal in more general alignments using equalization.) The filter asymptotic slopes are

$$|\text{filter slope}| = 20k\log_{10}(2)\text{ dB/octave}.$$

Since the filters can have linear phase, then the crossover frequency occurs when the attenuation is 6.02 dB. Ideally the composite response should sum to unity. However, in practice there is an error $\text{E(f)}_{\text{Butt}}$ in the response, which is equal to

$$E(f)_{\text{Butt}} = F_{\text{target}}(f) - \{\text{LPF(f)}_{\text{Butt}} + \text{HPF(f)}_{\text{Butt}}\}.$$

Normally the target response $F_{target}(f) = 1$, although in practice any desired frequency response can be defined, such as the filters described in Sections 5.1 and 5.2.

Modified filter responses can be defined that add corrective components to $LPF(f)_{Butt}$ and $HPF(f)_{Butt}$ to make $E(f)_{Butt} = 0$. To accommodate nonidentical and overlapping crossover frequencies in the Butterworth prototype filters, a raised-cosine window function win(n) can be used to smooth the transition band, that is,

$$ones(n) = \text{unit vector of length N elements}$$

$$win_l(n) = 1 \text{ for } 1 \leq n \leq (x_m - w/2)$$

$$win_l(n) = 0.5(1 + \cos(\pi(1{:}w)/w)), \text{ for } (n_{fo} - w/2) < n \leq (n_{fo} + w/2)$$

$$win_l(n) = 0 \text{ for } (n_{fo} + w/2) < n \leq N$$

$$win_h(n) = ones(n) - win_l(n)$$

where w is the window length and n_{fo} the element corresponding to the geometric mean $(f_{ol} * f_{oh})^{0.5}$. The amplitude-corrected filters $LPF_c(f)$ and $HPF_c(f)$ can then be defined as

$$LPF_c(f) = LPF(f)_{Butt} + win_l(n).*\{F_{target}(f) - LPF(f)_{Butt} - HPF(f)_{Butt}\}$$

$$HPF_c(f) = HPF(f)_{Butt} + win_h(n).*\{F_{target}(f) - LPF(f)_{Butt} - HPF(f)_{Butt}\}$$

such that

$$LPF_c(f) + HPF_c(f) = F_{target}(f) .$$

This method enables the high-pass and low-pass crossover frequencies to be independently defined and then corrected to give a desired target response while not compromising the rate of attenuation in the stopbands. Also, because the digital filter is not constrained by inductive and capacitive elements, the slope parameter *k* does not have to be integer, thus allowing for fractional slopes. FIR filters can then be designed to match the required amplitude responses together with a linear-phase response by using impulse symmetry and equal-length filters. FIR filters generally yield a response with only finite attenuation. Thus two additional requirements can be imposed to force zero dc gain for $HPF_c(f)$ and zero ac gain at $f_s/2$ for $LPF_c(f)$:

1) For $LPF_c(f)$ make the sum of the coefficients zero when every other coefficient is inverted (forces zero gain at $f_s/2$).

2) For $HPF_c(f)$ make the sum of the coefficients zero (forces zero gain at dc).

Fig. 13 shows an overlapping crossover design where the prototype low-pass 6-dB break frequency is 4 kHz, the prototype high-pass 6-dB break frequency is 2 kHz, and the raised-cosine window is 50 samples (2048 total). Both attenuation slopes are symmetrical with 30-dB per octave asymptotes. In this example two response dips are evident because of the choice of prototype break frequencies and window. However, this illustrates the innate flexibility available in digital filter design in choosing the crossover alignment. The upper curve in the display corresponds to the response env_{des}. It is discussed in more detail in Section 5.4 and is useful for determining off-axis behavior when additional phase shift is introduced between filter responses.

5.3.2 Minimum-Phase Derived Crossover Filters

Having obtained functions that meet the amplitude response criteria, minimum-phase processing can be applied to either the Butterworth filters LPF(f), HPF(f) or the corrected filters $HPF_c(f)$, $HPF_c(f)$. However, there is no guarantee that the filter pairs will exhibit identical phase responses (compared to LR-4). So equalization of the composite response [16] is required to correct for

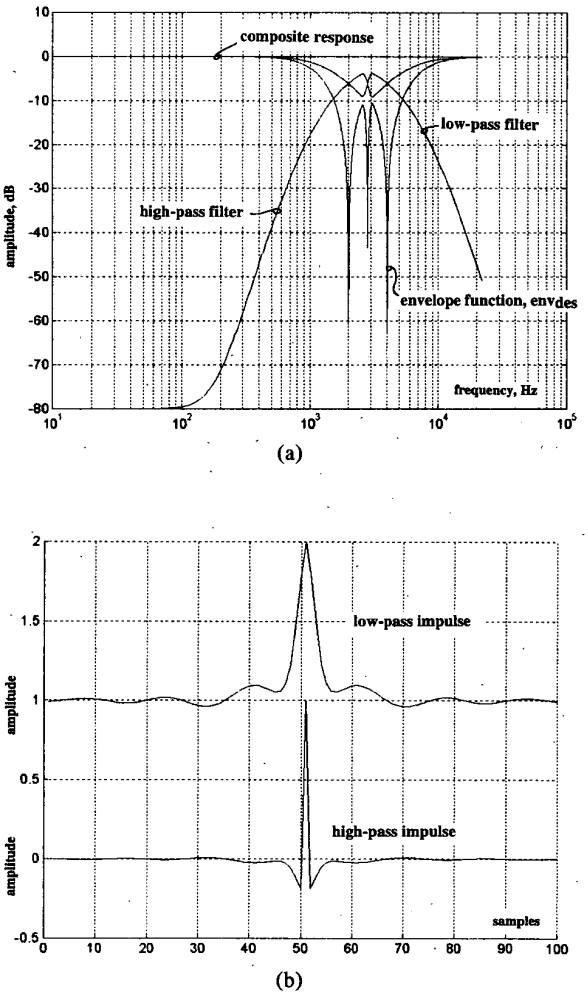


Fig. 13. (a) Overlapping linear-phase crossover design. (b) Linear-phase high-pass and low-pass impulse responses.

amplitude errors. Initially a minimum-phase computation is performed on each filter amplitude response to form the respective minimum-phase impulse responses $h_{lmin}(n)$ and $h_{hmin}(n)$,

$h_{lmin}(n)$ =

real(ifft(exp(conj(hilbert(log(abs(LPF(n)))))))))

$h_{hmin}(n)$ =

real(ifft(exp(conj(hilbert(log(abs(HPF(n)))))))) .

The composite response, either $\{h_{lmin}(n) + h_{hmin}(n)\}$ or $\{h_{lmin}(n) - h_{hmin}(n)\}$, is then computed, which in general does not have a constant magnitude response. An equalizer impulse response is then computed to correct for amplitude anomalies as follows. The amplitude response of equalizer EQ(n) (where ./ implies element-by-element division) is

$$EQ(n) = (ones(n))./abs(fft(h_{lmin}(n) \pm h_{hmin}(n)))$$

and the minimum-phase impulse response $E_{min}(n)$ of the equalizer is

$$E_{min}(n) = real(ifft(exp(conj(hilbert(log(abs(EQ(n)))))))) .$$

The modified filter responses $h'_{lmin}(n)$ and $h'_{hmin}(n)$ then follow by convolution,

$$h'_{lmin}(n) = conv(h_{lmin}(n), E_{min}(n))$$

$$h'_{hmin}(n) = conv(h_{hmin}(n), E_{min}(n)) .$$

Although there remains phase distortion in the composite response, it is all-pass within the constraints of realizability. We therefore define a generalized minimum-phase filter set which has the following properties:

1) Symmetrical filter functions about f_0 on dB versus log(frequency) scale. However, because of equalization of the composite response the low-pass and high-pass prototype filters do not have to have the same 6-dB crossover frequency.

2) Phase responses in low-pass and high-pass filters do not have to be identical.

3) All-pass phase distortion in composite response is minimized within the constraints of minimum-phase processing.

4) The technique can be readily generalized for any crossover filter amplitude responses, including asymmetric alignments [5], [17].

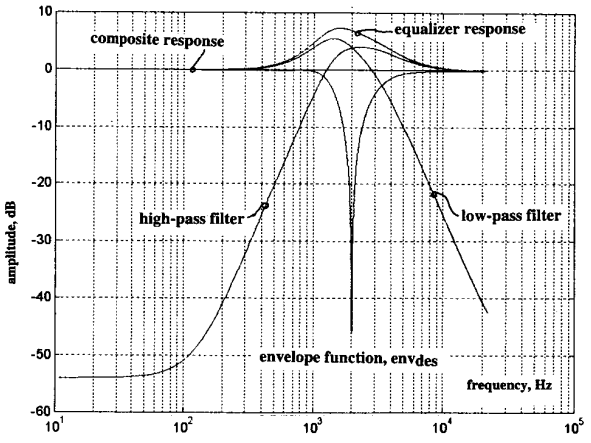
Fig. 14 gives an example minimum-phase design corresponding to f_0 = 2 kHz with 18-dB per octave slopes and in-phase filter addition. Fig. 14(a) shows the relevant frequency responses, including equalizer (see Section 5.4 for env_{add} and env_{des}), whereas Fig. 14(b) shows the corresponding low-pass and high-pass filter minimum-phase impulse responses.

5.4 Stochastic Interleave Crossover-Filter Alignment

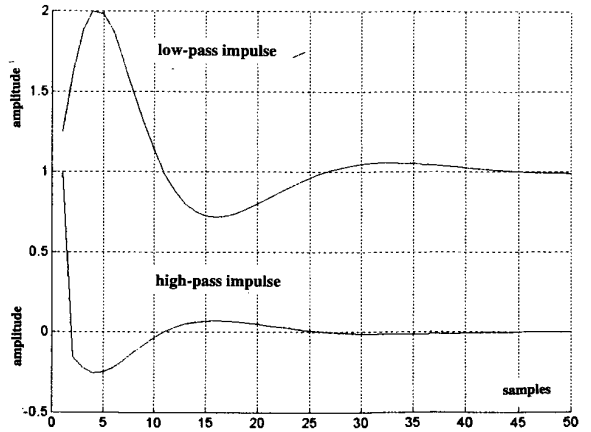
In this section the incorporation of a static random vector in the crossover filter response is explored as a means of reducing the subjective significance of polar response errors. This random sequence together with selected frequency shaping is termed the "interleave function."

A design requirement of a crossover filter is that the loudspeaker have a well-behaved off-axis frequency response. This is normally described in terms of the polar distribution, which is influenced by the frequency-dependent spatial characteristics of a drive unit, the physical spacing of the drive units, and the baffle size and profile. The frequency region in which the crossover has influence is defined in terms of the filter attenuation characteristics together with their relative phase responses. For example, linear-phase filters and the Linkwitz–Riley classes [6] of filter combine a constant composite magnitude response with identical phase responses for the high-pass and the low-pass filter. This technique produces a symmetrical lobing error.

The effect of noncoincident drive units is to introduce



(a)



(b)

Fig. 14. (a) Minimum-phase crossover design. (b) Linear-phase high-pass and low-pass impulse responses.

a time delay between the sound radiated from each drive unit. This in turn can be represented by a frequency-dependent phase difference which creates interference. Theoretically it follows that for a general crossover alignment defined by LPF(f, φ), HPF(f, φ) additive and destructive interference is bounded by the envelopes env_{add} and env_{des}, where

$$env_{add} = |LPF(f, \varphi)| + |HPF(f, \varphi)|$$

$$env_{des} = |LPF(f, \varphi)| - |HPF(f, \varphi)| .$$

The off-axis angle φ shows that in general the filter response must include the directional characteristics of the driver, although in this discussion it is assumed that the transfer functions include both the on-axis crossover filter and driver response, and that φ dependence is weak in the crossover transition region. The envelope parameters are useful indicators as they give a measure of the worst-case interference patterns that can occur for a given filter alignment, even though in practice with appropriate choice of crossover frequency and drive-unit spacing, the lower bound env_{des} should not be reached.

To implement a stochastic crossover alignment, a zero-mean random vector rd(n) is generated, which consists of a unit vector with superimposed noise sequence. An overall weighting factor λ is included to set the noise level together with an option for bandlimiting the noise spectrum formed by the destructive envelope bound $|LPF(f)_{Butt} - HPF(f)_{Butt}|$ of the prototype linear-phase Butterworth filters $LPF(f)_{Butt}$, $HPF(f)_{Butt}$. The modified Butterworth filters $LPF(f)'_{Butt}$, $HPF(f)'_{Butt}$ are then defined,

$$rd(n) = rand(n) , \qquad rd(n) = rd(n) - mean(rd(n))$$

$$LPF(f)'_{Butt} = LPF(f)_{Butt} \cdot * \{ones(n) + \lambda * (|LPF(f)_{Butt} - HPF(f)_{Butt}| - rd)\}/2$$

$$HPF(f)'_{Butt} = HPF(f)_{Butt} \cdot * \{ones(n) + \lambda * (|LPF(f)_{Butt} - HPF(f)_{Butt}| + rd)\}/2 .$$

The processing described in Section 3 can then be applied to these modified prototype filters, where it is emphasized that the target responses in both linear-phase and minimum-phase processes are noiseless. Consequently the composite response does not exhibit a noise structure. Full details can be observed in the listing in Appendix 2.

The effect of this process is that the composite response is unmodified, but the difference response env_{des}, which determines the off-axis lower interference bound, is noiselike. Consequently interdrive-unit interference that occurs in the filter transition band is distributed more broadly over the band with narrow bands of constructive and destructive interference.

It is conjectured that with music signals or any nonperiodic signal, which must have a broader spectrum, there are beneficial effects to be gained from this randomization. It should be noted that the proposed noise additions are static, and consequently the resulting changes in frequency response are also static and do not result in actual noise signals. To demonstrate the process, a linear-phase filter alignment is shown in Fig. 15, where complementary randomization in both the high-pass and the low-pass filters should be observed together with the destructive envelope env_{des}. Observe how the noise component amplitude has been weighted by the function env_{des}, but the composite function is unmodified. The effective "well" formed by the lower envelope function env_{des} is partially filled with the static noise, thus dispersing its effect.

5.5 Sinusoidal Frequency Interleaved Crossover-Filter Alignment

An alternative filter structure to that described in Section 5.4 is to use a sinusoidal interleave function to generate a regular array of peaks and dips in the low-pass and high-pass filter responses. The process is similar to the inclusion of a static noise function, except that the random vectors are replaced with complementary sinusoidal functions, where two alternatives are proposed:

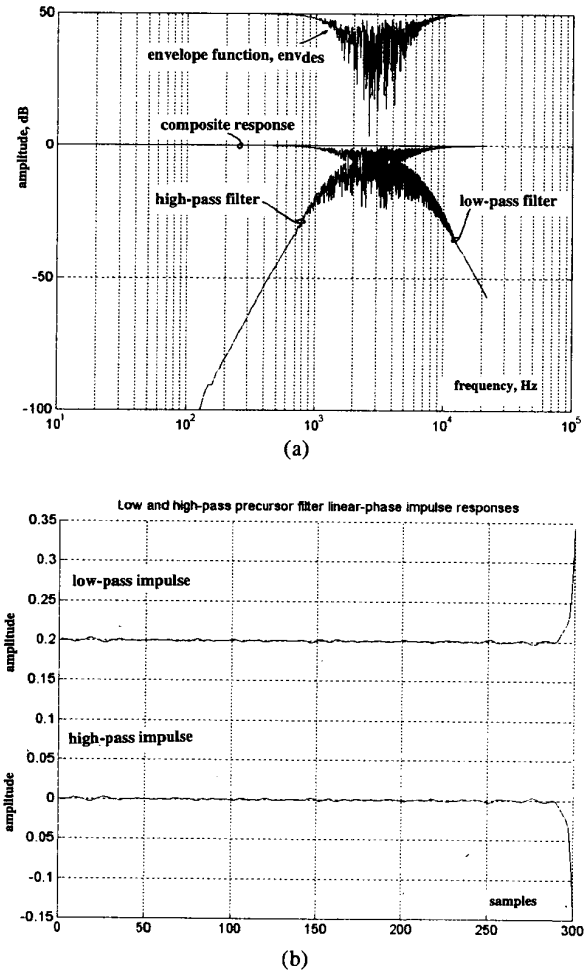


Fig. 15. (a) Linear-phase stochastic interleave crossover alignment. (b) Precursor impulse response of LPF and HPF.

1) An interleave function that is sinusoidal against linear frequency

2) An interleave function that is sinusoidal against logarithmic frequency.

In both cases the amplitude of the interleave function can be weighted according to the polar interference region defined by env_{des} with overall equalization to ensure that the composite response is unity. The logarithmic frequency function is considered the better choice as there is greater time dispersion with minimal discrete echo effect, which results from a linear comb filter. Also the calculation of filters suggests that linear-phase alignments are a better choice. Especially where the interleave function is of high amplitude, minimum-phase functions find it problematic to track the extreme ripples in the response.

Fig. 16 shows an example of a linear-phase sinusoidal interleave alignment. It is interesting to observe the upper envelope of the low-pass and high-pass functions as these give a better measure of the attenuation characteristic. The low-pass and high-pass prototype filters had -6-dB break frequencies of 2 kHz and 4 kHz yielding a crossover nominally at $\sqrt{8}$ kHz with respective filter slopes of -30 dB per octave and 30 dB per octave.

6 CONCLUSION

This paper has explored a number of examples of linear-phase and minimum-phase signal processing. The aim has been to demonstrate the tools and transforms that can be used to compute exact impulse sequences directly from their amplitude response descriptions using minimum-phase theory. The results were presented as time-domain sequences, CDS displays, and ETC responses.

The use of two-dimensional convolution masks was shown to improve the presentation of a CDS. Orders up to seven have been calculated, and their effect on a typical CDS is described. These modified CDS displays were applied to both directly measured and minimum-phase derived impulse responses. The method of presenting only the minimum-phase data was offered as an aid to analyzing loudspeaker performance as it biases the displayed data more closely to those performance attributes of subjective significance by effectively ignoring the excess-phase distortion. From measurements taken from a number of loudspeakers, it has been shown that much of the time-domain dispersion is phase distortion, which typically results from the incorporation of an analog all-pass crossover alignment. However, it is suggested that although the excess-phase information has been extracted, it should not necessarily be ignored, but considered as a separate performance measure and possibly observed on a dedicated CDS that excludes the minimum-phase distortion.

The ETC display was considered where the normally occurring precursor response was illustrated. It was shown that minimum-phase processing applied to the ETC could maintain the same spectral envelope while eliminating the apparent noncausal behavior. The process therefore improves the cosmetic appearance of the display and offers a response that is more realistic and acceptable to interpretation.

Two examples of frequency response weighting were presented and the linear and minimum-phase responses illustrated by way of examples. It is suggested that these two forms of equalization are particularly effective at correcting minor defects in music recordings.

Finally, a section was presented on loudspeaker crossover design, where a generalized design technique was followed. In particular, two approaches based upon zero-phase Butterworth prototype filters were presented and both linear-phase and minimum-phase realizations discussed. It was shown with DSP that arbitrary slope filters can be implemented and that with the use of a raised-cosine waiting function and overlapping filter responses, high rates of attenuation in the stopband could be retained if desired. Finally, consideration was given to a new class of interleave crossover alignments, where either a stochastic or a sinusoidal interleave sequence was superimposed upon the lower crossover interference bound such that the cancellation effects in the transition region are dispersed over a wider bandwidth. It is believed that this will offer advantages for broad-band sig-

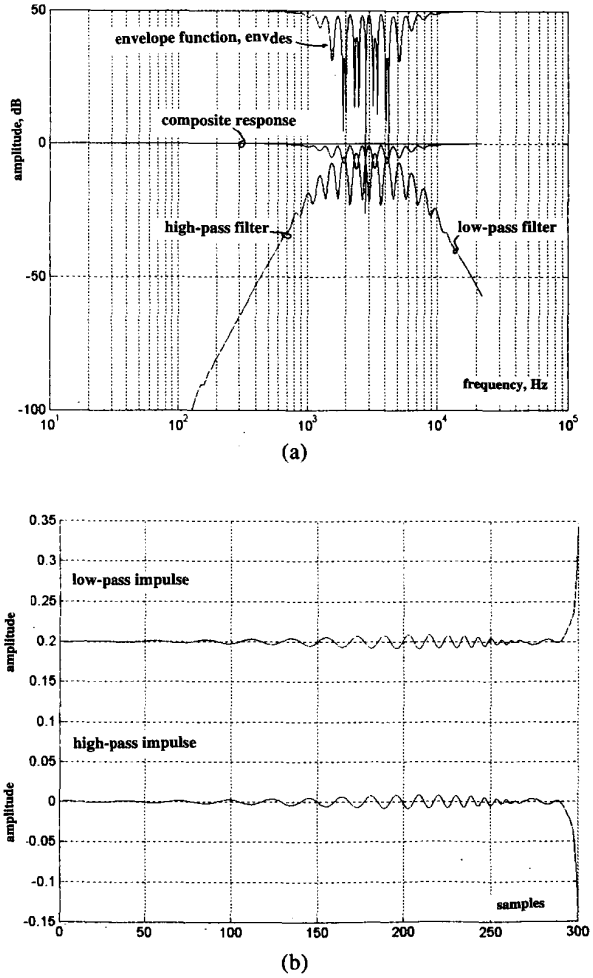


Fig. 16. (a) Linear-phase sinusoidal interleave crossover alignment. (b) Precursor impulse response of LPF and HPF.

nals such as music, especially as the on-axis target response is unaffected.

7 ACKNOWLEDGMENT

I wish to thank Joachim Gerhard of Audio Physic, Germany, for introducing me to the psychoacoustic characteristic described in Section 5.2. Also thanks go to Brian Elliott, Palo Alto, CA, for his many hours of encouragement and insight into the finer details of loudspeaker measurement and active loudspeaker systems.

8 REFERENCES

[1] J. Borwick, *Loudspeaker and Headphone Handbook*, 2nd ed. (Butterworth-Heinemann, 1994).

[2] S. P. Lipshitz and J. Vanderkooy, "In-Phase Crossover Network Design," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 969 (1983 Dec.), preprint 2051.

[3] R. H. Small, "Constant-Voltage Crossover Network Design," *J. Audio Eng. Soc.*, vol. 19, pp. 12–19 (1971 Jan.).

[4] P. Garde, "All-Pass Crossover Systems," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 28, pp. 575–584 (1980 Sept.).

[5] M. O. J. Hawksford, "Asymmetric All-Pass Crossover Alignments," *J. Audio Eng. Soc. (Abstracts)*, vol. 41, pp. 123–134 (1993 Mar.).

[6] S. H. Linkwitz, "Active Crossover Networks for Noncoincident Drivers," *J. Audio Eng. Soc.*, vol. 24, pp. 2–8 (1976 Jan./Feb.).

[7] R. Greenfield and M. O. J. Hawksford, "Efficient Filter Design for Loudspeaker Equalization," *J. Audio Eng. Soc.*, vol. 39, pp. 739–751 (1991 Oct.).

[8] J. A. Deer, P. J. Bloom, and D. Preis, "Perception of Phase Distortion in All-Pass Filters," *J. Audio Eng. Soc.*, vol. 33, pp. 782–786 (1985 Oct.).

[9] R. Greenfield and M. O. J. Hawksford, "The Audibility of Loudspeaker Phase Distortion," presented at the 88th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 38, p. 384 (1990 May), preprint 2927.

[10] B. Theiβ, J. Gerhard, and M. O. J. Hawksford, "Loudspeaker Placement for Optimised Phantom Source Reproduction," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 647 (1996 July/Aug.), preprint 4246.

[11] R. M. Heylen and M. O. J. Hawksford, "Interpolation between Minimum-Phase and Linear-Phase Frequency Responses," presented at the 98th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 43, p. 399 (1995 May), preprint 3994.

[12] A. Rimell and M. O. J. Hawksford, "Digital-Crossover Design Strategy for Drive Units with Impaired and Noncoincident Polar Characteristics," presented at the 95th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 41, p. 1065 (1993 Dec.), preprint 3750.

[13] J. Blauert, *Räumliches Hören* (Hirzel Verlag, Stuttgart, Germany, 1985. English version, *Spatial Hearing* MIT Press, Cambridge, MA, 1983).

[14] R. Chalupa, "A Subtractive Implementation of Linkwitz–Riley Crossover Design," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 34, pp. 556–559 (1986 July/Aug.).

[15] M. O. J. Hawksford, "A Family of Circuit Topologies for the Linkwitz–Riley (LR-4) Crossover Alignment," presented at the 82nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 35, p. 391 (1987 May), preprint 2468.

[16] S. P. Lipshitz and J. Vanderkooy, "Use of Frequency Overlap and Equalization to Produce High-Slope Linear-Phase Loudspeaker Crossover Networks," *J. Audio Eng. Soc.*, vol. 33, pp. 114–126 (1985 Mar.).

[17] B. Hillerich, "Acoustic Alignment of Loudspeaker Drivers by Nonsymmetrical Crossovers of Different Orders," *J. Audio Eng. Soc.*, vol. 37, pp. 691–699 (1989 Sept.).

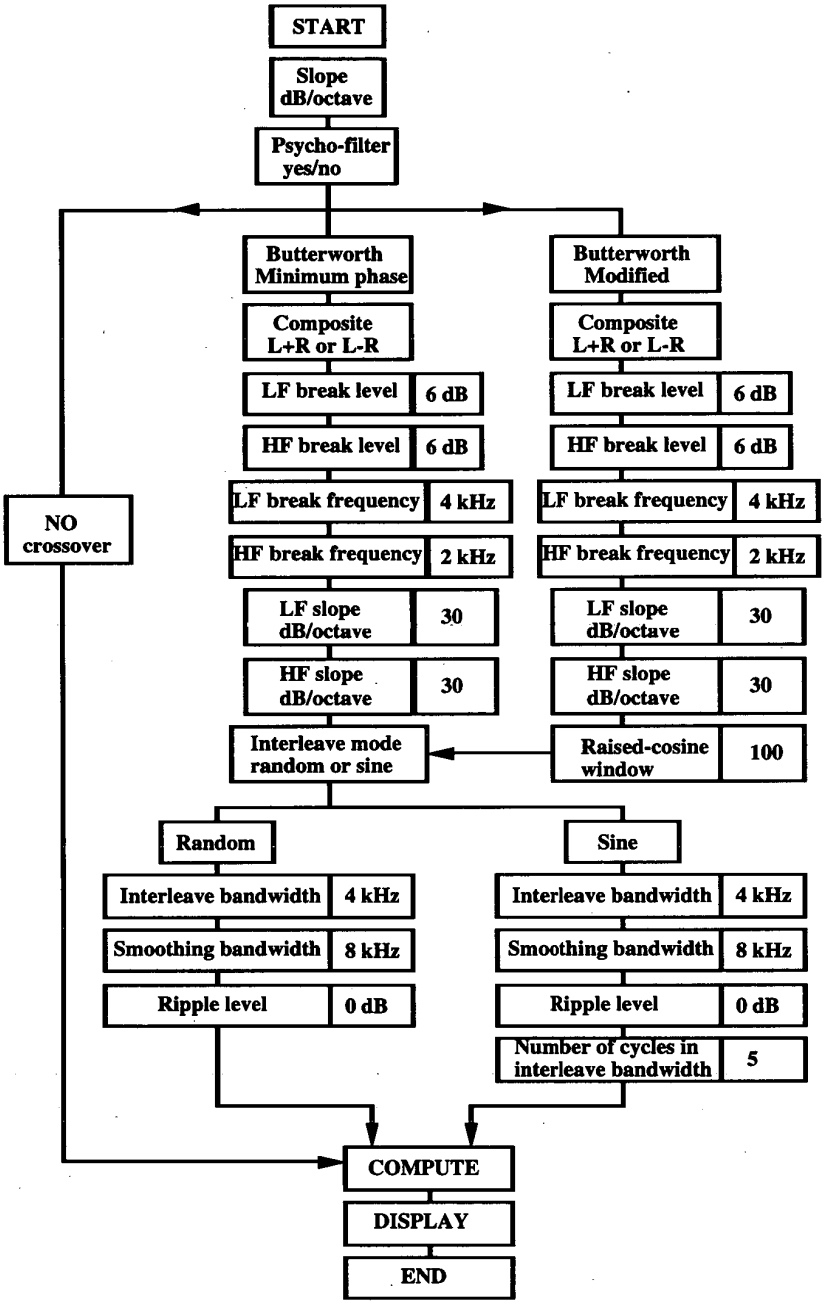
[18] S. P. Lipshitz and J. Vanderkooy, "A Family of Linear-Phase Crossover Networks of High Slope Derived by Time Delay," *J. Audio Eng. Soc.*, vol. 31, pp. 2–20 (1983 Jan./Feb.).

APPENDIX 1
LIST OF MATHEMATICAL OPERATORS

The following mathematical operators are used in the text and derived from the MATLAB family:

abs	absolute value of a vector
angle	vector representing the phase of a complex vector
conj	generates a complex conjugate vector
conv	one-dimensional convolution of two vectors
conv2	two-dimensional convolution of two square matrices
exp	returns vector representing a complex exponential of a vector
fft	fast Fourier transform of a vector
hankel	creates a Hankel matrix
hilbert	returns Hilbert transform: real part input vector, imaginary part Hilbert transform
ifft	inverse fast Fourier transform
imag	returns a vector that reads the imaginary part of a complex vector
mesh	produces a three-dimensional plot of a two-dimensional matrix
log	returns a vector that is the natural logarithm of a vector
ones(n)	unit vector of length N elements
zeros(n)	zero vector of length N elements
./	element-by-element division of two vectors
.*	element-by-element multiplication of two vectors
$x.\hat{}y$	elements of vector x raised to a power y

APPENDIX 2
MATLAB CROSSOVER DESIGN PROGRAM LISTING




```
% Loudspeaker digital crossover design program
home; clear; close;
fprintf('LOUDSPEAKER CROSSOVER DESIGN PROGRAM:\n');
fprintf('(hints given as guide for parameter values):\n');

% fs, sampling rate
fs=44100;

% mv  CDS orientation
mv=[30,-30];

% cm, ch  CDS matrix size
cm=150; ch=80;
```

```
% m, vector length
m=4096;
mm=1:m; m2=m/2; ml=fs/m; mh=m2*ml; mx=ml:ml:mh; rx=1:m2;


% number of samples in impulse plots
n=50;
nn=1:n; pq=1;


% cm3: 3 by 3 convolution mask
cm1=1;
cm2=[1 1;1 1]/4;
cm3=conv2(cm2,cm2); cm3=cm3/(sum(sum(cm3)));
cm4=conv2(cm3,cm3); cm4=cm4/(sum(sum(cm4)));
cm5=conv2(cm4,cm4); cm5=cm5/(sum(sum(cm5)));
cm6=conv2(cm5,cm5); cm6=cm6/(sum(sum(cm6)));
cm7=conv2(cm6,cm6); cm7=cm7/(sum(sum(cm7)));


% Filter tailoring
fprintf('\nTAILOR FILTER RESPONSE\n');
% Modify equalizer slope as N dB/octave / crossover filter
fprintf('\nEnter modified equalization characteristic in dB/octave');
Nx=input('Enter slope: ');
nj=Nx/(20*log10(2));  fj=(ml*mh)^.5;
a=(mx/fj).^nj; al=a/2; ah=a/2; xm=1;
fprintf('\nEnter: 0 no psycho filter, 1 psycho filter: ');
filt=input('Select additional filter type: ');
np=ml*fix(3000/ml);
if filt==1
aj=[-2.5*log10((ml:ml:np)/3000) 6.4*log10((np+ml:ml:mh)/3000)];
aj=aj-(aj(m2)+min(aj))/2;
aj=(10*ones(size(1:m2))).^(aj/20);
a=a.*aj; clear aj;
end;
fprintf('\nXover: 0 none, 1 Butt (min phase), 2 modified Butt');
filt=input('Select mode: '); filt=abs(filt);
if filt>0
fprintf('\nComposite minimum phase: Enter 1 for (L + H), 2 for (1 - H) )');
qq=input('Select mode: ');
if qq==2
pq=-1;
end
home
fprintf('\nDefine LF fndB break level: (eg n = 3 means f3dB)');
ncl=input('LF:  n dB (eg 6 dB): '); ncl=10^(abs(ncl)/20);
fprintf('\nDefine HF fndB break level: (eg n = 3 means f3dB)');
nch=input('HF:  n dB (eg 6 dB): '); nch=10^(abs(nch)/20);
fprintf('\nEnter prototype low-pass Butterworth fndB break frequency:');
xl=input('Frequency in kHz (eg 4 kHz): ');
xol=fix(1000*xl/ml);
fprintf('\nEnter prototype high-pass Butterworth fndB break frequency:');
xh=input('Frequency in kHz (eg 2 kHz): ');
xoh=fix(1000*xh/ml);
fprintf('\nEnter LF xover-filter slope in dB / octave:');
ordl=input('Filter slope (eg 30 dB / octave): '); ordl=abs(ordl/(20*log10(2)));
fprintf('\nEnter HF xover-filter slope in dB / octave:');
ordh=input('Filter slope (eg 30 dB / octave): '); ordh=abs(ordh/(20*log10(2)));
xm=fix((xol*xoh)^.5);
xofl=ml*xol/(ncl^(2/ordl)-1)^.5;
xofh=ml*xoh*(nch^(2/ordh)-1)^.5;
if filt ==2
fprintf('\nEnter raised-cosine transition window');
win=input('Number (even) of samples (eg 100): '); win=2*abs(fix(win/2));
if (xm-win/2)<1
win=win-(xm-win/2)-1;
end; end
al=a./(1+(mx/xofl).^2).^(ordl/2);
ah=a.*(((mx/xofh).^2)./(1+(mx/xofh).^2)).^(ordh/2);
home
```

```
fprintf('FILTER INTERLEAVE PARAMETERS\n');
fprintf('Enter 1 for random, 2 for sinusoidal interleave mode\n');
im=input('Interleave mode: '); im=abs(im);
fprintf('\nTotal interleave bandwidth, kHz\n');
bw=input('Bandwidth (eg 4 kHz): '); bw=abs(1000*bw/ml); centre=(xol*xoh)^.5;
ibh=.5*(bw+(bw^2+4*centre^2)^.5); ibl=centre^2/ibh;
ibl=ceil(ibl); ibh=ceil(ibh);
if ibh>m2
ibh=m2;
end
fprintf('\nTotal smoothing bandwidth for interleave function kHz\n');
wm=input('Bandwidth (eg 8 kHz): '); wm=abs(1000*wm/ml);
c1=wm+ibh-ibl; c2=ibl*ibh;
h=.5*(c1+(c1^2+4*c2)^.5); l=c2/h; h=ceil(h-.5); l=ceil(l);
if h>m2
h=m2;
end
wml=ibl-l; wmh=h-ibh; wu=(1:wml)/(wml+1); wd=(wmh:-1:1)/(wmh+1);
winr=[zeros(size(1:l-1)) wu ones(size(ibl:ibh)) wd zeros(size(h:m2-1))];
home
if im==1
fprintf('RANDOM INTERLEAVE CROSSOVER PARAMETERS\n');
fprintf('\nSpectral noise ripple in dB relative to 0 dB:');
fprintf('\n(Enter negative number for reduced level)');
rip=input('Noise ripple dB (eg 0 dB): '); nrip=10^(rip/20)*winr;
rd=rand(1,m2); rd=rd-mean(rd); rd=rd/max(abs(rd));
nl=.5*abs(ones(size(rx))+nrip.*(abs(al-ah)-rd));
nh=.5*abs(ones(size(rx))+nrip.*(abs(al-ah)+rd));
else
fprintf('SINUSOIDAL INTERLEAVE CROSSOVER PARAMETERS\n');
fprintf('\nSinusoidal amplitude ripple in dB relative to 0 dB:');
fprintf('\n(Enter negative number for reduced level)');
rip=input('Spectral sinusoidal ripple dB (eg 0 dB): '); nrip=10^(rip/20)*winr;
fprintf('\nEnter number of cycles in interleave bandwidth:');
nc=input('Number of cycles (eg 5 cycles): ');
nc=ceil(abs(nc)/log10(ibh/ibl)*log10(h/l)/log10(ibh/ibl)-.5)/log10(h/l)*log10(ibh/ibl);
sd=sin(2*pi*nc*log10(rx/ibl));
nl=.5*abs(ones(size(rx))+nrip.*(abs(al-ah)-sd));
nh=.5*abs(ones(size(rx))+nrip.*(abs(al-ah)+sd));
end
al=al.*nl; ah=ah.*nh;
if filt == 2
wl=[ones(size(1:xm-win/2)), .5*(1+cos(pi*(1:win)/win)),
zeros(size(1+xm+win/2:m2))];
wh=ones(size(1:m2))-wl;
mal=al+(a-al-ah).*wl; mah=ah+(a-al-ah).*wh;
al=mal; ah=mah; clear mal mah;
end, end
if filt==0
al=a; ah=a;
end
a=[a fliplr(a)]; al=[al fliplr(al)]; ah=[ah fliplr(ah)];
clc; home
fprintf('COMPUTING DATA');
if filt>0
if qq==2
fprintf('\nComposite minimum phase = {low - high}')
else
fprintf('\nComposite minimum phase = {low + high}')
end, end

% linear-phase impulse response
linl=real(ifft(al));
linh=real(ifft(ah));
lina=real(ifft(a));
lsct=linl+linh;

% minimum-phase impulse responses, non-composite equalised
minl=real(ifft(exp(conj(hilbert(log(al))))));
minh=pq*real(ifft(exp(conj(hilbert(log(ah))))));
mina=real(ifft(exp(conj(hilbert(log(a))))));
```

```
% minimum-phase composite equalised impulse responses
meqf=a./abs(fft(minl+minh));
meqt=real(ifft(exp(conj(hilbert(log(meqf))))));
minl=[conv(minl(1:m2),meqt(1:m2)),0];
minh=[conv(minh(1:m2),meqt(1:m2)),0];
msct=minl+minh;

% linear-phase composite sum amplitude response
lsum=abs(fft(linl+linh));

% linear-phase composite difference amplitude response
ldif=abs(fft(linl-linh));

% minimum-phase equalized amplitude responses
msum=abs(fft(minl+minh));
mal=abs(fft(minl));
mah=abs(fft(minh));

% Commence plot routine
% select plot routines
nn=1:n; ds=0;
while ds<1
home
fprintf('SELECT GRAPHICS:\n');
fprintf('\nLF linear-phase  impulse response                    1 ');
fprintf('\nHF linear-phase  impulse response                    2 ');
fprintf('\nLF minimum-phase impulse response                    3 ');
fprintf('\nHF minimum-phase impulse response                    4 ');
fprintf('\nLinear-phase  composite impulse response             5 ');
fprintf('\nMinimum-phase composite impulse response             6 ');
fprintf('\nLinear-phase  energy-time curve                      7 ');
fprintf('\nMinimum-phase energy-time curve                      8 ');
fprintf('\nCDS LF minimum-phase impulse (logarithmic)           9 ');
fprintf('\nCDS HF minimum-phase impulse (logarithmic)           10 ');
fprintf('\nCDS Composite minimum-phase impulse (logarithmic)    11 ');
fprintf('\nCDS LF linear-phase impulse (logarithmic)            12 ');
fprintf('\nCDS HF linear-phase impulse (logarithmic)            13 ');
fprintf('\nCDS Composite linear-phase impulse (logarithmic)     14 ');
fprintf('\nMinimum-phase fourier transform                      15 ');
fprintf('\nLinear-phase fourier transform                       16 ');
fprintf('\nComposite frequency response with polar scan         17 ');
fprintf('\nTerminate program                                    18 ');
ds=input('Select display option number 1 to 18: ds = ');
ds=abs(ds); home

% fprintf('\nVARIABLES: xl, xh, a, al, ah, linl, linh, mina, minl, minh');
% fprintf('\nVARIABLES: lsct, msct, meqf, meqt, a, lina, lsum, ldif, msum, mdif\n');

% LF linear-phase impulse response
if ds==1
fprintf('Composite frequency response\n')
n=input('Enter number of samples in display (eg 200): '); n=ceil(abs(n))/2; nn=1:n;
hold on; grid on
lins=[linl(m-n+1:m),linl(1:n)];
plot(1:2*n,lins(1:2*n))
title('Low-pass filter linear-phase impulse response')
hold off
pause; close; home
end

% HF linear-phase impulse response
if ds==2
fprintf('Composite frequency response\n')
n=input('Enter number of samples in display (eg 200): '); n=ceil(abs(n))/2; nn=1:n;
hold on; grid on
lins=[linh(m-n+1:m),linh(1:n)];
plot(1:2*n,lins(1:2*n))
title('High-pass filter linear-phase impulse response')
hold off; pause; close; home
end
```

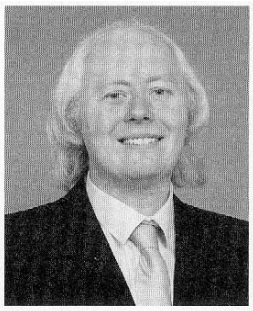
```
% LF minimum-phase impulse response
if ds==3
fprintf('Composite frequency response\n')
n=input('Enter number of samples in display (eg 200): '); n=ceil(abs(n)); nn=1:n;
hold on; grid on
plot(nn,minl(nn))
title('Low-pass filter minimum-phase impulse response')
hold off
pause; close; home
end

% HF minimum-phase impulse response
if ds==4
fprintf('Composite frequency response\n')
n=input('Enter number of samples in display (eg 200): '); n=ceil(abs(n)); nn=1:n;
hold on; grid on
plot(nn,minh(nn))
title('High-pass filter minimum-phase impulse response')
hold off; pause; close; home
end

% Linear-phase composite impulse response
if ds==5
fprintf('Composite frequency response\n')
n=input('Enter number of samples in display (eg 200): '); n=ceil(abs(n)); nn=1:n;
hold on; grid on
lins=[lsct(m-n+1:m),lsct(1:n)];
plot(1:2*n,lins(1:2*n))
title('Linear-phase composite impulse response')
hold off; pause; close; home
end

% Minimum-phase composite impulse response
if ds==6
fprintf('Composite frequency response\n')
n=input('Enter number of samples in display (eg 200): '); n=ceil(abs(n)); nn=1:n;
hold on; grid on
plot(nn,msct(nn))
title('Minimum-phase composite impulse response')
hold off; pause; close; home
end

% Linear-phase energy-time curve
if ds==7
etn=input('Enter number of samples to display: '); etn=ceil(abs(etn));
if etn>m2
etn=m2;
end
dis=input('Enter: 1 for LPF, 2 for HPF or 3 for composite: ');
if dis==1
ett=hilbert(linl);
elseif dis==2
ett=hilbert(linh);
else
ett=hilbert(lsct);
end; home
off=30;
ett=ett./max(abs(ett));
[p1,p2]=max(abs(ett)); of=off-p2;
if of<1
off=p2+1;
of=1;
end;
et(1:of+1)=ett(m-of:m); et(of+2:m)=ett(1:m-of-1); clear ett p1 p2 of;
% take minimum phase of envelope of energy-time curve
fa=abs(fft(abs(et)));
ett=real(ifft(exp(conj(hilbert(log(fa))))));
ett=ett./max(abs(ett)); clear fa;
[p1,p2]=max(abs(ett)); of=off-p2;
if of<1
of=1;
end;
```

```
etm(1:of+1)=ett(m-of:m); etm(of+2:m)=ett(1:m-of-1); clear ett p1 p2 of;
hold on
if dis==1
title('LPF (linear phase): ET-C (red), min. phase of ET-C (green)')
elseif dis==2
title('HPF (linear phase): ET-C (red), min. phase of ET-C (green)')
else
title('Composite (linear phase): ET-C (red), min. phase of ET-C (green)')
end
plot(1:etn,abs(et(1:etn))/max(et)+ones(size(1:etn)),'r')
plot(1:etn,abs(etm(1:etn))/max(etm),'g')
grid; hold on; pause; close
end


% Minimum-phase energy-time curve
if ds==8
etn=input('Enter number of samples to display: '); etn=ceil(abs(etn));
if etn>m2
etn=m2;
end
dis=input('Enter: 1 for LPF, 2 for HPF or 3 for composite: ');
if dis==1
ett=hilbert(minl);
elseif dis==2
ett=hilbert(minh);
else
ett=hilbert(msct);
end; home
off=30;
ett=hilbert(msct);
ett=ett./max(abs(ett));
[p1,p2]=max(abs(ett)); of=off-p2;
if of<1
off=p2+1;
of=1;
end;
et(1:of+1)=ett(m-of:m); et(of+2:m)=ett(1:m-of-1); clear ett p1 p2 of;
% take minimum phase of envelope of energy-time curve
fa=abs(fft(abs(et)));
ett=real(ifft(exp(conj(hilbert(log(fa))))));
ett=ett./max(abs(ett)); clear fa;
[p1,p2]=max(abs(ett)); of=off-p2;
if of<1
of=1;
end;
etm(1:of+1)=ett(m-of:m); etm(of+2:m)=ett(1:m-of-1); clear ett p1 p2 of;
hold on
if dis==1
title('LPF (minimum phase): ET-C (red), min. phase of ET-C (green)')
elseif dis==2
title('HPF (minimum phase): ET-C (red), min. phase of ET-C (green)')
else
title('Composite (minimum phase): ET-C (red), min. phase of ET-C (green)')
end
plot(1:etn,abs(et(1:etn))/max(et)+ones(size(1:etn)),'r')
plot(1:etn,abs(etm(1:etn))/max(etm),'g')
grid; hold off; pause; close
end


% set order of 2-D filter mask
if ds>8
if ds<15
cx=input('Enter 1 to 7 for order of 2-D filter mask: ');
if cx==2
cmx=cm2;
elseif cx==3
cmx=cm3;
elseif cx==4
cmx=cm4;
```

```
elseif cx==5
cmx=cm5;
elseif cx==6
cmx=cm6;
elseif cx==7
cmx=cm7;
else
cmx=cm1;
end; end; end;

% CDS LF minimum-phase impulse (logarithmic)
if ds==9
ra=input('Enter vertical range (nearest decade) in dB: '); ra=10*ceil(abs(ra)/10);
cd=20*log10(abs(fft(hankel(minl(1:cm)))));
cx=conv2(cd,cmx);
cx=cx-max(max(cx))+ra;
cx=.5*(cx+sign(cx).*cx)-ra;
mesh(cx(1:ch,1:cm)',mv)
title('CDS LF minimum-phase impulse (logarithmic)')
pause; close; home
clear cd cx cmx;
end

% CDS HF minimum-phase impulse (logarithmic)
if ds==10
ra=input('Enter vertical range (nearest decade) in dB: '); ra=10*ceil(abs(ra)/10);
cd=20*log10(abs(fft(hankel(minh(1:cm)))));
cx=conv2(cd,cmx);
cx=cx max(max(cx))+ra;
cx=.5*(cx+sign(cx).*cx)-ra;
mesh(cx(1:ch,1:cm)',mv)
title('CDS HF minimum-phase impulse (logarithmic)')
pause; close; home
clear cd cx cmx;
end

% CDS Composite minimum-phase impulse (logarithmic)
if ds==11
ra=input('Enter vertical range (nearest decade) in dB: '); ra=10*ceil(abs(ra)/10);
cd=20*log10(abs(fft(hankel(msct(1:cm)))));
cx=conv2(cd,cmx);
cx=cx-max(max(cx))+ra;
cx=.5*(cx+sign(cx).*cx)-ra;
mesh(cx(1:ch,1:cm)',mv)
title('CDS Composite minimum-phase impulse (logarithmic)')
pause; close; home
clear cd cx cmx;
end

% CDS LF linear-phase impulse (logarithmic)
if ds==12
ra=input('Enter vertical range (nearest decade) in dB: '); ra=10*ceil(abs(ra)/10);
cd=20*log10(abs(fft(hankel(linl(1:cm)))));
cx=conv2(cd,cmx);
cx=cx-max(max(cx))+ra;
cx=.5*(cx+sign(cx).*cx)-ra;
mesh(cx(1:ch,1:cm)',mv)
title('CDS LF linear-phase impulse (logarithmic)')
pause; close; home
clear cd cx cmx;
end

% CDS HF linear-phase impulse (logarithmic)
if ds==13
ra=input('Enter vertical range (nearest decade) in dB: '); ra=10*ceil(abs(ra)/10);
cd=20*log10(abs(fft(hankel(linh(1:cm)))));
cx=conv2(cd,cmx);
cx=cx-max(max(cx))+ra;
cx=.5*(cx+sign(cx).*cx)-ra;
mesh(cx(1:ch,1:cm)',mv)
```

```
title('CDS HF linear-phase impulse (logarithmic)')
pause; close; home
clear cd cx cmx;
end

% CDS Composite linear-phase impulse (logarithmic)
if ds==14
ra=input('Enter vertical range (nearest decade) in dB: '); ra=10*ceil(abs(ra)/10);
cd=20*log10(abs(fft(hankel(lsct(1:cm)))));
cx=conv2(cd,cmx);
cx=cx-max(max(cx))+ra;
cx=.5*(cx+sign(cx).*cx)-ra;
mesh(cx(1:ch,1:cm)',mv)
title('CDS Composite linear-phase impulse (logarithmic)')
pause; close; home
clear cd cx cmx;
end

% Minimum-phase fourier transform
if ds==15
semilogx(mx,20*log10(.00001*ceil(100000*mal(rx))),'g',mx,20*log10(.00001*ceil(1
00000*mah(rx))),'r')
hold on
semilogx(mx,20*log10(meqf(rx)),'w',mx,20*log10(mal(rx)-
mah(rx))+50,'b',mx,20*log10(msum(rx)),'y')
grid
if qq==2
title('FT(minimum phase): y |low-high|, b |low|+|high|, w |EQ|, g |low|, r |high|')
else
title('FT(minimum phase): y |low+high|, b |low|-|high|, w |EQ|, g |low|, r |high|')
end
hold off; pause; close; home
end

% Linear-phase fourier transform
if ds==16
semilogx(mx,20*log10(.00001*ceil(100000*al(rx))),'g',mx,20*log10(.00001*ceil(10
0000*ah(rx))),'r')
hold on
semilogx(mx,20*log10(ldif(rx))+50,'b',mx,20*log10(lsum(rx)),'y')
grid
title('FT(linear phase): y |low+high|, b |low|-|high|, g |low|, r |high|')
hold off; pause; close; home
end

% Composite frequency response with polar offset
if ds==17
fprintf('Composite frequency response with polar offset\n')
tx=input('Enter time-delay offset between drivers in ms (eg 0 to 1 ms): '); tx=tx*1E-3;
semilogx(mx,20*log10(a(rx)),'r')
title('Composite frequency response with polar offset')
hold on; grid
lpolar=abs(exp(i*pi*mm*tx).*fft(linl)+exp(-i*pi*mm*tx).*fft(linh));
semilogx(mx,20*log10(lpolar(rx)),'y')
hold off; pause; close; home
end

% Program terminate
if ds==18
ds=input('Select 0 for termination, 1 to continue: ');
ds=abs(ds);
if ds==0
home
fprintf('\nProgram termination:\n\n')
return
end; end; ds=0; end;

return
%
*****************************************************************
```

THE AUTHOR



Malcolm Omar Hawksford is director of the Centre for Audio Research and Engineering, a professor in the Department of Electronic Systems Engineering at the University of Essex and postgraduate scheme director, where his interests encompass audio engineering, electronic circuit design, and signal processing. Professor Hawksford studied electrical engineering at the University of Aston in Birmingham where he gained a First Class Honours B.Sc. and Ph.D. His Ph.D. program, which was sponsored by a BBC Research Scholarship, investigated delta modulation and delta-sigma modulation (now commonly known as bitstream coding) for color television and also the development of a time-compression/time-multiplex system for combining luminance and chrominance signals (a forerunner of the MAC/DMAC system). While at Essex University, he has undertaken research principally in the fields of analog amplifiers, digital signal processing, and loudspeaker systems. Since 1982 research into digital crossover networks and equalization for loudspeakers has been pursued which has culminated in an advanced digital and active loudspeaker system being designed within the university. Research topics have also encompassed oversampling and noise shaping techniques applied to analog-to-digital and digital-to-analog conversion, the linearization of PWM encoders, and 3-dimensional spatial audio and telepresence including multichannel sound reproduction.

Professor Hawksford has published in the *Journal of the Audio Engineering Society* on topics that include error correction in amplifiers, oversampling techniques, and MLS techniques. His supplementary activities include writing contributions for *Hi-Fi News and Record Review* and *Stereophile* magazine as well as designing high-end analog and digital audio equipment. He is a chartered engineer and is a fellow of the AES, the Institution of Electrical Engineers, and the Institute of Acoustics. He is a member of the technical committee of Acoustic Renaissance for Audio (ARA), a group currently promoting a system for storing multichannel, high-definition audio signals on high-capacity DVD optical disks. He is also technical adviser to *HFN and Record Review* and a technical consultant to LFD Audio, UK.

MATLAB Program for Loudspeaker Equalization and Crossover Design*

M. O. J. HAWKSFORD, *AES Fellow*

Centre for Audio Research and Engineering, University of Essex, UK CO4 3SQ

A digital design filter program is presented written in the MATLAB environment. Impulse response time editing is implemented together with various options for spectral domain processing. The filter inputs time-domain impulse response data and outputs filter coefficients for both FIR and IIR implementations. Comprehensive display options are incorporated, including minimum-phase processing and CDS, and consideration is given to both specular and diffuse loudspeakers. Applications include the design of digital equalization filters and digital crossover filters for loudspeaker systems.

0 INTRODUCTION

Digital filters enable accurate loudspeaker equalization and the implementation of near-theoretic crossover alignments [1], [2]. As such it is possible to implement loudspeaker systems which yield extremely accurate overall frequency responses by taking full account of the drive units' inherent response irregularities. This technique is not only attractive for digital and active loudspeaker systems, but it also applies to passive loudspeakers that are preceded by a digital filter to fine-tune the frequency response.

To develop crossover and equalizer filters, design tools are required to compute the digital filter coefficients that can accommodate a range of target responses, including those of the crossover. This engineering report describes a filter design program that is written within the MATLAB[1] environment (version 5.2). The program inputs data in the form of a loudspeaker's impulse response and can embed specific equalization target functions, including frequency response shaping and crossover design. A flexible design is offered, where either an FIR or an IIR filter structure can be selected and the number of coefficients in the numerator and the denominator specified. An excess-phase equalizer is also computed to enable correction of phase distortion. This allows the user to change the number of coefficients and then to observe the effect on the overall equalization error.

The filter design can take account of a drive unit's individual response (which is entered as an impulse response) and matches this to a user-specified target frequency response. In its simplest form this target is constant with frequency, but can be adapted to include a number of equalization functions. The target function can also include a crossover filter, and a procedure is described in Section 2.3 (see [3] for more detail).

The program incorporates a range of data display options including time-domain data, frequency domain and cumulative decay spectrum (CDS) together with the cumulative decay root (CDR) spectrum variant reported earlier [3], and the energy–time curve (ETC).

Also, a modified linear-phase Butterworth crossover filter is incorporated as well as "tilt" and psychoacoustic filters [4]. Pivotal to the design procedure is minimum-phase signal processing, which is used in both the filter synthesis and the data display options, including impulse and CDS. As such the program can be used to design digital filters for two-way digital and active loudspeakers [5].

The background to the signal processing techniques was described in a recent paper [3], and it is suggested that this be read as a companion to the present text. The program listing is available on request,[2] while this report presents an overview of the computation procedures. To illustrate the capabilities of the program two loudspeaker examples are presented. The first is a conventional small two-way system using moving-coil drive units, whereas the second is a distributed-mode loudspeaker (DML) [6].

* Presented at the 105th Convention of the Audio Engineering Society, San Francisco, CA, 1998 September 26–29; revised 1999 June 29.

[1] MATLAB is the trade name of a commercial matrix-based processing language.

[2] E-mail address: mjh@essex.ac.uk.

These widely differing designs were chosen specifically to enable a deeper appreciation of the characteristics of a DML when compared against conventional specular radiating devices. However, it should be observed that a DML is a spatiotemporal dispersive radiator. As such the example presented here is specific to one measurement location. In practice there is stochastic variation in the polar response.

1 OVERVIEW OF DIGITAL EQUALIZATION FILTER DESIGN

The equalization filter design procedure is reviewed briefly in this section.

1) *Impulse response measurement.* A direct measure of the loudspeaker impulse response is captured using, for example, a maximum-length sequence (MLS) excitation.

2) *Time-domain editing of measured impulse response*

a) *Preresponse editing.* Redundant samples prior to the main impulse response can be removed using on-screen editing. However, any prering due to antialiasing filters must be considered in this truncation process.

b) *Postresponse editing.* First and subsequent boundary reflections from within the measurement environment can be removed together with noise that often contaminates the low-level "tail" of an extended impulse response measurement. This truncation includes a short raised-cosine window to smooth the endpoint data.

3) *Frequency domain editing of the transformed, time-edited impulse response.* Because the loudspeaker impulse response has been captured over a finite time window and at a finite sampling rate, the high-frequency and low-frequency measured responses are in error. Also, in designing an equalization filter the extrema of the high-frequency and low-frequency responses may be defined so that the equalizer is not required to provide excessive and inappropriate gain. A degree of modification can be performed again by on-screen editing of the displayed amplitude–frequency response.

a) *Low-frequency editing of amplitude–frequency response.* Two edit points are selected on screen (using the mouse), and a curve is computed using a quadratic approximation. Consequently the inverse Fourier transform of the impulse response now extends in length beyond the original truncated response. For example, where an impulse response is truncated in 2) to 256 samples and subsequently represented for processing by 4096 samples, the additional samples computed to match the edited frequency response are virtually noiseless.

b) *High-frequency editing of amplitude–frequency response.* A similar two-point selection editing process is performed at high frequency, allowing the selected high-frequency curve to be replaced by a quadratic approximation.

4) *Linear-phase target amplitude–frequency response.* The following four options have been included in the program for modifying the overall frequency response. They are performed only on the amplitude response. Consequently their native form results in a linear-phase response.

a) Constant gain with frequency (no additional equalization).

b) Constant slope filter (dB gain against logarithmic frequency scale) of N dB per octave (both positive and negative slopes enabled). This modification has been found to be a useful tool in matching loudspeakers to different room acoustics.

c) Linear-phase loudspeaker crossover response using a modified Butterworth amplitude alignment. Both low-pass and high-pass options are available, and crossover frequency is specified at −6-dB gain. However, the Butterworth response is modified to guarantee that high-pass filter and low-pass filter sum to a unity-amplitude composite response. (See Section 2.3 for more detail, as well as [3].)

d) Psychoacoustic weighted frequency response, available for subjective tailoring in the midrange frequency region where the ear is most sensitive. This characteristic has found favor in Germany [4].

The program allows these options to be selected individually and mixed if required.

5) *Calculation of equalizer amplitude–frequency response.* The equalizer amplitude response is calculated by inverting the product of the loudspeaker amplitude–frequency response [which has been both time-domain and frequency-domain edited, see 2) and 3)] and the target equalizer amplitude frequency response. This is performed only on the magnitude response.

6) *Calculation of minimum-phase impulse response of equalizer.* Using the Hilbert transform, the minimum-phase impulse response of the equalizer is evaluated.

7) *Digital filter design.* The equalizer's impulse response is divided into two regions within the program. The first region defines an FIR filter, whereas an IIR filter represents the second region where coefficients are calculated using either the Prony method[3] [7] or a least-mean-square (LMS) technique. The method that yields the lower error can then be selected. A general digital filter is then defined with both numerator and denominator coefficients.

8) *Excess-phase equalization of loudspeaker.* Using the Hilbert transform, the truncated loudspeaker impulse response is decomposed into minimum- and excess-phase components and then described in the time domain [3]. Using truncation and time reversal of the excess-phase impulse response, the filter coefficients required for phase correction are estimated. The phase equalizer can then be implemented either as an additional cascaded filter or convolved with the numerator coefficients derived in 7) to form a single filter.

2 PROGRAM FUNCTIONALITY

This section highlights some of the mathematical procedures used within the main program and describes the functionality in pseudo-MATLAB code.

[3] The Prony function was authored by L. Shure and is written into the MATLAB signal processing suite as a function. Details can be found in [7].

2.1 Minimum-Phase Processing

The minimum-phase signal-processing algorithm used in the program has been described previously [3]. The algorithm enables a straightforward method of computing the peak-level normalized minimum-phase impulse response eit from a magnitude-frequency response xefa (which is a symmetrical Fourier transform about $f_s/2$, where f_s is the sampling frequency), that is,

```
eit = real(ifft(exp(conj(hilbert)log(xefa))))));
eit = eit./max(abs(eit));
```

2.2 Low-Frequency and High-Frequency Editing of Amplitude–Frequency Response

This editing routine modifies the amplitude–frequency response at both low frequency and high frequency. Two options are analyzed, a cubic fit and a quadratic fit.

2.2.1 Cubic Fit Approximation

In each of these frequency regions two edit points are marked and a cubic-approximation curve is generated to replace the actual response. As shown in Fig. 1 the curve is matched in both level and first derivative by forcing a fit to two adjacent samples, $x = m$ and $x = m + 1$, to smoothly link with the amplitude response. For low frequency, the first derivative is also set to zero for sample $x = 1$, whereas at high frequency this is performed at $f_s/2$. By way of example, the procedure is shown for the low-frequency cubic fit.

Assume that the cubic curve has a value $y(1)$ at $x = 1$ and that the first derivative is zero. Also, the curve must match in level at two adjacent samples $x = m$ with a corresponding amplitude $y(m)$ and $x = m + 1$ with a corresponding amplitude $y(m + 1)$.

Now for a cubic curve,

$$y = ax^3 + bx^2 + cx + d .$$

Then

$$\frac{dy}{dx} = 3ax^2 + 2bx + c$$

where for $x = 1$, $dy/dx = 0$,

$$0 = 3a + 2b + c .$$

In addition, matching levels at $x = m$ and $x = m + 1$ generates four equations,

$$\begin{bmatrix} 0 \\ y(m+1) \\ y(m) \\ y(1) \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 & 0 \\ (m+1)^3 & (m+1)^2 & (m+1) & 1 \\ m^3 & m^2 & m & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} .$$

Hence matrix inversion allows the coefficients $[a \;\; b \;\; c \;\; d]$ to be calculated, and the cubic curve is fully defined. A similar calculation procedure can be performed at high frequency.

2.2.2 Quadratic Fit Approximation

A similar procedure can be followed for a quadratic approximation, except that there is one less degree of freedom. Thus here, as shown in Fig. 2, the derivative can be matched only at one of the extreme ends of the spectrum.

Now for a quadratic curve,

$$y = ax^2 + bx + c .$$

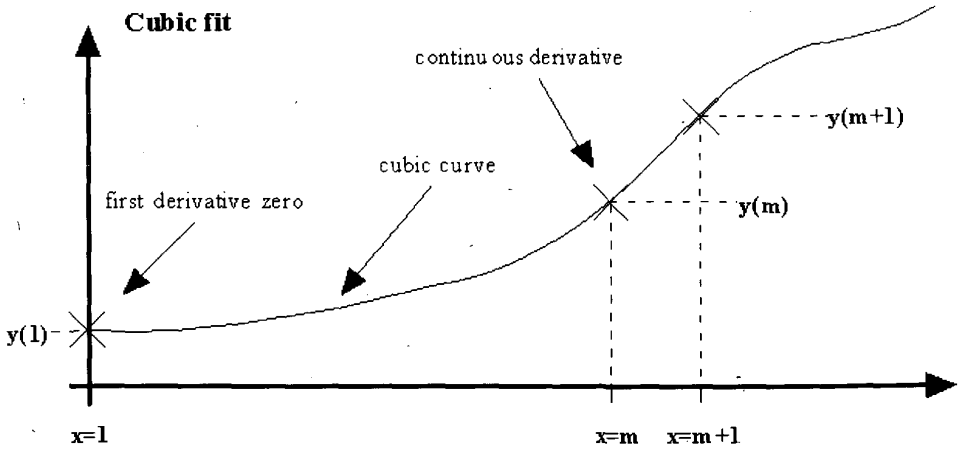


Fig. 1. Low-frequency cubic substitution (continuous derivative at $x = m$).

Then,

$$\frac{dy}{dx} = 2ax + b$$

where for $x = 1$, $dy/dx = 0$,

$$0 = 2a + b .$$

Two options now exist. Either the derivative at the extreme frequency range can be equated to zero, or the function can be matched to the derivative of the signal by again forcing a fit to two adjacent points at $x = m$ and $x = m + 1$. This gives rise to two matrix equations, which can be solved for $[a \quad b \quad c]$ by matrix inversion, that is, either for a zero first derivative at low-frequency,

$$\begin{bmatrix} 0 \\ y(m) \\ y(1) \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 \\ m^2 & m & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} .$$

or for a matching first derivative of the actual frequency response,

$$\begin{bmatrix} y(m+1) \\ y(m) \\ y(1) \end{bmatrix} = \begin{bmatrix} (m+1)^2 & (m+1) & 1 \\ m^2 & m & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} .$$

Matching at high or low frequency is similar, although there are detailed changes of variables in the equations. In practice the quadratic fit with a forced zero first derivative at the frequency extremes was found to give the best overall results and has been adopted in the current version of the program.

2.3 Pseudo-Butterworth Linear-Phase Crossover Alignment

The program includes a routine to embed either a low-pass or a high-pass filter in the target equalization characteristic aj to enable a two-way digital crossover filter to be designed and made compatible with aj.

The frequency range over which the computation is performed is defined by a frequency vector mx with an element range {1:m2}, where

```
mx(1:m2) = ml*(1:m2);
```

Here ml is the lowest frequency in the vector whereas m2*ml = $f_s/2$, where f_s is the sampling rate.

The process commences by calculating an amplitude–frequency response based on a Butterworth magnitude filter template.

The crossover frequency xof (scaled by ml to form a corresponding element number in the vector mx) and the asymptotic attenuation slope of the filter N dB per octave (N need not be an integer as there are no "analog" restrictions on the filter) are specified as input to the design. The order ord of the Butterworth filter is then determined,

```
ord = abs(N/(20*log10(2)));
```

from which the 3-dB break frequencies xofl and xofh of the low-pass and high-pass filters are calculated,

```
xofl = ml*xof/(2^(2/ord)-1)^.5;

xofh = ml*xof*(.5^(-2/ord)-1)^.5;
```

The amplitude–frequency responses of the low-pass and high-pass filters ajl and ajh based on the Butterworth magnitude frequency response then follow as

```
ajl = aj./(1+(mx/xofl).^2).^(ord/2);

ajh = aj.*(((mx/xofh).^2)./(1+(mx/xofh).^2)).^(ord/2);
```

However, the composite zero-phase summation (ajl + ajh) does not generally sum to the target response aj, which itself may not be flat due to other frequency shaping characteristics selected in the program. Consequently a symmetrical (about the crossover frequency

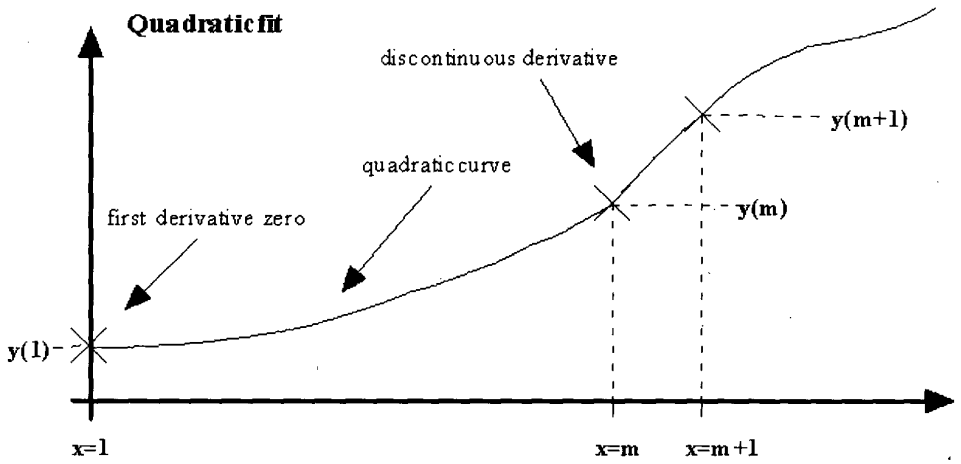


Fig. 2. Low-frequency quadratic substitution (discontinuous derivative at $x = m$).

xof) compensation to ajl and ajh must be made.

For ajl, the response is modified in the frequency range 1 to xof,

```
ajl(1:xof) = ajl(1:xof) + aj(1:xof) - ajl(1:xof
             - ajh(1:xof);
```

while for ajh, the response is modified in the frequency range xof to m2,

```
ajh(xof:m2) = ajh(xof:m2) + aj(xof:m2)
              - ajl(xof:m2) - ajh(xof:m2);
```

Now when the summation (ajl + ajh) is formed over {1:m2}, the composite amplitude response is aj, which is the target response of the equalizer.

2.4 Tilt Filter

A simple frequency tilt filter is included where the target equalization response aj can be tilted at a constant slope of *Nx* dB per octave (as described in a logarithmic space). The response is normalized to unity gain at the geometric frequency mean fj and computed as

```
aj = (mx/fj). ^(Nx/(20*log10(2)));
```

2.5 Psychoacoustic Equalizer

A psychoacoustic filter was described earlier [3], [4]. The filter adds a mild reduction in the midband frequency region where the ear is most sensitive. This can prove a subjectively useful modification for some (poorer) recordings. This filter is included in the present program as an option and is defined by the routine

```
np = ml*fix(3000/ml);

ajj = [-2.5*log10((ml:ml:np)/3000) 6.4*log10(np + ml:ml:mh)/3000)];

ajj = ajj - (ajj(m2) + min(ajj))/2;

ajj = (10*ones(size(1:m2))). ^(ajj/20);

aj = aj.*ajj;
```

where again aj is the target amplitude–frequency response.

2.6 Two-Stage FIR–IIR Filter Using Prony's Method

Once the amplitude–frequency responses of the target aj and of the loudspeaker c are determined, the exact equalizer amplitude–frequency response xefa can be calculated,

```
xefa = aj./c;
```

from which the normalized minimum-phase equalizer impulse response eit follows,

```
eit = real(ifft(exp(conj(hilbert(log(xefa))))));

eit = eit./max(abs(eit));
```

The impulse response eit can form a digital filter directly, although in general it represents an excessive number of coefficients. In the present program the impulse response is subdivided into two regions. The first region of length en forms the first-stage FIR filter directly, whereas the remainder of the impulse response is represented by a second-stage FIR–IIR filter and is designed using either the Prony method or an LMS method. Both processes are supported in MATLAB. Intuitively this appears a logical subdivision as the FIR response carries much of the fine detail required of the equalizer. However, the low-level tail of the impulse response is primarily a result of bandwidth constraints such as the natural low-frequency rolloff of a loudspeaker. As such it may be surmised to have a simpler form more amenable to a low coefficient representation. However, because en can be user selected together with the second-stage filter numerator coefficients nx and denominator coefficients nd required to represent the tail, a wide range of filters can be chosen extending from pure FIR to pure IIR. As the second-stage filter is an FIR–IIR form, the total number of numerator coefficients in the overall filter is (en + nx). However, it turns out that when the tail response is calculated, it is not generally minimum phase. Consequently additional numerator coefficients nx can be specified. However, the number of coefficients in each of the three categories remains under user control, and through the use of time and frequency displays the impact of a given design selection can be observed. Also, after en is specified, the program can be selected to interrogate the tail response and to recalculate en so that the highest value in the initially selected tail impulse becomes the new first sample of the tail. The stage-one filter length en is then reassigned. An autoselect procedure can be selected to facilitate this process. However, if only an FIR filter is required, then en remains fixed and a short raised-cosine window is applied to the end of the FIR filter.

The tail response is selected according to

```
tail = eit(en + 1:m2);
```

to which both a Prony [7] and an LMS filter design procedure (the latter using the MATLAB invfreq function[4]) are applied to determine an FIR–IIR filter with nx numerator and nd denominator coefficients. In the LMS procedure an option is included to both bandlimit and power-weight the tail spectrum, where appropriate parameters can be user selected. The MATLAB functions have a form

```
[tnp tdp] = prony(tail(1:tmx),nx,nd);

[tnl tdl] = invfreqz(fft(tail) ,nx - 1,nd,);
```

[4] Written as a function in MATLAB, coauthored by J. N. Little, J. O. Smith, Lennart Ljung, and T. Krauss.

After selecting the desired tail approximation using a normalized LMS error calculation, the FIR and IIR filter sections can be spliced to form a complete filter. In the program up to three vectors are outputted to describe the overall filter. The first filter $N_1(z)$ of length en represents the first-stage FIR filter while the second recursive filter has a numerator polynomial $N_2(z)$ of length nx and a denominator polynomial $D_2(z)$ of length nd (the first coefficient being unity and not included in nd). The complete z-domain equalizer response EQ(z) is then calculated as

$$\mathrm{EQ}(z) = N_1(z) + \frac{N_2(z)}{D_2(z)} z^{-(\mathrm{en}+1)}$$

where a tap delay of length {en + 1} is included in the second-stage filter to locate the approximated tail in its correct position within the overall impulse response.

2.7 CDS Display

The generation of a CDS has been described [3] in MATLAB where either a linear or a logarithmic magnitude display can be formed. The option of applying a two-dimensional Gaussian filter mask is also included where fine detail may require smoothing for presentational reasons. In the program, options exist for including minimum-phase processing based on the algorithm described in Section 2.1. This is a useful tool as it gives often a better description of a loudspeaker's performance stripped of the excess-phase distortion which generally has lower subjective significance yet may introduce dominant features in the impulse response displayed in the CDS. The CDS is described by a two-dimensional matrix cd, where

```
cd = 20*log10(abs(fft(hankel(e(1:cm)))));
```

The MATLAB "mesh" function then translates this matrix into a corresponding three-dimensional display forming the CDS.

2.8 CDR Display

The CDR [3] is used here to display the equalizer impulse response ei. First the roots rt of ei are calculated and sorted into rank order. Then they are edited and converted into a corresponding magnitude–frequency response. Finally minimum-phase impulse responses are calculated from the set of magnitude responses and converted into corresponding ETC (see Section 2.9), from which a two-dimensional matrix is assembled to describe the CDR. During processing a frequency response approximation is displayed to check the validity of the CDR after the roots are sorted and edited.

2.9 ETC Display with Excess-Phase Response Correction

The ETC et is calculated directly from the magnitude envelope of the Hilbert transform of an impulse response function a, where

```
et = abs(hilbert(a));
```

However, in the program the practice reported earlier [3] of eliminating the noncausal distortion by minimum-phase processing is taken. That is, a vector is formed representing the magnitude Fourier transform of et,

```
fa = abs(fft(abs(et)));
```

from which a minimum-phase impulse response is computed which has the same amplitude spectrum as fa,

```
ett = real(ifft(exp(conj(hilbert(log(fa))))));
```

The option for displaying the ETC and the minimum-phase ETC is provided together with an excess-phase corrected ETC [3], where the magnitude of ett is convolved with the excess-phase impulse response f, that is,

```
ete = abs(conv(abs(ett),f));
```

The function ete then has a similar precursive response to the envelope of the impulse response by including both minimum-phase and excess-phase attributes.

3 FILTER DESIGN EXAMPLES FOR LOUDSPEAKERS

This section presents example filter designs for two classes of loudspeaker, the broad-band passive system and the digital and active loudspeaker system [5]. Discussion as to the advantages of digital and active loudspeaker technology is also included.

3.1 Broad-Band Equalization Filters for Full-Range Passive Loudspeakers

This section compares two filter examples with the results are shown in Figs. 3–12. The (a) parts are for a small two-way conventional drive unit loudspeaker whereas the (b) parts are for a single DML. All curves were generated from within the MATLAB design program and for comparison 150 coefficients are force selected in the first FIR stage with 50 denominator and 30 numerator coefficients in the second-stage tail filter. The same filter size was used for both two-way and DML loudspeakers.

The target response for filter set (a) includes a tilt of 0.5 dB per octave, whereas in set (b) it is flat. For the DML its native response persists for well over 1000 samples, including diffuse room reflections. However, this was truncated to about 250 samples for processing, which is not strictly representative of this class of loudspeaker. For proper analysis of a DML, large-space measurements are required to extend the first reflection arrival time together with the option for spatial averaging data derived from the long impulse response and spatially diffuse character of these transducers.

3.2 Digital Equalization Filters Including Crossover Alignment

A two-way digital and active system is shown in Fig. 13. Here the digital-to-analog converters (DACs) are

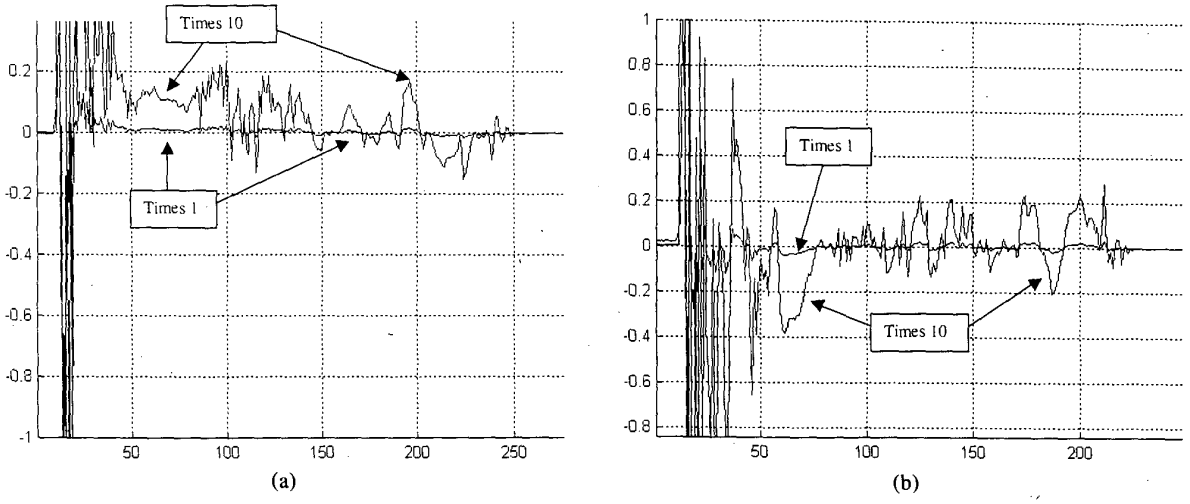


Fig. 3. Time-edited impulse response derived using MLS measurement. (a) Two-way dynamic loudspeaker. (b) DML.

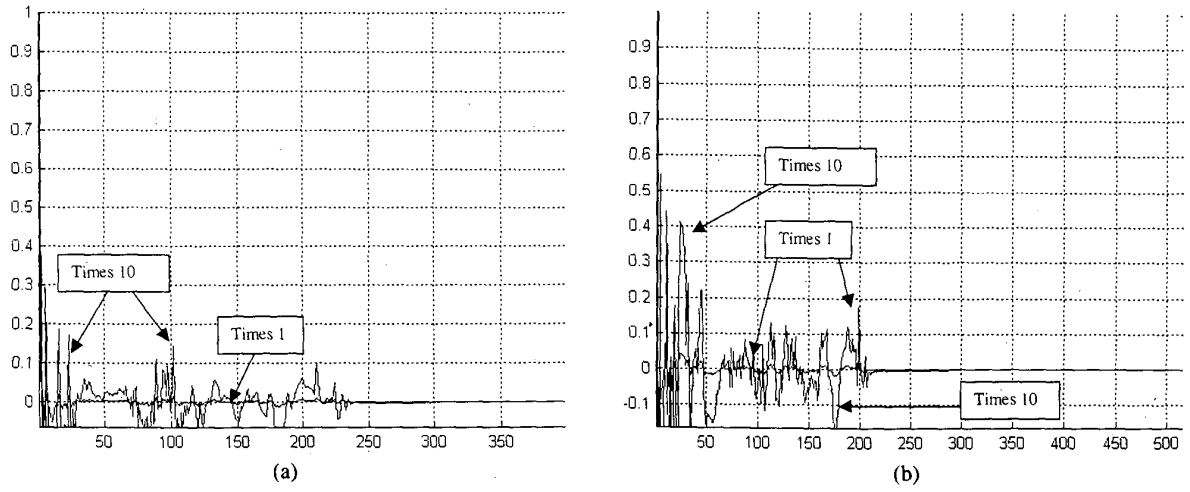


Fig. 4. Minimum-phase impulse response. (a) Two-way dynamic loudspeaker. (b) DML.

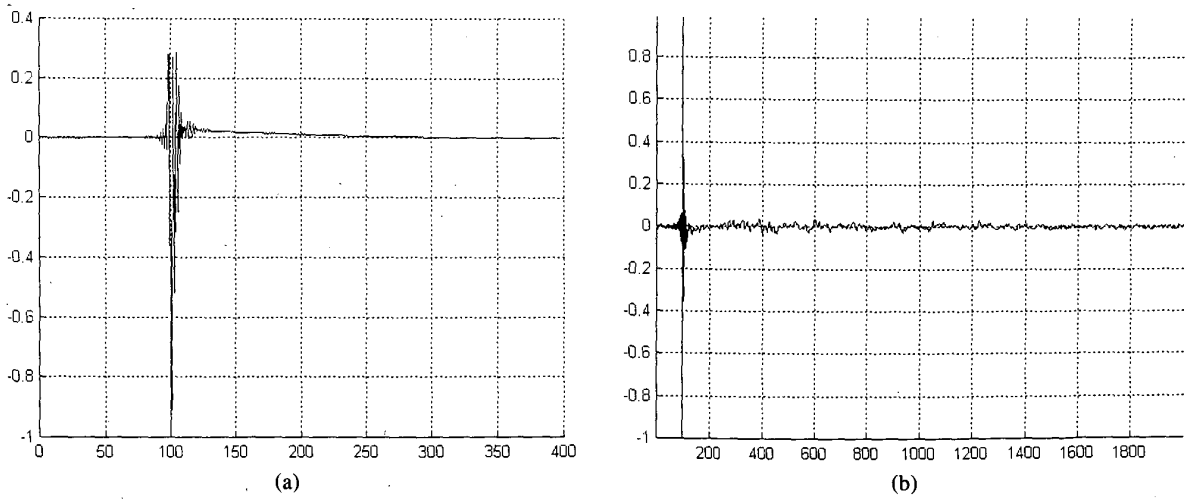
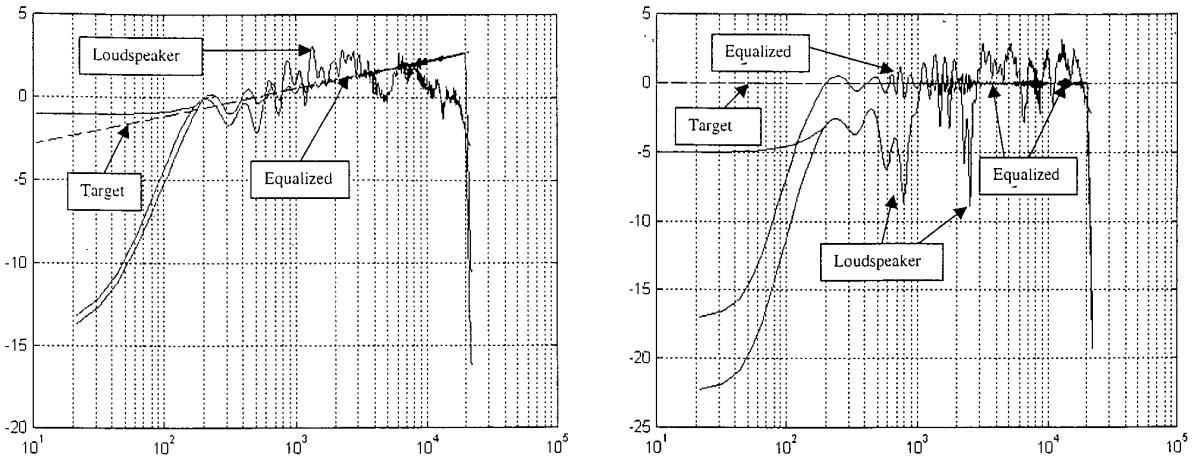


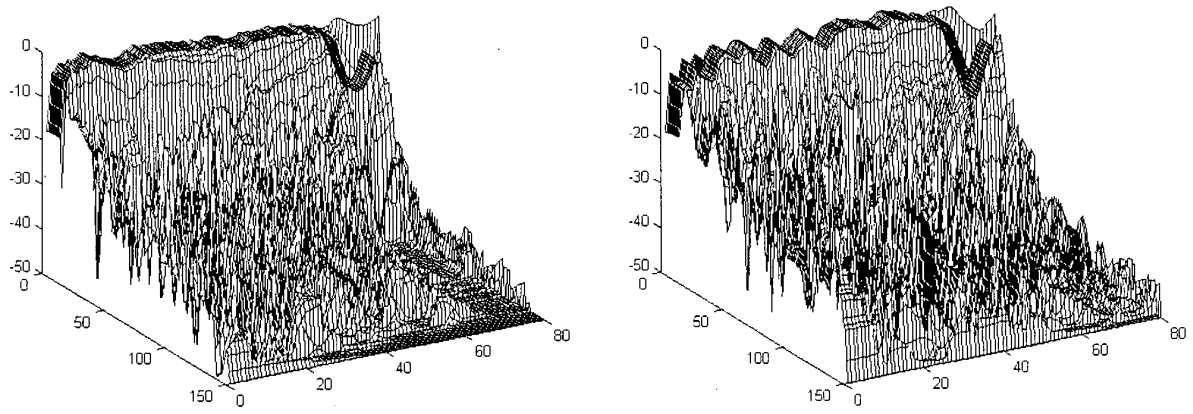
Fig. 5. Excess-phase impulse response. (a) Two-way dynamic loudspeaker. (b) DML.

interfaced directly to the power amplifiers, resulting in a welcome reduction of analog signal processing. Using digital signal processing (DSP) it is straightforward to integrate near-ideal crossover filters together with accurate drive unit equalization. Although analog crossover filters can be synthesized, a more precise strategy is to define the low-pass filter $A_L(z)$ and the high-pass filter $A_H(z)$ as delay derived, as this eliminates the nonlinear



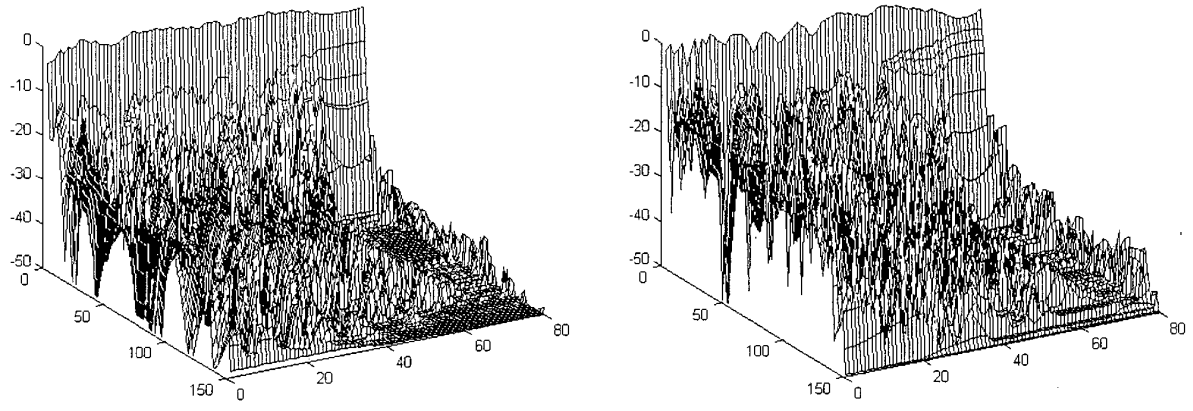
(a) (b)

Fig. 6. Target, loudspeaker, and equalized frequency responses. (a) Two-way dynamic loudspeaker. (b) DML.



(a) (b)

Fig. 7. CDS. (a) Two-way dynamic loudspeaker. (b) DML.



(a) (b)

Fig. 8. Minimum-phase CDS. (a) Two-way dynamic loudspeaker. (b) DML.

phase distortion of an all-pass crossover alignment, where

$$A_L(z) + A_H(z) = z^{-\lambda}$$

where $z^{-\lambda}$ is an appropriate delay.

Section 2.3 described a pseudo-Butterworth design that matches this criterion and is included within the filter design program as an integral part of the target function generator.

Some performance and system attributes of a digital and active loudspeaker are summarized as follows.

1) DSP enables the crossover target transfer functions, including amplitude and phase compensation of each drive unit together with enclosure characteristics, to be specified to a high degree of accuracy. Thus the polar response can be optimized and the overall on-axis frequency response can approach a constant-amplitude, linear-phase characteristic.

2) Crossover filters can include options for either sharp frequency transition bands or more gentle frequency response slopes, including all-pass analog alignments. The new class of interleave crossover alignments can be incorporated [3]. Multiple filters offer guaranteed synchronization and exact replication.

3) Delay compensation to align on-axis acoustic centers of drive units is straightforward to implement using a memory-based digital delay.

4) Simple adjustment for individual drive unit sensitivities without wasting power in passive elements of a crossover, thus enabling amplifier power to be used efficiently. Also, use of digital gain control (with dither) guarantees exact tracking of each filter channel.

5) Because individual power amplifiers are directly coupled to each drive unit, there is a corresponding division of signal power. This reduces the voltage and current demand placed on an individual amplifier, where usually an individual drive unit has a benign terminal impedance compared with passive systems with complicated crossovers.

6) Momentary clipping of a single power amplifier can be softened by preprocessing and has only a localized, in the sense of frequency, impact on subjective performance. Hence the system is more overload tolerant.

7) Close coupling of amplifiers and drive units (compared to passive systems) results in tighter control of the speech coil and lower dependence on nonlinearity in the drive unit impedance.

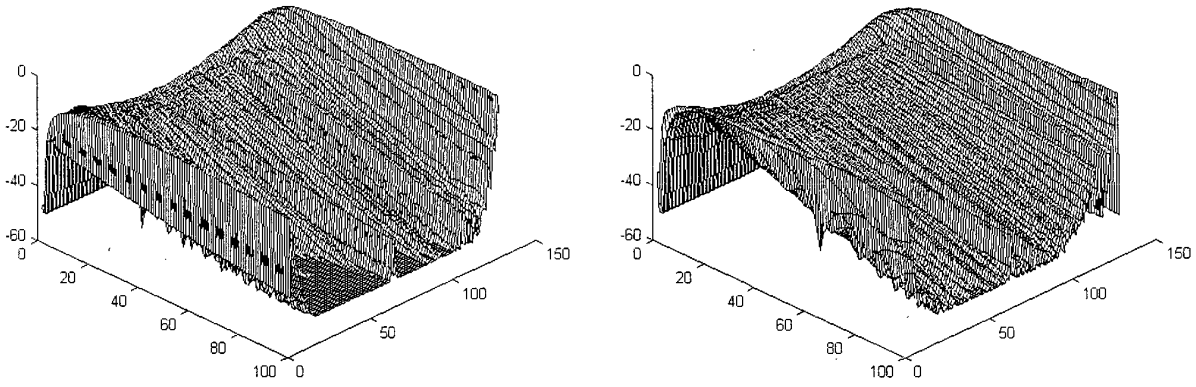


Fig. 9. CDR of equalization filter. (a) Two-way dynamic loudspeaker. (b) DML.

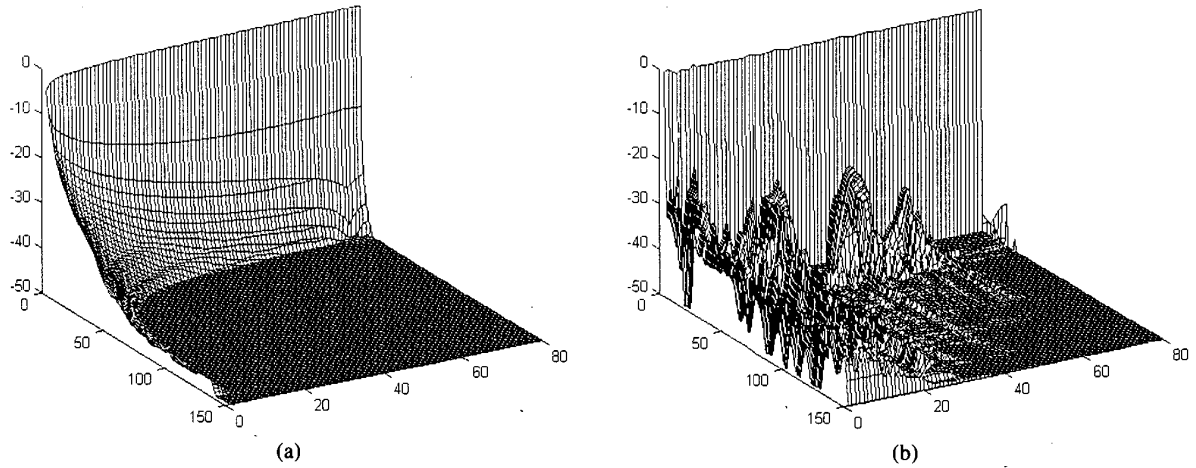


Fig. 10. CDS, minimum-phase correction only. (a) Equalized two-way dynamic loudspeaker. (b) Equalized DML.

8) Use of current-drive technology or mixed current-drive–voltage-drive systems can enhance system performance.

9) Each DAC only handles a band-limited audio signal, thus reducing intermodulation distortion and lowering the probability overload in drive-unit frequency-response correction processes.

10) Signals can be routed to an active loudspeaker system via either an optical or an electrical digital interface with central commands such as standby mode, volume level, equalization programs being remotely downloaded from a central (or indeed distributed) control center. There is the option here to define a local-area network for digital or video distribution systems.

In Fig. 14 an example crossover target filter design is presented based on the pseudo-Butterworth algorithm of Section 2.3, whereas Figs. 15 and 16 show the corresponding low-pass and high-pass filter responses, including drive-unit response equalization. Each filter used a total of 285 coefficients.

Finally, the new class of interleave-crossover alignment [3] is illustrated by way of example in Fig. 17, which is also incorporated in the program. The purpose of this alignment is to distribute and randomize in frequency the interference patterns that result in the off-axis polar response [8]. Both low-pass and high-pass filters for a stochastic-interleave function are shown that includes a mild equalization tilt function. Fig. 18 shows the resulting composite frequency response when there is a time offset between drive units. A complicated interference spectrum is formed in the crossover region that is in some respects analogous to the behavior of a DML [6]. Also, by comparing the inverted connection composite responses shown in the top curves of Figs. 14 and

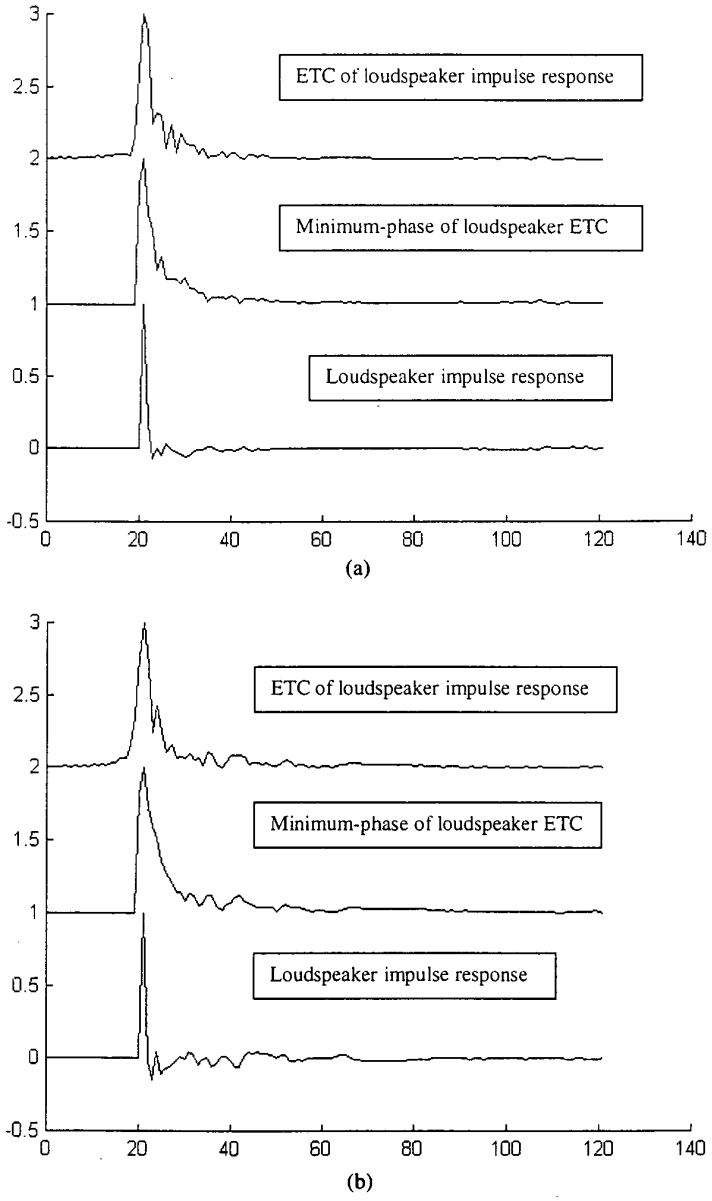
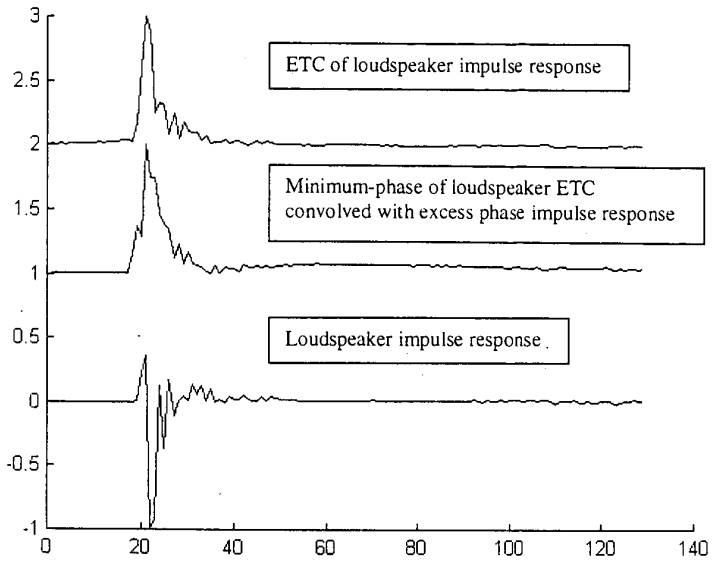
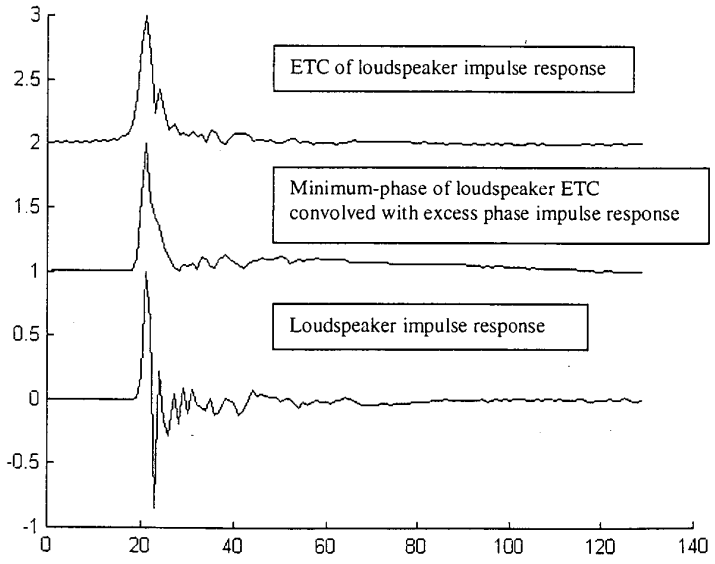


Fig. 11. Minimum-phase ETC. (a) Two-way dynamic loudspeaker. (b) DML.



(a)



(b)

Fig. 12. Excess-phase corrected ETC. (a) Two-way dynamic loudspeaker. (b) DML.

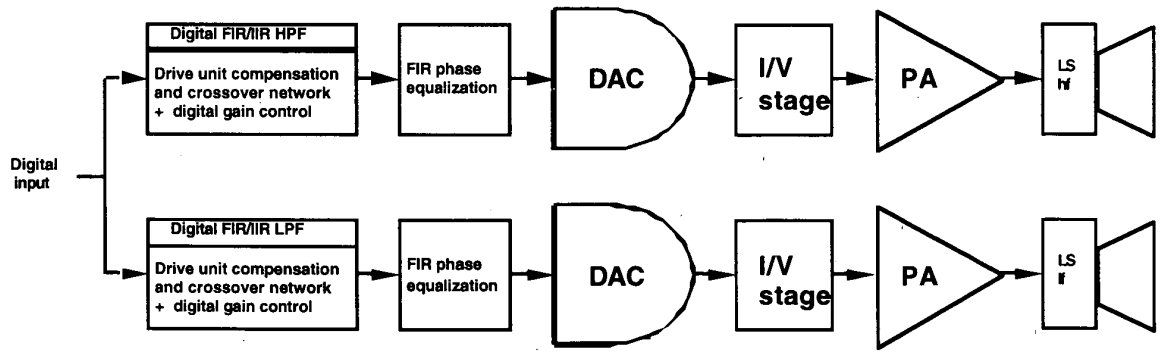


Fig. 13. Two-way loudspeaker system using digital crossover filters with amplitude and phase correction.

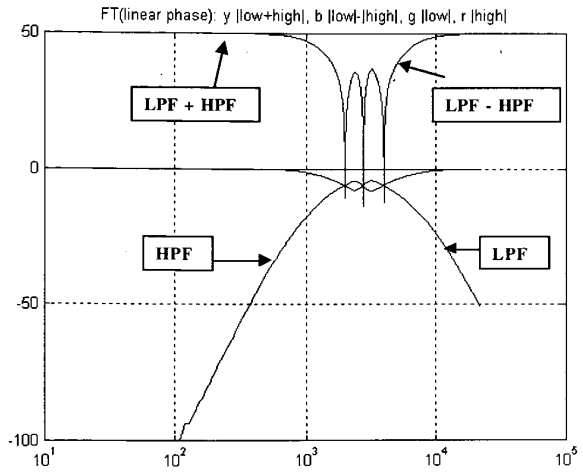


Fig. 14. Low-pass and high-pass pseudo-Butterworth target responses.

17 (with and without interleave function, respectively), the effect of a stochastic interleave function is dramatically illustrated. In studying these results it should be recalled [3] that the interference patterns are bounded by the envelopes $|A_L + A_H|$ and $|A_L - A_H|$.

4 CONCLUSION

A MATLAB program to facilitate the design of loudspeaker equalization and crossover FIR–IIR filters has been described. Two equalization examples were presented without crossover filters and one example with a crossover filter together with examples of computed output data. In particular, the minimum-phase and excess-phase data of a DML reveal interesting detail about this class of drive unit. The main differences observed are in the impulse response, which has an extended duration and "noiselike" character. This makes equalization less straightforward. Because of the spatially diffuse form of a DML's frequency response, it is better to perform

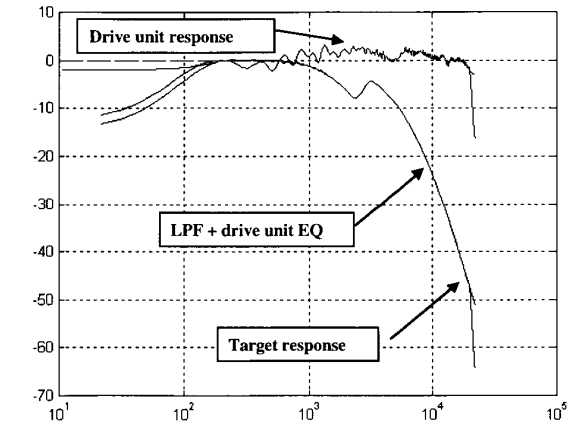


Fig. 15. Low-pass filter response including drive-unit equalization.

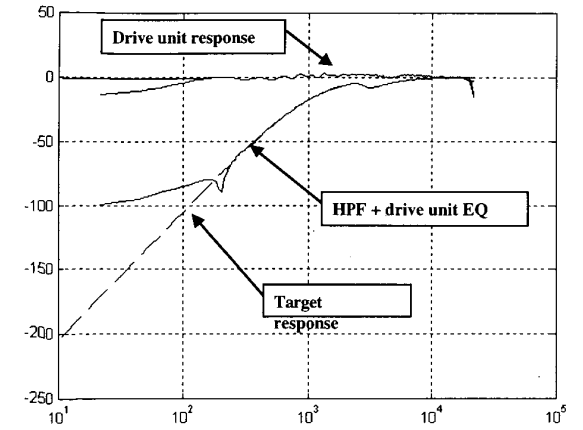


Fig. 16. High-pass filter response including drive-unit equalization.

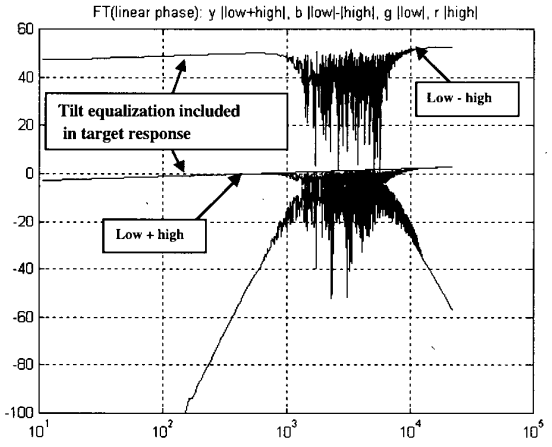


Fig. 17. Example target response of stochastic interleave alignment. (Compare with noninterleaved response of Fig. 14.)

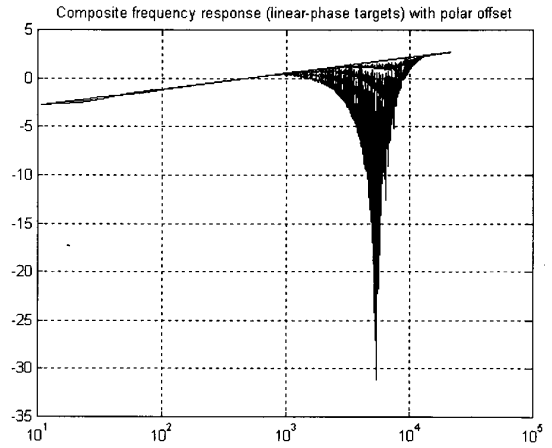


Fig. 18. Family of four composite responses time offsets between drive units.

equalization to a spatially averaged frequency response. In practice it is recommended that the spatial average be performed using measurements taken over a sampled hemisphere. Failing this a truncated impulse can be used together with a more limited number of coefficients in the equalization filter. It is important to appreciate that spatial averaging is an intrinsic requirement of DML measurement, especially when considering equalization, as these transducers offer an almost omnidirectional polar response together with diffuse properties. (One could argue that this new class of polar response is omnidiffuse.) However, the contrary is true for most conventional loudspeakers, where these generally show much wider variations in frequency response over a measurement sphere. Consequently any spatial averaging windows have to be applied over a more restricted measurement area with corresponding limitations on listening postion and associated room interaction problems. Experience suggests that the on-axis response forms the best data in well-designed, conventional loudspeakers.

The method of filter design has proved efficient in terms of computational time, and using an IIR implementation to describe the impulse response tail would appear intuitively to offer advantage. However, in practice, comparing a pure FIR design with a pure IIR design showed little overall advantage where for an N-coefficient filter a good fit to the minimum-phase impulse response is achieved over about N samples in each case. It appears that the total number of coefficients, irrespective of their distribution between the two filter sections, is the more critical factor. As reported in earlier work [1], [2], the results demonstrate that by using a digital equalization filter, extremely accurate system performance is possible, and that a wide range of target frequency responses can be easily accommodated.

The two loudspeaker examples selected attempt to highlight some performance differences between a specular radiator and a DML. In particular, the DML minimum-phase impulse response showed good form, although the excess-phase impulse response revealed an extended noiselike structure in addition to a well-focused central response, confirming its temporally diffuse nature. It is therefore suggested that DMLs probably exhibit low spatial variation in both the initial period of the minimum-phase impulse response and the central period of the excess-phase impulse response (where these responses also show low noise structure). However, a much greater intersample variation in the pre- and postresponse of the excess-phase impulse response would be anticipated as the measured window encircles the transducer. This is implied by the transducer being spatially diffuse. It is in these areas where fundamental performance differences between specular radiators and DMLs are observed, from which a number of attributes can be deduced, for example, those relating to imaging when multiple arrays are employed.

The embedded crossover design function was also demonstrated and a brief discussion given on the merits of digital and active loudspeaker systems. This was used to illustrate the stochastic interleave-crossover alignment which, it is suggested, disperses the interference patterns in the off-axis frequency response, thus lowering polar-related coloration in the crossover region.

5 REFERENCES

[1] R. Greenfield and M. O. J. Hawksford, "Efficient Filter Design for Loudspeaker Equalization," *J. Audio Eng. Soc.*, vol. 39, pp. 739–751 (1991 Oct.).

[2] M. O. J. Hawksford and R. G. Greenfield, "A Comparative Study of FIR and IIR Digital Equalization Techniques for Loudspeaker Systems," *Proc. Inst. Acoust.*, vol. 12, pt. 8, pp. 77–86 (1990).

[3] M. O. J. Hawksford, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," *J. Audio Eng. Soc.*, vol. 45, pp. 37–62 (1997 Jan./Feb.), Correction, ibid., vol. 45, p. 497) (1997 June).

[4] B. Stark, "Ein Fazit," in *Das Lautsprecher Jahrbuch 86/87*, M. Gaedke, Ed. (Hifisound, Münster, Germany, 1986).

[5] M. O. J. Hawksford, "Digital and Active Loudspeaker Systems for High-Quality Monitoring," in *Proc. Active 95* (1995 Int. Symp. on Active Control of Sound and Vibration, Newport Beach, CA, 1995 July 6–8), pp. 1247–1258.

[6] "NXT," white paper, Huntingdon, UK.

[7] T. W. Parks and C. S. Burrus, *Digital Filter Design* (Wiley, New York, 1987), p. 226.

[8] A. Rimell and M. O. J. Hawksford, "Reduction of Loudspeaker Polar Response Aberrations through the Application of Psychoacoustic Error Concealment," *IEE Proc. Vision Image Signal Process.*, vol. 145, pp. 11–18 (1998 Feb.).

THE AUTHOR

Malcolm Hawksford is director of the Centre for Audio Research and Engineering, a professor in the Department of Electronic Systems Engineering at the University of Essex, and Postgraduate Scheme director, where his interests encompass audio engineering, electronic circuit design, and signal processing. Professor Hawksford studied electrical engineering at the University of Aston in Birmingham where he gained a First Class Honours B.Sc. and Ph.D. His Ph.D. program, which was sponsored by a BBC Research Scholarship, investigated delta modulation and delta-sigma modulation (now commonly known as "bitstream" coding) for color television and the development of a time-compression/time-multiplex system for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

While at Essex University, he has undertaken research principally in the fields of analog amplifiers, digital signal processing, and loudspeaker systems. Since 1982 research on digital crossover networks and equalization

for loudspeakers has culminated in an advanced digital and active loudspeaker system being designed within the university. This was one of the first systems of its type and in 1986 a prototype was demonstrated at the Canon Research Centre in Tokyo. Research topics have also encompassed oversampling and noise shaping techniques applied to analog-to-digital and digital-to-analog conversion, the linearization of PWM encoders, and 3-dimensional spatial audio and telepresence including multichannel sound reproduction.

Since 1971 and throughout his career at Essex University, Professor Hawksford has lectured at university level in electronics and in particular the undergraduate scheme in audio engineering. He has supervised numerous Ph.D. and M.Sc. students, many of whom are now employed within the audio industry. Malcolm is particularly proud of the fact that with his encouragement, two former research students have established the Singapore Section of the AES.

Professor Hawksford has published in the *Journal of the Audio Engineering Society* and at the Society's conventions on topics that include error correction in amplifiers, oversampling techniques, jitter and MLS techniques, and loudspeaker crossover systems. His supplementary activities include writing contributions for *Hi-Fi News and Record Review* (a magazine at which he is a technical adviser) and *Stereophile* magazine as well as designing high-end analog and digital audio equipment. He is a chartered engineer and is a fellow of the AES, the Institution of Electrical Engineers, and the Institute of Acoustics. Professor Hawksford is a member of the technical committee of Acoustic Renaissance for Audio (ARA), a group that has been instrumental in promoting multichannel, high-definition audio signals on high-capacity DVD optical disks.

PAPERS

Smart Digital Loudspeaker Arrays*

M. O. J. HAWKSFORD, *AES Fellow*

Centre for Audio Research and Engineering, University of Essex, Colchester, CO4 3SQ, UK

A theory of smart loudspeaker arrays is described where a modified Fourier technique yields complex filter coefficients to determine the broad-band radiation characteristics of a uniform array of micro drive units. Beamwidth and direction are individually programmable over a 180° arc, where multiple agile and steerable beams carrying dissimilar signals can be accommodated. A novel method of stochastic filter design is also presented, which endows the directional array with diffuse radiation properties.

0 INTRODUCTION

This paper considers from a theoretical stance the fundamental requirements of a programmable polar response, digital loudspeaker array, or smart digital loudspeaker array (SDLA), which consists of either one-dimensional or a two-dimensional array of micro radiating elements. The principal problem addressed here is the design of a set of digital filters which together with a uniform array of small drive units, achieve a well-defined directional beam that can both be steered over a 180° arc and be specified in terms of beamwidth such that it remains constant with the steering angle. Intrinsic and critical to the SDLA is the requirement that the beam parameters remain stable over a broad frequency range. In addition to addressing the problem of coherent radiation, the theory is extended to include the synthesis of directionally controllable diffuse radiation that is similar although not identical to the class of sound field produced by a distributed-mode loudspeaker (DML) [1], [2]. DML behavior can be emulated using an array of discrete radiating elements with excitation signals calculated to model panel surface wave propagation and boundary reflections using techniques such as finite element vibration analysis [3]. However, for the SDLA a different approach is taken where each element drive signal is derived by convolution of the input signal with an element-specific but stochastically independent temporally diffuse impulse response (TDI) [4]. Each TDI is calculated to have a constant-magnitude response but a unique random-phase response, where for loudspeaker applications it is formed asymmetrically to have a rapid initial response and a decaying "tail" exhibiting a noise-like character.

The conceptual structure of an SDLA is shown in Fig. 1, where each mircodriver within the array is addressed directly by a digital signal that has been filtered adaptively to allow the polar response to be specified and controlled dynamically. It is proposed to configure the transducer array with a large number of nominally identical acoustic radiating elements, where the overall array size and interelement spacing (referred to here as interspacing)

Presented at the 110th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2001 May 12–15; revised 2003 September 29. This study was undertaken for NXT Transducers plc, UK.

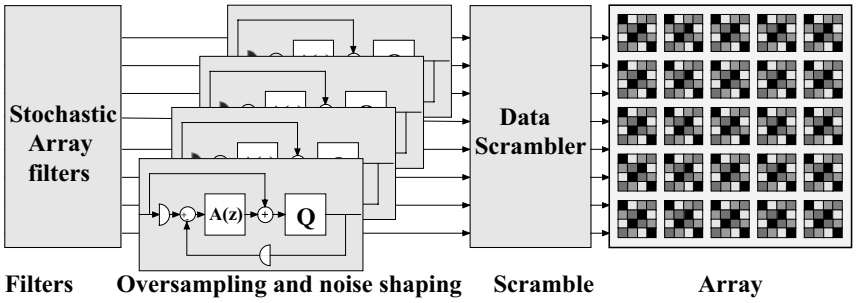


Fig. 1. Conceptual model of smart digital loudspeaker array (SDLA).

determines the usable bandwidth over which the polar response can be controlled. The size of each element must be sufficiently small so as to launch a hemispherical wavefront within the audio band up to the highest operating frequency. However, it is suggested that each element in the array could itself be a set or microarray of, say, 16 or possibly 64 microradiators. Each such microelement within the set could then carry equal acoustic weight and be driven from the output of an individual thermometer-style quantizer [5], [6] that is embedded within a noise-shaping loop. The reason for choosing the thermometer structure is that the multiple binary outputs build uniformly and progressively in discrete steps and can drive the individual elements since they also carry equal weight in the conversion. Also, because of the equal weights the connections between the thermometer DACs and the microradiators can be scrambled dynamically to decorrelate systematic errors in the reconstructed acoustic output using techniques similar to those developed by Adams and coworkers [5], [6]. As such, this structure constitutes a multilevel digitally addressed transducer, where using upsampling and multiple distributed noise-shaping loops enables in principle the required acoustic-signal resolution to be achieved. Possible suggested implementations for the elemental radiators could exploit ceramic piezoelectric technology,[1] the technology behind micromirror video projection [7], or, alternatively, conventional miniature moving-coil (MC) drive units could be used. However, the principal objective of this paper is to establish a framework for controlling acoustic radiation that not only enables the beam characteristics in terms of angle and width to be specified but also its directional correlation function, hence diffuse characterization.

An original contribution of the paper is the introduction of stochastic filters within the array to control the polar response while simultaneously achieving diffuse acoustic radiation. For this reason reference to the DML must be made as this class of loudspeaker is characterized by diffuse radiation [1], [4], although unlike the system under discussion here, the polar response is normally nondirectional. Preceding the multiple noise-shaping loops, a filter bank containing stochastically derived, frequency-dependent coefficients controls both the array polar shape and its direction, embeds diffuse sound-field characterization, and achieves a much more even distribution of power across the array elements. An even power distribution is important, especially when a narrow-beamwidth polar response is formed, where with conventional coherent arrays the power tends to cluster toward only a limited number of elements, which consequently limits the array power output.

An SDLA requires a number of technological developments mainly in the fabrication of large arrays with complicated microradiating structures together with extensive signal-processing circuits to perform the filtering and noise shaping on a very large scale. Such a technology potentially offers a new class of loudspeaker with smart control of its directional characteristics. It could, for example, create multiple and dynamically steerable beams from a single array which adapt to the environment, track individual targets for message delivery, or create a new method of large-scale multimedia presentation, possibly conveying different audio information in different areas of the reproduction space. Applications in large-scale immersive virtual reality are also conceivable together with possible methods for achieving large-venue three-dimensional sound reproduction.

ABBREVIATIONS USED

ADC	analog-to-digital conversion
β_x	angle beam x makes with the normal, rad
DAC	digital-to-analog conversion
DETn	digital elemental transducer with n-bit resolution
DML	distributed-mode loudspeaker
ECTF	element-channel transfer function
FIR	finite-impulse response
LPCM	linear pulse-code modulation
L_x	width of beam x, rad
MC	moving-coil (loudspeaker)
PWM	pulse-width modulation
SDLA	smart digital loudspeaker array
SDM	sigma–delta modulation
TDI	temporally diffuse impulse response

1 DIGITAL TRANSDUCER INCORPORATING UPSAMPLING AND NOISE-SHAPING PROCESSING

In this section a discrete array of digital elemental transducers DETn is introduced, where each element has an n-bit amplitude resolution rather than 1-bit resolution. Two conceptual examples of elemental transducer arrays are shown in Figs. 2 and 3, where for illustrative purpose, the construction is presented as a hybrid integrated circuit. Each integrated circuit could house a subarray of the complete loudspeaker array that can be tiled into a two-dimensional structure to form any required array size. Associated digital signal processing could then be integrated within the back plate to facilitate a modular and expandable construction and retain short path lengths for critical signals.

Earlier elemental transducers have incorporated only binary activation while MC drive units with multitap voice coils [8], [9] have also been described, offering the advantage of multilevel signals and a coherent sound source that radiates from a conventional cone. However, it is conjectured here that digital transducer elements should be capable of more than two levels to achieve closer synergy with linear pulse-code modulation (LPCM) and especially to reduce high-frequency noise in any noise-shaping scheme used to enhance the audio-band dynamic range. This implies that the radiating element must either be capable of a range of linear displacements or use an area-modulation technique where each microarea of the element has a binary weight. However, inevitably such elements will have a limited digital dynamic range and therefore require noise shaping to

[1]"Helimorph," a helically wound PZT ceramic actuator, see www.1limited.com for a description.

extend the final output dynamic range to meet audio requirements. In the limit a binary element could be used with a serial digital code produced, for example, with sigma–delta modulation (SDM) [11]. Alternatively a multiarea device would accommodate limited-resolution LPCM and offer the advantage of reduced high-frequency noise.

Noise shaping has been researched in depth for a range of applications, which, include analog-to-digital conversion (ADC), digital-to-analog conversion (DAC), pulse-width modulation (PWM), and signal requantization. Also, it is well established that noise shaping with uniform quantization and optimal dither facilitates an exchange between amplitude resolution and sample rate [11]–[14] while linear performance is retained. By way of illustration, a sample amplitude resolution converter is shown in Fig. 4, which uses a high-order noise shaper, possibly with perceptual weighting, presented in an SDM configuration. The use of noise shaping is important in this application because the class of transducer being described is only able to support a digital dynamic range well below that required for high-quality audio applications. It is assumed here that the source information with a sampling rate of f_s Hz is LPCM and that the upsampled rate is Rf_s Hz, where typically $R >> 1$. A desirable characteristic of the noise-shaper topology shown in Fig. 4 is that its signal transfer function is unity. This is achieved here by including a feedforward path applied directly to the input of the quantizer [15, ref. to path *x*] and also by delaying the main input by one sample period in order to compensate for the unit sample delay required in the feedback path. This process is demonstrated in the following analysis.

The output sequence $O(z)$ is expressed in terms of the input sequence $I(z)$, the forward filter transfer function

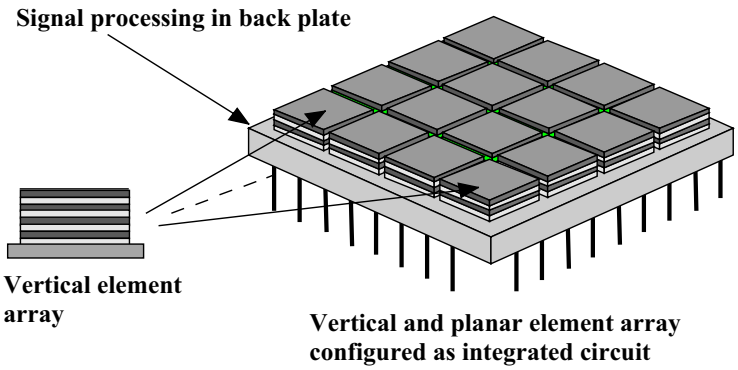


Fig. 2. Planar elemental array of multilevel DET*n*.

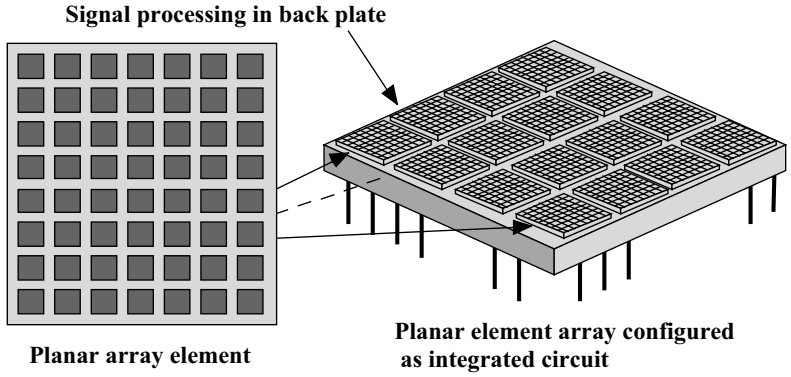


Fig. 3. Planar elemental array of planar segmented DET*n*.

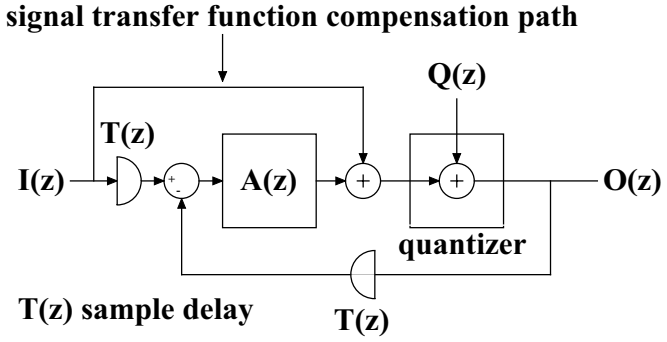


Fig. 4. Multilevel sigma–delta modulator noise shaper.

$A(z)$, the quantization noise $Q(z)$ dither, and delay $T(z)$ as

$$O(z) = I(z) + Q(z) + \text{dither} + A(z)\left[I(z)T(z) - O(z)T(z)\right].$$

Rearranging and substituting $T(z) = z^{-1}$,

$$O(z) = I(z) + \frac{Q(z) + \text{dither}}{1 + z^{-1}A(z)}. \quad (1)$$

Eq. (1) confirms that the overall signal transfer function is unity while the noise-shaping transfer function has the form $[1 + z^{-1}A(z)]^{-1}$. The output signal of the SDM normally has a reduced word length compared to that of the input signal, yet the in-audio band resolution can be almost completely maintained. The output code, for example, could be 4 bit, implying that each transducer requires only 16 levels. A multilayer transducer, as shown in Fig. 2, would then have 16 binary controlled layers. In suggesting the use of area modulation techniques the array loudspeaker approximates an active surface, where it is important that the dimensions of each element be small compared to that of the wavelength of sound at the highest audible frequency. The use of binary-weighted elements and the addition of scrambling as shown in Fig. 1 introduce spatial decorrelation of the sound field, making the element appear as a small diffuse source.

2 PRINCIPLES OF POLAR RESPONSE FORMATION IN ONE-DIMENSIONAL ARRAY LOUDSPEAKERS

The challenge is to endow the loudspeaker array with the means to control fully the polar response and to allow both coherent and diffuse radiation options, including a mixed option for simultaneous diffuse and coherent radiation from a single array. However, our study commences by examining from first principles how an array can achieve broad-band directional radiation, irrespective of whether diffuse processing is included. The parameters that require control are beam angle, beam width, and the number and type of beams, each of which should be specified independently.

Consider a directional array operating initially at a single frequency f Hz, where polar formation is shown to be a process of frequency-domain filtering. Fig. 5 illustrates a line array of N uniformly spaced elements with an interspacing of g meters. The transducer elements are fed with individual signals weighted by a set of complex coefficients $\{a_r + jb_r\}$, for $r = 0, \ldots, N - 1$, that modify the amplitude and phase of the input signal, where the basic structure is shown in Fig. 6. Although it will be shown that $\{a\}$ and $\{b\}$ coefficients have to be frequency dependent, for a single frequency the coefficients appear as constants therefore for that frequency the signal received appears to have been processed by a finite-impulse response (FIR)

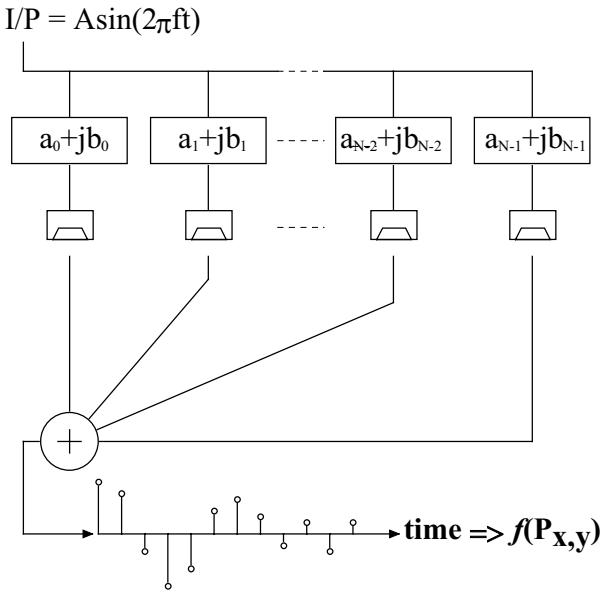


Fig. 6. FIR filter representation of line array.

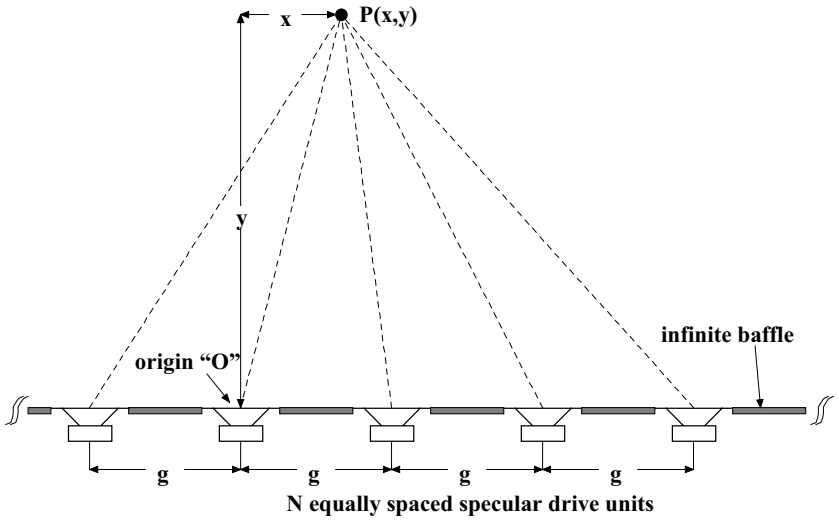


Fig. 5. Line source of N small transducers with interspacing g meters.

filter, where the sampling frequency is a function of the geometry and velocity of sound in air. Hence at the listening point $P(x, y)$ an impulse response is observed that is a combination of the selected filter coefficients and the individual paths between transducers and $P(x, y)$, which for the special case of $P(x, y)$ being in the far field, map directly to a set of time delays.

Let the received acoustic impulse response be representative of a low-pass filter with a windowed sinc function impulse response. The cutoff frequency of this filtered response depends on the time scale of the impulse response, which in turn is dependent on the location of $P(x, y)$. As $P(x, y)$ moves along an arc, the time scale shifts in proportion to the sine of the angle made with the normal to the array. For example, for an increase in angle the sinc function is stretched correspondingly in time because of the greater propagation time across the array; consequently the filter cutoff frequency is lowered. Hence for a signal frequency of f Hz there is an angular region where this signal falls within the passband of the filter whereas for larger angles, the signal falls within the stopband. This illustrates how a directional array can be formed. It also exposes the mechanism that controls the polar rate of attenuation with angle, which is a function of the attenuation rate of the low-pass filter determined by the array length (that is, the number of elements) and coefficients.

Time scaling is illustrated is Fig. 7 for both large and small angles to the array normal, and the corresponding inverse trend in the passband response is also shown. Both time-domain traces maintain the same shape; it is just their time scales that differ. The selection of the impulse response controls the shape of the frequency-domain plots, although because of the finite length of the array it is not possible to achieve a theoretical brickwall response.

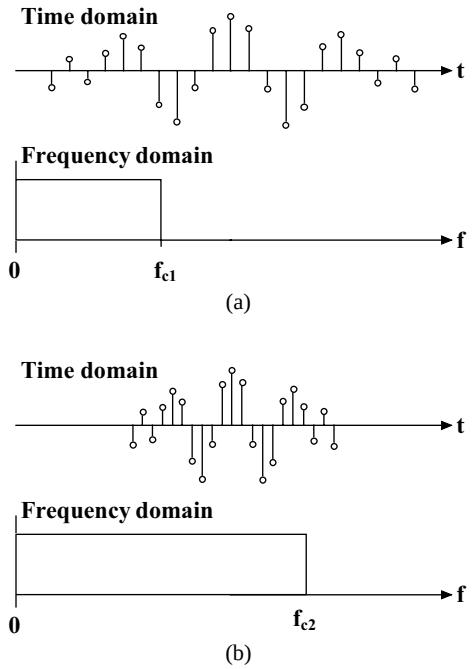


Fig. 7. Two example of time scaling of filter response with observation angle.

Nevertheless it will be shown that windowed sinc functions achieve satisfactory responses for practical loudspeakers, especially when secondary factors such as reflection, diffraction, and driver imperfections are considered.

This argument has so far been applied only to a single input frequency (equivalent to the carrier in the radio-frequency case) where the frequency determines the spatial low-pass filter response. Consequently, when the input frequency is changed, the filter inherent in the array requires a modified cutoff frequency if the polar characteristic is to remain similar. A set of filter coefficients specific to discrete signal frequencies therefore has to be calculated where for a broad-band input signal the array coefficients form a set of complex functions of frequency. Hence for the rth channel the function AR_r is

$$AR_r = a_r(f) + jb_r(f). \tag{2}$$

In the design process a discrete number of low-pass filters are calculated over the operating band to match the required polar response. Interpolation, such as spline [16] interpolation, is then used to obtain a finer frequency resolution. The interpolated complex functions AR_r contain broad-band amplitude and phase information, which can be applied directly to the input signal spectrum. Alternatively these functions can be transformed into the time domain to realize a set of impulse responses, which, following convolution with the input signal, yield the input signals for the elemental array transducers.

Caveat: The polar response depends on the distance from the array to the listener. Consider a FIR low-pass filter formed by a uniformly spaced array, where the number of taps equals the number of transducers. Observed at a large distance, the FIR response has a uniformly sampled impulse response with an effective uniform sampling frequency that is a function of the angle. However, when the listening distance is reduced, the angles associated with each transducer and the normal to the array differ progressively and introduce nonuniform sampling while variations in path length create attenuation differentials that alter the relative tap weights. These two effects together modify the filter transfer function and cause the target polar response to deteriorate. Consequently polar response control is normally applied only to the far field.

3 ARRAY SIZE AND ELEMENT INTERSPACING IN POLAR RESPONSE FORMATION

In this section the relationship between polar response bandwidth, array size, number of transducers, and transducer interspacing is analyzed. It is shown that the bandwidth over which the polar response can be controlled must have a bandpass characteristic that is related fundamentally to the array size and the number of elements. Once these global parameters are estimated from the array specification, the design of the digital filter bank can be performed and is presented in Section 4.

An N-element uniform line array of transducers with an interspacing of g meters have an overall span width, in

meters, of

$$\text{width} = (N - 1)g \ . \tag{3}$$

The array span determines ultimately the low-frequency polar bandwidth while the interspacing determines the upper polar bandwidth, where above this frequency spatial aliasing distortion will occur. Observe that spatial aliasing does not imply non linear signal distortion. It manifests itself as high-frequency spatial replication of the polar lobes with the consequence that typically high-frequency signal components leak into the polar stopband region. The upper frequency limit has an exact figure whereas the lower frequency bound is less precise because there is more gradual degradation of the FIR filter attenuation rate for a given filter length as its cutoff frequency is lowered. The upper bound follows directly from Nyquist sampling theory applied in the spatial domain and can be determined by inspecting the line array and the corresponding filter topology shown in Fig. 5 and 6. Assume here that the filter coefficients are all real and a unit impulse is applied to the input of the array processor. Each transducer in the array then produces a coherent impulse. Observed along the normal (at a distance that is large compared to the overall array width), all the individual impulses arrive simultaneously, implying an infinite sampling frequency. However, as the angle of θ rad to the normal increases, the individual impulses arrive at progressively greater incremental times such that the observed sampling interval T_θ is

$$T_\theta = \frac{g \sin\theta}{c} \tag{4}$$

where c is the velocity of sound in air. From Eq. (4) the observed sampling frequency is shown to be lowest at the maximum polar angle of $\theta = \pi/2$ rad, where from Nyquist's sampling frequency the maximum signal frequency f_{max} that just avoids the onset of aliasing distortion is one-half the sampling frequency, that is,

$$f_{max} = \frac{c}{2g \sin\theta} \ . \tag{5}$$

For an array to form a correctly shaped beam that can be directed over $\theta = -\pi/2$ to $\pi/2$ and where the maximum time-domain signal frequency is f_{high}, then by substituting $f_{max} = f_{high}$ and $|\theta| = \pi/2$ in Eq. (5) it follows that the optimum interspacing g_{opt} has an upper limit,

$$g_{opt} = \frac{c}{2f_{high}} \tag{6}$$

although in practice the actual interspacing g must be selected such that

$$g \le g_{opt} \ . \tag{7}$$

This is a fundamental limit on g, which if exceeded results in spatial aliasing distortion, which creates false frequency-dependent lobes within the polar response. However, if the interspacing is too small, then for a given number of elements N the array is unnecessarily narrow, causing the low-frequency polar response to degrade. Consequently, selecting the array size and number of elements to match the signal is paramount to maximizing the frequency range of a properly formed beam pattern.

4 FILTER DESIGN: DETERMINATION OF ELEMENT CHANNEL TRANSFER FUNCTIONS (ECTFs)

To control beamwidth and beam angle, updateable digital filters are located between the input and each radiating element of the array. This section analyzes the transfer function requirements. Initially in Section 4.1 a beam is considered formed symmetrically about the normal to the array whereas in Section 4.2 the analysis is extended to include offset beams over an arc of π rad. Each filter establishes an individual element channel transfer function (ECTF), which can also take account of the radiating element transfer function.

The core filter design concept is to interpret the line array as the FIR filter shown in Fig. 6. Normally in FIR filter design the sampling rate is constant and the input frequency is a variable. However, in this scenario the signal frequency is considered constant whereas the effective sampling rate is variable, being determined by the observation angle θ to the array normal, as expressed in Eq. (4). Changes in the effective sampling rate thus scale the filter's low-pass cutoff frequency, which together with the FIR filter taps defines the shape and beamwidth of the polar response. The calculation of filter taps must be applied to discrete frequencies taken over the full operating range set by the array size and the element interspacing. As a consequence the individual filter taps are themselves functions of frequency and are required in each element path to emulate the variation of FIR filter taps with signal frequency. It should be observed that because the effective sampling rate changes symmetrically either side of the normal, it follows that the polar response is symmetrical about the normal, although Section 4.2 describes a method of beam steering while allowing the polar response shape to remain invariant.

Assume that at a discrete signal frequency the FIR filter formed by the array and tap weights is designed to have a low-pass filter characteristic. Because the observation angle determines the effective sampling rate, the corresponding scaled low-pass cutoff frequency then defines the beam's width and the shape of the angular transition region. This reveals an intimate relationship between element location, FIR filter coefficients, and signal frequency for a given beamwidth. To alter the beamwidth all filter cutoff frequencies must be modified at each discrete design frequency. Consequently a low-pass filter set is required, designed to match the desired beamwidth over a range of discrete frequencies. The taps of these filters when expressed as a function frequency then define the ECTFs, which may be augmented further by interpolation to achieve a finer frequency resolution, as described in Section 4.3. To complete the ECTF

design process, discrete impulse responses are derived by Fourier transformation. Although this design process must be applied to each individual beam, precalculation and memory can be used to simplify smart operation, requiring only that the digital filter coefficients be updated so as to allow smooth morphing from one beam characteristic to another.

Consider an array capable of producing multiple beams where for polar beam x in an M-beam system, there are three principal parameters to consider in designing the set of digital filters:

- Beamwidth L_x, rad
- Angle the beam makes with the normal β_x, rad
- Rate of polar response attenuation with angle as the spatial stopband region is entered (that is, polar transition region).

In the following analysis a rapid polar response transition region is assumed, although at lower frequencies there is an inevitable degradation in the rate of attenuation with the angle because of the finite length of the FIR filter response determined by the number of array elements.

4.1 Far-Field Polar Width Relationship to FIR Filter Cutoff Frequency for $\beta_x = 0$

Initially a far-field beam that is symmetric about the normal is considered, where $\beta_x = 0$. The cutoff frequency of each input-signal frequency-specific low-pass filter affects only the width of the beam and not the angle it makes with the normal. To prove this observation, assume that a set of filter coefficients has been chosen for a given operating frequency. As the monitoring position moves away from the normal, the filter impulse response $h(nT_\theta)$ spreads as a function of the angle θ, such that the impulse response has a total observed time duration T_{fd} given by

$$T_{fd} = (N-1)\frac{g\sin\theta}{c}. \tag{8}$$

Reversing the angle θ leads to a time-reversed impulse response $h(-nT_\theta)$ as the observed coefficients now appear in reverse order. Consequently the filter has an identical-magnitude frequency response for $\theta = -\theta$, because

$$\text{abs}\Big(\text{fft}\big(h\left(nT_\theta\right)\big)\Big) = \text{abs}\Big(\text{fft}\big(h\left(-nT_\theta\right)\big)\Big)$$

where abs(…) implies the magnitude operator and fft(…) the Fourier transform. This implies a symmetrical polar response about the normal to the array.

A beam of width L_x rad symmetrical about the normal is shown in Fig. 8. Consider the relationship between the beamwidth, the low-pass filter cutoff and the signal frequency of f Hz of an N-element line array with an actual transducer spacing of g meters, where the upper frequency limit of the array is f_{high} Hz. To achieve this polar response, the cutoff frequency of the array FIR filter observed at angles $\theta = -L_x/2$ and $\theta = L_x/2$ must equal the signal frequency.

From Eq. (4) the effective sampling frequency of the FIR filter is $f_s = 1/T_\theta$. Assuming for an angle $\theta = L_x/2$, that $f_{sx} = f_s$, then

$$f_{sx} = \frac{c}{g\sin\left(L_x/2\right)}$$

Substituting for c from Eq. (6),

$$f_{sx} = \frac{2f_{high}}{\sin\left(L_x/2\right)}\frac{g_{opt}}{g}. \tag{9}$$

The FIR filter has N coefficients, so taking the Fourier transform of a vector of length N, the fundamental frequency of the spectrum is f_{high}/N. Hence if the input signal frequency is f Hz, then for a (unrealizable) brickwall response the spectrum must terminate its passband at harmonic ncf, where

$$\text{ncf} = \text{ceil}\left(\frac{f}{f_{sx}}N\right) = \text{ceil}\left[0.5N\frac{f}{f_{high}}\frac{g}{g_{opt}}\sin\left(\frac{L_x}{2}\right)\right] \tag{10}$$

where ceil is a rounding function.[2] Inevitably using a discrete Fourier transform there must be an integer number of harmonics that restricts the choice of cutoff frequency, although to improve the resolution, linear interpolation can be applied to the harmonics either side of the nominal cutoff frequency. By employing a rectangular window in the frequency domain, a direct synthesis method has been adopted that uses a sinc function with a low-pass filter cutoff frequency observed at the edge of the polar response set to match the signal frequency of f Hz. The function is sampled effectively at f_{sx}, whereas to form a finite filter length and to smooth the frequency response within the polar transition region, a window function is applied subsequently. That is,

$$h(t) = \frac{\sin(2\pi ft)}{2\pi ft} = \text{sinc}(2\pi ft).$$

Sampling at a rate f_{sx} that corresponds to a polar response

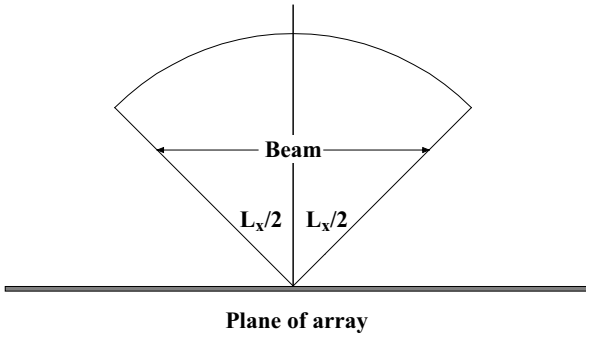


Fig. 8. Far-field symmetrical beam about normal, spanning $-L_x/2$ to $L_x/2$.

[2] Ceil is a matlab function implying an integer roundup.

angle of L_x, the rth tap is

$$\text{tap}|_r = w_r \,\text{sinc}\left(\frac{2\pi f r}{f_{sx}}\right) \tag{11a}$$

where f_{sx} is given by Eq. (9) and w_r is the rth coefficient of a finite-length window function used also to smooth the FIR filter frequency response within the corresponding polar transition region. In this study a rectangular window with raised-cosine termination was used to weight the elements at the line array ends. However, to maximize power handling it is desirable to have most transducer elements contribute to the whole sound field. Therefore the number of low-valued coefficients should be minimized. Consequently the cosine functions selected span typically wc = 5 elements at each end of the $N = 64$ array. The window vector win of which w_r is the rth element is given in Matlab[3] notation as

$$\begin{aligned}\text{win} = \Big[&\Big(\big(1 + \cos\big(\text{pi} * \big((\text{wc} - 1):-1:0\big)/\text{wc}\big)\big)/2 \\ &\text{ones}\big(1, (N - 2 * \text{wc})\big) \\ &\big(1 + \cos\big(\text{pi} * \big(0:(\text{wc} - 1)\big)/\text{wc}\big)\big)/2\Big].\end{aligned} \tag{11b}$$

It is shown in Section 7 that the direct synthesis technique can readily be adapted to embed diffuse processing into the filters, where a method based on the Fourier transform is well matched to this task.

[3]Matlab is a trade name of MatWorks Inc.

4.2 Beam Steering with Offset Angle β_x

In this section the case is considered where a far-field beam of width L_x is offset from the normal to the array by an angle β_x, as shown in Fig. 9. The principal means used to achieve beam steering is to introduce progressive time delays into each transducer feed so that the beam becomes offset from its symmetrical position about the normal. The delays can be calculated directly from the path lengths d_{-2}, d_{-1}, d_1, d_2, …, as shown in Fig. 10, where four transducers are illustrated out of an N-element array.

For reasons of symmetry, half the elements are drawn to the left of the array center and half are drawn to the right.

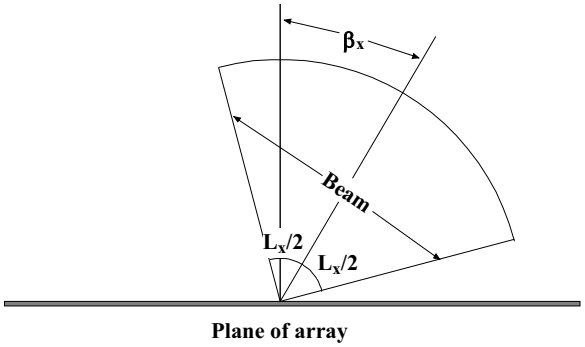


Fig. 9. Far-field polar beam of width L_x with offset angle β_x.

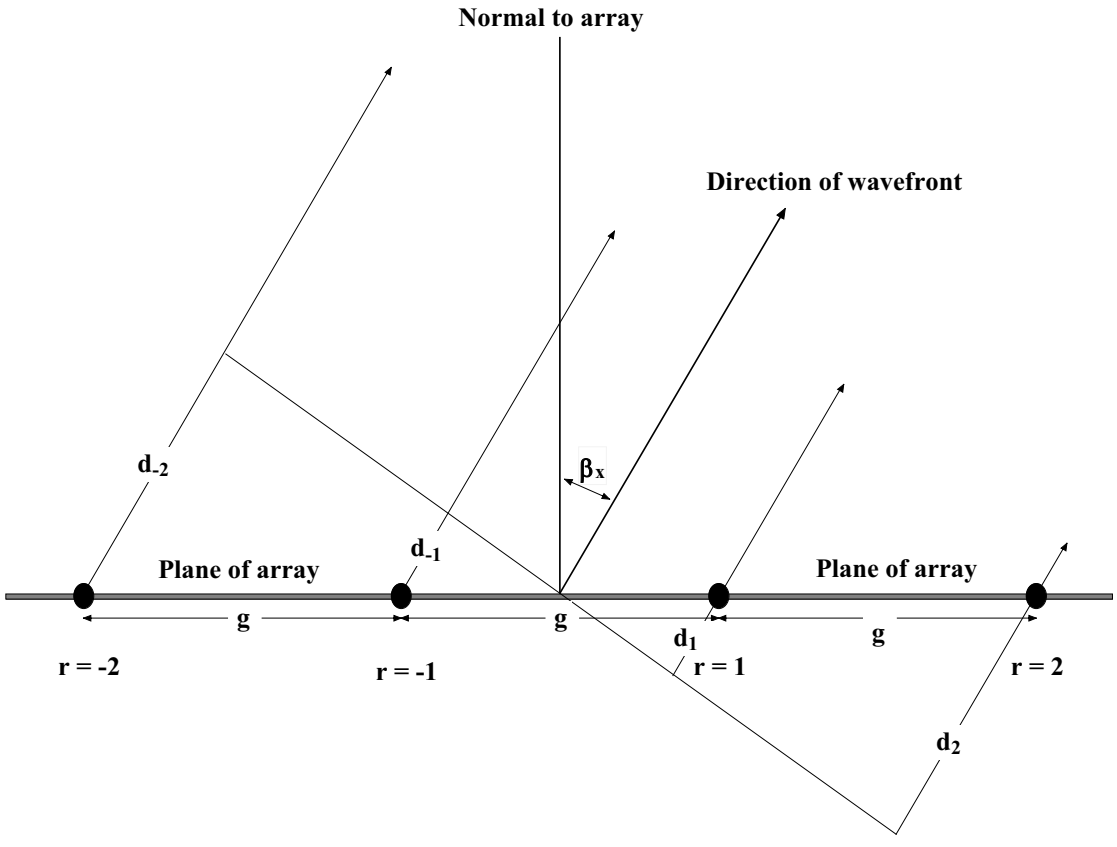


Fig. 10. Delay paths for each transducer for beam offset β_x.

Consider the right-hand rth element, where the corresponding path length d_r is given by

$$d_r = g(0.5 + r)\sin\beta_x .$$

If c is the velocity of sound in air, then the time delay T_r corresponding to d_r is

$$T_r = \frac{g(0.5 + r)\sin\beta_x}{c}$$

resulting in respective frequency-domain transfer functions delay_{rl} and delay_{rr} for the left- and right-hand rth elements of the array, where

$$\text{delay}_{rl} = \exp\left[+\text{j}\frac{2\pi f}{c}(0.5 + r)\sin\beta_x\right] \quad (12)$$

$$\text{delay}_{rr} = \exp\left[-\text{j}\frac{2\pi f}{c}(0.5 + r)\sin\beta_x\right] . \quad (13)$$

However, just introducing a set of time delays is a necessary but not a sufficient condition to steer the beam away from the normal. When the beam is offset by an angle β_x from the normal, the apparent interspacing observed along the center line of the beam is reduced to ($g \cos \beta_x$). Consequently Eq. (10) is modified as

$$\text{ncf} = \text{ceil}\left(\frac{f}{f_{sx}}N\right) = \text{ceil}\left[0.5N\frac{f}{f_{\text{high}}}\frac{g\cos\beta_x}{g_{\text{opt}}}\sin\left(\frac{L_x}{2}\right)\right] \quad (14)$$

where $\{|\beta_x| + |L_x/2|\} < \pi/2$ since the whole beam must be contained within an arc $-\pi/2$ to $\pi/2$. The interspacing compensation applied to Eq. (14) is therefore critical to the polar response performance. Otherwise the beamwidth would change with the offset angle.

4.3 Completion of ECTF Design Procedure

Each unique filter design corresponds to a given polar response that is matched to a single signal frequency taken from a discrete preselected range. Both frequency- and time-domain windows are incorporated in the process together with the discrete Fourier transform to derive the ECTF filters. Initially an ideal brickwall filter response is represented by a rectangular frequency-selective mask function scf as defined by Eq. (15) and illustrated in Fig. 11. The mask is defined here in terms of the cutoff harmonic ncf calculated in Section 4.2 and is written in Matlab notation as

$$\begin{aligned}\text{scf} = \big[&\text{ones}\big(\text{size}(1{:}\text{ncf})\big)\\ &\text{zeros}\big(\text{size}(1{:}N - \text{ncf}*2 + 1)\big)\\ &\text{ones}\big(\text{size}(1{:}\text{ncf} - 1)\big)\big] .\end{aligned} \quad (15)$$

Next the Fourier transform tmp of a unit impulse is calculated as

$$\begin{aligned}\text{tmp} = \text{fft}\Big(\big[&\text{zeros}\big(\text{size}(1{:}N/2 - 1)\big)\\ &1\,\text{zeros}\big(\text{size}(1{:}N/2)\big)\big)\Big] .\end{aligned} \quad (16)$$

The filtered array coefficients that correspond to a single frequency are contained within a finite impulse response imp as

$$\text{imp} = \text{win}.* \ \text{ifft}\big(\text{scf}.* \ (\text{tmp}./\text{abs}(\text{tmp}))\big) \quad (17)$$

where fft and ifft are the Fourier transform and the inverse Fourier transform, respectively, and win is the time-domain window function defined by Eq. (11b) and used to truncate the coefficients to match the array size. The coefficients within the vector imp are subsequently normalized by the sum of the coefficient set to guarantee a unity-gain transfer function along the axis of symmetry of the polar response, that is,

$$\text{imp} \Rightarrow \text{imp}\Big/\sum_{x=1}^{n}\text{imp}(x) . \quad (18)$$

To implement the polar offset β_x, and thus steer the beam, the two sets of phase functions delay_{rl} and delay_{rr} defined by Eq. (12) and (13) are multiplied with their corresponding array coefficients in the vector imp. Each array

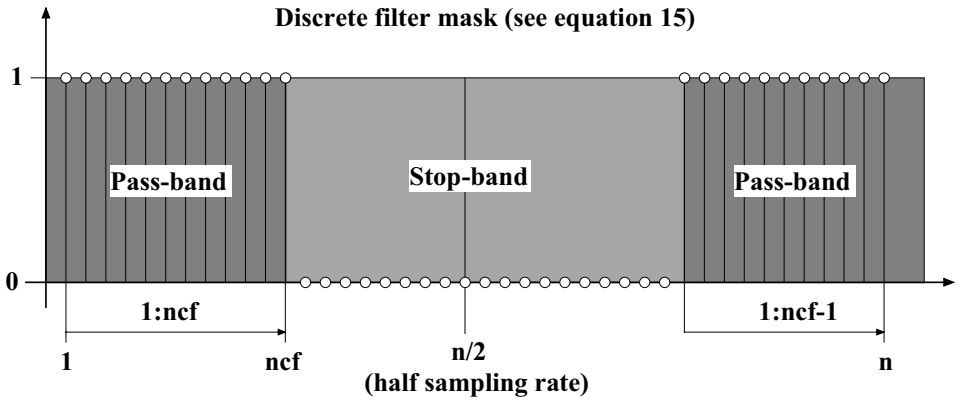


Fig. 11. Filter mask defined by Eq. (15).

coefficient is now a complex number, which introduces both amplitude weighting and phase shift appropriate to the signal frequency *f*.

The filter design process is then repeated for discrete frequencies taken over the array passband in order to determine the channel transfer functions in each element feed. The process can be performed either directly to the required frequency resolution, or alternatively to a lower resolution with spline interpolation used in the frequency domain to extend the number of discrete frequencies, as illustrated in Fig. 12. The cubic spline [16] was chosen here as it produces smooth interpolation without ringing or discontinuity. The spline function is created by triple convolution of a rectangular function although in the program derived to generate example polar responses, the Matlab spline function was employed. Using interpolation reduces the number of filter designs and speeds up the design process. Also, spline interpolation must be applied separately to both the real and the imaginary parts of the coefficient sets for each channel in the array. Consequently a set of complex transfer functions is produced, where if desired these can be transformed into the time domain to realize a set of impulse responses, which can be used directly within a digital processor.

By way of example (see polar example 1 in Section 5 for computational data), Fig. 13 shows an actual frequency-dependent function for channel 20 in a 64-element linear array. The dotted lines show the discrete calculated coefficients as functions of frequency, whereas the enveloping curves show the result of spline interpolation. The real and imaginary parts of the function are depicted separately, as indicated, and the interpolation ratio is 1:8.

5 SIMULATION AND SYSTEM VERIFICATION FOR DUAL COHERENT POLAR RESPONSE EXAMPLES

The performance of a coherent array together with the ECTF signal processing described in Section 4 can be verified by simulating the far-field pressure calculated over an arc of π rad. Only the line array shown in Fig. 5 is considered, although by using two-dimensional transforms the techniques can be adapted to a planar array where the number of elements and signal processors inevitably increases. The simulation assumes that each transducer element of the one-dimensional array is sufficiently small that within the array operating band they radiate hemispherically into 2π-steradian space with a pressure wave that decays as an inverse function of distance. Normally the polar response is specified in the far field where in presenting the results it is assumed that the monitoring distance is large compared to the array dimensions. This implies that the angles each transducer makes with the normal are virtually identical while differences in attenuation with distance are negligible. However, the simulation program also allows for a finite distance, so that any degradation in the polar response can be explored.

A direct method of pressure summation is used at the observation point *P* (see Fig. 5), where for the acoustic domain the pressure is weighted as an inverse of the transducer distance from *P* and appropriate transfer functions are calculated to take into account the propagation time delays from each element. The program also calculates individual ECTFs using the methods described in Section 4 and embeds them into each transducer feed. Consequently the simulation takes account of channel signal processing, array dimensions, and anechoic acoustics. The program can output two formats of polar response:

- A circular coordinate presentation computed for discrete frequencies taken over the passband of the array, where radial length corresponds to sound pressure level expressed in dB and angle maps directly to direction over the arc $-\pi/2$ to $\pi/2$. The contours for each discrete signal frequency are superimposed on the display.
- A three-dimensional rectilinear plot where sound pressure level expressed in dB is plotted vertically and the two horizontal axes are observation angle and frequency, respectively. This presentation allows a clearer impression, especially as to how the polar response varies as a function of frequency.

In addition, the frequency-dependent ECTF coefficient maps and polar correlation plots are also generated:

- It is constructive to observe how the input signal energy for a given polar response is distributed across the elements of the array. This can be revealed by displaying the frequency-dependent coefficient maps of the array, which in effect are the ECTFs used in each element chan-

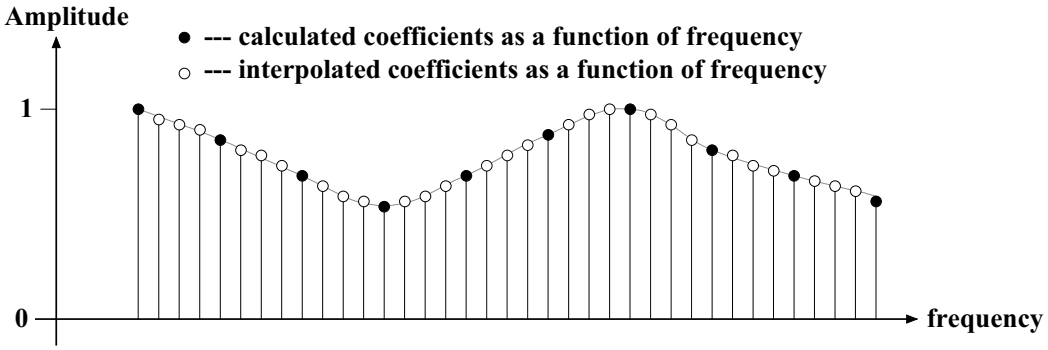
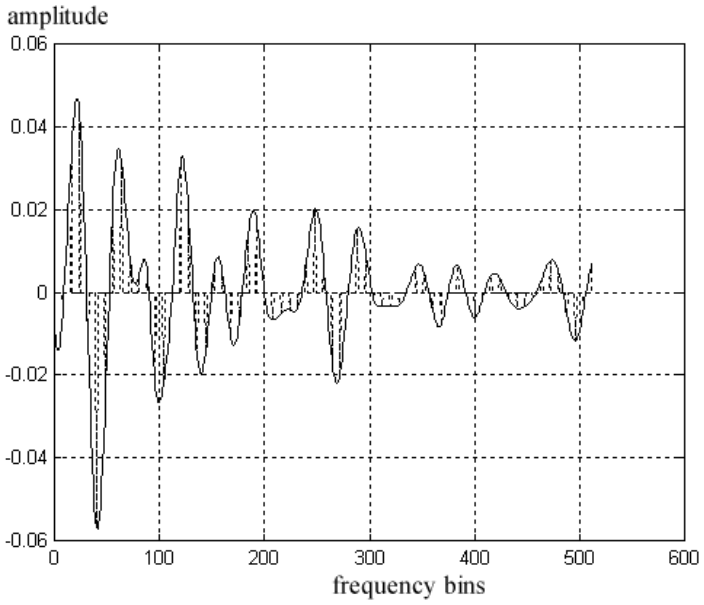


Fig. 12. Interpolation applied to coefficients.

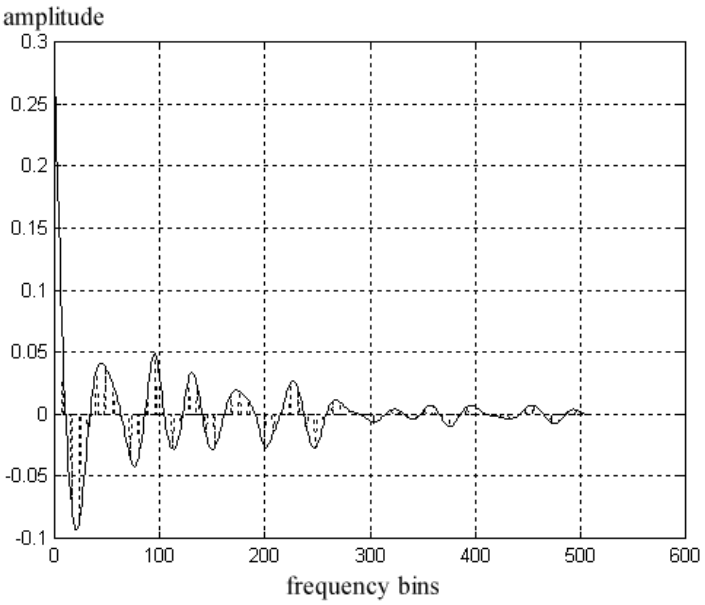
nel. The maps form three-dimensional surfaces where the vertical axis is linear magnitude and the horizontal axes are element number and frequency. Hence by observing the map the distribution of the signal as a function of frequency is shown for each element of the array.

- A three-dimensional map, which has gained importance with the emergence of DML technology, is the cross-correlation display of the polar response. However, unlike the polar frequency response that shows the sound pressure level of the loudspeaker as a function of angle and frequency, the cross-correlation plot presents the cross-correlation function of the off-axis impulse response with the on-axis impulse response as a function of polar angle. In forming the cross-correlation functions the impulse response at the center of the beam is taken as the reference. The vertical axis of the display is therefore the amplitude of the cross-correlation function, whereas the horizontal axes are time and angle. With a coherent design as described in Section 4 the value of the correlation when the impulses are optimally aligned should be maintained over the whole width of the beam. However, this is not the case with the diffuse array described in Section 6.

It is shown that both the coefficient map and the cross-correlation map are useful metrics in quantifying the per-



(a)



(b)

Fig. 13. Example of spline interpolation applied to frequency-dependent function, shown for coefficient 20 of a 64-element array. (a) Real part of SDS function. (b) Imaginary part of SDS function --- actual coefficients.

formance differences of coherent and diffuse loudspeaker arrays, whereas the traditional polar response is rather less discriminating. The range of frequencies used in all output data displays is selected to match the polar response passband of the array. At high frequency the upper limit is set by spatial aliasing distortion, which has been shown in Section 3 to be related to the interspacing of the radiating elements. However, at low frequency the limit is set principally by the overall array width, where the response tends to become omnidirectional due to the effective FIR filters having insufficient length.

A Matlab program[4] was written based on the analysis of Section 4 to implement the coherent array design process, where the beamwidth and the offset angle can be set arbitrarily. The program can simulate a dual-beam design formed by the superposition of independently derived sets of ECTFs, where three polar examples are presented for coherent beam formation and use the following sets of input parameters (β_x and L_x are shown for clarity in degrees):

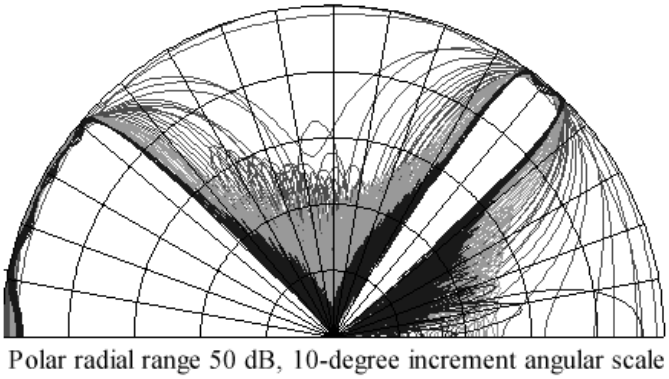
Polar example 1:
N = 64 $\beta_1 = -60°$ $L_1 = 30°$ $\beta_2 = 40°$ $L_2 = 10°$

Polar example 2:
N = 64 $\beta_1 = 0°$ $L_1 = 45°$ $\beta_2 = 0°$ $L_2 = 0°$

Polar example 3:
N = 64 $\beta_1 = 0°$ $L_1 = 180°$ $\beta_2 = 0°$ $L_2 = 0°$

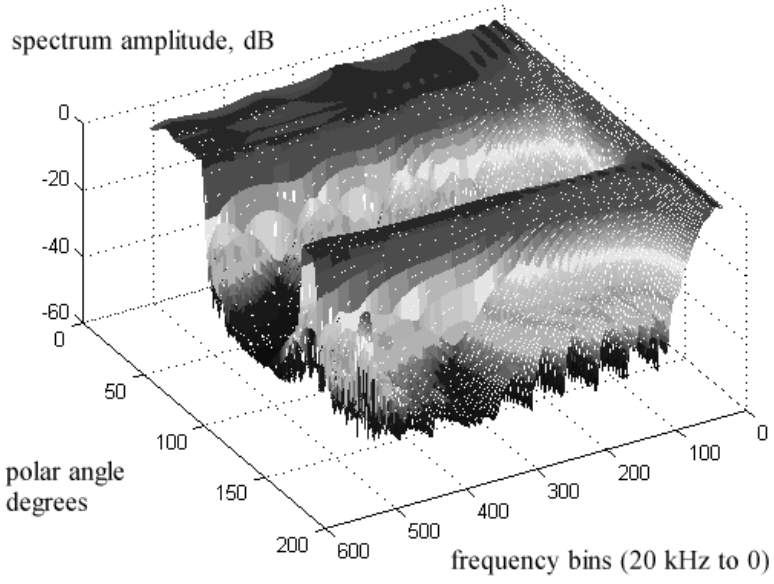
The corresponding simulation results are shown in Fig. 14–16, where plots (a) correspond to cylindrical polar displays (b) to three-dimensional polar displays, (c) to frequency-dependent coefficient maps, and (d) to the polar cross-correlation plots. Observe how the plots reveal well-formed, near frequency-independent polar displays, whereas the cross-correlation plots confirm the coherent nature of each radiated beam.

[4]Available on request.



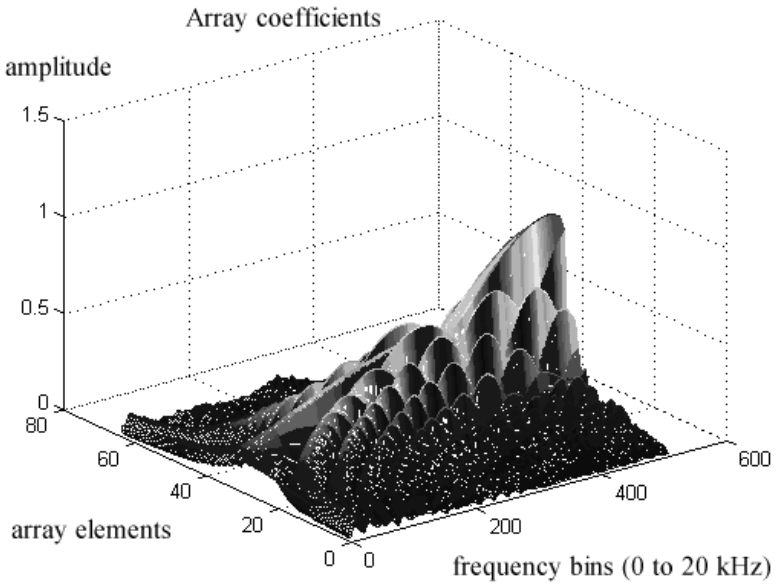
(a)

Fig. 14. (a) Cylindrical polar plot, polar example 1.



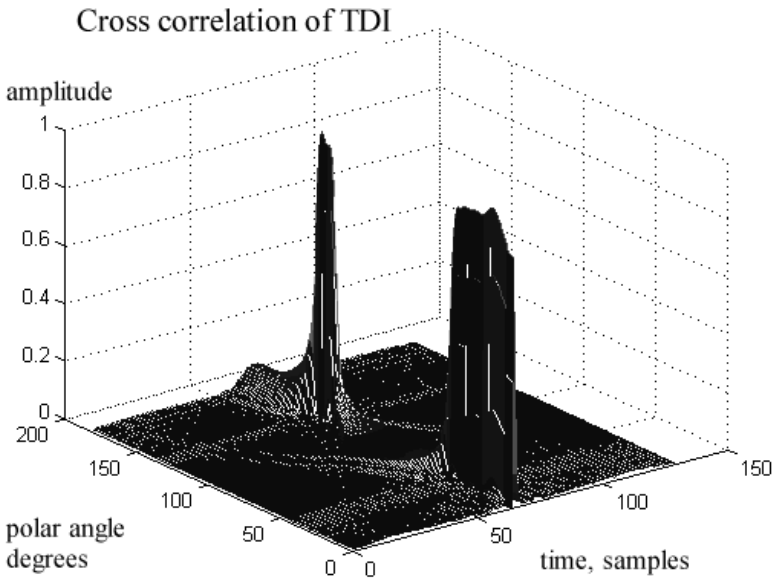
(b)

Fig. 14. (b) Three-dimensional polar plot, polar example 1.



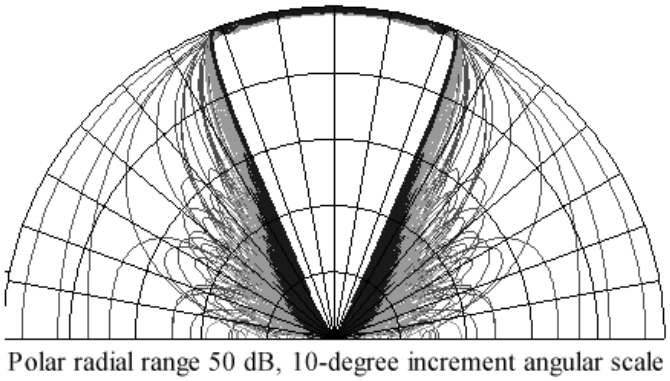
(c)

Fig. 14. (c) Three-dimensional array coefficient plot versus frequency, polar example 1.



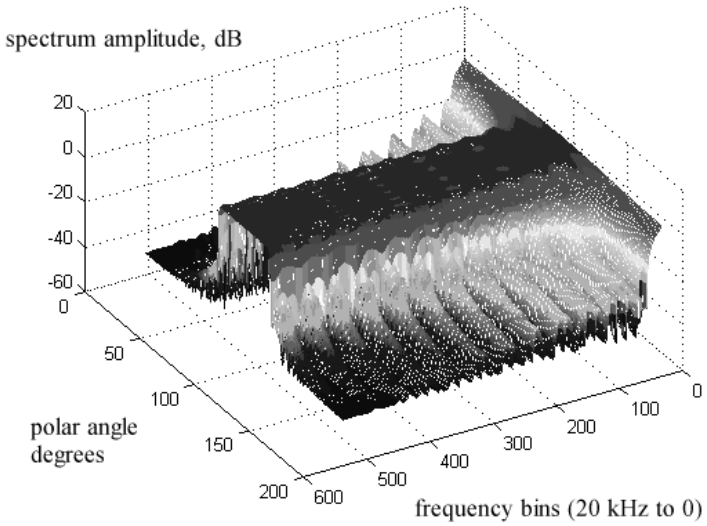
(d)

Fig. 14. (d) Three-dimensional cross-correlation function, polar example 1.



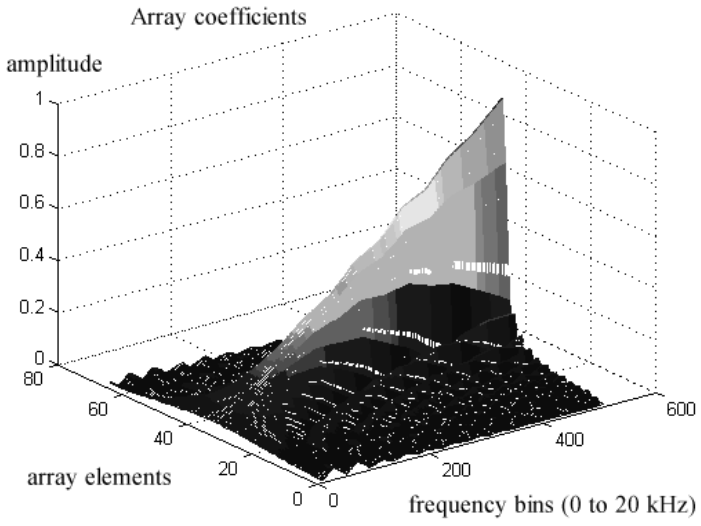
(a)

Fig. 15 (a) Cylindrical polar plot, polar example 2.



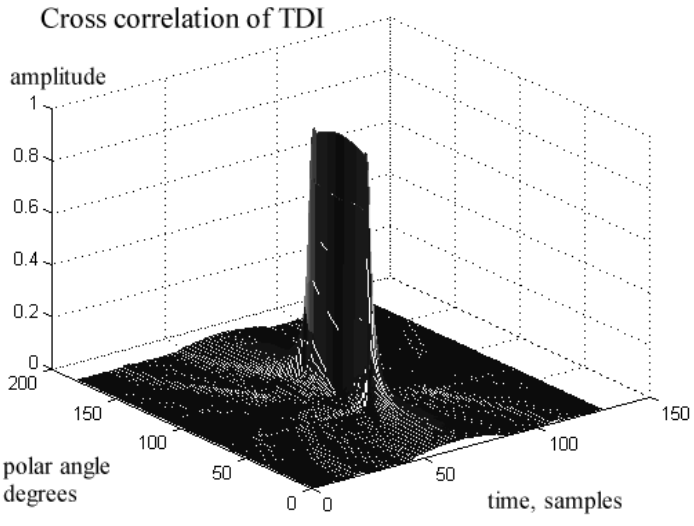
(b)

Fig. 15. (b) Three-dimensional polar plot, polar example 2.



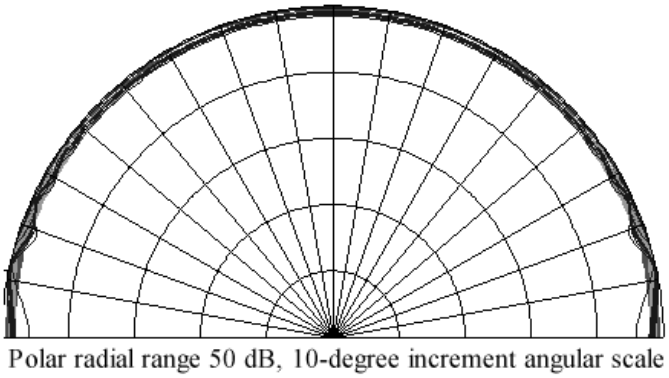
(c)

Fig. 15 (c) Three-dimensional array coefficients plot versus frequency, polar example 2.



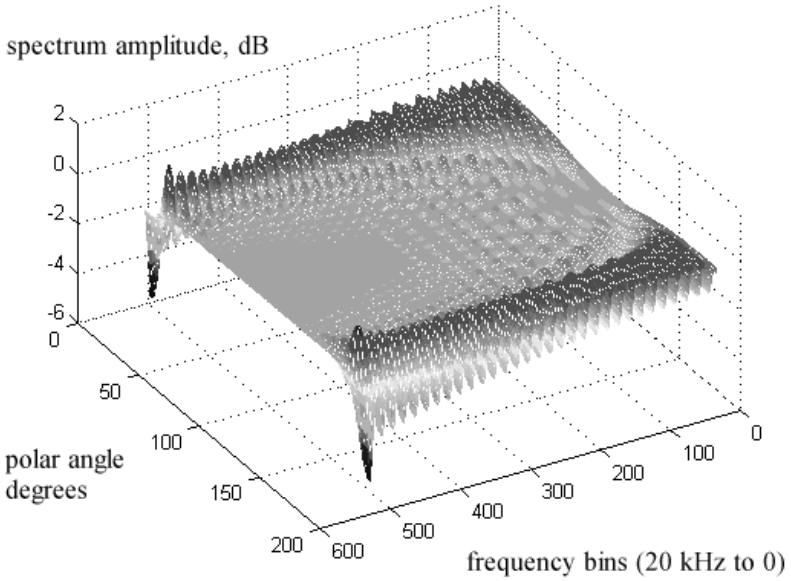
(d)

Fig. 15. (d) Three-dimensional cross-correlation function, polar example 2.



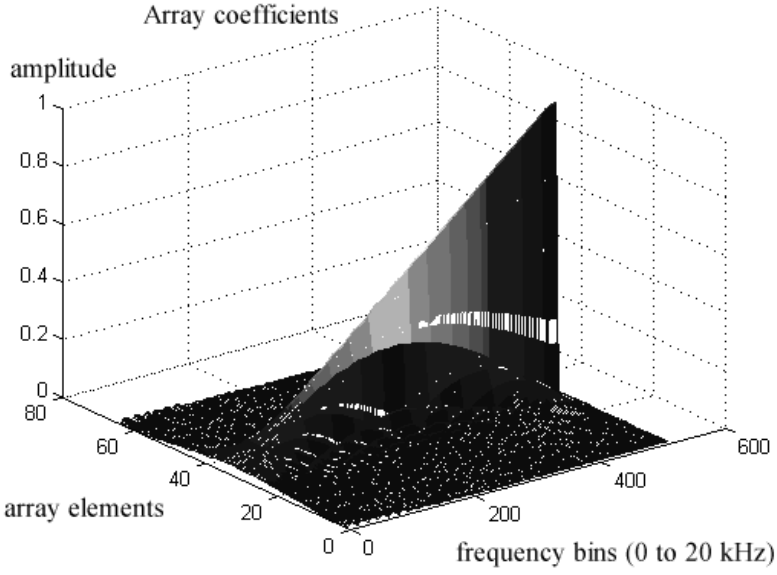
(a)

Fig. 16 (a) Cylindrical polar plot, polar example 3.



(b)

Fig. 16. (b) Three-dimensional polar plot, polar example 3.



(c)

Fig. 16. (c) Three-dimensional array coefficients plot versus frequency, polar example 3.

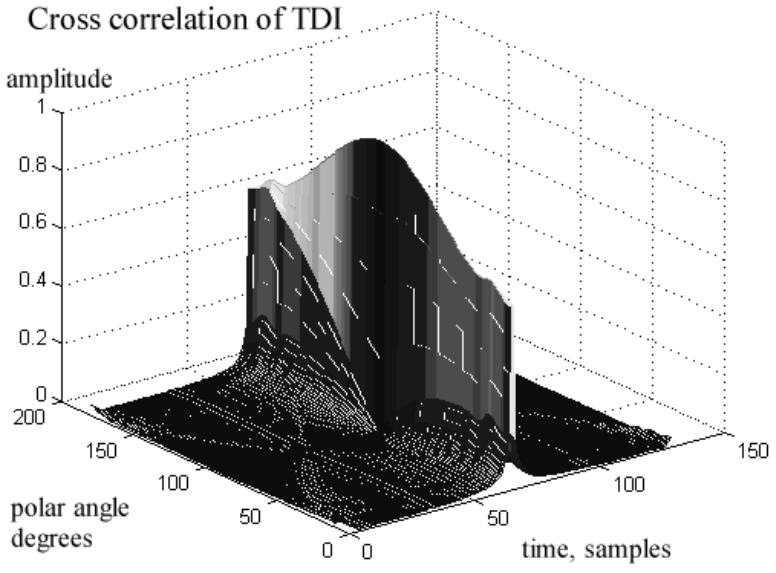
6 DIFFUSE RADIATION FIELDS AND STEERABLE ARRAYS

The results presented in Section 5 apply to a coherent source and therefore relate to specular radiation. In particular, the plots of coefficient maps related to directional beams show typical clustering toward the center of the array with a significant taper toward the array boundary. This implies that signal energy is poorly distributed and array elements are inefficiently used, which has implications on power handling and peak signal levels. Since each elemental transducer within the array is physically small and therefore limited in its power handling, the array has to rely ideally upon an even distribution of energy across the available elemental transducers. The clustering of energy is a direct consequence of the coherent sinc functions used in the derivation of the frequency-dependent FIR filters described in Section 4.

The solution to poor signal distribution across the array is to implement what is defined here as a stochastic filter while retaining appropriate intertransducer phase relationships to achieve a controlled polar response. The use of stochastic filters can also be viewed as a form of diffuse signal processing, which is shown to give the radiated field from the array a diffuse characterization. The key to the technique lies in the form of the sinc generating function used in the derivation of the FIR filters, which are calculated at each discrete signal frequency. In Section 4.3 the analysis used a Fourier transform of a unit impulse multiplied by a filter mask scf as defined by Eq. (15), a method that, after application of a time-domain window, produced a weighted sinc function segment defining an FIR filter at a specific signal frequency. However, rather than using a unit impulse, a vector is selected that in the frequency domain retains a constant-magnitude response but exhibits a random-noise phase response.

Such a function can be derived directly by using a complex exponential function, where the phase is a random-noise vector. When this phase random exponential function is filtered by the same mask scf and normalized subsequently by the absolute value of the noise spectrum, a magnitude spectrum identical to that of the sinc function is produced, but with a random phase response. The FIR filter coefficients can then be calculated as before, using Eq. (17), but where tmp is now the filtered frequency-domain noise function. The random phase attributes of this approach endow the array system with the ability to produce a diffuse acoustic field, which can be confirmed through applying the cross-correlation display map.

To illustrate this process a Matlab program fragment is presented in the Appendix. In this program the binary variable λ controls the program polar-type selection, where for $\lambda = 0$, a coherent sinc function is generated and for $\lambda = 1$, a random function generator is substituted. Complex functions are also allowed. An additional feature (see hereafter and Section 7 for further explanation) is the "for loop" containing the parameter search, which is set typically to 50. This helps select well-formed random functions, which achieve a better polar response shape, especially in the attenuation region. Because there is a random phase function in the filter design process, the resulting polar stopband attenuation is not uniquely defined, especially in the lower frequency region where the limitations of FIR filtering become more evident. Inevitably some randomly selected functions give better out-of-band attenuation than others, so a search loop is used to sift through a number of designs. Filter designs (determined by the parameter search) are performed within a loop, and the stopband mean square error is evaluated in each case. The filter design exhibiting the lowest stopband error is then selected. A typical search loop is 50, although this number may be increased substantially if a final design is being sought for practical implementation. Note that this



(d)

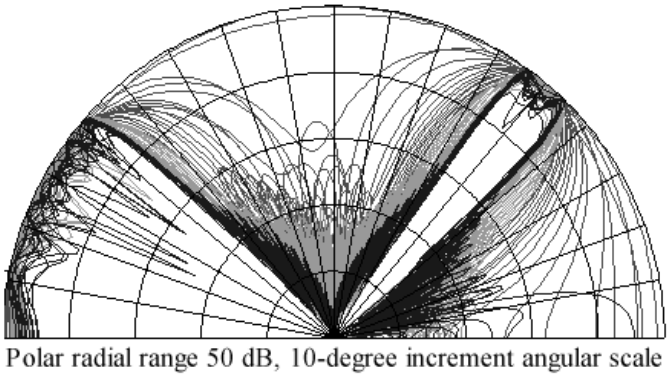
Fig. 16. (d) Three-dimensional cross-correlation function, polar example 3.

process is only relevant to the random-noise vector and offers no advantage for the unit impulse function because of its deterministic form. The process of spline interpolation is also retained, although this has implications for the resolution of the diffuse properties of the array. Here the greater the number of discrete frequencies at which diffuse filters are calculated, the greater will be the diffuse-frequency characterization.

The filter design process for both coherent and diffuse synthesis used the discrete Fourier transform, which is a circular transform, whereas the array forms a finite set of coefficients without circularity. So to address this problem two strategies are used.

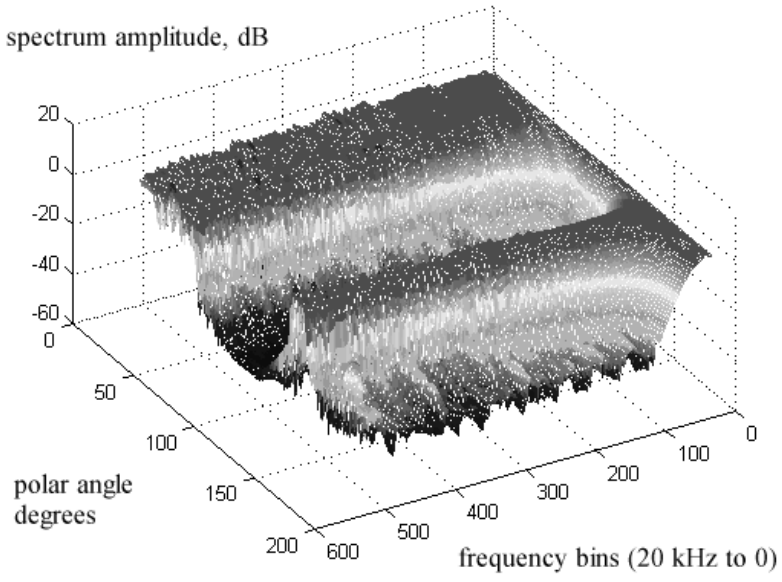
- First, as described in Section 4.1, the coefficient set represented by the vector imp is windowed by the function win to attenuate the contribution from the extreme ends of the array.
- The windowed array function is then stuffed with zeros to reduce circularity effects and to increase the number of frequency bins when the polar frequency response is calculated.

To show the comparison with the coherent design, a set of results similar to those presented in Section 5 is computed using the same design in polar examples 1, 2, and 3 but employing a random function generator. These results are shown in Figs. 17–19 and correspond directly to those of the coherent designs in Figs. 14–16. The simulations reveal that similar polar response formation is possible, although there is now a fine noiselike characterization etched into the polar plots, which is also confirmed by the cross-correlation maps. The cross-correlation functions now show only coherence at the center of each polar beam, whereas responses in different directions reveal the diffuse behavior normally associated with DML. The coefficient maps also show a noiselike characterization, but with the critical attribute that the values are now much more evenly distributed. This bodes well for improved power handling, as energy is no longer clustered towards the central elements. However, of greatest significance is the confirmation that the polar response can be both directional and diffuse within the passband region, that is, a diffuse sound field is not just the domain of loudspeakers with a near omnidirectional polar response.



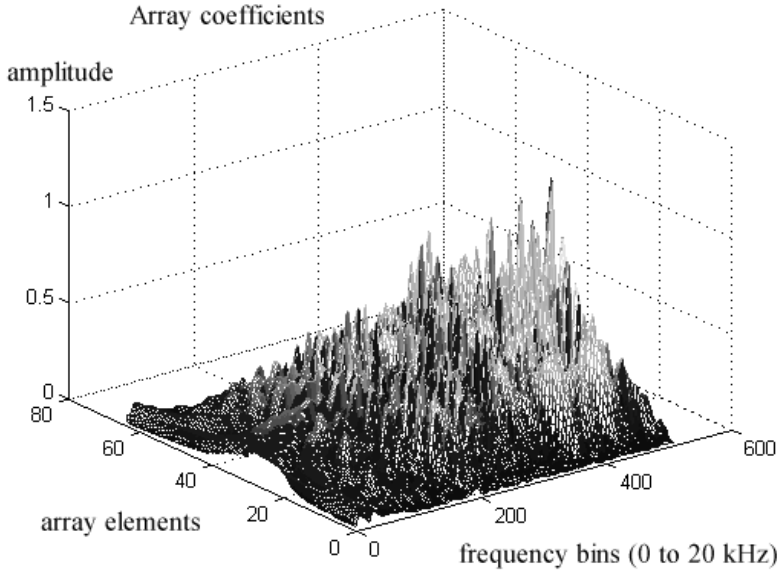
(a)

Fig. 17. (a) Cylindrical polar plot, polar example 1.



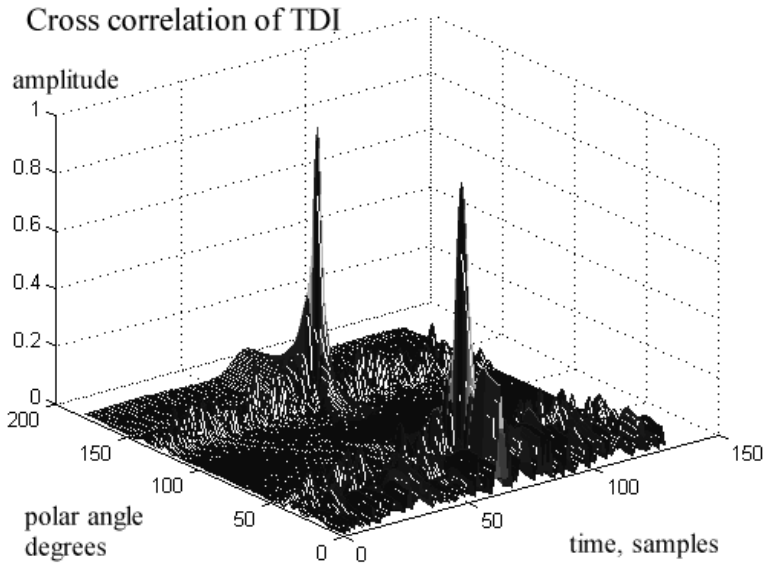
(b)

Fig. 17. (b) Three-dimensional polar plot, polar example 1.



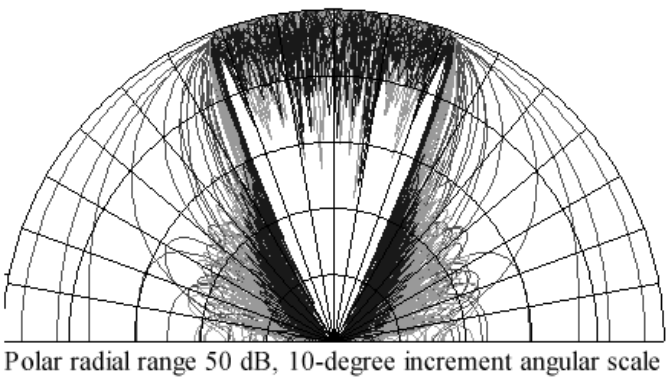
(c)

Fig. 17. (c) Three-dimensional coefficient plot versus frequency, polar example 1.



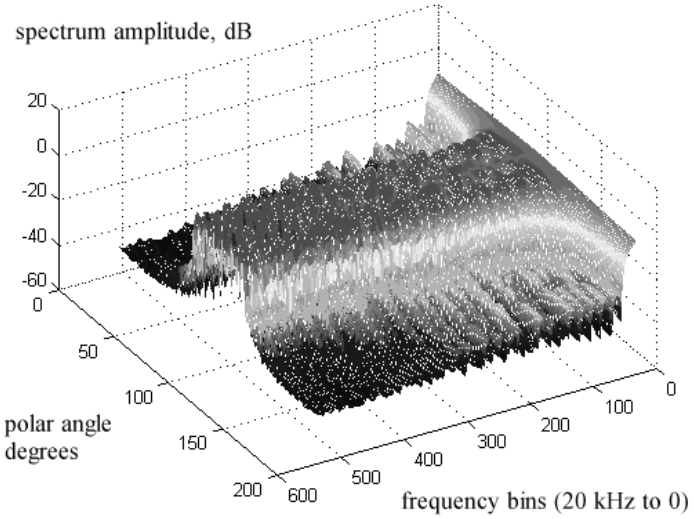
(d)

Fig. 17. (d) Three-dimensional cross correlation, β_1 reference response, polar example 1.



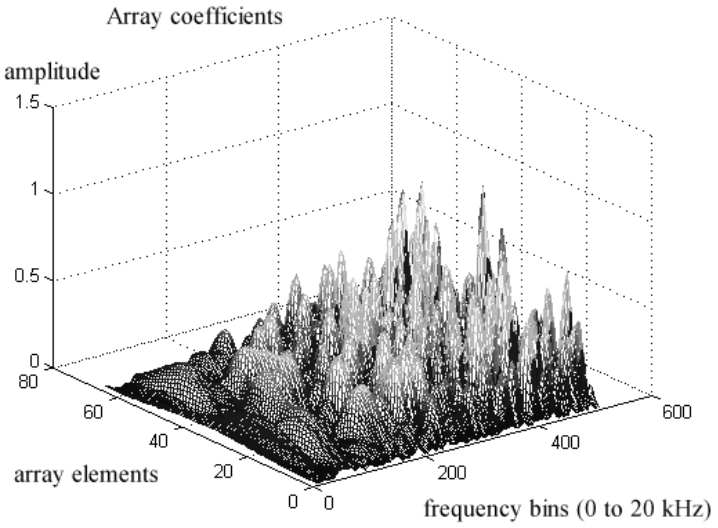
(a)

Fig. 18. (a) Cylindrical polar plot, polar example 2.



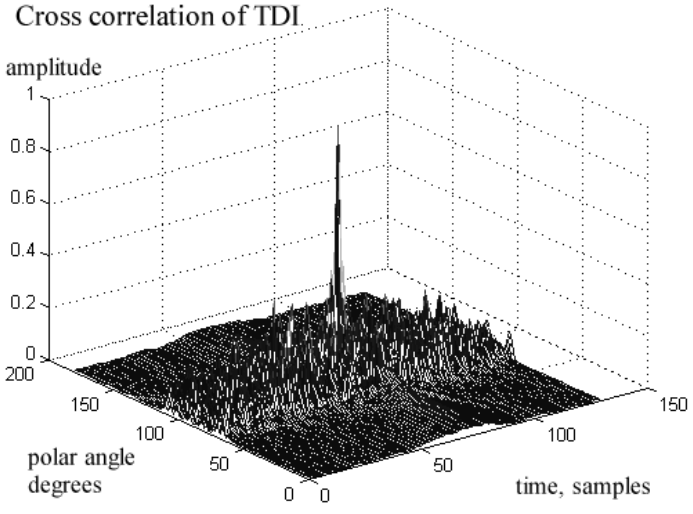
(b)

Fig. 18. (b) Three-dimensional polar plot, polar example 2.



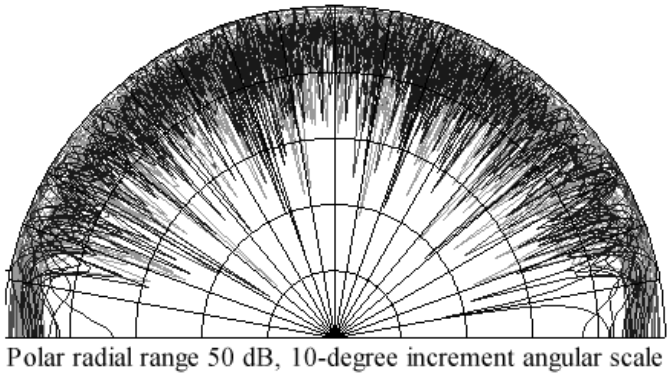
(c)

Fig. 18. (c) Three-dimensional array coefficient plot versus frequency, polar example 2.



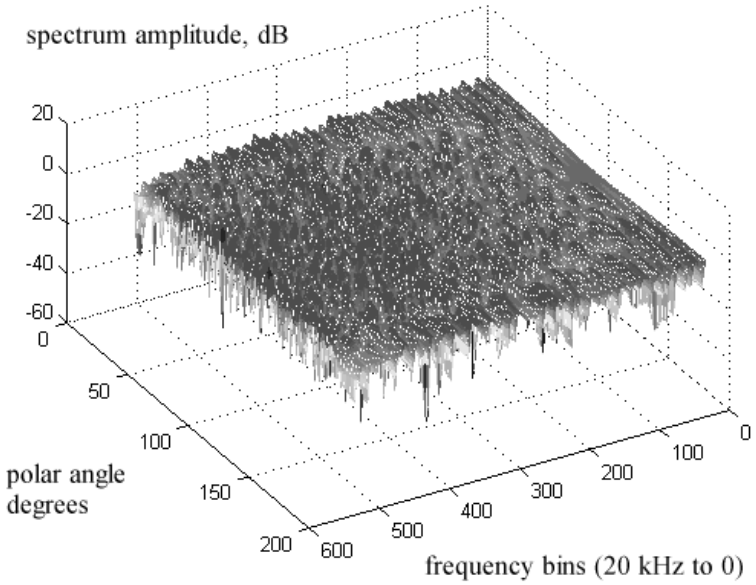
(d)

Fig. 18 (d) Three-dimensional cross correlation, β_1 reference response, polar example 2.



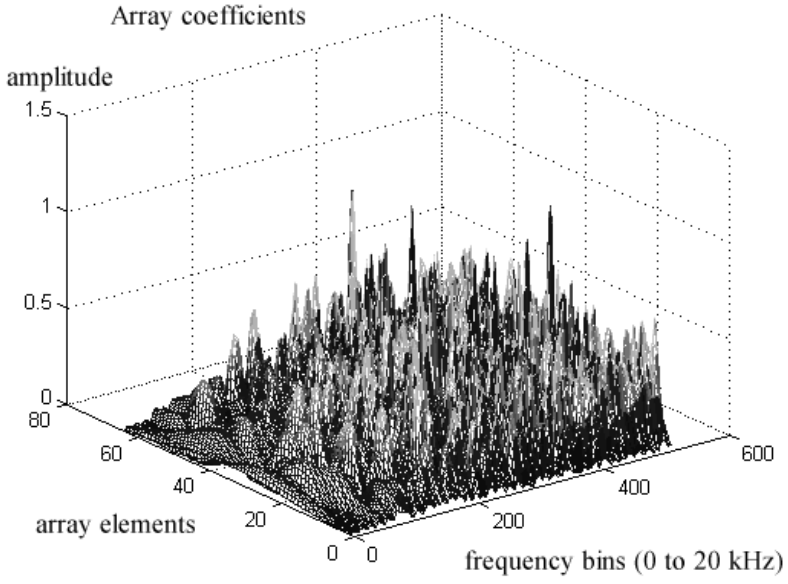
(a)

Fig. 19. (a) Cylindrical polar plot, polar example 3.



(b)

Fig. 19. (b) Three-dimensional polar plot, polar example 3.



(c)

Fig. 19. (c) Three-dimensional array coefficient plot versus frequency, polar example 3.

7 OBSERVATIONS WITHIN STOCHASTIC FILTER DESIGN

This section illustrates the effect of changing the search size in computing stochastic filters. It also examines a possible equalization scheme to correct for low-frequency polar response broadening when listening in closed spaces. Finally, some thoughts are given to the diffusivity of the radiated sound field and its relationship to the trade-off made between the number of discrete stochastic FIR filters and the use of spline interpolation.

7.1 Search Cycles in Stochastic Filter Selection

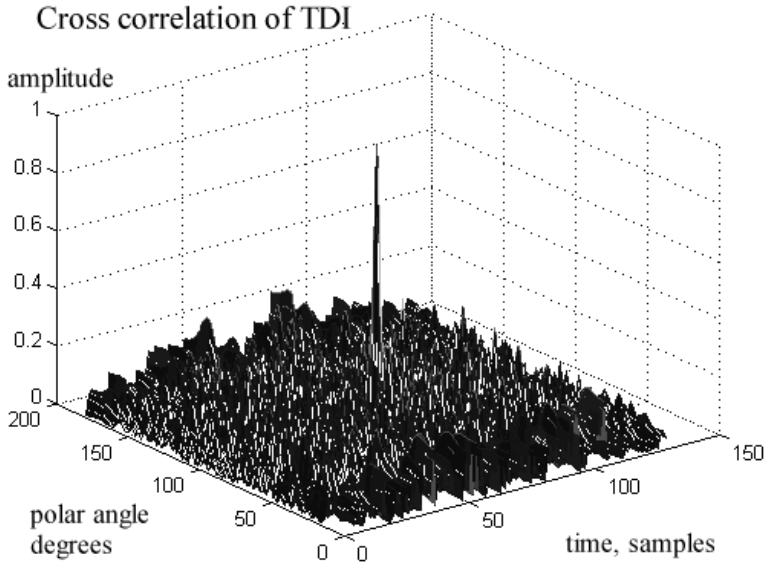
The stochastic method applied to filter design leads to an infinite range of filters with similar but subtly different characteristics. In Section 6 a loop with search cycles was described in order to identify filters with better attenuation in the stopband, implying a greater attenuation in the polar response. To illustrate this feature a polar response example is computed with the corresponding data plotted in Figs. 20–22 using the stochastic routine for search = 1, 50, and 1000, respectively. Also, to illustrate the effect of using more elements in the array, then $N = 256$. The parameters selected are

Polar example 4:
N = 256 $\beta_1 = -45°$ $L_1 = 180°$ $\beta_2 = 60°$ $L_2 = 10°$

The results reveal that in going from search = 1 to search = 50, there is a useful increase in attenuation in the stopband region of the polar response, but in further searches to 1000, the improvement proved to be marginal. Consequently a search of 50 cycles is a good compromise. On close inspection of the three-dimensional polar plots, a small degree of polar-aliasing distortion occurs at extreme high frequencies (see bottom left-hand corner of the responses) and is just evident in the stopband region—probably an artifact of spline interpolation.

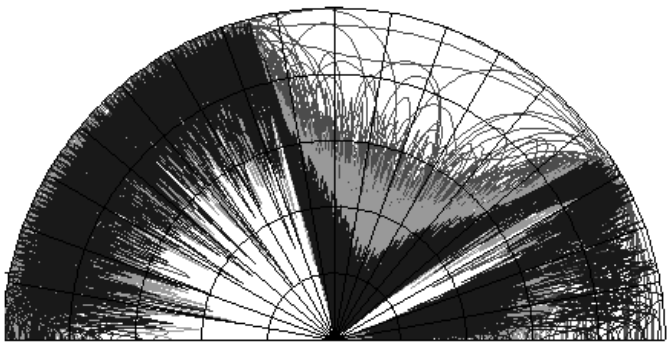
7.2 Equalization as a Function of Polar Selectivity

The polar responses show exceptional performance at mid to high frequency, but inevitably because of the finite



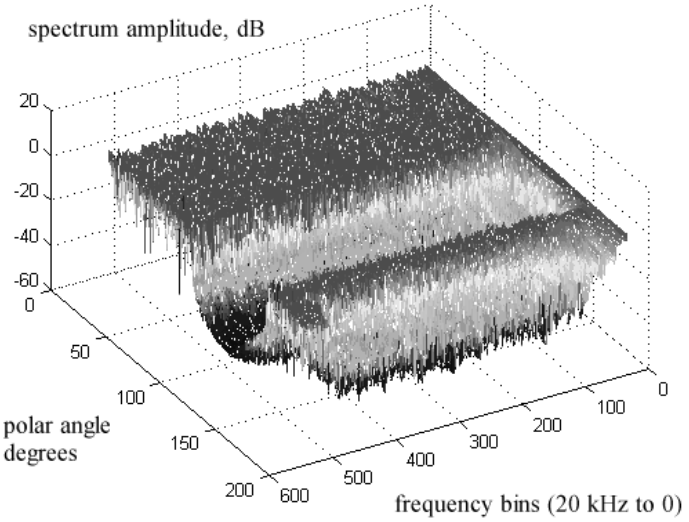
(d)

Fig. 19. (b) Three-dimensional polar plot, polar example 3.



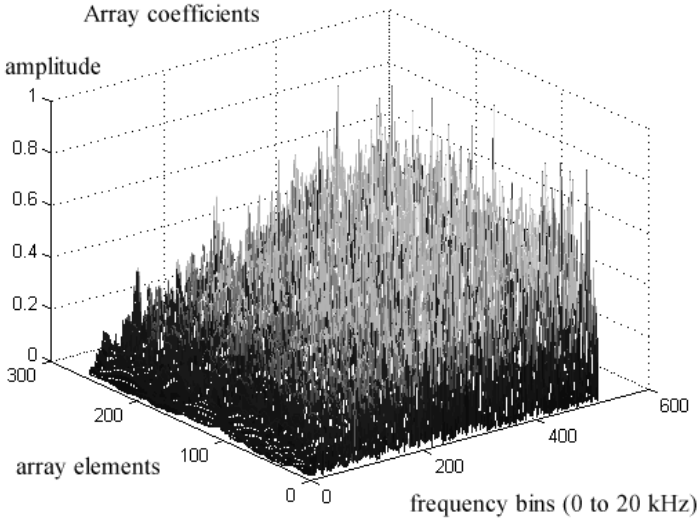
(a)

Fig. 20. (a) Cylindrical polar plot, search = 1.



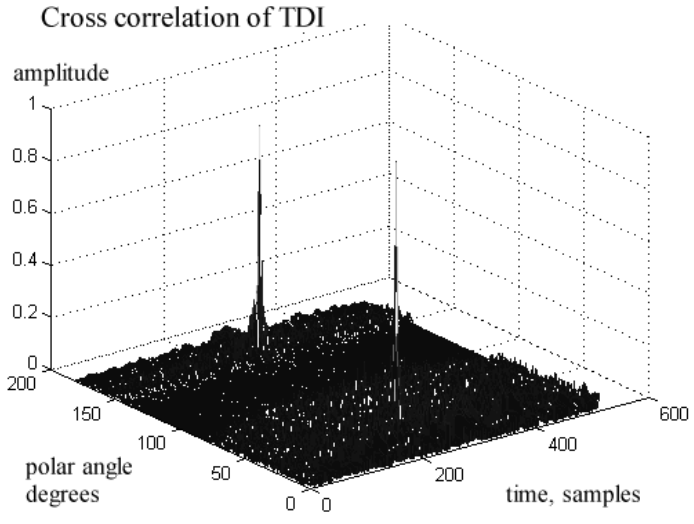
(b)

Fig. 20. (b) Three-dimensional polar plot, search = 1.



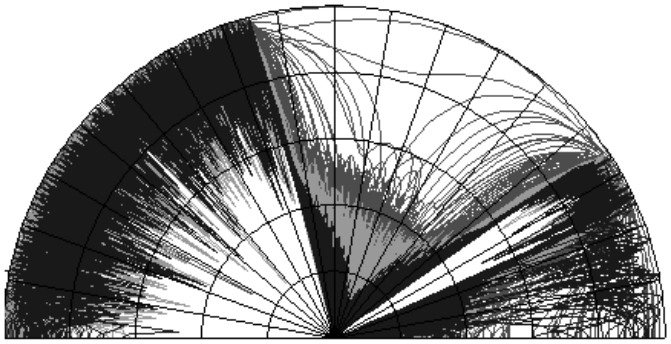
(c)

Fig. 20. (c) Three-dimensional coefficient plot versus frequency, search = 1.



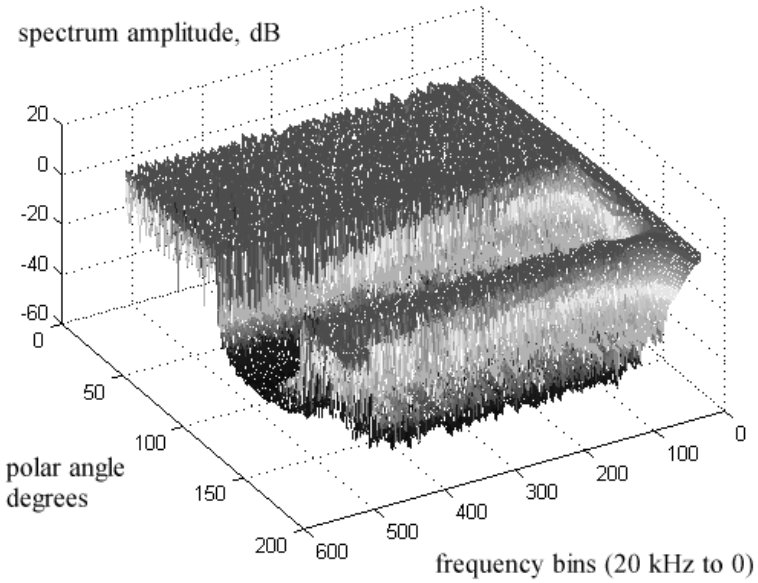
(d)

Fig. 20. (d) Three-dimensional cross correlation, β_1 reference response, search = 1.



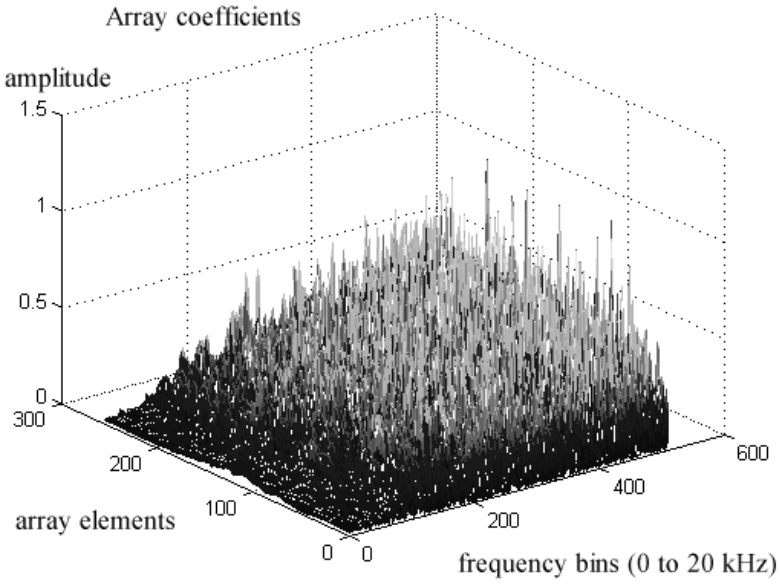
(a)

Fig. 21. (a) Cylindrical polar plot, search = 50.



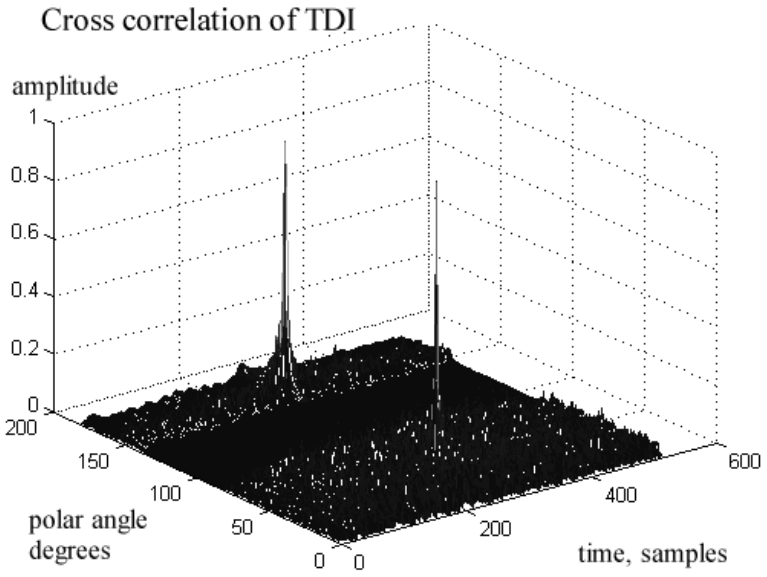
(b)

Fig. 21. (b) Three-dimensional polar plot, search = 50.



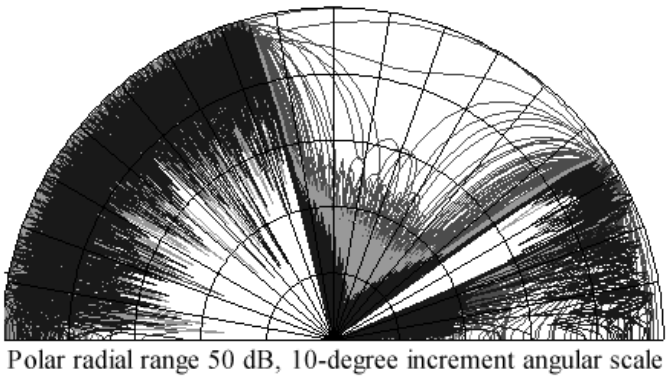
(c)

Fig. 21. (c) Three-dimensional coefficient plot versus frequency, search = 50.



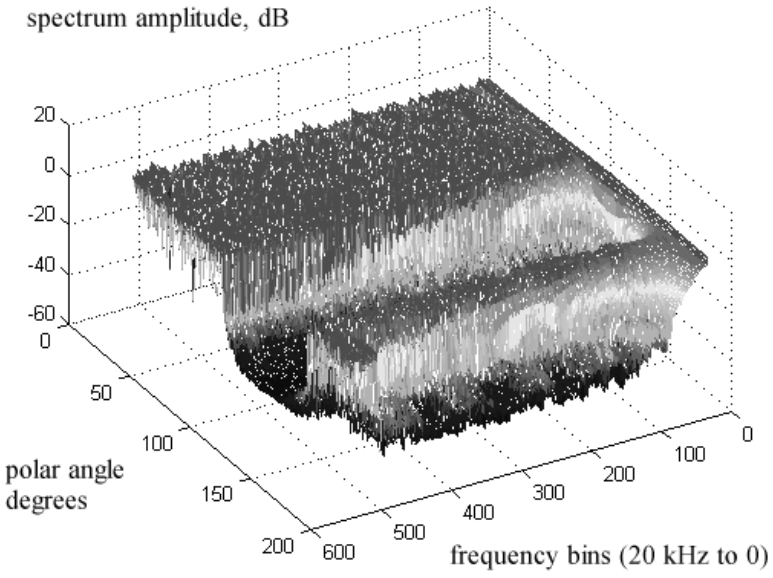
(d)

Fig. 21. (d) Three-dimensional cross correlation, β_1 reference response, search = 50.



(a)

Fig. 22. (a) Cylindrical polar plot, search = 1000.



(b)

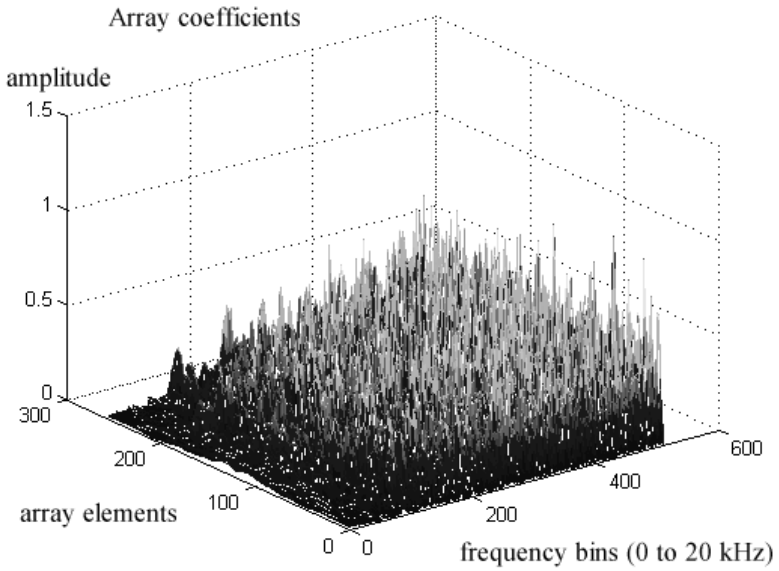
Fig. 22. (b) Three-dimensional polar plot, search = 1000.

array size the polar resolution degrades at low frequency. This implies that if arrays are to be driven in the lower frequency range, then a greater total energy is radiated into the listening space because of the broadening in polar response. Two possible strategies are either to bandlimit the input signal to avoid this problem area or to introduce equalization designed for constant-energy output, even though the on-beam response is compromised. On balance, equalization is the preferred option when an array is used in a closed space where the listener would hear significant low-frequency reflected contribution from the boundaries.

At each discrete frequency f prior to spline interpolation where the ECTFs are determined, a pressure calculation is performed at 1° polar intervals and at discrete design frequencies to estimate the cylindrical polar response. The total pressure response is then normalized by the maximum absolute pressure so the maximum pressure is unity irrespective of the monitoring distance. A weighting function can be estimated by calculating the root mean square of the pressure response σf, performed at each discrete design frequency f and integrated over a hemisphere, that is,

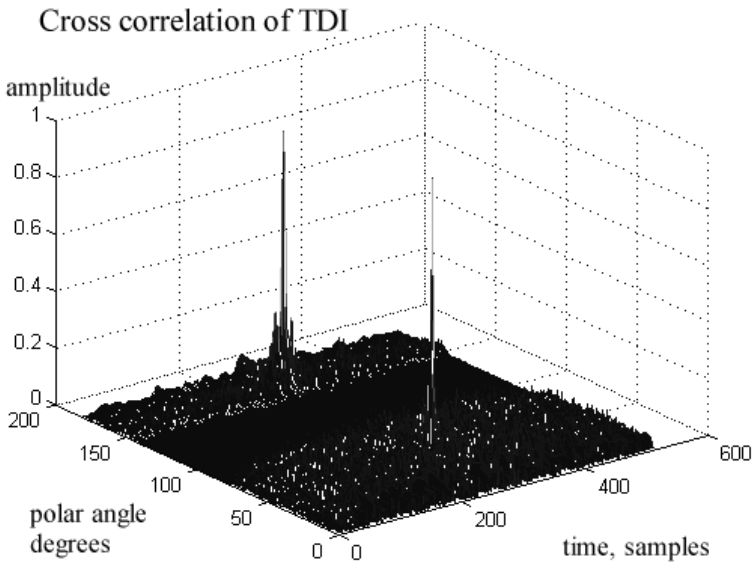
$$\sigma_f = \left[\frac{1}{181}\sum_{r=0}^{180}\mathrm{abs}\left(p(r)\right)^2\right]^{0.5}. \tag{19}$$

The sets of array coefficients corresponding to the frequency of f are then weighted by $1/\sigma_f$ to form the equalized weighting functions. To illustrate a typical equalization process, polar example 2 is repeated and the equalization characteristics are calculated for both coherent and stochastic designs, and the corresponding equal-



(c)

Fig. 22. (c) Three-dimensional coefficient plot versus frequency, search = 1000.



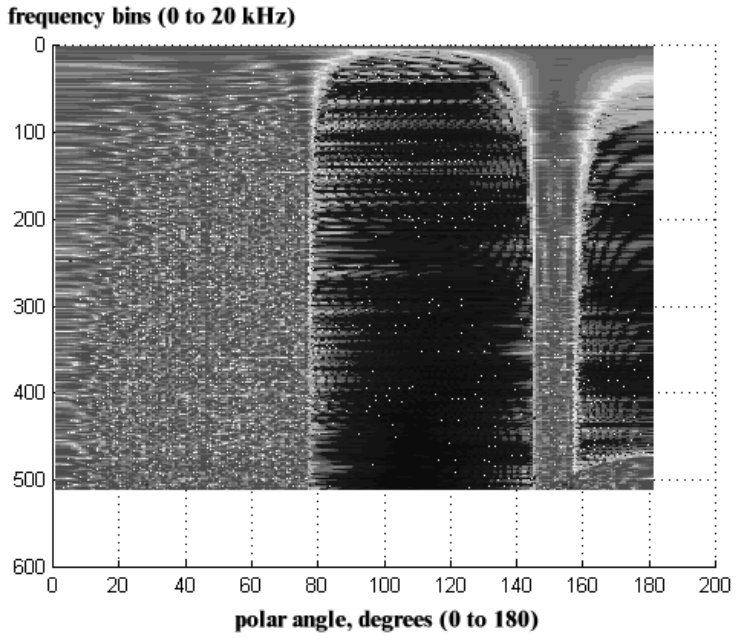
(d)

Fig. 22. (d). Three-dimensional cross correlation, β_1 reference response, search = 1000.

ization characteristics are shown in Figs. 23 and 24. Fig. 24 reveals a significant degree of high-frequency variation in the integrated response that is a consequence of the stochastic filters. The application of equalization to stochastic array filters is shown in the three-dimensional polar plot of Fig. 25. The plots appear similar to the nonequalized example, although a slight attenuation at low frequency is just evident on close inspection, which compensates for the broadening in the polar response. Finally the cross-correlation response if shown in Fig. 26, which should be compared with Fig. 18(d).

7.3 Frequency-Domain Structure and Diffusivity

The design process as implemented calculates N FIR filters, where N is also equal to the number of array elements. Consequently each array element is characterized by a frequency-dependent function with N frequency bins such that the array is represented as an $N \times N$ matrix. To enhance the frequency resolution prior to the time-domain transformation used to determine a set of TDIs [4], spline interpolation was used. Alternatively a greater number of filters could have been calculated, which would have



(e)

Fig. 22. (e) Two-dimensional polar plot [as Fig. 22 (c), top view], search = 1000.

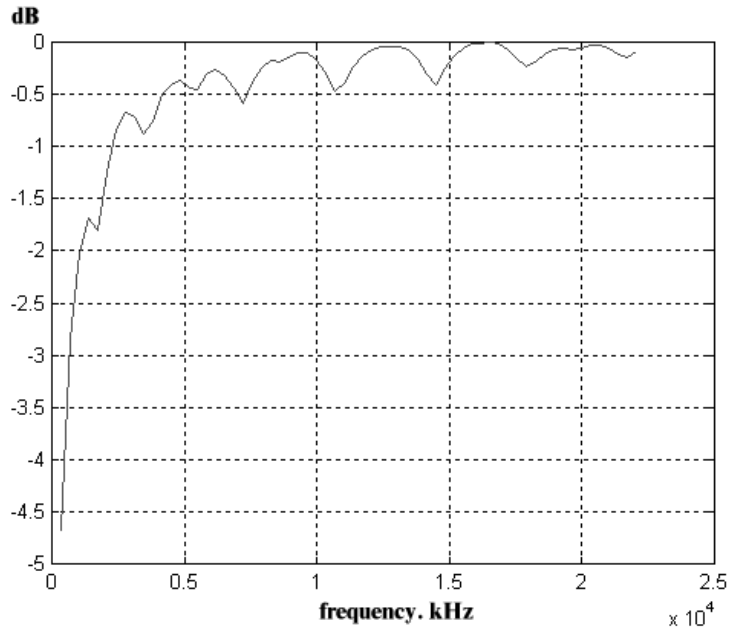


Fig. 23. Integrated power polar equalization, coherent design.

impaired the design time significantly. For coherent filter design, interpolation proves an efficient strategy whereas for the diffuse array an increase in the number of filters, each with a unique random characterization, is the better choice. This has the advantage of improved frequency-domain diffusivity [17] and leads to more focused cross-correlation functions with the corresponding improvements in the fineness of the noiselike structure exhibited in the polar response plots. This tradeoff is not explored in detail within the current version of the design program. However, some insight can be obtained by comparing earlier polar results for $N = 64$ with those of polar example 4, where $N = 256$.

8 CONCLUSION

The primary objective of this paper has been to develop a processing strategy to obtain directional radiation from an array of elemental transducers. A Fourier transform method was adopted for audio frequency operation in order to maintain the polar response shape over the signal frequency band. Transform methods have been used to design antenna systems, but generally these applications have very narrow passbands compared to their operating frequency. In a loudspeaker application operation over many octaves is required, thus a single transform is no longer sufficient. The current analysis is restricted to a

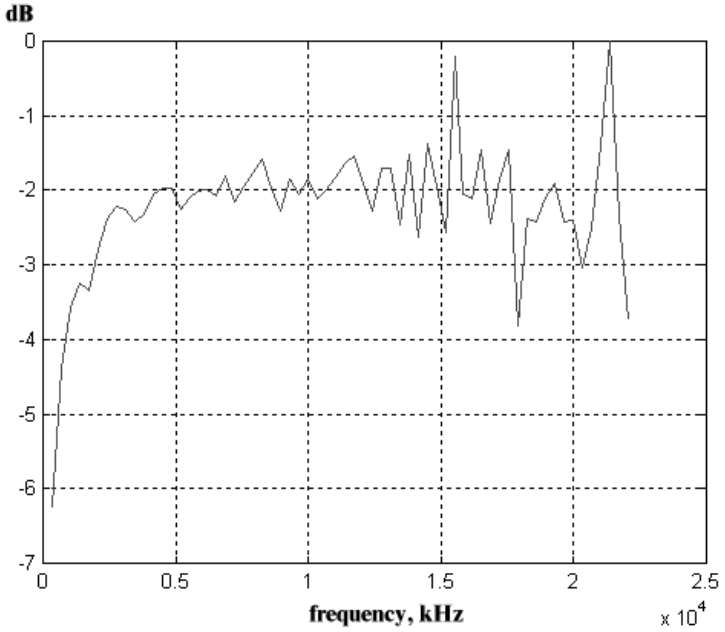


Fig. 24. Integrated power polar equalization, stochastic design.

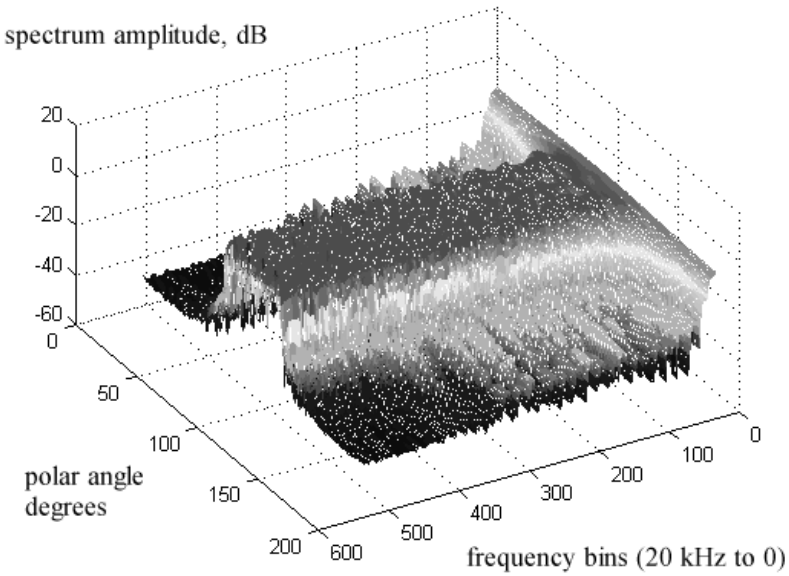


Fig. 25. Polar-frequency response with power-integrated equalization.

one-dimensional array, although the techniques are readily extendable to two dimensions, using two-dimensional filtering with two-dimensional Fourier transform. However, no account was made for any edge diffraction or structural reflections that may occur and cause interference with the main radiation.

The analysis has shown how a wide-band polar response can be achieved where beamwidth and beam direction can be freely programmed over a 180° arc. By using superposition, multiple beams can be defined and controlled independently, subject of course to processing power. Also coherent and stochastic filters can be mixed, allowing a composition of coherent and diffuse beams radiated simultaneously from a single array. In principle, the acoustic radiation could have part of its bandwidth diffuse and part coherent, should that be required, and could prove advantageous in specialized spatial audio applications. The sets of processor associated with individual beams could be driven either by the same signal or by different signals, allowing a single array to beam several signals in different directions. Once a set of ECTFs for a given beam is defined, they can be updated, morphed, or otherwise modulated to produce complicated time-varying acoustic fields. There is inevitably a heavy signal-processing penalty for such functionality, although technology evolution should accommodate this complexity in the medium-term future. Altogether this structure and approach to signal processing enables innovative applications to be conceived for smart array loudspeakers using multiple dynamic beam steering technology.

The study showed that element interspacing determines directly the upper frequency of the array, assuming that a maximum beamwidth of 180° is required. The lower frequency limit is ultimately a function of the array dimensions, although this is less well defined and exhibits significant yet more gradual deterioration with decreasing frequency in the lower frequency range. An equalization strategy was suggested in Section 7.2 to compensate for changing beamwidth at low frequency in order that the radiating power response is controlled. This is an important consideration within enclosed nonanechoic listening spaces.

Stochastic processing in a multifilter synthesis process was investigated, where the advantage of distributing signal loading more evenly across the array elements when compared to a coherent sinc function method was demonstrated using simulation. This can be observed by comparing the frequency-dependent array functions in a number of examples presented, such as Figs. 14(c)/17(c); 15(c)/18(c); 16(d)/19(d). Also, stochastic processing was shown to introduce a diffuse characterization similar to that of a DML, but with the inclusion of a programmable directional polar response. Spline interpolation and inverse Fourier transform methods were then used to forge a set of TDIs that can be embedded directly into an array of digital filters.

Cross-correlation analysis confirmed that the stochastic design method yielded a diffuse polar response, even when significant polar directionality was required. The only peaks exposed in the cross-correlation functions were the central on-axis functions associated with each beam, which is anticipated given that each beam is equalized to have an exactly flat on-axis central response. Nevertheless, the correlation plots show rapid decorrelation with the angle, where it should be observed that the angular resolution is set here to 1° intervals. In all diffuse two-beam examples, independent stochastic filters were used for each beam synthesis and the composite beam pattern was formed by superposition.

Finally, brief consideration was given to the general form of array element structure. It was suggested that each element could be fabricated from a multifaceted structure that can include both depth and area elements. This implies that each element is a multibit digital-to-acoustic converter, requiring the drive signals to be both noise shaped and randomized to decorrelate and disperse distor-

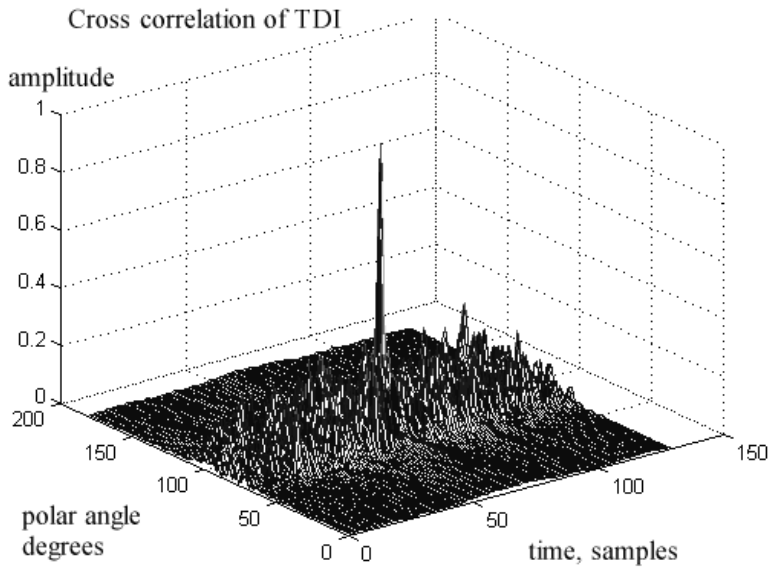


Fig. 26. Three-dimensional cross correlation, with equalization applied to functions.

tion over frequency. It is important that the physical dimensions of each element be sufficiently small that they radiate hemispherically. Earlier work has proposed that an array element should be addressed in binary to allow the terminology "digital loudspeaker" to be used correctly. This is a fuzzy area as it can be argued, for example, that a conventional loudspeaker fed directly with a 1-bit SDM data stream is a digital loudspeaker. Here the transducer itself must filter high-frequency noise components, although problems of heating and ultrasonic induced intermodulation distortion requires an additional low-pass filter, implying that transducer action is then definitely analog. More realistic and closer to the digital philosophy is the concept that each element is a multilevel device, as discussed in Section 1.

There are many challenges ahead with respect to fabrication to combine both nanomechanisms together with integrated drive electronics, especially if multiple inputs and dynamic beam control are to be included. However, to form an SDLA that can be associated with the concept of a digital loudspeaker to describe its mode of transduction, such challenges will have to be met.

9 ACKNOWLEDGEMENT

The author wishes to thank Henry Azima of NXT for permission to publish this study.

10 REFERENCES

[1] G. Bank and N. Harris, "The Distributed Mode Loudspeaker—Theory and Practice," presented at the AES UK Conf. "The Ins and Outs of Audio (London, 1998 Mar.).

[2] N. Harris and M. O. J. Hawksford, "The Distributed-Mode Loudspeaker (DML) as a Broad-Band Acoustic Radiator," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 1007 (1997 Nov.) preprint 4526.

[3] M. Petyt, *Introduction to Finite Element Vibration Analysis* (Cambridge University Press, Cambridge, UK, 1998).

[4] M. O. J. Hawksford and N. Harris, "Diffuse Signal Processing and Acoustic Source Characterization for Applications in Synthetic Loudspeaker Array," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 50, pp. 511–512 (2002 June), preprint 5612.

[5] R. Adams, "Unusual Applications of Noise-Shaping Principles," presented at the 101st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 1166 (1996 Dec.), preprint 4356.

[6] R. Adams, K. Nguyen, and K. A. Sweetland, "A 112-dB Oversampling DAC with Segmented Noise-Shaped Scrambling," presented at the 105th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 46, p. 1027 (1998 Nov.), preprint 4774.

[7] S. G. Kim, K. H. Hwang, and M. K. Koo, "Thin-Film Micromirror Array (TMA) for High Luminance and Cost-Competitive Information Display Systems," *SPIE Proc.*, vol. 3634, pp. 207–216 (1999 Jan.).

[8] K. Inanaga and M. Nishimura, "The Acoustic Characterization of Moving-Coil Type PCM Digital Loudspeakers," in *Proc. Spring Conf. of Acoustic Soc. of Japan* (1982), pp. 647–648.

[9] Y. Huang, S. C. Busbridge, and P. A. Fryer, "Interactions in a Multiple-Voiced-Coil Digital Loudspeaker," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 48, pp. 545–552 (2000 June).

[10] H. Takahashi and A. Nishio, "Investigation of Practical 1-bit Delta–Sigma Conversion for Professional Audio Applications," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 49, p. 544 (2001 June), preprint 5392.

[11] H. A. Spang and P. M. Schultheiss, "Reduction of Quantizing Noise by Use of Feedback," *IRE Trans. Commun. Sys.*, pp. 373–380 (1962 Dec.).

[12] S. K. Tewksbury and R. W. Hallock, "Oversampled Liner Predictive and Noise Shaping Coders of Order $N > 1$," *IEEE Trans. Circuits and Sys.*, vol. CAS-25, pp. 437–447 (1978 June).

[13] M. O. J. Hawksford, "Chaos, Oversampling and Noise Shaping in Digital-to-Analog Conversion," *J. Audio Eng. Soc.*, vol. 37, pp. 980–1001 (1989 Dec.).

[14] J. R. Stuart and R. J. Wilson, "Dynamic Range Enhancement Using Noise-Shaped Dither at 44.1, 48, and 96 kHz," presented at the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (Abstracts), vol. 44, p. 646 (1996 July/Aug.), preprint 4236.

[15] J. Vanderkooy and M. O. J. Hawksford, "Relationship between Noise Shaping and Nested Differentiating Feedback Loops," *J. Audio Eng. Soc.*, vol. 47, pp. 1054–1060 (1999 Dec.).

[16] P. H. Kraght, "A Linear Phase Digital Equalizer with Cubic-Spline Frequency Response," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 40, pp. 403–414 (1992 May).

[17] V. P. Gontcharov and N. P. R. Hill, "Diffusivity Properties of Distributed Mode Loudspeakers," presented at the 108th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 48, p. 350 (2000 Apr.), preprint 5095.

APPENDIX

The following Matlab program fragment can be used to search for a stochastic filter that achieves the lowest polar stopband error within a finite vector search range:

```
% commerce search loop x=1 to "search" for filter giving lowest error in polar stopband region
for x=1:search
% Fourier transform of noise vector with constant magnitude and random phase or coherent impulse
tmp=fft(λ*exp(i*phw*(round(rand(1,nc))−.5))+(1−λ)*[zeros(size(1:nc2−1))1+im*i zeros(size(1:nc2))]);
% frequency domain low-pass filter with mask scf
tmp=scf.*(tmp./abs(tmp));
% take inverse Fourier transform and time window to form FIR filter
```

```
imp=win.*ifft(tmp);
% normalize coefficients of filter to give maximum of unity
imp=imp/max(abs(imp));
% zero-stuff FIR vector to reduce effects due to transform circulatory
imp=[imp zeros(size(nc+1:m))];
% calculate magnitude spectrum from finite length impulse response
fimp=abs(fft(imp));
% determine standard deviation of frequency domain vector masked for only polar stopband region
err=std(fimp(cutoff+1:m-cutoff));
% sift for lowest mean-square error in the polar stopband taken over random search range
if err<err0
al(y,1:nc)=imp(1:nc)
err0=err;
end; end
```

THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and he studied delta modulation and sigma–delta modulation (SDM) for color television applications. During this period he also invented a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, Colchester, UK, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems, with a strong emphasis on audio systems including signal processing and loudspeaker technology. Since 1982 his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. The first one was developed in 1986 for a prototype system to be demonstrated at the Canon Research Centre and was sponsored by a research contract from Canon. Much of this work has appeared in *JAES*, together with a substantial number of contributions at AES conventions. He is a recipient of the AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," for the best contribution by an author of any age for *JAES*, volumes 45 and 46.

Dr. Hawksford's research has encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with special emphasis on SDM and its application to SACD technology. In addition, his research has included the linearization of PWM encoders, diffuse loudspeaker technology, array loudspeaker systems, and three-dimensional spatial audio and telepresence including scalable multichannel sound reproduction.

Dr. Hawksford is a chartered engineer and a fellow of the AES, IEE, and IOA. He is currently chairman of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA). He is also a technical consultant for NXT, UK and LFD Audio UK and a technical adviser for *Hi-Fi News* and *Record Review*.

PAPERS

Spatial Distribution of Distortion and Spectrally Shaped Quantization Noise in Digital Micro-Array Loudspeakers*

MALCOLM HAWKSFORD, *AES Fellow*
(mjh@essex.ac.uk)

University of Essex, Department of Electronic Systems Engineering, Colchester, Essex, CO4 3SQ, UK

A concept for a digital loudspeaker array, composed of clusters of micro-radiating elements that form individual digital-to-acoustic converters, is studied. In this scheme a large-scale array is composed of subgroups of micro clusters. To accommodate the finite resolution of each cluster, noise shaping is proposed and parallels are drawn with the processes used in digital-to-analog converters. Various elemental array geometries for each micro cluster are investigated by mapping transduction output into three-dimensional space to reveal the spatial distribution of both noise and distortion, which result from noncoincident and quantized digital-to-acoustic elements.

0 INTRODUCTION

Earlier work [1] has proposed a smart array loudspeaker composed of elemental radiators that form the individual transducers of the large-scale array. In the present study we consider in more detail the theoretical relationship between the geometric structures of the array elements and the resulting performance. Each element of the array may itself be composed of micro radiators which, for our purposes, will be considered as individual quantized digital radiators having limited amplitude resolution. Other pioneering work in this area by the team at B and W Loudspeakers and Brighton University has been presented using moving-coil drive units [2]–[5], although not specifically directed at small clusters of digital elements. However, it is believed that the cluster approach may well become feasible with the advent of nanotechnology and the physical integration of electronics and nanomachinery.

A cluster of micro radiators forms a digital-to-acoustic converter which by its nature is amplitude quantized. This structure is similar in concept to digital-to-analog converters (DACs), which use a conversion stage with limited amplitude resolution [6]. For example, such a stage may have binary weighted elements, or it may have a number of elements with equal weights. In these schemes, as proposed earlier [1], oversampling and noise shaping are used in order to achieve an adequate signal-to-noise ratio at low frequency. In 2004 Tatlas and Mourjopoulos [7] studied an array of idealized micro elements driven by 1-bit code generated using a sigma–delta modulator (SDM), where novel means were described for mapping bit patterns to both a one- and a two-dimensional cluster of elements. This work correctly identified the problem of off-axis distortion that can result in digital arrays of finite dimension. However, the proportionality between far-field pressure and radiator acceleration was not taken into account which, as shown in this paper, presents further complications for signal processing and especially the required volume displacement of an array at lower frequencies.

In the present study similar techniques are proposed, but here, unlike with the electrical DAC, there can be a spatial separation between the effective elements of the converter, which produce small time delays that are related to the polar coordinates of the monitoring location. Ultimately the size of the cluster determines the path differences between elements and a point in space, so there is inevitable interference and geometric dependent filtering that modifies the spatial distribution of the quantization noise, especially at high frequency. Also, noncoincidence of the elements will affect the linearity of the conversion process as this depends on the various path lengths and therefore the monitoring position in space. The spatial separation of micro radiators within a cluster is potentially both a problem and an asset as, effectively, it allows a degree of

*Presented at the 120th Convention of the Audio Engineering Society, Paris, France, 2006 May 20–23. Manuscript received 2006 April 28; revised 2006 November 8.

filtering to be achieved at higher frequencies. Also of specific interest is the use of decorrelation techniques [8] in the converter and how this relates to a cluster array that is small but not infinitesimal.

The paper examines the basic signal processing requirements and evaluates examples of the spatial distribution of the quantization noise and distortion. Matlab simulations of various clusters are performed, which includes both the physical geometry of a cluster and signal processing in terms of noise shaping and decorrelation. Fig. 1 shows the basic concept of a digital loudspeaker array, where each radiating surface is itself composed of clusters of micro elements. The concept, as reported earlier [1], presents these clusters as discrete subsystems, integrating both micro actuators and electronics within a common structure. It has been shown that in principle several clusters can be combined into an array to produce a loudspeaker with steerable attributes, where signal processing enables beam forming in terms of direction and width. In this scheme the interspacing between clusters established the upper spatial bandwidth of the array while total array size determined the low-frequency limit. Between these limits signal processing could then control beam width and direction over several octaves.

However, in this paper the focus is on the quantization noise and distortion performance of a single cluster in terms of its size, micro-element interspacing, and mode of transduction in relation to signal processing and noise shaping. In effect the problem of specifying the resolution of a digitally addressed surface is considered in order to achieve an acceptable signal-to-noise ratio. This problem is fundamental and theoretically distinct from the actual mechanics of transduction. Also, this study is of interest because where a transducer is conceived of discrete elements that are not spatially coincident, the spatial distribution with frequency of the radiated quantization distortion can exhibit far-field polar variations. However, the study here does not consider practical imperfections of micro elements other than their geometric distribution and mode of signal quantization. The aim therefore is to establish bounds on the principal characteristics and geometry of a cluster.

1 MODE OF OPERATION OF A SINGLE CLUSTER

The underlying structure of a digitally addressed cluster is an array of micro radiators, where each element has either a specific digital weight or, if multilevel, a restricted amplitude resolution. The loudspeaker is assumed to be fed a uniformly sampled digital signal. Thus for each sample a micro radiator creates an impulsive acoustic signal related here to displacement. The amplitude of the impulsive signal depends on the nature of the conversion process, where to gain insight into the requirements it is helpful to make comparisons with electrical DAC systems. Because, for practical reasons, it is desirable to limit the total number of quanta in a cluster to be less than the amplitude resolution of a sample of linear pulse-code modulation, it is expedient to exploit noise-shaping and oversampling techniques. Such techniques are well known and widely practiced in DAC systems [6], [8]. Two main options exist, one where the elements are given binary weights and the second where they are given equal weights. The equal-weight topology is considered more logical as each element then has the same physical form as

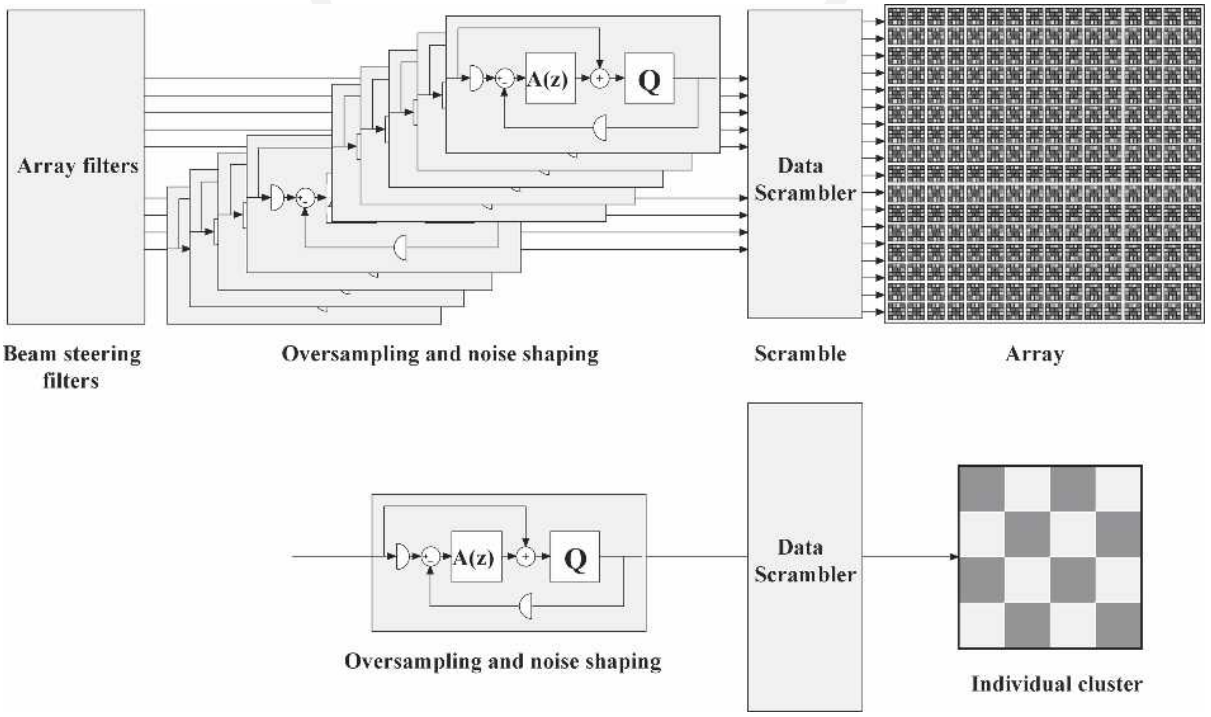


Fig. 1. Concept of digital loudspeaker composed of clusters of micro-radiating elements [1].

represented conceptually by the "checker board" in Fig. 1.

In considering the geometry and resolution of a cluster used at the core of a digitally controlled radiating surface, the acoustic relationship between far-field pressure and surface displacement must be taken into account. It is well known from conventional analog loudspeaker theory that the far-field acoustic pressure is proportional to the acceleration of a radiating surface and not to its displacement. Consequently this relationship must be embedded into the signal processing as the current proposal is to control element displacement in terms of a digital word. Fig. 2 shows a conventional moving-coil drive unit placed in a sealed enclosure which in normal usage is voltage controlled. Also shown is a basic low-frequency equivalent circuit that relates to the case where the radiating diaphragm can be considered pistonic. Defining:

$V_{in}(f)$ Input voltage to speech coil
$P_f(f)$ Far-field output pressure at remote monitoring point
$I(f)$ Input current in speech coil
$U(f)$ Diaphragm velocity
r_s Speech-coil resistance
L_s Static inductance of speech coil
r_{ma} Effective air-load resistance
$\overline{\beta\ell}$ Force factor averaged over length of speech coil in magnetic circuit
m_{ma} Effective mass of air load
m_{mc} Mass of cone and coil assembly
r_{ms} Resistive loss of suspension
c_{mT} Total compliance of suspension, spider, and enclosure
f Frequency
ω Angular frequency, $=2\pi f$
λ Constant of proportionality.

Equating far-field acoustic pressure to input driving voltage, taking into account the relationship between far-field pressure and diaphragm acceleration, and ignoring speech coil inductance, it follows that

$$\frac{P_f(f)}{V_{in}(f)} \approx \frac{\lambda(j\omega)^2 \dfrac{c_{mT}\overline{\beta\ell}}{r_s}}{1 + j\omega c_{mT}\left[r_{ma} + r_{ms} + \dfrac{(\overline{\beta\ell})^2}{r_s} \right] + (j\omega)^2 (m_{mc} + m_{ma}) c_{mT}} . \quad (1)$$

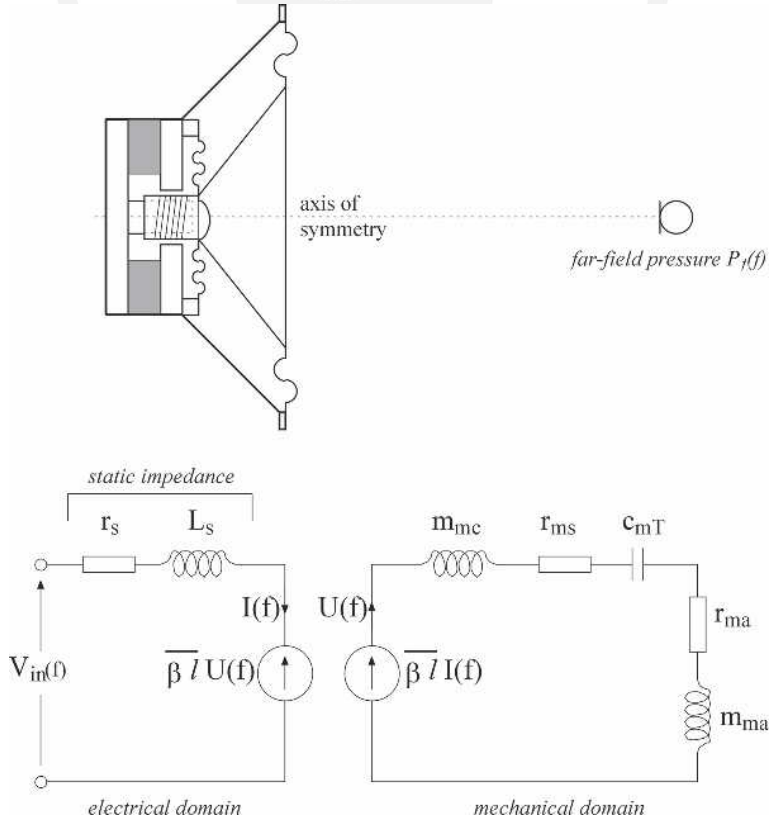


Fig. 2. Moving-coil drive unit and electrical equivalent circuit.

This transfer function can be compared directly with the standard form of a second-order high-pass filter, where

$$\frac{P_f(f)}{V_{in}(f)} = \frac{K\left(\frac{j\omega}{\omega_n}\right)^2}{1 + \frac{j\omega}{Q\omega_n} + \left(\frac{j\omega}{\omega_n}\right)^2} \tag{2a}$$

with K being the high-frequency far-field pressure to input-voltage gain, Q the quality factor controlling damping, and ω_n the undamped natural resonance. Comparing Eqs. (1) and (2a) it follows that

$$K = \frac{\lambda\overline{\beta\ell}}{r_s(m_{mc} + m_{ma})} \tag{2b}$$

$$\omega_n = \frac{1}{\sqrt{(m_{mc} + m_{ma})c_{mT}}} \tag{2c}$$

$$Q = \frac{c_{mT}\sqrt{m_{mc} + m_{ma}}}{r_{ma} + r_{ms} + (\overline{\beta\ell})^2/r_s} \tag{2d}$$

Eqs. (1) and (2) follow the well-known relationship that the diaphragm motion of an idealized moving-coil drive unit yields the precise dynamics required for sound reproduction, where within the second-order filter passband, diaphragm acceleration is controlled by the input voltage such that the far-field pressure is proportional to the applied voltage. Also, the drive unit's innate filter response provides a natural low-frequency limit, which for the moving-coil drive unit is determined by its fundamental resonance. Such behavior is a fundamental requirement as otherwise there would be no constraint on diaphragm displacement, which then would increase progressively at lower frequencies for a given sound pressure level.

This comparison with analog technology is made as similar physical constraints must be imposed upon a digitally addressed cluster, especially as it is the displacement of the radiating surface which is under the direct control of the input. Consequently quantization distortion present in the displacement of an element cluster must be processed by a second-order differential equation in order to derive an expression in the time domain for acoustic far-field pressure $\widetilde{p_f}(t)$, where for displacement $\tilde{x}(t)$,

$$\widetilde{p_f}(t) = K_x \frac{d^2\tilde{x}(t)}{dt^2} \tag{3}$$

with K_x being a constant. Because of Eq. (3) it is necessary for a quantized displacement-driven cluster to have an embedded complementary double integrator function in the input-stage processor. However, to impose a finite bound on transducer displacement at low frequency while in the cluster passband for the input signal to relate to acceleration when the output signal informs element displacement, a second-order low-pass filter is required. It is proposed in this study to use a second-order Butterworth filter alignment, where the z-transform process is shown in Fig. 3. Assume in the z domain that in(z) represents the linear pulse-code modulation source signal, out(z) the digital signal addressing the displacement of cluster elements, $m(z)$ the output of the input difference stage, $p_f(z)$ the far-field pressure, and P_1, P_2 the respective integrator and air propagation constants.

Analyzing the filter shown in Fig. 3, then relating out(z) to $m(z)$,

$$m(z) = P_1^{-2}(1 - z^{-1})^2\,\text{out}(z). \tag{4}$$

Considering signals at the input difference stage,

$$m(z) = \text{in}(z) - \text{out}(z)z^{-1}[1 + \sqrt{2}\,P_1^{-1}(1 - z^{-1})]$$

whereby eliminating $m(z)$ from Eq. (4) yields the second-order low-pass filter response,

$$\frac{\text{out}(z)}{\text{in}(z)} = \frac{1}{P_1^{-2}(1 - z^{-1})^2 + \sqrt{2}\,z^{-1}P_1^{-1}(1 - z^{-1}) + z^{-1}}. \tag{5}$$

Noting that the differential operation $1 - z^{-1} \equiv 1 - e^{-j2\pi f/f_s}$, where f_s is the sampling rate, then for $f \ll f_s$,

$$1 - z^{-1} \equiv 1 - \cos(2\pi f/f_s) + j\,\sin(2\pi f/f_s) \Rightarrow j2\pi f/f_s.$$

Consequently the filter expressed as a continuous time function $G(f)$ becomes

$$G(f) = \frac{1}{1 + \sqrt{2}\,j2\pi P_1^{-1}f/f_s + (j2\pi P_1^{-1}f/f_s)^2} \tag{6}$$

where Eq. (6) describes the standard second-order Butterworth low-pass filter.

Fig. 3 also includes a second-order digital differentiator to model the relationship between element displacement and far-field pressure. Since signal out(z) represents the element displacement driving function, the far-field pressure $p_f(z)$ is

$$p_f(z) = K_x\left(\frac{1 - z^{-1}}{P_2}\right)^2 \text{out}(z). \tag{7}$$

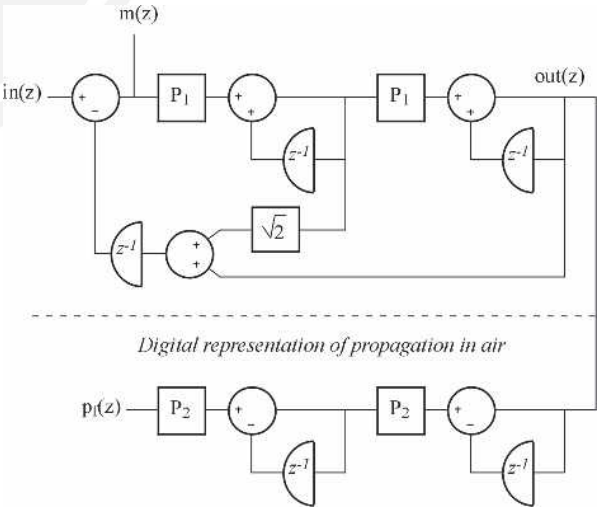


Fig. 3. Second-order filter to compensate for displacement to far-field pressure transduction.

Hence from Eq. (4) the relationship between $m(z)$ and out(z) is

$$p_f(z) = K_x\left(\frac{P_1}{P_2}\right)^2 m(z). \tag{8}$$

Eq. (8) reveals that the input error signal in the second-order filter shown in Fig. 3 corresponds directly to the far-field pressure, whereby From Eqs. (4) and (5) it is observed that

$$\frac{m(z)}{\text{in}(z)} = \frac{P_1^{-2}(1-z^{-1})^2}{P_1^{-2}(1-z^{-1})^2 + \sqrt{2}\,P_1^{-1}(1-z^{-1}) + 1} \tag{9}$$

or alternatively in terms of $p_f(z)$,

$$\frac{p_f(z)}{\text{in}(z)} = \frac{K_x P_2^{-2}(1-z^{-1})^2}{P_1^{-2}(1-z^{-1})^2 + \sqrt{2}\,P_1^{-1}(1-z^{-1}) + 1}. \tag{10}$$

Eq. (10) reveals that far-field pressure expressed as a function of the input signal is characterized by a second-order high-pass filter function and thus exhibits a response similar to that of a moving-coil drive unit when mounted in a sealed enclosure. A principal advantage of the filter shown in Fig. 3 is that it allows the correct pressure response to be established in the far field while also controlling the dc conditions of the integration process. The performance compromise is a reduction in output at lower frequencies. However, when the practical displacement constraints of each driving element are taken into account, then such a response is mandatory. This technique produces a practical solution, which also emulates the response of conventional analog transducers.

Because the elements forming the cluster are assumed both quantized and to have finite amplitude resolution, the signals processed by the high-pass filter must also be quantized appropriately and constrained in amplitude to match the properties of the cluster. Because it has been shown that signal $m(z)$ in Fig. 3 is proportional to far-field pressure, it is proposed to introduce multilevel SDM at this location. The SDM is designed to have a limited amplitude range and to produce a uniformly quantized output with a unit quantum. However, because the output of the SDM is processed subsequently by the two digital integrators, it then follows that the low-pass filter output is also uniformly quantized though the quantum is scaled by a factor P_1^2. Integer quanta are thus restored by scaling the output by P_1^{-2}. The scaled output out$_s(z)$ then follows directly from Eq. (4),

$$\text{out}_s(z) = \frac{1}{(1-z^{-1})^2}\,m(z). \tag{11}$$

In the next signal processing stage the quantized and noise-shaped signal out$_s(z)$ is coded and applied to the multielement cluster, where the conceptual system that includes the multilevel SDM embedded within the second-order low-pass filter is shown in Fig. 4.

The coding module that is used to map the signal out$_s(z)$ to the cluster together with an example of cluster geometry is described in Section 3. However, the architecture in Fig. 4 is sufficiently flexible such that by specifying the code converter stage, the driver weights, and the interelement time delays a wide range of geometry can be accommodated together with the calculation of the corresponding polar response.

2 MULTILEVEL SDM

The SDM topology selected for this study is shown in Fig. 5. It uses a parametric noise-shaping architecture [9], [10] where the core structure of principal integrators (here designated I_1 to I_5 is supplemented by a series-connected cascade of second-order biquadratic parametric equalizers. The parametric stages can augment the noise-shaping transfer function (NSTF) by embedding both band-pass and low-pass frequency-selective enhancement to the closed-loop gain of the overall noise shaper. It is critical to be able to control the NSTF beyond that achievable by just the principal integrators. It is also desirable to limit the oversampling ratio. Otherwise the range of signals output

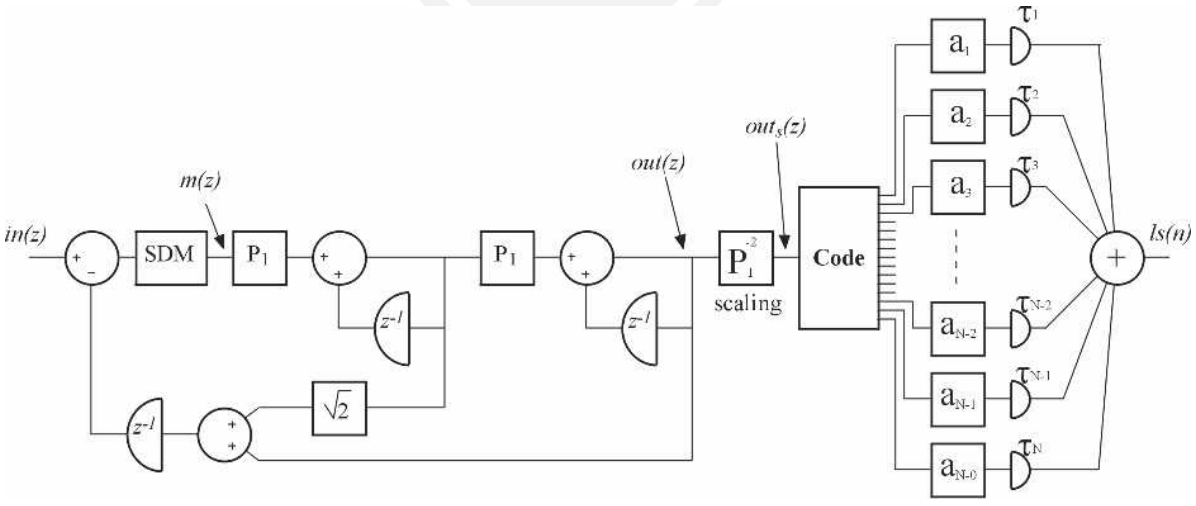


Fig. 4. Complete system structure with multilevel SDM, second-order high filter, and array driver.

from the system shown in Fig. 4 can become excessively large, thus demanding an extremely high number of elements in each cluster. Because the oversampling ratio is constrained, then in order to achieve an acceptable signal-to-noise ratio, the SDM loop must employ an M-bit multilevel quantizer Q, which also offers the advantage of relaxing the stability conditions compared to binary SDM.

The biquadratic transfer function $\mathrm{PF}_r(z)$ of the r^{th} parametric filter stage is given as

$$\mathrm{PF}_r(z) = \frac{g_1(r)p^{-1}(r)k(r)(1-z^{-1}) + g_2(r)}{p^{-2}(r)(1-z^{-1})^2 + k(r)p^{-1}(r)z^{-1}(1-z^{-1}) + z^{-1}} \tag{12}$$

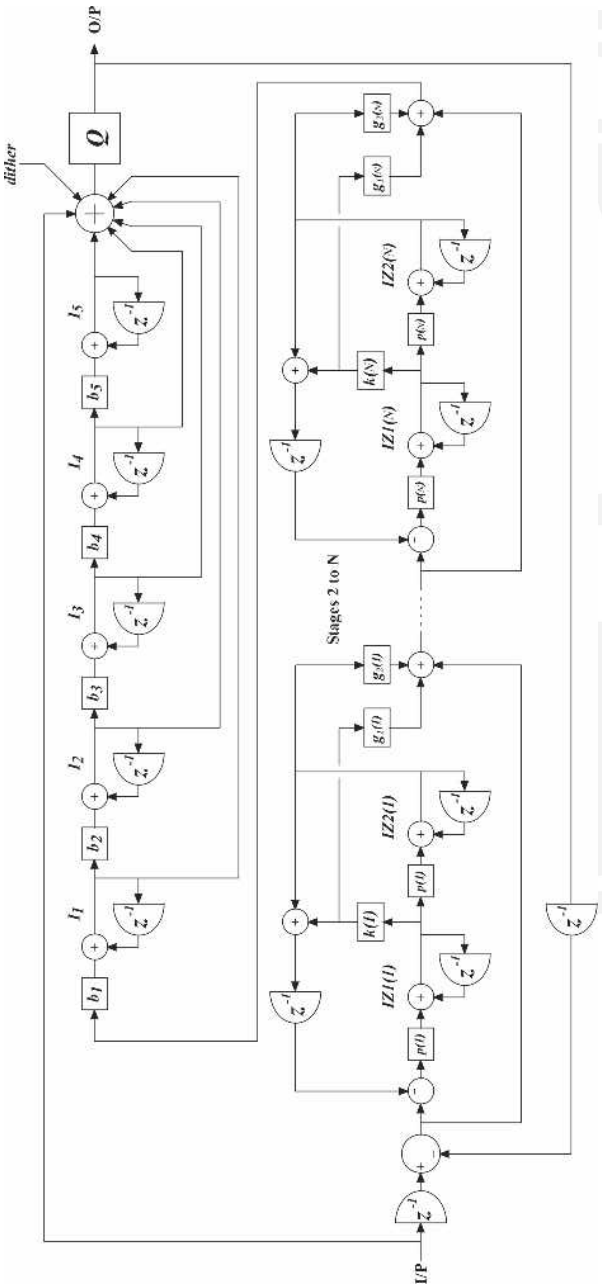


Fig. 5. Parametric SDM topology showing N parametric stages and five principal integrators.

where the scale factor $p(r)$ sets the resonant frequency, $k(r)$ sets the Q factor, and $g_1(r)$ and $g_2(r)$ are the insertion gains of the band-pass and low-pass parametric responses, respectively. The overall transfer function of the N-stage series parametric equalizer $\mathrm{PT}_N(z)$ is

$$\mathrm{PT}_N(z) = \prod_{r=1}^{N} [1 + P_r(z)]. \tag{13}$$

The transfer function of the R principal integrators is

$$\mathrm{AP}_R(z) = \sum_{h=1}^{R} \left[\prod_{s=1}^{h} b_s \left(\frac{1}{1-z^{-1}} \right)^s \right] \tag{14}$$

and the NSTF of the overall system is

$$\mathrm{NSTF} = \frac{1}{1 + z^{-1}\mathrm{AP}_R(z)\mathrm{PT}_N(z)}. \tag{15}$$

To illustrate the performance of parametric multilevel SDMs, a simulation is shown in Fig. 6 for a system employing 16 times Nyquist oversampling with three principal integrators and a three-stage parametric equalizer. The 24-bit quantized input signal consists of two sine waves, each with peak amplitudes of 0.5 and respective frequencies of 15 and 17 kHz. The uniform quantizer in this simulation has a quantum interval of 1 and is unbounded so that the natural output signal range of the SDM can be observed. Fig. 6(a) shows the output spectrum without the parametric filters, whereas Fig. 6(b) shows an example with the parametric filters adjusted to give enhanced noise reduction up to about 20 kHz. Fig 6(c) and (d) displays the corresponding output signal histograms with and without the parametric equalizer. In both examples the principal integrator weightings were set to $b_1 = 1$, $b_2 = 0.5$, and $b_3 = 0.25$, following conventional SDM practice [11].

The inclusion of three parametric stages within the SDM loop allows a substantial improvement in noise-shaping performance, where tailoring of the loop parameters has enabled a 24-bit resolution to be achieved up to a frequency of about 20 kHz. Also included in Fig. 6(a) and (b) is a reference spectrum of the input signal that has been quantized to 24 bit and also band-limited to 22 kHz. The parametric noise shaper (see Fig. 5) is defined by the following parameters:

$$\begin{array}{llll} f(1) = 18750 & k(1) = 0.005 & g_1(1) = 100 & g_2(1) = 16 \\ f(2) = 11250 & k(2) = 0.001 & g_1(2) = 100 & g_2(2) = 32 \\ f(3) = 7500 & k(3) = 0.01 & g_1(3) = 500 & g_2(3) = 64 \end{array}$$

where for an SDM sampling rate f_s the two identical integrator constants within the rth parametric stage are

$$p(r) = 2\pi \frac{f(r)}{f_s}. \tag{16}$$

Applying these parameters to Eqs. (12)–(16) the theoretical NSTF shown in Fig. 7 is formed. Inclusion of the parametric stages produces increased quantizer activity, as revealed in the histograms of Fig. 6(c) and (d). However,

the reward is a significant improvement in the signal-to-noise ratio within the band of 0–20 kHz. The increase in quantizer activity is to be anticipated as information theory applied to noise-shaping theory, as shown by Gerzon and Craven [12], requires the quantization noise power to be conserved.

The next stage in system development is to embed the SDM within the second-order high-pass filter architecture

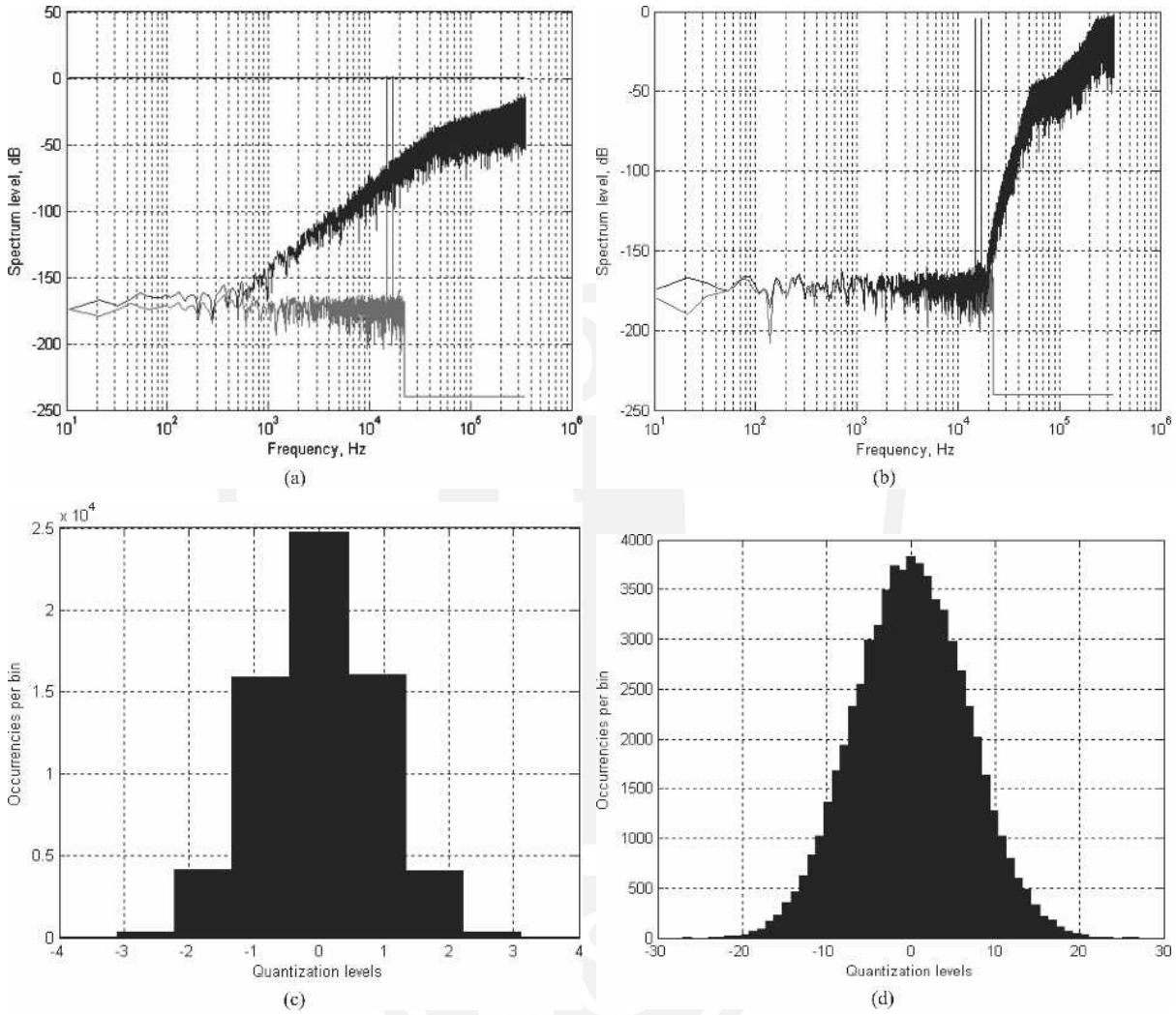


Fig. 6. SDM spectra and histograms. (a), (c) No parametric NSTF. (b), (d) Parametric NSTF.

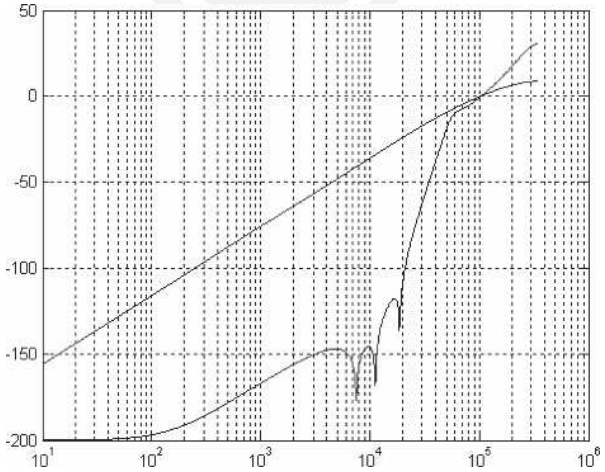


Fig. 7. Theoretical NSTF, both with and without parametric stages.

shown in Fig. 4. Initially the output is taken from the second integrator (ignore for the moment the coder and element drive stage) as this is a signal proportional to driver displacement. Then linking this displacement-defining signal out(z) by Eq. (7) to the far-field pressure $p_f(z)$, the overall response of SDM and the second-order filter can be determined. By way of illustration, Fig. 8(a) and (b) shows two output spectra computed for the far-field pressure where the high-pass filter is set to have the respective low-frequency bandwidths of 500 Hz and 5 kHz. In Fig. 8(c) a histogram of out(z) is shown in order to illustrate the range of amplitudes that the transducer cluster must be able to handle. Simulations revealed that the output histogram did not vary significantly as a function of the low-frequency filter bandwidth.

So far the results appear to support that very high resolution approaching 24 bit in the audio band can be obtained using a displacement-driven cluster with relatively low resolution, as Fig. 8(c) suggests an amplitude range of only about −64 to 64. However, practical acoustics as encapsulated by Eq. (7) demand that at lower signal frequencies, volume displacement must increase in order to maintain the corresponding pressure levels. To illustrate this observation, the simulation used to produce the results in Fig. 8 is repeated (but only for a low-frequency filter bandwidth of 500 Hz) with the input signal frequencies lowered by a factor of ten to 1.5 and 1.7 kHz, but with the input levels of the two sinusoidal inputs maintained at 0.5. The corresponding far-field spectrum and displacement histogram are shown in Fig. 9. An interesting characteristic of the displacement histogram is revealed both here and in Fig. 8(c) in that for a given signal, the signal amplitude boundaries of the histogram are relatively abrupt and do not show a classic Gaussian-like distribution.

Fig. 9(b) reveals that there is about a 100-fold increase in displacement, that is, recall displacement and far-field pressure are linked by double integration due to acceleration dependence. This implies that in order to maintain a broad-band pressure capability the resolution of the quantized displacement-driven cluster must collectively embrace at least several thousand quanta. The conclusion therefore is that a cluster must have either a commensurate number of binary-driven elements or a more limited number of elements must be capable of multilevel displacement transduction. At this point the conceptual vision of the cluster needs to be revisited as the required dynamic

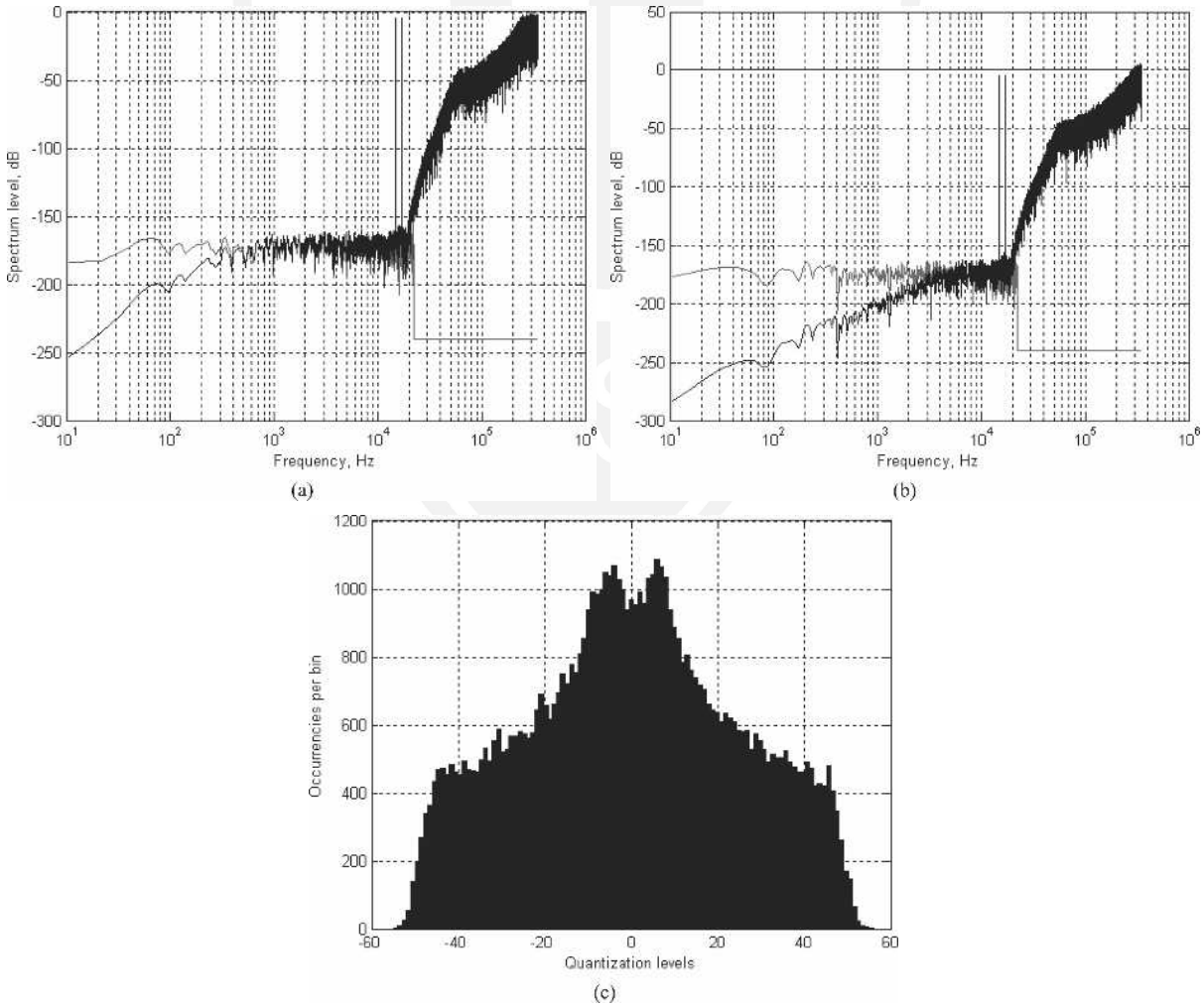


Fig. 8. (a) Far-field SDM spectrum; 500 Hz. (b) Far-field SDM spectrum; 5 kHz. (c) Histogram of displacement out (z).

range dictates that a small acoustically radiating surface must be populated by a large number of micro elements, each of which is capable of being driven over a limited yet quantized amplitude range. Given that nanotechnology is the principal contender for engineering a solution to this problem, at a theoretical level this requirement, though daunting with conventional technology, is not necessarily unrealizable. Consequently the study moves forward on the assumption that such a class of radiator with a large number of controlled micro elements will become feasible. Thus our motivation is to understand the requirements of such a structure and to shed light on possible system configurations as well as potential innate limitations. Nevertheless, the concepts and simulations presented in this section set a foundation for the signal processing that demonstrates noise shaping to be a potential means for lowering the amplitude resolution of a cluster while revealing how embedding SDM within a second-order high-pass filter can offer a signal processing solution to the electromechanical-acoustical requirements of transduction. In Section 3 SDM and the second-order high-pass filter are combined with a cluster driving stage in order to explore the limitations imposed by cluster geometry and code-book design.

3 MODELING CLUSTER GEOMETRY AND EXAMPLE CONFIGURATIONS

In mapping the displacement function derived in Section 2 to a specific array geometry a generalized approach is taken, as depicted in Fig. 4, in terms of element weights and their relative time delays as a function of the location of the listening position. Initially a single element $P(p, q, 0)$ is considered using the geometry shown in Fig. 10.

Consider a single radiating element located in a Cartesian coordinate system at point $P(p, q, 0)$ which resides in the xy plane defined here as the plane of a two-dimensional cluster. The soundfield is monitored at a location $M(r, \theta, \phi)$, where (r, θ, ϕ) are the spherical coordinates as defined in Fig. 10. However, for far-field monitoring as assumed in this analysis take $r \to \infty$. Drawing vector $\overrightarrow{RP}$, which is perpendicular to vector $\overrightarrow{OR}$, the magnitude of $\overrightarrow{OR}$, designated Δ, represents the path difference in propagation terms between an element located at $O(0, 0, 0)$ and at $P(p, q, 0)$ for the far-field monitoring location $M(r, \theta, \phi)$.

Applying vector addition to triangle OPR in Fig. 10,

$$\overrightarrow{RP} = \overrightarrow{OP} - \overrightarrow{OR}.$$

Defining $\hat{i}, \hat{j}, \hat{k}$ as unit factors in Cartesian space, then for line $O(0, 0, 0)$ to $R(x, y, z)$,

$$\overrightarrow{OR} = \hat{i}x + \hat{j}y + \hat{k}z$$

and for line $O(0, 0, 0)$ to $P(p, q, 0)$,

$$\overrightarrow{OP} = \hat{i}p + \hat{j}q + \hat{k}0$$

giving

$$\overrightarrow{RP} = (\hat{i}p + \hat{j}q + \hat{k}0) - (\hat{i}x + \hat{j}y + \hat{k}z).$$

For the right-angle triangle OPR,

$$\left|\overrightarrow{OR}\right|^2 = \left|\overrightarrow{OP}\right|^2 - \left|\overrightarrow{RP}\right|^2$$

and substituting for $\overrightarrow{OR}$, $\overrightarrow{RP}$ and putting $\left|\overrightarrow{OR}\right| = \Delta$,

$$\Delta^2 = (p^2 + q^2 + 0^2) - [(p - x)^2 + (q - y)^2 + z^2] \qquad (17)$$

which reduces to

$$\Delta^2 = 2px + 2qy - x^2 - y^2 - z^2.$$

From the spherical coordinates of $M(r, \theta, \phi)$, the Cartesian coordinates of $R(x, y, z)$ follow,

$$x = \Delta \sin(\phi) \sin(\theta)$$

$$y = \Delta \sin(\phi) \sin(\theta)$$

$$z = \Delta \cos(\phi).$$

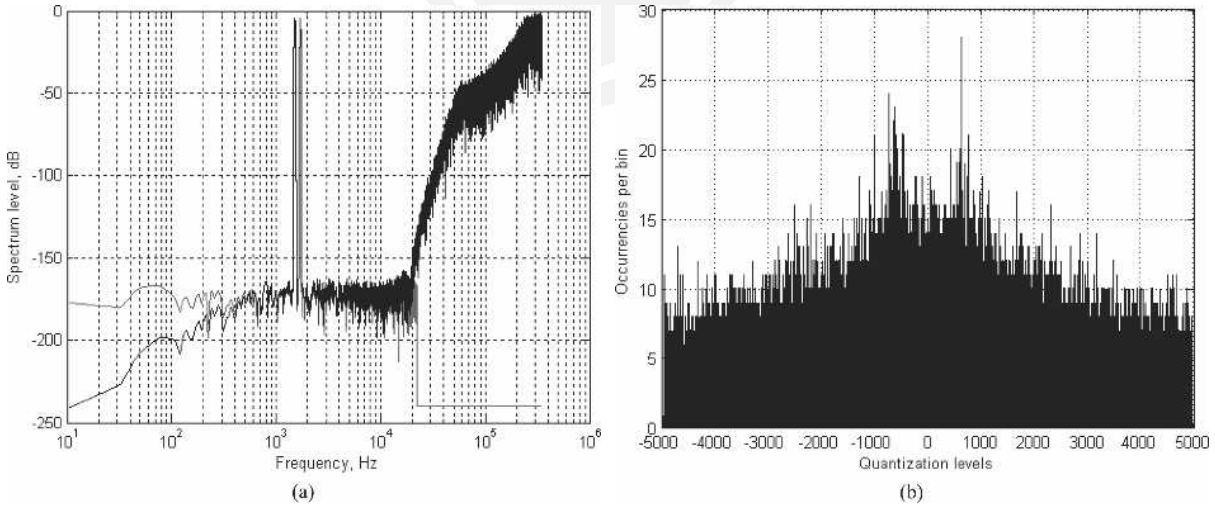


Fig. 9. (a) Far-field SDM spectrum; 500 Hz. (b) Histogram of displacement out (z).

Hence by substitution,

$$\Delta = \sin(\phi)[p\cos(\theta) + q\sin(\theta)]. \quad (18)$$

Consequently knowing the spherical coordinates of the listening location $M(r, \theta, \phi)$ and the coordinates of a general element $P(p, q, 0)$, the interelement time difference τ_p with respect to an element located at the origin $O(0, 0, 0)$ can be determined from Eq. (18) as

$$\tau_p = \frac{\Delta}{c} = \frac{1}{c}\sin(\phi)[p\cos(\theta) + q\sin(\theta)] \quad (19)$$

where c is the velocity of propagation of sound in air. Consequently if element number p has weight a_p, then for an N-element cluster, the time-domain output $\mathrm{ls}(n/f_s)$, as depicted in Fig. 4, takes the form

$$\mathrm{ls}\left(\frac{n}{f_s}\right) = \sum_{p=1}^{N} a_p \mathrm{cout}_p\left(\frac{n}{f_s} - \tau_p\right) \quad (20)$$

where $\mathrm{cout}_p(n/f_s - \tau_p)$ is the element-coded and time-delayed drive signal for sample n formed at the output of the code converter (see Fig. 4). Consequently by applying Eq. (20) the noncoincidence of each element in a cluster can be taken into account.

In this study the embedded SDM within the second-order high-pass filter shown in Fig. 4 is simulated to produce a time-domain output. This signal is then applied to the code converter, which in turn outputs N parallel signals to drive the individual elements of a cluster. The number of elements and the amplitude range of the input to the code converter depend on whether the drive signals are binary, tristate, or multilevel. The simulation allowed specification of the total number of elements and then automatically determined whether the drive signals required multilevel capability. As such the system did not have to be redesigned as a function of amplitude range but could simply identify the requirements of an individual element in terms of required amplitude resolution. Also the physical locations of the N elements could be specified to establish specific array geometries and to accommodate the overall size of the cluster. The code converter could also be configured to change the signal assignment within the cluster, a process that can be deterministic, geometry related, or random, the latter to enable decorrelation techniques to be explored. Finally the cluster geometry and the far-field listening position were accounted for through Eq. (18), which was used to calculate the path differences for each element. This technique allowed the noise-shaping spectrum to be expressed in terms of a polar response. This is an especially critical performance aspect as an off-axis monitoring location moderates the interelement time delays as described by Eq. (19), thus potentially causing degradation in noise-shaping performance that is akin to time jitter in more conventional digital systems.

To illustrate degradation in noise-shaping performance as a function of the far-field monitoring position an initial example is considered where a square cluster is formed from a rectilinear array of $W \times W$ elements, as shown in Fig. 11. The dimension of cluster S is also specified. The system allows each element to be multilevel with a quantization interval of unity such that the levels are {. . . −3 −2 −1 0 1 2 3 . . . }. This implies that a tristate element {−1 0 1} could accommodate a signal range of $-W^2$ to W^2. In the

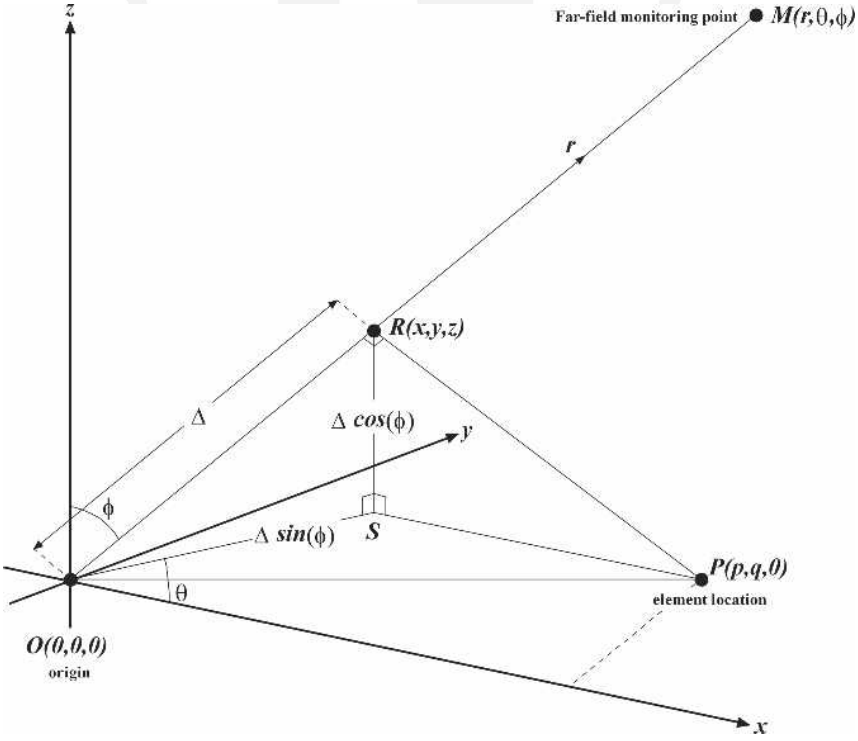


Fig. 10. Geometry to determine path difference Δ for single element.

first example the code converter stage takes the quantized output of the high-pass filter (Fig. 4) and applies it in a regular manner to the array using a two-dimensional "thermometer" style code [8]. This means that if for sample n the sample value is ls(n), then the element values must sum to ls(n), thus conserving the sample value. Also initially groups of 1 or −1 are clustered.

Since the coordinates of each element are known, the interelement time delays [reference to the origin $O(0, 0, 0)$] can be calculated using Eq. (19). A simulation using the same SDM parameter regime as described in Section 2 was performed for an array size of $S = 10$ mm with $W = 8$. The input was again two sine waves of amplitude 0.5 with frequencies of 15 and 17 kHz in order keep the signal range small for this exploratory stage of the study.

Simulations revealed a spectral response identical to that of Fig. 8(a) for an on-axis far-field monitoring location. However, shifting only 0.1° resulted in some degradation, as indicated in the spectral response in Fig. 12(a), while Fig. 12(b) shows the histogram of the interelement time delays, where the histogram reveals a time-delay range of about ±25 ns. The latter spectrum reveals that there is significant intermodulation distortion, which arises because of the noncoincidence of the elements within the cluster, effectively causing different signal levels to have slightly different time delays. In an attempt to decorrelate this distortion from the signal, a randomization function was introduced into the code module to associate individual drive signals and elements randomly. The resulting spectrum is shown in Fig. 12(c) and reveals that the distortion has become noiselike.

The simulations for the rectilinear array geometry show that there is rapid deterioration of the noise-shaping performance as a function of the polar angle, where even at 0.1° off axis the degradation is significant such that in practical terms there is little advantage in using a parametric SDM to gain improvements in high-frequency noise shaping. A second geometry is therefore explored, which is based on concentric rings, each consisting of an

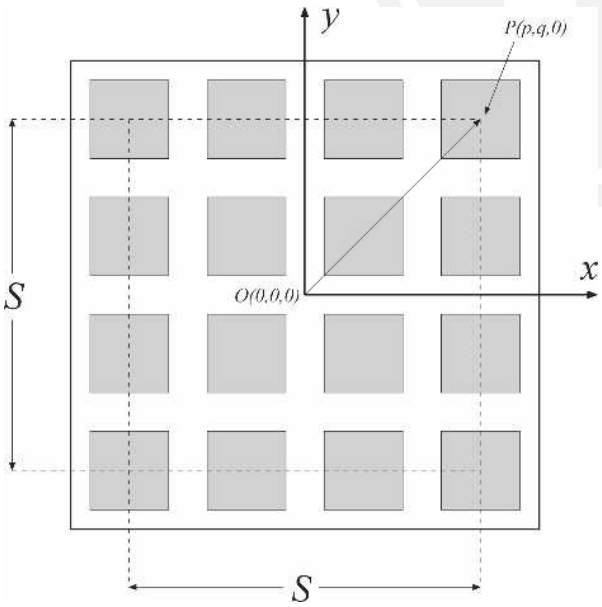


Fig. 11. Rectilinear cluster geometry; $W \times W$ elements.

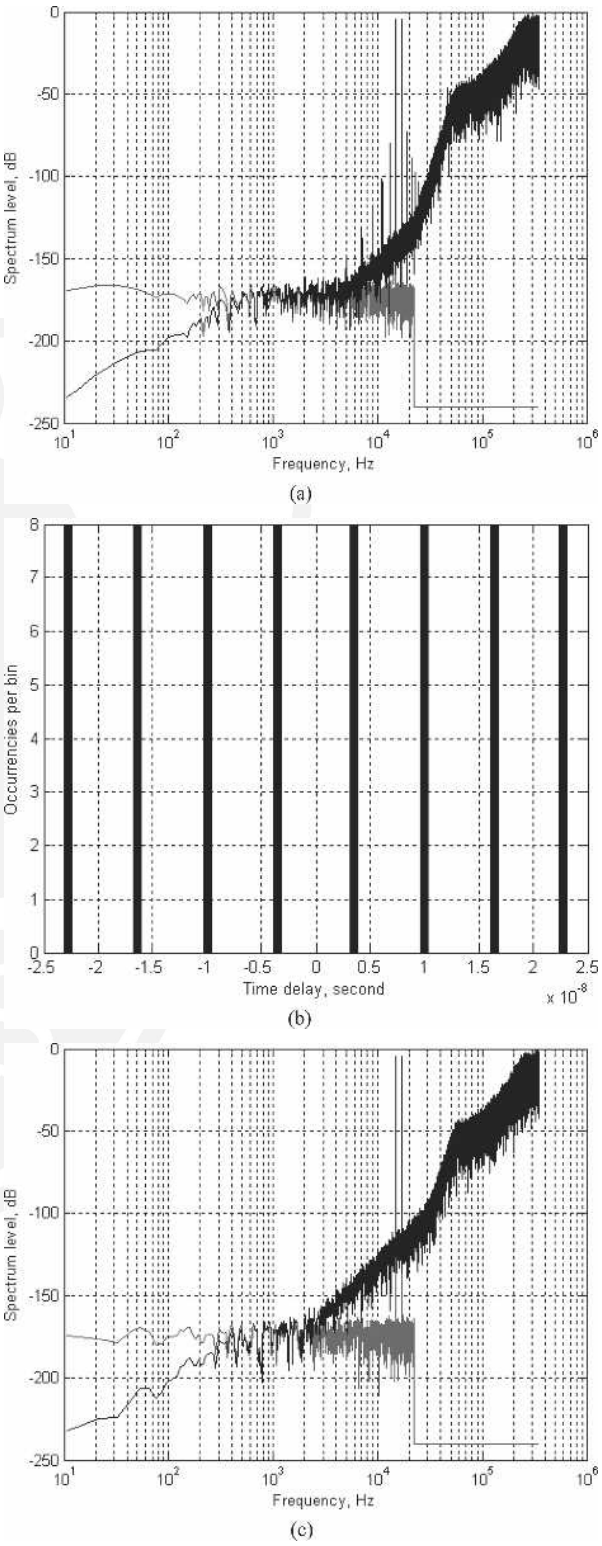


Fig. 12. (a) Far-field spectrum at $\phi = 0.1°$. (b) Histogram of interelement time delays, far field at $\phi = 0.1°$ off axis. (c) Far-field spectrum at $\phi = 0.1°$ using randomization in code stage.

equal number of symmetrically distributed elements to endow each ring with equal weight. Hence the volume displacement is matched. The circular geometry is shown in Fig. 13, where in this example each of the five concentric rings is composed of eight uniformly spaced elements. Initially the drive electronics are configured to drive the cluster progressively from the center, again using thermometer-style coding. Thus level 1 excites the first ring, level 2 both the first and second rings, and so forth, on to higher levels. The number of concentric rings is finite such that if a sample value exceeds the total number of rings, then the drive reverts to multilevel displacement of each element.

Simulations were performed using eight elements per concentric ring, all driven in parallel. Thus each ring can be considered effectively as a single element. The resulting far-field spectrum (using an input identical to earlier simulations) at 5° off axis is shown in Fig. 14 both with and without drive signal randomization in order to demonstrate the decorrelation of signal-related distortion into noise.

There is now less degradation in the high-frequency spectrum compared to the rectilinear geometry because the circular configuration, with the elements distributed symmetrically about the origin, has effectively both positive and negative time delays with respect to the center of the cluster. This implies that different rings contribute only magnitude spectral errors with zero phase distortion. To demonstrate the elimination of phase modulation as a function of ring radius, consider a ring of radius r with N_r identical elements driven in parallel, where the far-field monitoring location has spherical coordinates $M(r, \theta, \phi)$, as defined in Fig. 10. The coordinates of $P(p, q, 0)$ for element s out of N_r symmetrically distributed elements residing on the circumference of a circle of radius r with rotational offset angle ψ are

$$p(s \text{ of } N_r) = r\cos\left[2\pi\frac{(s-1)}{N_r} + \psi\right] \tag{21}$$

$$q(s \text{ of } N_r) = r\sin\left[2\pi\frac{(s-1)}{N_r} + \psi\right]. \tag{22}$$

Hence the time delay $\tau_p(s)$ for element s with respect to the center of the cluster follows from Eq. (19) as

$$\tau_p(s) = \frac{r}{c}\sin(\phi)\left\{\cos\left[2\pi\frac{(s-1)}{N_r} + \psi\right]\cos(\theta) + \sin\left[2\pi\frac{(s-1)}{N_r} + \psi\right]\sin(\theta)\right\}.$$

Simplifying gives

$$\tau_p(s) = \frac{r}{c}\sin(\phi)\cos\left[2\pi\frac{(s-1)}{N_r} + \psi - \theta\right]. \tag{23}$$

Since all elements on a ring have equal weight and are driven with the same signal, the ring transfer function

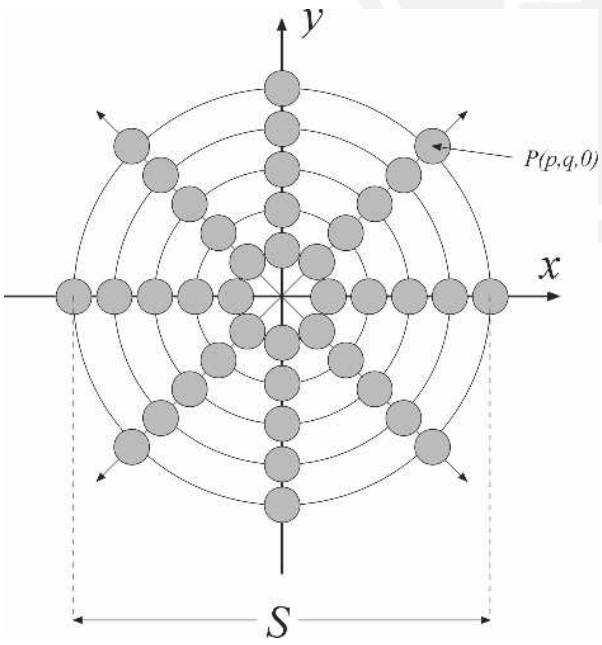


Fig. 13. Concentric cluster geometry, multiple rings.

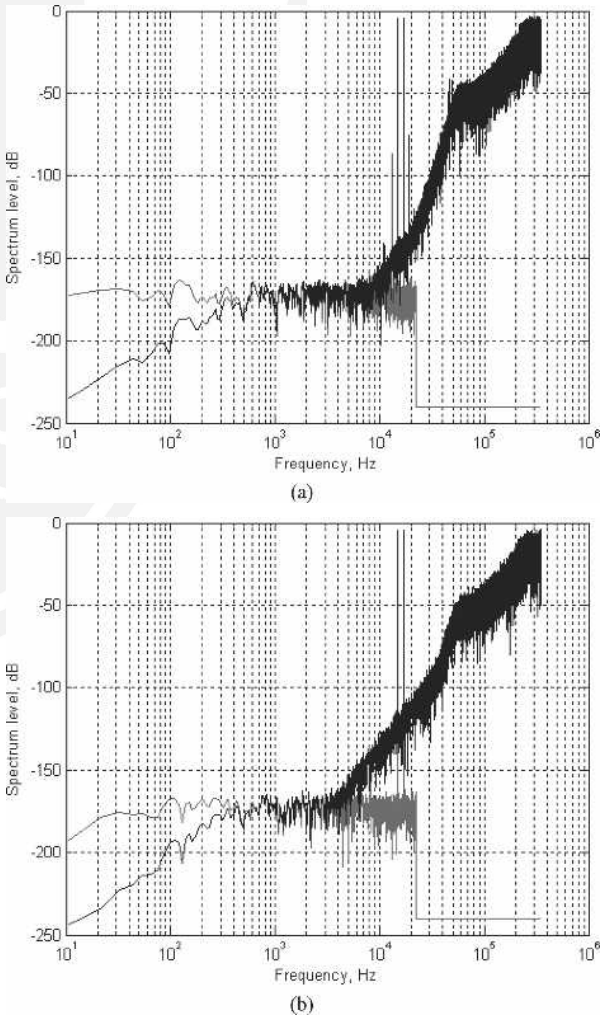


Fig. 14. (a) Far-field spectrum at $\phi = 5°$. (b) Far-field spectrum at $\phi = 5°$ using randomization in code stage.

RT(f, N_r) with respect to the far-field monitoring location $M(r, \theta, \phi)$ for all N_r elements follows, using a summation of exponential delay functions,

$$\mathrm{RT}(f, N_r) = \sum_{s=1}^{N_r} \mathrm{e}^{-\mathrm{j}2\pi f \tau_p(s)} \tag{24}$$

where f is the signal frequency. Eq. (23) shows for even N_r with diametrically opposed elements that

$$\tau_p(s)|_{s=1\ldots 0.5N_r} = -\tau_p\left(s + \frac{N_r}{2}\right). \tag{25}$$

Thus from Eqs. (23) and (24),

$$\begin{aligned}\mathrm{RT}(f, N_r) &= \sum_{s=1}^{N_r/2} [\mathrm{e}^{-\mathrm{j}2\pi f \tau_p(s)} + \mathrm{e}^{\mathrm{j}2\pi f \tau_p(s)}] \\ &= 2 \sum_{s=1}^{N_r/2} \cos\left\{2\pi f \frac{r}{c} \sin(\phi) \cos\left[2\pi \frac{(s-1)}{N_r} + \psi - \theta\right]\right\}.\end{aligned} \tag{26}$$

Eq. (26) confirms that there is only a magnitude error in the transfer function for a symmetrical cluster of N_r identical elements driven in parallel. Now employing the cluster geometry shown in Fig. 13 and applying Eq. (26), the variation in the ring transfer function can be determined as a function of signal amplitude, noting that rings become progressively active with the signal amplitude using a thermometer-style coding. Thus with a nonrandomized drive process, higher amplitude signals excite circular arrays of greater radius, noting that each ring has the same number of elements. Fig. 15(a) and (b) displays a set of superimposed ring transfer functions at $\phi = 5$ and 60° off axis for $N_r = 8$, with the number of discrete amplitude levels restricted here to $W = 32$. The same information is replotted in the three-dimensional graph shown in Fig. 15(c) and (d), although the frequency axis is now linear and restricted to a range of 0.25 of the sampling rate.

Fig. 15(a) and (b) indicates that as the radius of the ring increases, the low-frequency modulation of the transfer function migrates downward in frequency to where at a radius of 0.01 m there is spectral variation to below 10 kHz. Comparing Figs. 14 and 15 reveals sufficient similarity of form to predict the potential degradation in the

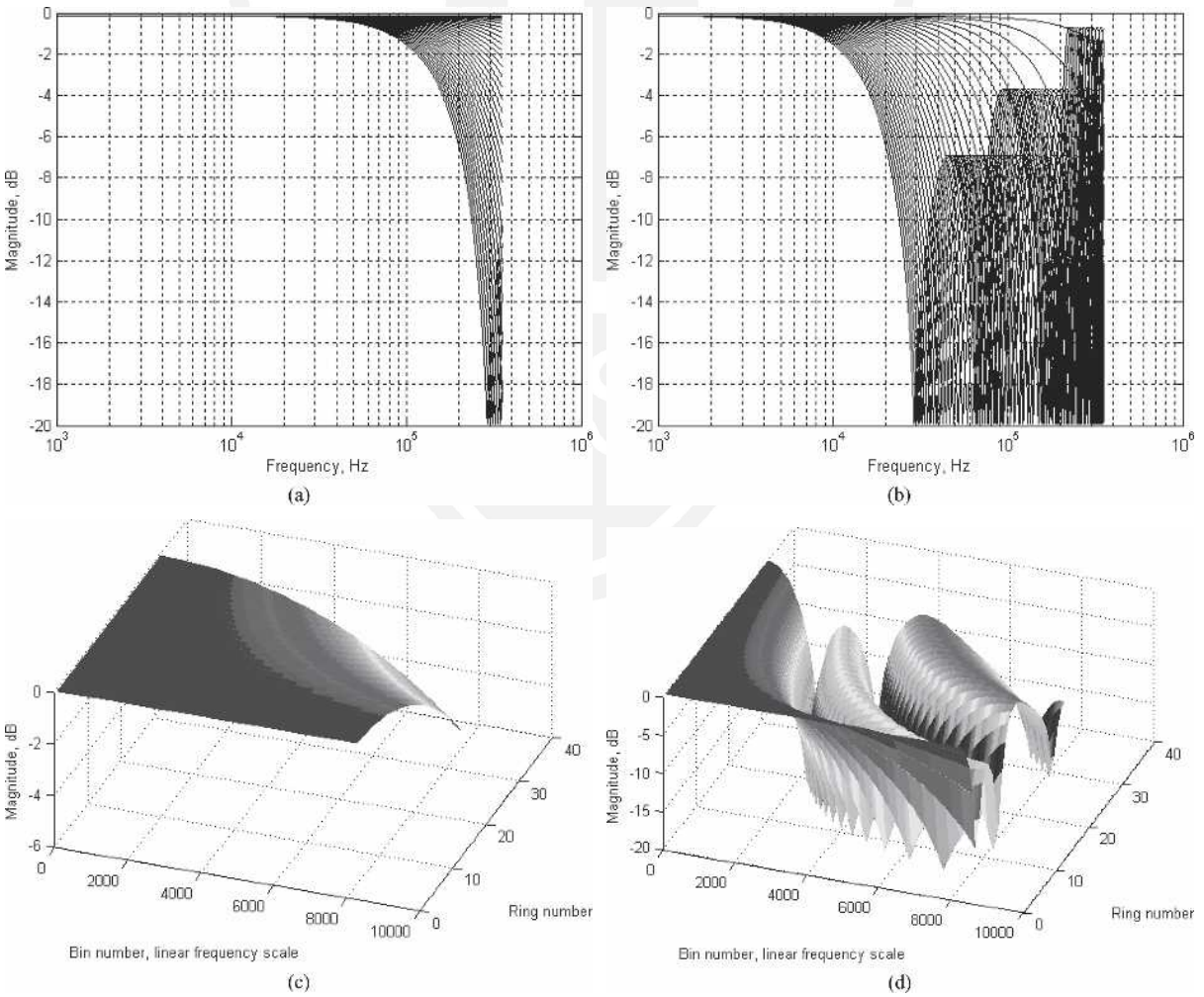


Fig. 15. Family of magnitude transfer functions; and three-dimensional plots $N_r = 8$; $W = 32$. (a), (c) At $\phi = 5°$. (b), (d) At $\phi = 60°$.

ring transfer function determined at the far-field measurement point. Consequently although spectral degradation increases with increasing off-axis angle, there is still a fundamental theoretical constraint on performance.

The results shown in Fig. 15 do, however, suggest that further gains in performance can be achieved by not modulating the ring radius as a function of signal level and using the constant radius array geometry shown in Fig. 16. Again symmetrical and even numbers of elements are used, which are linked together in subgroups of N_r that are again progressively driven as the signal level is increased. However, the W subgroups are now displaced by a constant angular offset increment α rather than being offset by a radius r. Hence all elements now lie on the same circle of diameter S m.

Fig. 16 illustrates two subgroups of eight elements with an offset angle α. For the case of W subgroups of N_r elements the offset angle is then defined,

$$\alpha = \frac{2\pi}{N_r W}. \tag{27}$$

Thus for a sample of amplitude 1, only one subgroup of N_r elements is driven, for a sample of amplitude 2 two subgroups are driven, and so on. To illustrate the relationship between this circular ring geometry and the ring transfer functions with increasing signal level, Figs 17–19 show the corresponding magnitutde transfer functions for N_r = 2, 4, and 8, respectively. A single ring of 20-mm-diameter is used where for Figs. 17 and 18 the off-axis angle is 5°, while for Fig. 19 it is increased to 60°. Using Eq. (27) the totality of elements is N_rW, and the cluster is circularly symmetric.

Fig. 17 reveals for N_r = 2 that even at 5° off axis there is severe transfer function modulation, whereas Fig. 18 shows that modulation is still evident for N_r = 4. However, for N_r = 8 virtually no modulation is found up to one-half the sampling frequency, while when increasing the off-axis angle to 60° Fig. 19 reveals a similar trend in-band, although the modulation at higher frequencies is now much greater. Consequently for N_r = 8 it is evident that a low degree of ring transfer function modulation can be achieved within the audio band, which bodes well in terms of reduced deterioration in high-frequency quantization noise at an off-axis angle. These results suggest that for a 20-mm-diameter cluster N_r = 8 offers an appropriate compromise, as further increases in the subgroup number would compromise the resolution of the cluster for a given total number of elements.

Using the information derived from this study of ring transfer function modulation, an SDM-driven array was simulated using simulations similar to those reported earlier, except that the cluster is now formed on a single ring with subgroups rotationally displaced [see Eq. (27)]. Spectral results were determined for both 5° and 60° off-axis monitoring locations for the three cases of N_r = 2, 4, and 8, corresponding to the transfer function results discussed

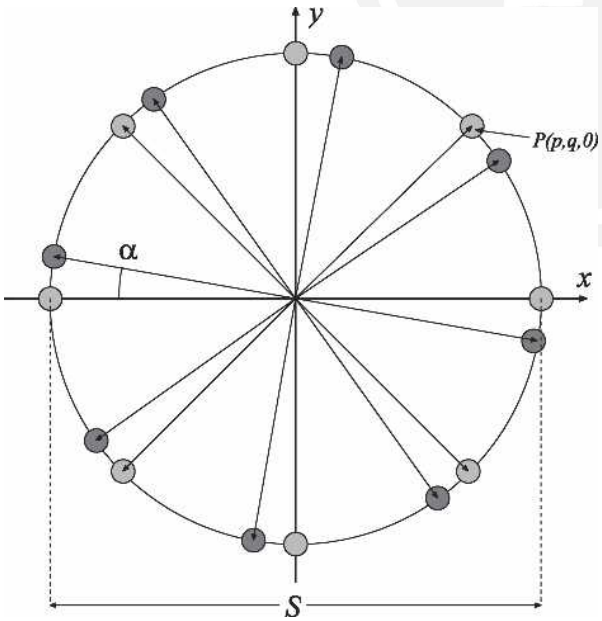


Fig. 16. Single-ring cluster geometry with rotationally displaced elements.

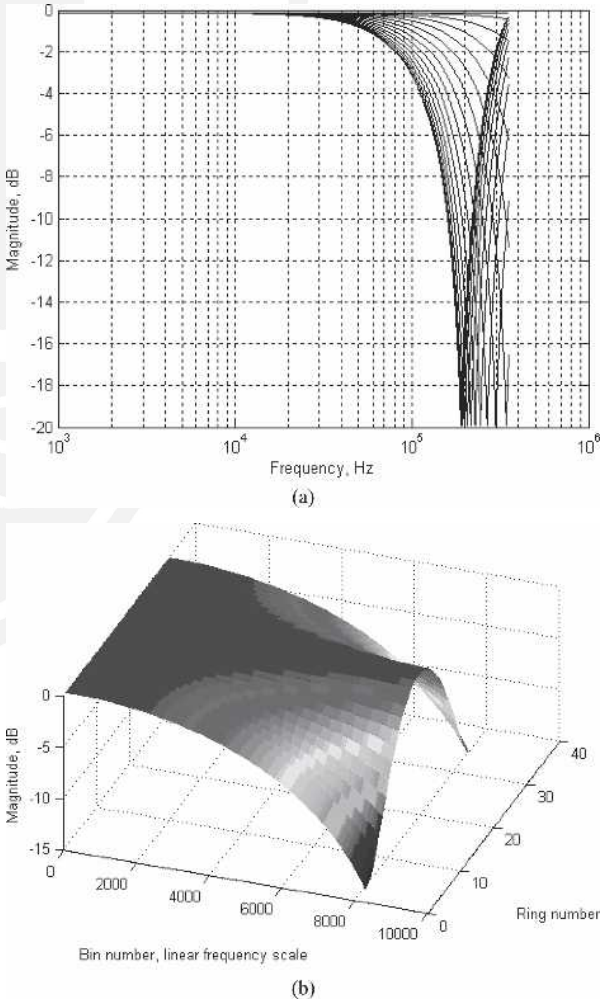


Fig. 17. (a) Family of magnitude transfer functions at ϕ = 5° for ring of N_r = 2 elements, single ring with W = 32, ring diameter 20 mm. (b) Three-dimensional plot.

previously. The sets of results were repeated for direct and randomized signal assignment in terms of driving the elements in each subgroup. The computed spectra and associated data are summarized in Table 1.

The spectral results shown in Figs. 20 and 21 should be compared against the corresponding ring transfer functions shown in Figs. 17–19, as indicated in column 1 of Table 1. At 5° off-axis simulation [results not shown as there is minimal difference when compared to the on-axis location in Fig. 9(a)] confirmed that for $N_r = 8$ element subgroups, the degree of spectral modulation is so low that 24-bit resolution within the audio band is theoretically possible for an idealized array, while for a 60° off-axis angle Fig. 21 reveals almost identical results. It is shown that applying the randomization function breaks down correlated distortion and translates intermodulation distortion into a noiselike residue that can be maintained at a very low level, even to 20 kHz. However, the size of the cluster is critical, as with any high-frequency drive unit that does not operate in a diffuse mode [13], [14]. At the selected diameter of 20 mm the results show that at an off-axis angle of 60° the frequency response errors are just beginning to become evident due to interference effects. In this respect the cluster is similar to that of an analog pistonic drive unit, indicating that in a practical system a slightly larger diameter could be used if some minor off-axis response errors were acceptable.

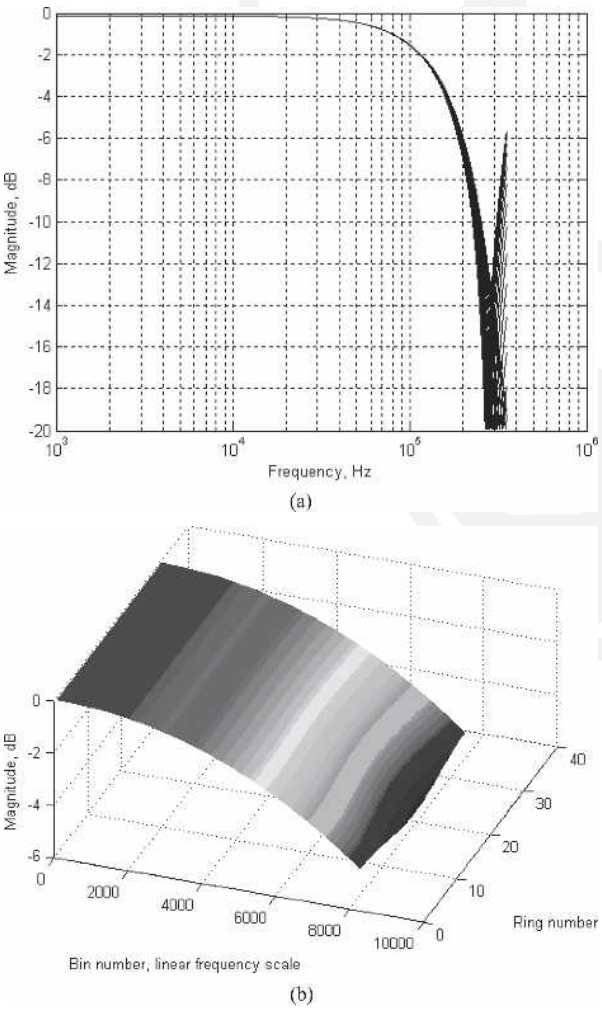


Fig. 18. (a) Family of magnitude transfer functions at $\phi = 5°$ for ring of $N_r = 4$ elements, single ring with $W = 32$, ring diameter 20 mm. (b) Three dimensional plot.

4 OPTIMUM CLUSTER GEOMETRY

The results produced in Section 3 suggest that a circular geometry is optimum for a ring cluster. Obviously the

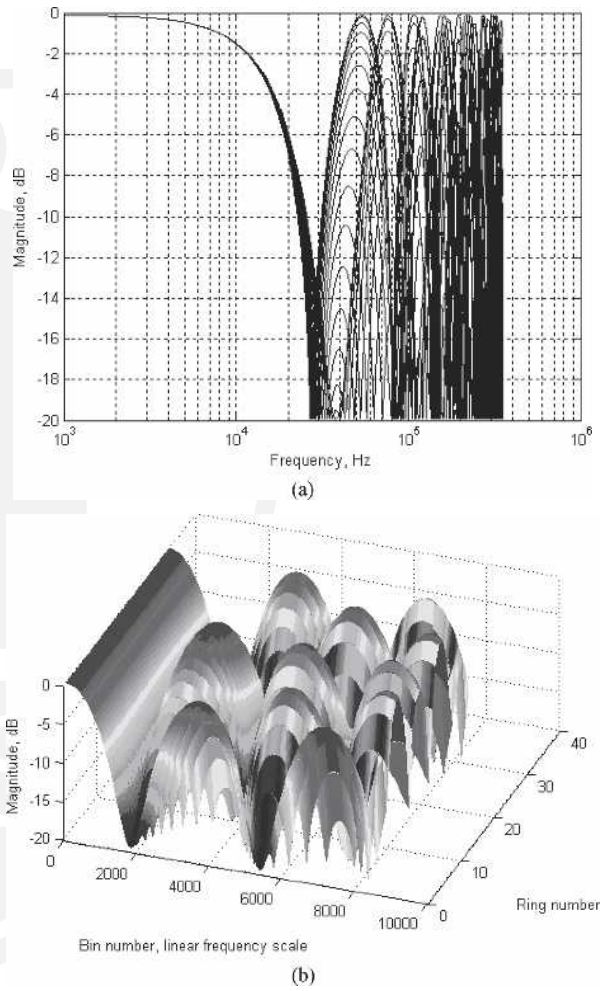


Fig. 19. (a) Family of magnitude transfer functions at $\phi = 60°$ for ring of $N_r = 8$ elements, single ring with $W = 32$, ring diameter 20 mm. (b) Three-dimensional plot.

Table 1. Simulation data for SDM far-field spectra.

Figure Number	N_r	W	Random Assignment	Off-Axis Angle ϕ
17–20(a)	2	32	No	5°
17–20(b)	2	32	Yes	5°
18–20(c)	4	32	No	5°
18–20(d)	4	32	Yes	5°
19–21(a)	8	32	No	60°
19–21(b)	8	32	Yes	60°

family of ellipses is relevant as this geometry applies to an off-axis point of observation, although to force axial symmetry in the polar response, the circular on-axis geometry is required. To test this hypothesis four examples are considered where the circular geometry of each subcluster is modified. In each case a basis cluster pattern is defined,

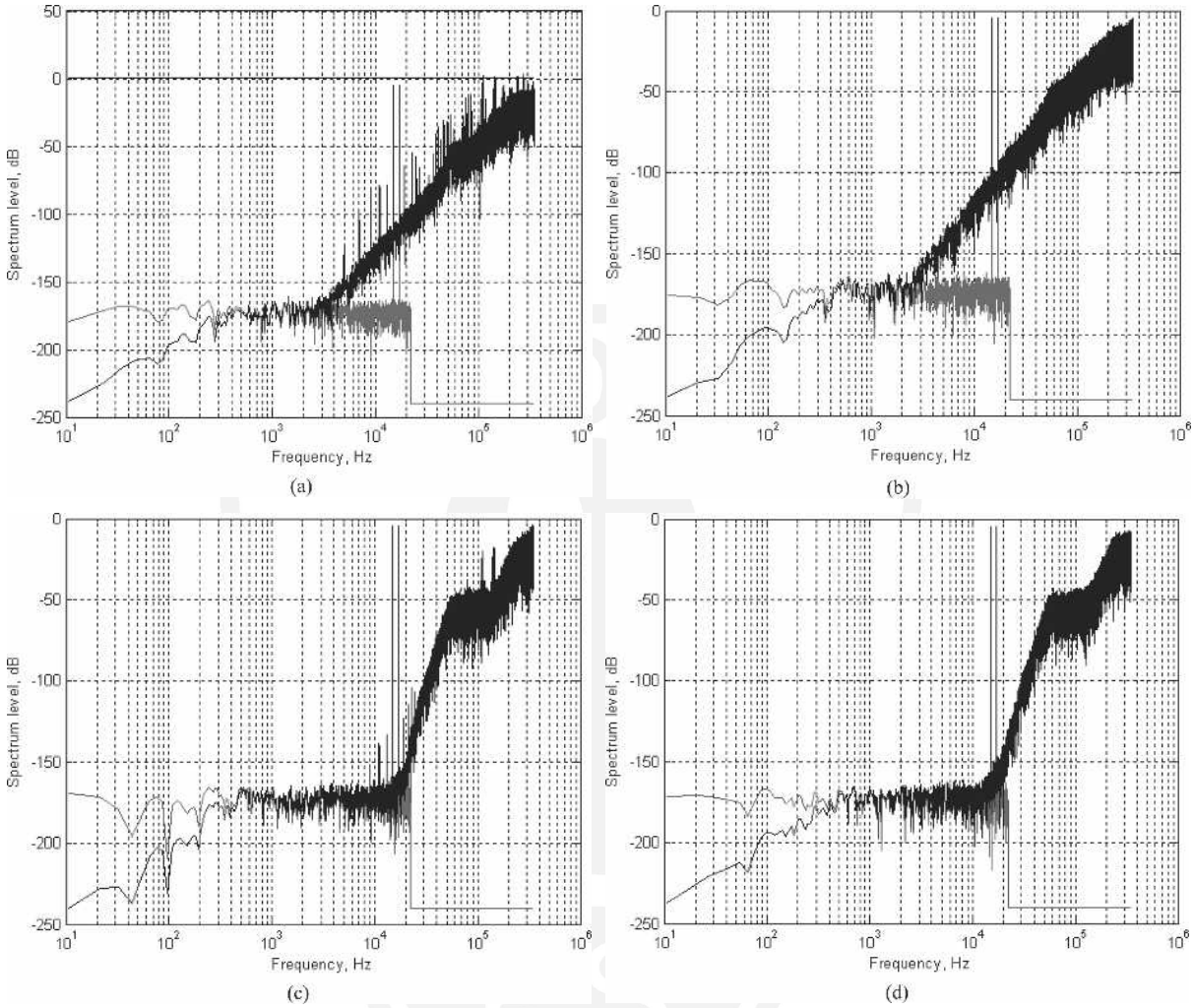


Fig. 20. $W = 32$; $\phi = 5°$. (a) $N_r = 2$; random. (b) $N_r = 2$; nonrandom. (c) $N_r = 4$; nonrandom. (d) $N_r = 4$; random.

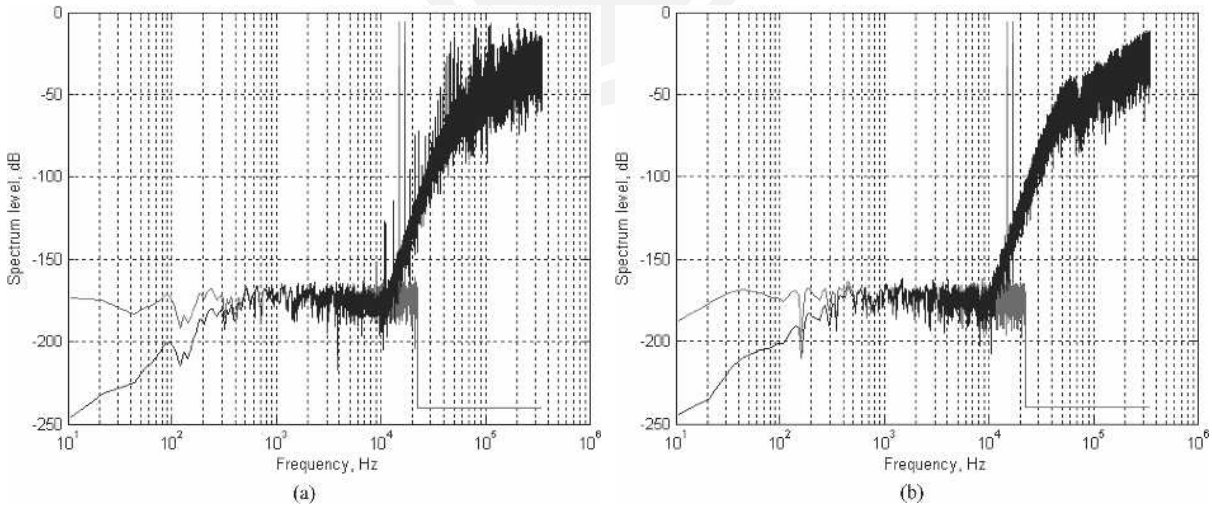


Fig. 21. $N_r = 8$; $W = 32$; $\phi = 60°$. (a) Nonrandom. (b) Random.

replicated, and then progressively rotated about the central axis to form the set of subcluster elements used in the transducer. Off-axis noise-shaping spectra are then determined and benchmarked against the circular geometry using the same number of elements as selected for example 1. In each example the average cluster radius is 0.08 m, $N_r = 64$ elements per subgroup cluster, with each element spaced at equal angles over 2π, and the number of subgroups is $W = 4$.

The following four examples test whether a circular array is critical to minimizing quantization noise observed at a 60° off-axis far-field monitoring location. Every subcluster in each example has equal weight, and random assignment of digital signals is employed.

Example 1 Subcluster with pure circular symmetry used as benchmark. Fig. 22(a): overall cluster geometry; Fig. 22(b): off-axis spectrum.

Example 2 Pure circular symmetry, but each element in a subcluster is given a random radial displacement with a peak value of 0.0025 m. The pattern is then repeated to form an overall structure with axial symmetry. Fig. 23(a): overall cluster geometry; Fig. 23(b): off-axis spectrum.

Example 3 As example 2, but diagonally opposing pairs given the same radial offset to force center-symmetric symmetry and eliminate time modulation. Fig. 24(a): overall cluster geometry; Fig. 24(b): off-axis spectrum.

Example 4 Elements within the cluster are given a period offset from the pure circular array, so the first element is displaced by 0.0025 m the next by −0.0025 m, and so on. If a single cluster is observed, then the elements are split equally and thus lie on one of two concentric circular paths. Fig. 25(a): overall cluster geometry; Fig. 25(b): off-axis spectrum.

Observations The principal observation is that the single circular geometry of example 1 yields the lowest high-frequency noise at an off-axis listening position, with example 4 following closely behind, while random dis-

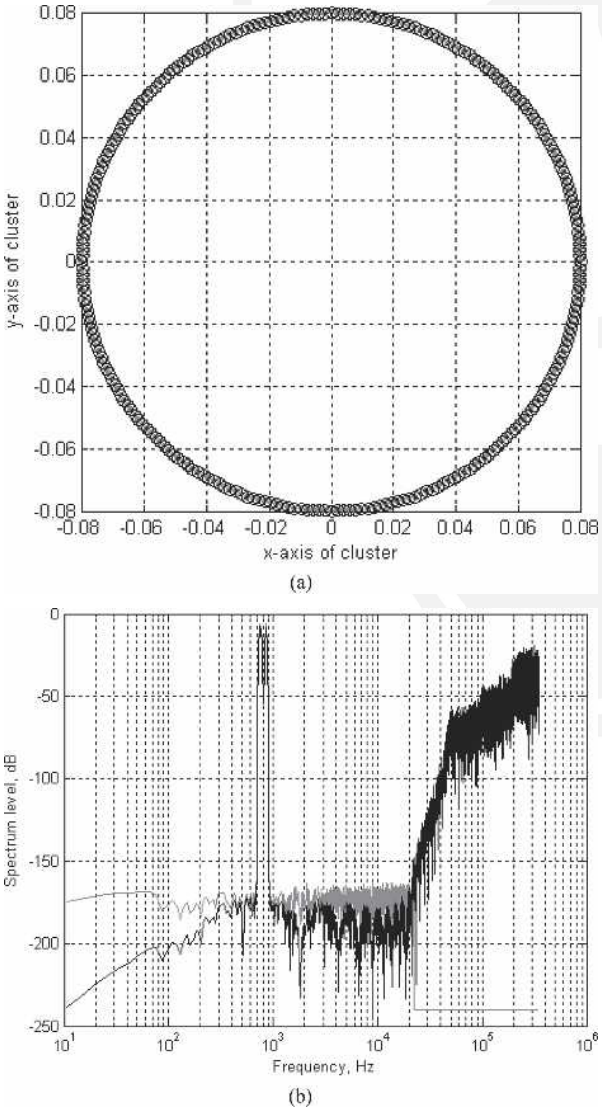


Fig. 22. Example 1. (a) Element layout; 0.08 m radius. (b) Derived off-axis spectrum; $N_r = 64$; $W = 4$; exact circular array.

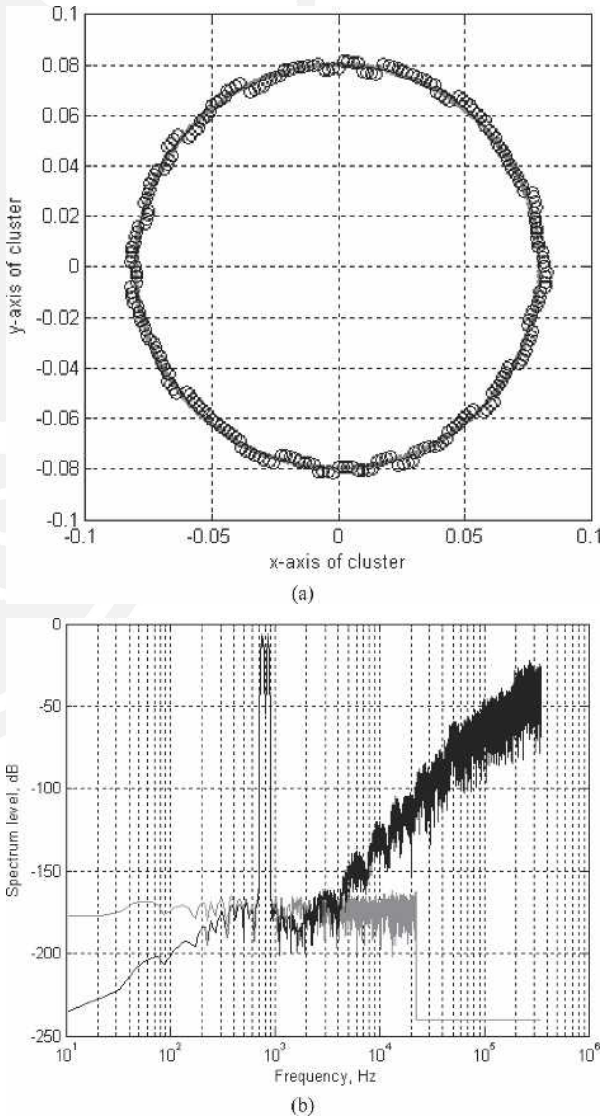


Fig. 23. Example 2. (a) Element layout; no diagonal symmetry; 0.08 m radius. (b) Derived off-axis spectrum, $N_r = 64$; $W = 4$.

placements, even with forced diagonal symmetry, do not perform well. Hence it may be concluded for a given number of elements, that with respect to transfer function modulation with off-axis angle, it is best to locate all elements at a constant radius. Also the need for the polar response of quantization noise to be axially symmetric implies that a circular pattern is the optimum choice. However, it will be shown in Section 5 that there can be some advantage in using multiple circles, as these can increase overall volume displacement as well as offering a degree of high-frequency filtering for off-axis monitoring locations while still retaining an axially symmetric performance.

5 MIDRANGE CLUSTER

The results so far have been applied mainly to high-frequency signals where SDM and filtering can achieve an acceptable dynamic range using a relatively small number of tristate elements. However, as the signal frequency is lowered, in order to maintain the far-field pressure, the volume displacement must increase as an inverse square law of frequency. This necessitates either driving each element over a wider amplitude range or using a greater number of elements if each is constrained to tristate operation. Consequently to retain tristate operation and to employ elements of identical size, the overall surface area must increase. However, any increase in area must take into account the potential degradation in off-axis quantization noise as well as wavelength-dependent interference that degrades the polar frequency response. To address these two fundamental issues, the approach adopted divides the audio band into a number of subbands and then associates each subband with an individual ring cluster of appropriate diameter. As a result signals falling within an individual subband excite only a ring cluster of a given

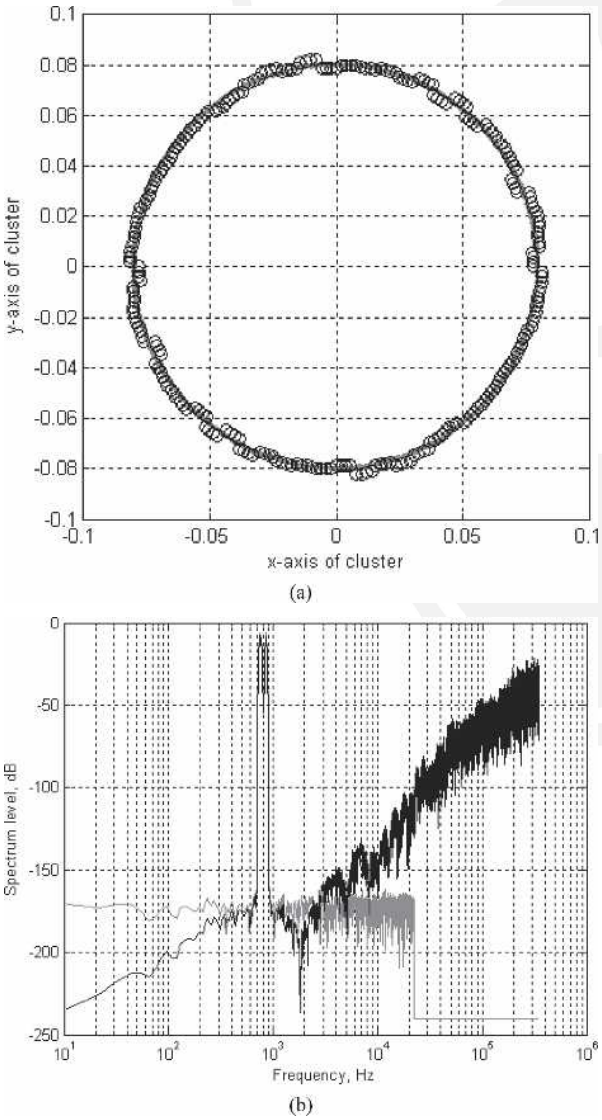


Fig. 24. Example 3. (a) Element layout; diagonal symmetry; 0.08 m radius. (b) Derived off-axis spectrum, $N_r = 64$; $W = 4$.

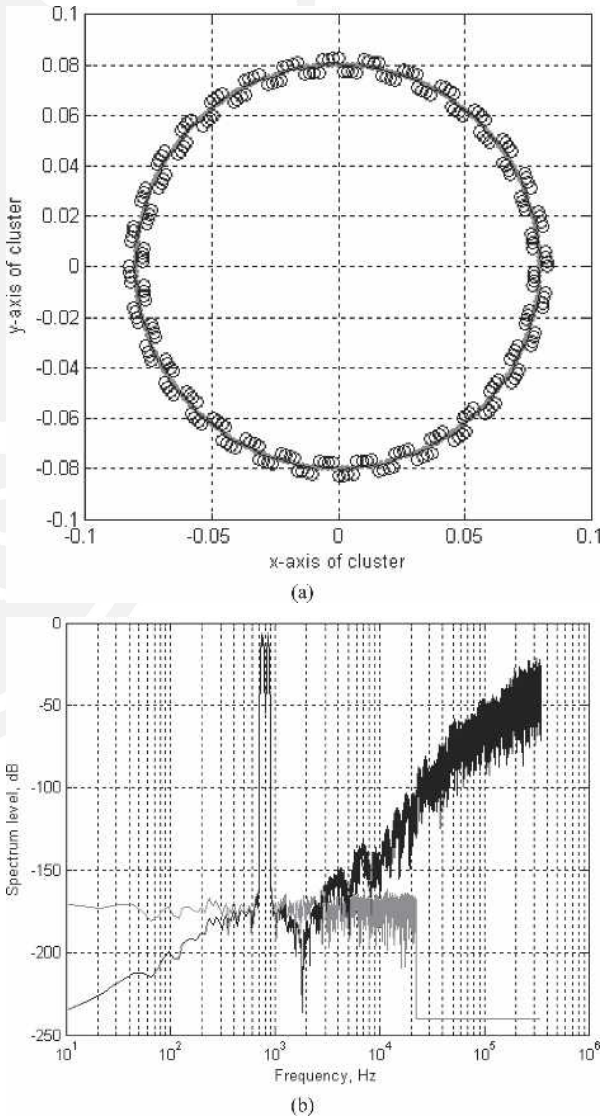


Fig. 25. Example 4. (a) Element layout; periodic symmetry, 0.08 m radius. (b) Derived off-axis spectrum, $N_r = 64$; $W = 4$.

radius, where each ring is designed to maintain both polar response and off-axis quantization noise performance. To explore the performance characteristics of a substantially larger cluster, a ring of 160-mm diameter is investigated, as it compares favorably with the diameter of a typical bass/midrange analog drive unit.

The conceptual model for the multiring cluster is shown in Fig. 26, which contains a subband filter bank that drives a parallel array of SDMs, which in turn drives a specific ring of elements on the loudspeaker array. Taking the example of a 160-mm-diameter ring and to ascertain the required subgroup size given the diameter is larger, simulations were performed for subgroup sizes $N_r = 16$, 32, and 64, all at an off-axis angle of 60°. In each case the input consisted of two sine waves of amplitude 0.5 with respective frequencies of 750 and 850 Hz, and the resulting far-field spectra are shown in Fig. 27 As the number of subgroups has increased substantially, the simulation had to accept multilevel elements because of memory constraints and the high number of Fourier transforms. Examining the displacement amplitude range from the histogram in Fig. 28 and knowing W, the number of discrete tristate elements can be inferred. Inevitably these numbers are high as simulations assume a dynamic range commensurate with 24-bit resolution.

Fig. 27 confirms the trend that as the number of circularly symmetric subgroup elements increase degradation in the high-frequency quantization noise is reduced in the

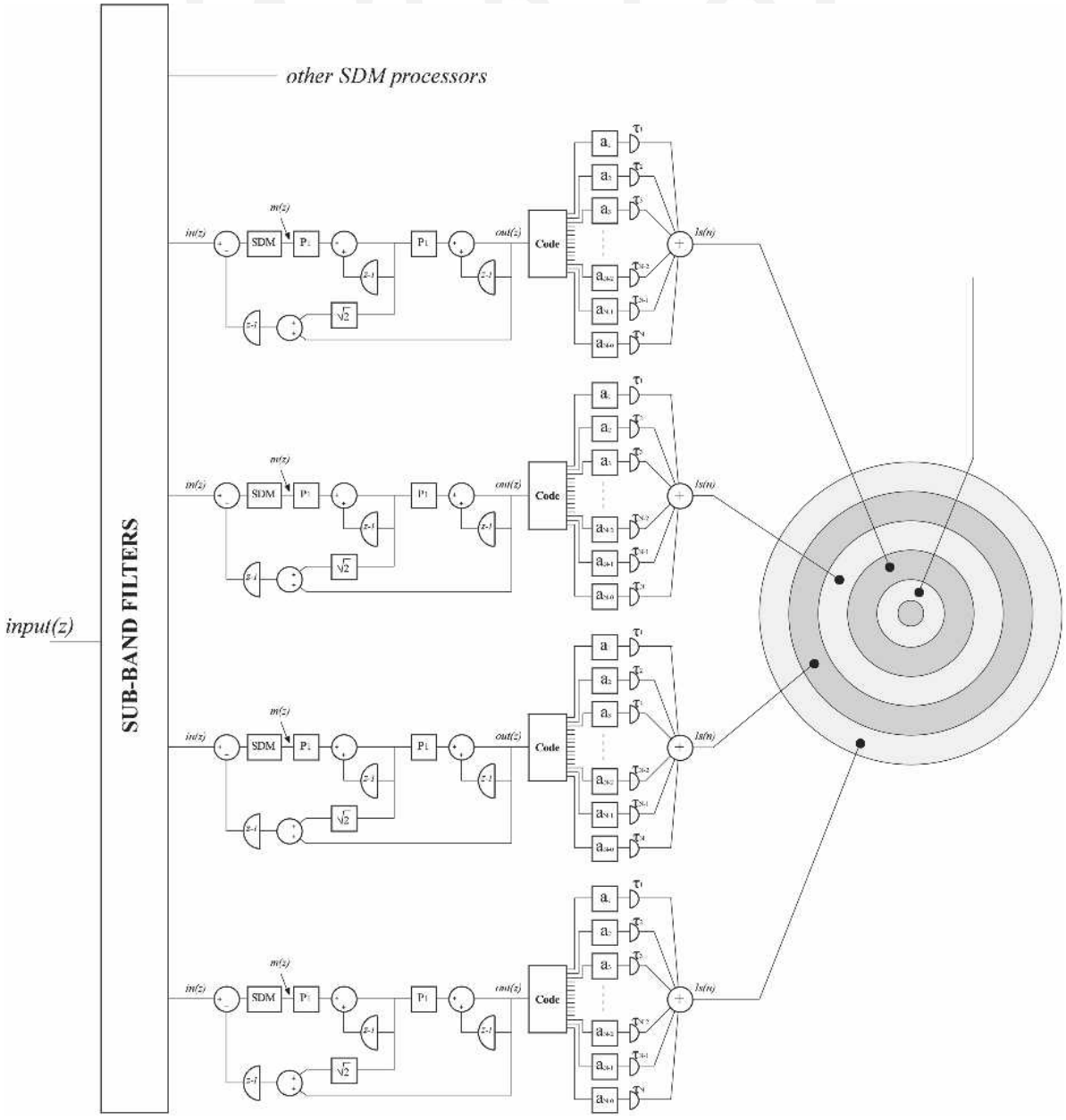


Fig. 26. Multiring cluster model with subband filters and parallel SDM.

band below 20 kHz, where at 60° off-axis Fig. 27(c) reveals almost ideal conditions for $N_r = 64$. This should be compared against a 20-mm-diameter ring with $N_r = 8$. Thus an eightfold increase in ring diameter requires a proportional increase in N_r. To substantiate this observation, spectra for $N_r = 16$ and 32 show significant levels of high-frequency quantization noise progressively entering the audio band.

Employing a larger diameter ring cluster inevitably limits the bandwidth of a coherently driven set of elements, where interference nulls can be observed in the spectrograms at around 2 kHz and above, depending on N_r. Consequently a 160-mm-diameter ring can only be used below about 1 kHz. Observing the histogram presented in Fig. 28 shows the peak signal to be close to 19 000 quanta, where in this example, because $W = 64$, it follows that each element must produce a peak amplitude of about 300. This increase in signal level can also be anticipated as the frequency has been reduced by a factor of 20, whereby the required volume displacement must increase by 400 in order to maintain the same far-field pressure level. As a confirmation, Fig. 29 displays the corresponding time-domain displacement output of the second-order high-pass filter (see Fig. 4), where a peak level of just under 300 can be observed.

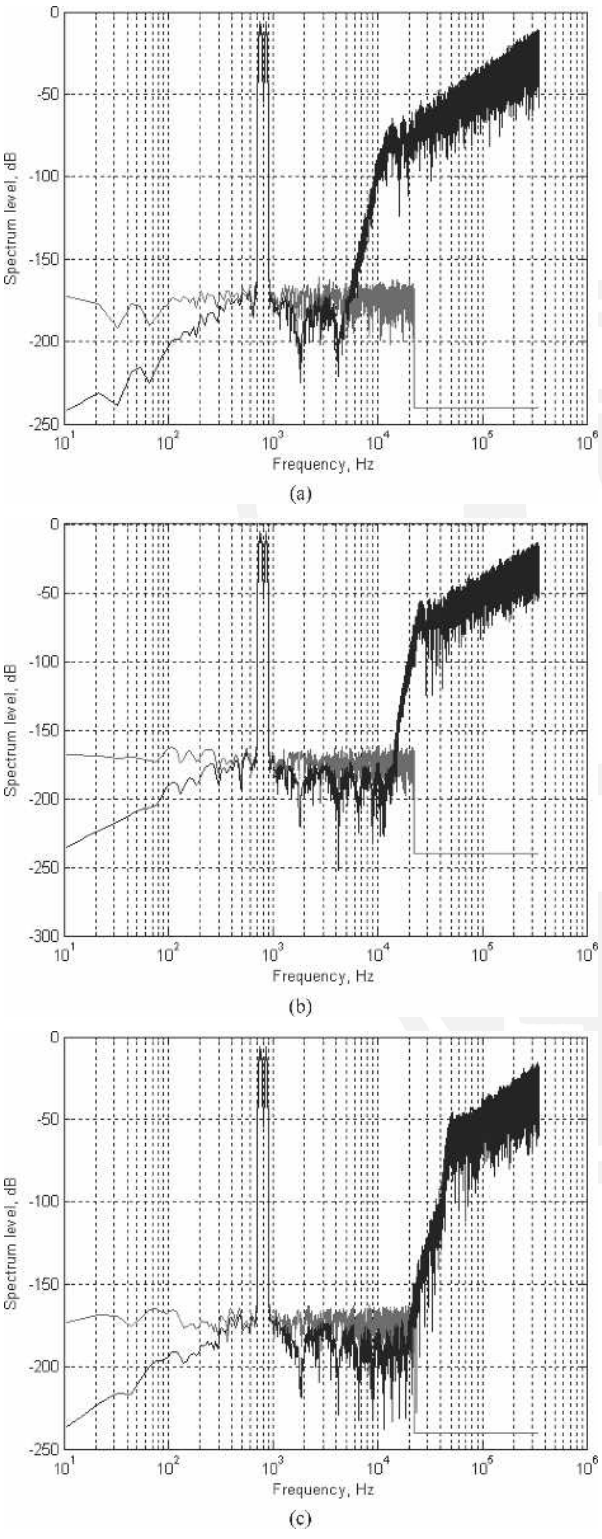


Fig. 27. (a) $N_r = 16$; $W = 64$; $\phi = 60°$; random. (b) $N_r = 32$; $W = 64$; $\phi = 60°$; random. (c) $N_r = 64$; $W = 64$; $\phi = 60°$; random.

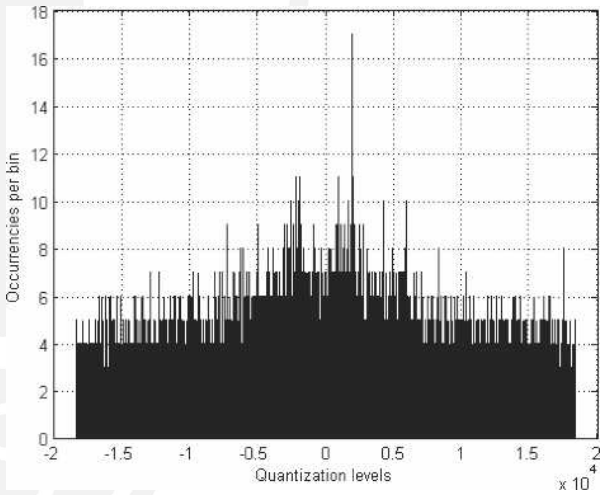


Fig. 28. Histogram of drive signals; $N_r = 64$.

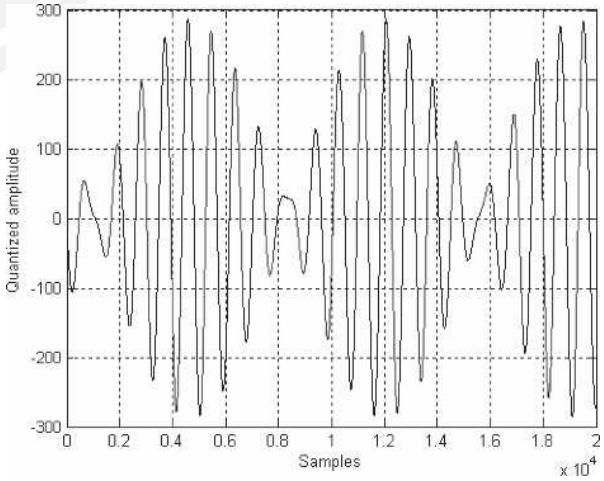


Fig. 29. Time-domain displacement sequence; $N_r = 64$.

The study has shown that the use of a circular array with circular and symmetrically distributed elements can limit degradation in the high-frequency noise with an increase in the off-axis angle. However, this performance is quite critical to the maintenance of accurate geometry. To demonstrate this sensitivity for a 160-mm cluster, a 60° off-axis response is computed, but with a 0.1% rectangular PDF random error added to the radial location of each of the $N_r = 64$ by $W = 64$ elements. The results shown in Fig. 30 reveal that at high frequency the noise performance, although usable, together with improvements at reduced off-axis angles, has fallen below that of a 16-bit resolution system.

To address the sensitivity to geometric errors and also to improve off-axis noise performance, the use of array filtering is investigated. Here, rather than configuring the array from just a single circle, an elongated element shape in the radial direction can be used, as illustrated in Fig. 31(a). Alternatively a multicircle array geometry can be used, as shown in Fig 31(b). These geometric changes can offer the advantage of increased volume displacement, which becomes more demanding at lower frequency, especially if adequate sound pressure levels are to be maintained. Although multiple circular arrays with individual filter weights could be used, here, in order to demonstrate the principle, only a simple three-circle symmetrical cluster is investigated, as shown in Fig. 31(b).

The central circle has a 160-mm radius and a weight of 0.5 while the additional inner and outer circles of respective radii of 150 and 170 mm are each given a weight of 0.25. Effectively each element on the 160-mm circle now has two additional satellite elements located symmetrically along a radius, with each of the three-element subclusters driven in parallel, but with the [0.25 0.50 0.25] weighting. Fig. 32 shows two 60° off-axis spectra with and without a 0.1% randomization of element location, but where triplet elements have been grouped to maintain their relative dimensions. Fig. 32(a) reveals a significant reduction in high-frequency noise where this attenuation with frequency actually becomes greater as the off-axis angle increases. When the 0.1% random geometric errors are introduced, there is still degradation, but not that great when compared to the single-circle results shown in Fig. 30. It is therefore evident that a variety of geometric and subcluster design strategies can be adopted to accommodate the required number of tristate elements, provided the underlying physics of the system in terms of size and specifically volume displacement are observed.

By decomposing the overall array into a number of frequency-selective subbands that relate to a specific cluster radius, and because of the spectral energy spread in-

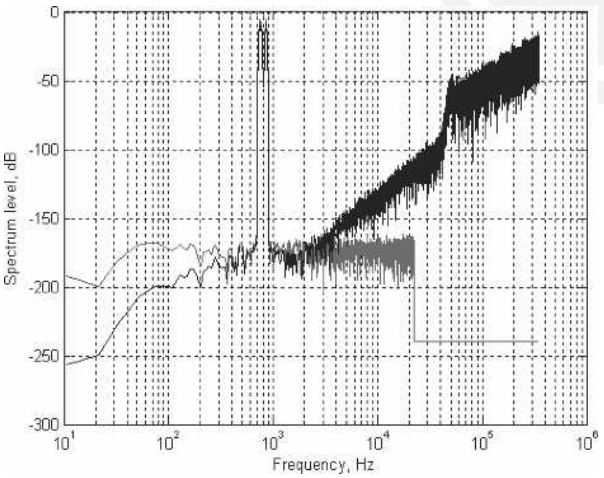


Fig. 30. 0.1% random radius error; 160 mm; $N_r = 64$; $W = 64$; $\phi = 60°$.

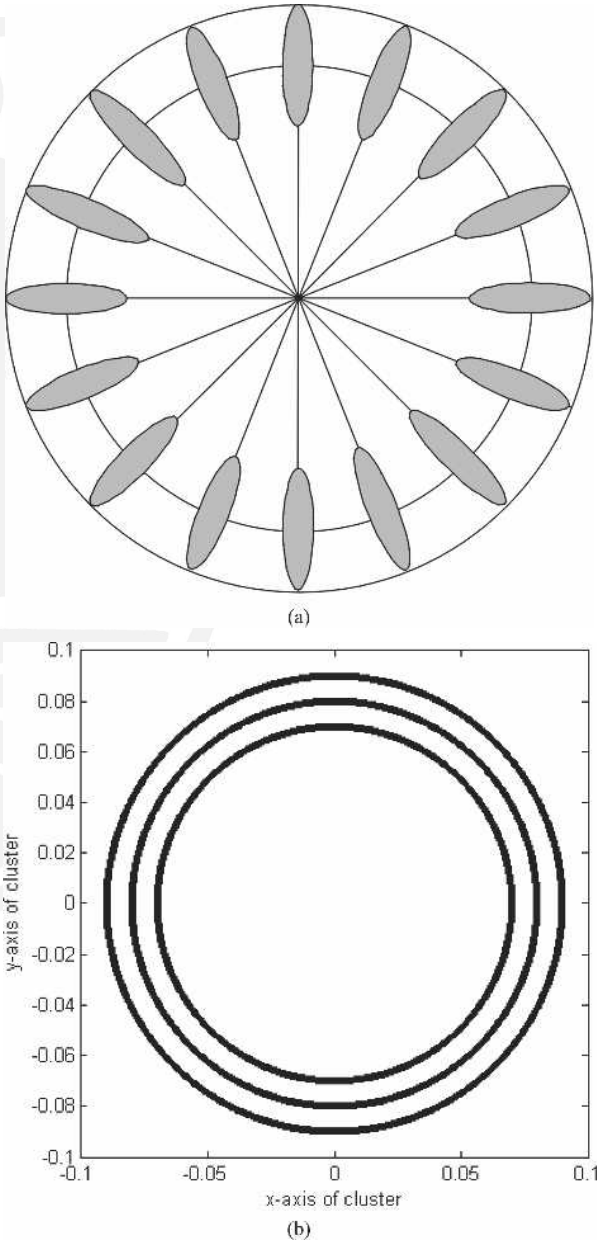


Fig. 31. (a) Example (exaggerated) of cluster with element extension along radial axis. (b) Three-circle (150, 160, and 170 mm) array geometry to improve off-axis high-frequency noise filtering; $N_r = 64$.

herent in audio signals, it follows that each band can have reduced amplitude resolution, which in turn reduces the total number of discrete elements. Also, associating dedicated areas of the array with specific frequency bands helps counter the generation of intermodulation distortion due to element nonlinearity. Therefore the overall array can achieve a broad polar response that is not frequency dependent and where high frequencies radiate from the center with lower frequencies radiating progressively from larger diameter ring clusters which, by symmetry, maintain the precise coherence of a single radiating object. Effectively the size of the array is frequency dependent, although unlike with its analog counterpart, there is also a distribution of signal across the radiating area, mainly in the circumferential direction, which is amplitude dependent, albeit with an embedded drive-signal randomization function to prevent correlated distortion products. For lower frequency subbands the use of radial elongated element shapes also appears attractive as this offers the advantage of introducing high-frequency filtering as the off-axis angle increases, a trend opposite to what occurred with geometries such as the rectangular "checkerboard" cluster. Also, the study has shown that because of the need for increased volume displacement at lower frequencies larger clusters with greater multilevel displacement capabilities are required.

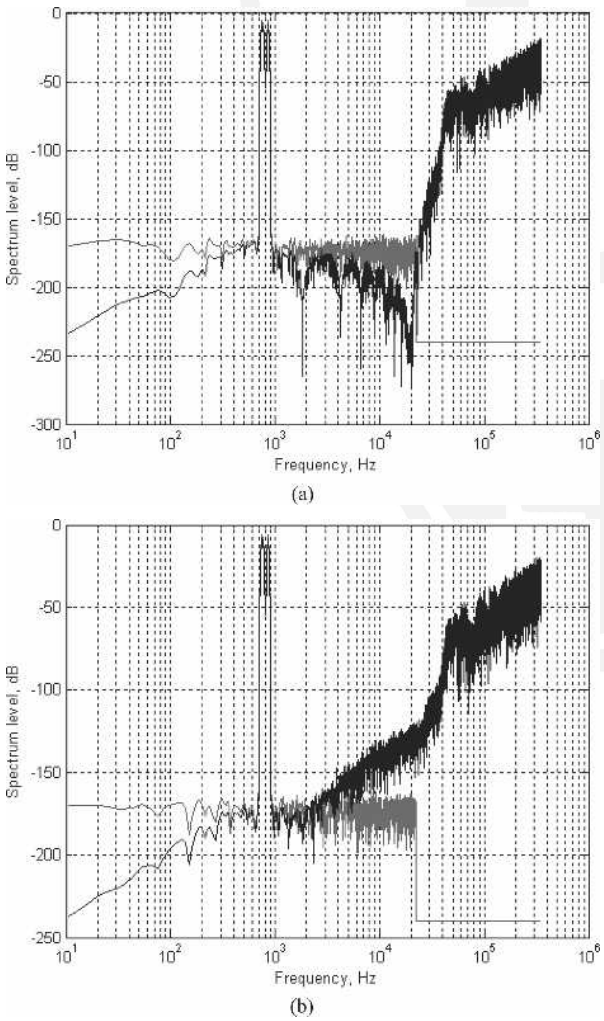


Fig. 32. (a) Three-ring geometry; no random radius error; 160 mm; $N_r = 64$; $W = 64$; $\phi = 60°$. (b) Three ring geometry; 0.1% random radius error; 160 mm; $N_r = 64$; $W = 64$; $\phi = 60°$.

6 CONCLUSIONS

This paper has described possible theoretical structures for a digitally addressed, displacement-driven multielement loudspeaker array where the aim has been to theorize on a conceptual system architecture that could be implemented by integrating a large number of discrete, ideally tristate, micro-radiating elements. It is assumed that such elements would have identical physical forms, although area scaling at low frequency is possible, and be implemented by some appropriate nanotechnology fabrication process that can be interfaced directly with signal processing electronics integrated into the back plane of the device.

A key conceptual requirement for a digital loudspeaker is that each radiating element produce an output related directly to the discrete nature of the digital input signal and not be a continuous-time filtered representation of the input data such that the notional boundary between digital loudspeaker and analog loudspeaker is transgressed. A simple example of this conceptual divide is to apply the 1-bit serial code from an SDM directly to a conventional moving-coil drive unit. As such the drive unit performs transduction from electrical to acoustical domains and also performs as a crude low-pass analog filter due to its limited bandwidth. However, such a system is not a digital loudspeaker, as the drive unit is still performing a purely analog function. It is suggested that in a strictly digital application the drive unit must produce an output in direct proportion to the digital code, be it mechanical output displacement, velocity, or acceleration.

Irrespective of how multiple radiating elements are configured and whatever means of code conversion is employed, whether it is binary-weighted pulse-code modulation, multilevel SDM, or just 1-bit SDM code, it follows that information is distributed across the radiating area according to both its amplitude and its frequency and that such a process is not linear. In an analog drive unit such signal-area mappings are normally only frequency dependent and therefore linear within the bounds of motor design. However, in a digital transducer a signal can become dispersed over an area in an amplitude-dependent manner, where there is no guarantee that this is a linear process. A simple example of this phenomenon is binary weighted pulse-code modulation, where the codes with higher weight address larger radiating areas. Hence by the noncoincident nature of the radiator the polar response will inevitably exhibit quantization noise and distortion patterns that vary with the off-axis monitoring angle.

However, this paper has shown that it is possible to configure clusters of nominally equally weighted radiating elements that can reduce the problem of polar-response dependency of the quantization distortion. Also by using a

modest degree of oversampling and noise shaping the amplitude resolution requirements and number of elements can be reduced, although the volume displacement commensurate with a given pressure level is a critical design factor. The study has concluded that a geometry based upon circular arrays offers a number of key advantages, especially how high-frequency noise is reproduced and distributed over the polar response. In reaching this conclusion the way quantization noise varied with polar angle was taken into account and a principal design objective selected where in-audio band noise degradation was minimized up to an off-axis angle of 60°. It was demonstrated by computing far-field spectra derived at off-axis locations that cluster symmetry was a key requirement and that a circular array when composed of a limited number of circumferentially distributed elements, all driven in parallel, could meet this objective. It was also shown that as the signal amplitude was increased, additional subclusters could be activated progressively, where their physical location was offset by constant angular increments around the central ring of the cluster. In assigning signals to these additional radiating element subclusters the use of random assignment was adopted as this was shown to offer an effective means of translating quantization distortion into a noiselike residue. This process mimicked a similar technique often employed in DAC systems, although for a cluster the effective spatial location was the principal parameter to be randomized in terms of signal assignment.

The multilevel SDM used in this study was derived from earlier work on parametric SDM, where such techniques have been shown to be capable of achieving extremely low levels of in-band quantization noise. The near 24-bit resolution up to 20 kHz offered by this class of system proved particularly useful in the study of the high-frequency noise performance in terms of polar angle. It was also a key factor in reducing the number of elements required. It should be noted that in a broad-band application with subband filtering using multiple ring clusters of differing radii together with individual SDMs, the high-frequency noise reproduced by each cluster is incoherent because uncorrelated dither sequences can be used in each SDM. Consequently each ring is a source of incoherent high-frequency quantization noise, so there is no spatial noise filtering across rings and therefore no polar-angle dependency. The intercluster high-frequency quantization noise is effectively diffuse.

Finally it should be noted that the problem of noise degradation as a function of polar angle can theoretically be avoided completely by associating each element with an individual SDM. However, given the large number of elements, this architecture requires massive parallelism, and although attractive it is considered prohibitive. Consequently the paper has concentrated on grouping elements into clusters with only limited numbers of subband filters and SDMs. Nevertheless the balance between signal processing and cluster design is a fruitful area for optimization if such systems are to be developed, but this is deferred for future study.

It is recognized that the present study is in no way a complete solution for a digitally controlled loudspeaker array. Nevertheless it is hoped that some of the observations made will prove useful, especially in terms of system topology and conceptualization, and thus will act as catalysts to stimulate further research in this field. A pivotal requirement is seen as the means of implementing low-cost, high-density arrays of micro-displacement transducers to control a radiating surface in a precise way rather than just leaving its motion to the whims of a distributed vibrating structure.

7 REFERENCES

[1] M. O. J. Hawksford, "Smart Digital Loudspeaker Arrays," *J. Audio Eng. Soc.,* vol. 51, pp. 1133–1162 (2003 Dec.).

[2] Y. Huang, S. C. Busbridge, and P. A. Fryer, "Interactions in a Multiple-Voice-Coil Digital Loudspeaker," *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 48, pp. 545–552 (2000 June).

[3] S. C. Busbridge, P. A. Fryer, and Y. Huang, "Digital Loudspeaker Technology: Current State and Future Developments," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, p. 500 (2002 June), convention paper 5518.

[4] Y. Huang, S. C. Busbridge, and D. S. Gill, "Distortion and Directivity in a Digital Transducer Array Loudspeaker," *J. Audio Eng. Soc.,* vol. 49, pp. 337–352 (2001 May).

[5] H. Zhang, S. C. Busbridge, and P. A. Fryer, "Bit Expansion in Digital Loudspeakers with Oversampling and Noise Shaping," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 52, p. 806 (2004 July/Aug.), convention paper 6094.

[6] M. O. J. Hawksford, "Chaos, Oversampling, and Noise Shaping in Digital-to-Analog Conversion," *J. Audio Eng. Soc.,* vol. 37, pp. 980–1001 (1989 Dec.).

[7] N. A. Tatlas and J. Mourjopoulos, "Digital Loudspeaker Arrays Driven by 1-Bit Signals," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 52, p. 791 (2004 July/Aug.), convention paper 6036.

[8] R. Adams, K. Nguyen, and K. Sweetland, "A 112-dB Oversampling DAC with Segmented Noise-Shaped Scrambling," presented at the 105th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts*), vol. 46, p. 1027 (1998 Nov.), preprint 4774.

[9] M. O. J. Hawksford, "Time-Quantized Frequency Modulation, Time-Domain, Dither, Dispersive Codes, and Parametrically Controlled Noise Shaping in SDM," *J. Audio Eng. Soc.,* vol. 52, pp. 587–617 (2004 June).

[10] M. O. J. Hawksford, "Parametrically Controlled Noise Shaping in Variable State-Step-Back Pseudo-Trellis SDM," *IEE Proc. Vis. Image Signal Process,* vol. 152, pp. 87–96 (2005 Feb.).

[11] H. Takahashi and A. Nishio, "Investigation of Practical 1-Bit Delta–Sigma Conversion for Professional Audio Applications," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 49, p. 544 (2001 June), convention paper 5392.

[12] M. Gerzon and P. G. Craven, "Optimal Noise Shaping and Dither of Digital Signals," presented at the 87th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 37, p. 1076 (1989 Dec.), preprint 2822.

[13] V. P. Gontcharov and N. P. R. Hill, "Diffusivity Properties of Distributed Mode Loudspeakers," presented at the 108th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 48, p. 350 (2000 Apr.), preprint 5095.

[14] M. O. J. Hawksford and N. Harris, "Diffuse Signal Processing and Acoustic Source Characterization for Applications in Synthetic Loudspeaker Arrays," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 50, pp. 511, 512 (2002 June), convention paper 5612.

THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and he studied delta modulation and sigma–delta modulation (SDM) for color television applications. During this period he also invented a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, Colchester, UK, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems, with a strong emphasis on audio systems including signal processing and loudspeaker technology. Since 1982 his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. The first one (believed to be unique at the time) was developed in 1986 for a prototype system to be demonstrated at the Canon Research Centre, Tokyo and was sponsored by a research contract from Canon. Much of this work has appeared in *JAES,* together with a substantial number of contributions at AES conventions. He is a recipient of the AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," for the best contribution by an author of any age for *JAES,* volumes 45 and 46; and he has been awarded the AES Silver Medal for "major contributions to engineering research in the advancement of audio reproduction."

Dr. Hawksford's research has encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with special emphasis on SDM and its application to SACD technology. In addition, his research has included the linearization of PWM encoders, diffuse loudspeaker technology, array loudspeaker systems, and three-dimensional spatial audio and telepresence including scalable multichannel sound reproduction.

Dr. Hawksford is a chartered engineer and a fellow of the AES, IEE, and IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA).

5 Perceptual and multi-channel audio systems

5-1 Performance assessment

5-1 COMMUNICATIONS IN NOISE - PERFORMANCE RANKING METRIC, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *BT Technology Journal*, vol.10, no.4, pp 109-115, October 1992

5-8 CHARACTERIZATION OF COMMUNICATION SYSTEMS USING A SPEECHLIKE TEST STIMULUS, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *JAES*, vol. 41, no. 12, pp 1008-1021, December 1993

5-22 ERROR ACTIVITY AND ERROR ENTROPY AS A MEASURE OF PSYCHOACOUSTIC SIGNIFICANCE IN THE PERCEPTUAL DOMAIN, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R., *Proc. IEE of Vision, Image and Signal Processing*, vol. 141, no. 3, pp 203-208, June 1994

5-28 ALGORITHMS FOR ASSESSING THE SUBJECTIVITY OF PERCEPTUALLY WEIGHTED AUDIBLE ERRORS, Hollier, M.P., Hawksford, M.O.J. and Guard, D.R. *JAES*, vol. 43, no. 12, pp 1041-1045, December 1995

5-2 Multi-channel audio

5-33 SCALABLE MULTICHANNEL CODING WITH HRTF ENHANCEMENT FOR DVD AND VIRTUAL SOUND SYSTEMS, Hawksford, M. O. J., *JAES*, vol. 50, no. 11, pp 894-913, November 2002

Communication in noise — a performance ranking metric

M P Hollier, D R Guard and M O Hawksford

Users (and people standing near users!) are very much aware that mobile handsets are not well suited to use in conditions of high ambient noise. This is because some parameters important to their use in such conditions have been neglected. This paper considers the performance of fixed and mobile handsets in noise and proposes a method for ranking noise performance based on field experience and a simulation of the human acoustic interface. The method is validated with experimental results for CT2 handsets, analogue Cellphones, and a noise cancelling payphone handset. It is shown that handsets can be efficiently ranked objectively, in agreement with simple subjective performance. Future refinements of the method are planned, utilising the output of a number of existing research activities.

1. Introduction

There is a substantial trend in the telecommunications world towards mobile communicators. The introduction of cheap, flexible radio systems, and increasing expectation of payphone performance, has served to highlight the problems of communicating in conditions of high background noise. Mobile phone users often want to place and receive calls over the network while located in busy streets, in motorway service areas, on railway station platforms or while inside their cars. In such locations background noise is commonly sufficient to impair, or even prevent, satisfactory communication over the network.

When new services have been introduced, some manufacturers have given insufficient regard to the acoustic properties required from the handsets in non-ideal environments. This has been exacerbated by the legal requirement for the handset to comply with standards based upon the normal office and domestic situations.

Existing measurement methods describe many telephony parameters which can be combined, using a variety of assumptions, to give some useful indicators of the likely performance of telephones in noisy conditions. The way in which these existing measurement methods are currently combined is not appropriate to noisy environments. Nevertheless it has been possible to predict the performance in noise of some communicators, in terms of MOS (mean opinion score), using CATNAP (computer aided telephone network assessment program) modelling [1] with relevant performance parameters for the instrument under test entered into the programme.

An efficient objective method of ranking the performance, in terms of noise, of different telephones, which would be in agreement with subjective ranking is not presently available and would be immediately useful. Such a method would allow performance assessment of prototypes and might form the basis of a product specification or standard.

This paper introduces a number of related ideas and proposes a simplified method for ranking the performance of telephones in conditions of high ambient noise with the following main properties:

- method efficiency — measurement of large numbers of telephony parameters, and their combination with attendant assumptions, is not required,
- certain types of nonlinearity in the telephone under test are accounted for,
- the structure of the method is such that it will readily embody further refinements to a simulation of the human acoustic interface.

The proposed method has been verified for telephones known to exhibit widely different performance in noise. Field experience, and simple subjective evaluation by the experimenter, was used to validate the experimental

results and it was concluded that the method developed here will readily produce the expected performance ranking.

2. Communication in noise

2.1 The human acoustic interface

The primary acoustic interactions at the human interface during communication in conditions of high background noise are shown in Fig 1.

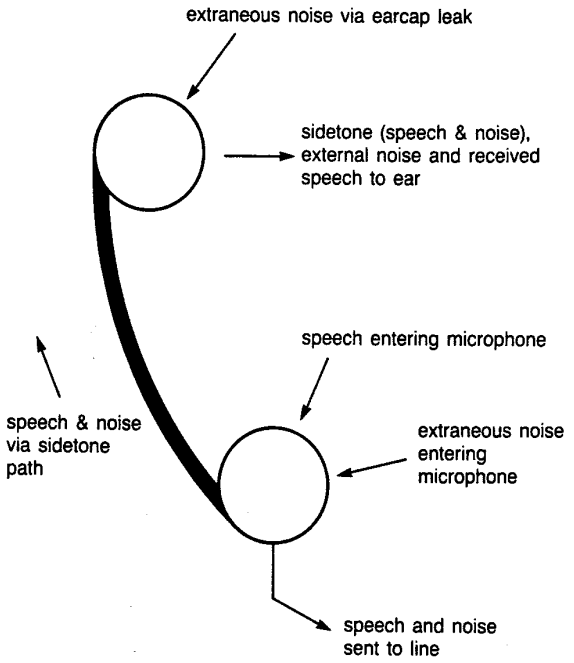


Fig 1 Primary interactions at the human interface.

The primary acoustic considerations are as follows:

- spectrum, level and temporal character of ambient noise,
- spectrum, level and temporal character of speech,
- obstacle effect of the user (including the obstacle effect of, and interaction with, the handset),
- radiating properties of the mouth,
- characteristics of the ear — in particular the 'compressed' ear, which must provide a representative ear-leak.

Some of these parameters can be simulated with existing test hardware.

2.2 Noise environment

The noise environments of a variety of mobile and payphone situations have been investigated during recent work. Measurements have concentrated on busy city centres, railway stations and motorway service area locations. It has become apparent from these studies that the average spectrum of the interfering noise has a similar characteristic in the majority of cases. An average of several typical noise spectra is shown in Fig 2, providing a useful frequency distribution for general analysis.

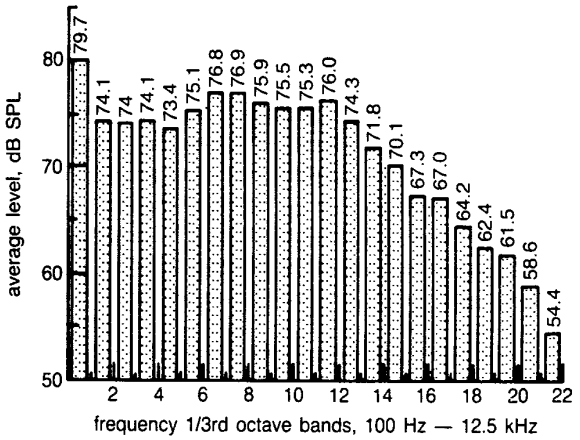


Fig 2 Average noise spectrum.

The power spectral density, $s(f)$, of the long-term average noise can be approximated by:

$$s(f) = 1377.94 - 1842.48(\log_{10} f) + 951.571(\log_{10} f)^2 - 214.785(\log_{10} f)^3 + 17.7415(\log_{10} f)^4$$

for a total SPL (sound pressure level) of 87 dB SPL. It is proposed that this spectrum is applicable over the range 70 to 90 dB SPL, i.e. the first term may be varied between 1360.94 and 1380.94.

The average noise spectrum is suitable for city-centre- and motorway service area locations where engine noise, vehicle movement and footfall are the most significant factors. It may also be used for railway station platforms which have, in general, the same average spectral distribution.

2.3 Speech characteristics

The power spectral density $s(f)$ of the long-term average of speech [2] can be modelled as:

$$s(f) = -282.44 + 465.439(\log_{10} f) - 157.745(\log_{10} f)^2 + 16.7124(\log_{10} f)^3$$

for a total SPL of 89.3 dB SPL, which is also applicable from 74.3 to 104.3 dB SPL, i.e. the first term may be varied between −276.44 and −297.44.

The increase in vocal effort which occurs in conditions of high background noise is well documented [3], but no account of any changes in voice spectra in noise has been considered in the above formula. Further development of the artificial voice is beyond the scope of this investigation. It may be noted that vocal effort

is also influenced by received speech level and sidetone conditions.

2.4 Overall performance

The dominant effect of increasing vocal effort, with increasing background noise, typically ensures that adequate speech signal is sent to line. In practice it is more common to find that it is inability to hear received speech which limits the environmental noise level in which communication remains possible.

It is interesting to note that by using the volume control on certain mobile handsets, it is possible to turn up the receive level to an unpleasant or even unbearable level. If this very high receive level is used, then the user tends to move the handset away from the ear breaking the earcap seal to the ear — this allows external ambient noise to enter the ear. Even in the highest background noise levels it is not beneficial to simply increase the receive level since:

- there is a limit to the bearable (and safe) receive level which may be employed,
- if the earcap seal to the ear is broken an increase in ambient noise entering the telephony ear will occur,
- the user may reduce his vocal effort and hence his transmitted speech level to the detriment of the far-end user.

The overall complexity of communications in noise should not be underestimated. The interactions of different telephony parameters can emphasise, or suppress, human speech and auditory mechanisms, and there are important mandatory limits on quantities such as maximum receive level [4] and noise dose [5].

3. Simulation of noise performance

It is proposed that a number of the items in the preceding discussion be combined to simulate telephone performance, in noise, which is representative of reality. In particular it should be possible to produce an objective measurement method capable of ranking the performance of telephones in conditions of high ambient noise which would be in agreement with their subjective ranking.

The audibility of the speech signal over the background noise will determine the viability of communication. By considering those locations where the average noise spectrum described above applies, and by utilising existing data on speech spectrum and vocal effort, it should be possible to predict the 'signal to noise' provided by different telephones in conditions of high background noise.

A simplified presentation of the results is achieved by applying perceptual weightings to the available signal spectrum to yield a single figure performance metric.

A simple test rig was constructed to verify the measurement method. The main assumptions for the experimental investigation are discussed below together with refinements which will result from known research programmes.

3.1 Handset orientation on the head

For the experimental investigation, the use of a particular handset on real people was observed and this orientation copied on to the head and torso simulator (HATS). This is a vital part of the method and great care was taken to obtain the best possible simulation of natural use, including an estimation of earcap pressure. A fixture for mounting handsets on to the HATS giving repeatable pressure on to the ear was devised by B&K for a CCITT Q12/XII (Question 12 study group 12) round-robin experiment. This fixture was employed for the experiment described in section 5; the HATS standard is still under development and therefore best endeavours were used to create the optimum simulation.

An on-going research programme at BT Laboratories is employing image processing techniques to measure the orientation of handsets on user's heads. The technique is non-intrusive and experiment conditions include high background noise. Completion of this activity and development of an appropriate means of simulating the modal position on a head and torso simulator is scheduled for 1992.

3.2 Artificial ear and representative ear-leak

Established artificial ear couplers which simulate the impedance of the human ear are available. Couplers complying with IEC 318 simulate the sealed ear condition [6] which is assumed for many telephonometry measurements, to reduce experiment complexity and variability. Couplers complying with IEC 711 simulate the occluded ear condition [7] and in association with an artificial Pinna and head obstacle, can be used to measure human hearing response in a free-field environment.

A key consideration for this method is representative ear-leak. The CCITT are developing a standard for a Telecom Pinna intended to simulate a representative ear-leak (Q12/XII). Preliminary tests showed that with great care the B&K flexible Pinna could be repeatably compressed to approximate an ear-leak. Hence this experiment employed an IEC 711 coupler located in a head and torso simulator with the flexible Pinna.

NOISE PERFORMANCE METRIC

Ongoing research into the modal position of the handset on the user's head will also investigate how ear-leakage varies with ambient noise levels and the effect of handset design on such mechanisms. Improvements in ear-leak simulation are likely to follow, particularly an adjustment to account for user reaction to background noise.

3.3 *Obstacle effect of the user*

The significance of the obstacle effect of the handset/head system when developing noise cancelling handsets has been previously demonstrated [8]. Established head and torso simulators are available and the CCITT are in the process of establishing a standard for such a simulator (Q12/XII). The B&K Type 4128 HATS complies with IEC 959 [9] and was employed for this experiment.

3.4 *Artificial mouth*

There are a small number of established artificial mouths used in telephonometry. The B&K Type 4128 HATS is fitted with a mouth simulator which was used for this experiment. The CCITT with BTL and others is researching further refinement for artificial mouth standards. The radiation pattern of the mouth and the near field characteristics are particularly relevant in conditions of high background noise when handset orientation may be closer to the mouth and noise-rejecting features, such as handset geometry and noise cancelling microphones, are increasingly used.

3.5 *Simulation of speech and background noise*

The long-term average spectra of speech and background noise were introduced above. As a first approximation these spectra can be used to shape broadband noise for the experiment. In this way it is possible to include the effect of certain simple system non-linearities. It has been shown that loudness ratings calculated using a pseudo-random noise stimulus are substantially better correlated with real speech results than estimates obtained with sinusoidal stimulus [10].

The need for speech and noise signals with the spectral, temporal, and amplitude characteristics of the original sources is apparent since the characterisation of nonlinear systems is stimulus dependent. The CCITT P.50 artificial speech signal may provide a suitable speech-like stimulus, although refinements are under consideration (Q14/XII).

4. Audibility of available SNR

The audibility of the speech signal is controlled by the complex interaction of the speech signal level, level and masking effect of the background noise (both of which have important temporal characteristics), and the cognitive nature of the speech signal. The potentially increased disturbance of a background noise signal with an intelligible content is acknowledged, but can be ignored since it would be expected to produce only detail differences in the way in which different telephones would be ranked under the same conditions.

Since the initial objective of the method is to rank the performance, in noise, of different telephones, it is apparent that a single figure performance metric is highly desirable. Recent papers have shown the use of band pass auditory filters, masking effects and perceptual weightings to calculate the loudness and hence the audibility of a particular signal or distortion [11]. The perceptual weightings employed are determined from binaural, free-field data and must be corrected prior to application to monaural telephony applications.

A simplification is possible by referring to existing monaural weighting functions, which were developed to enable the calculation of perceived loudness for a telephone connection. These weightings can be found in the CCITT P.79 Recommendation and are used routinely in telephonometry. The weighting characteristics for sending (W_{Sn}) and receiving (W_{Rn}) are shown in Fig 3.

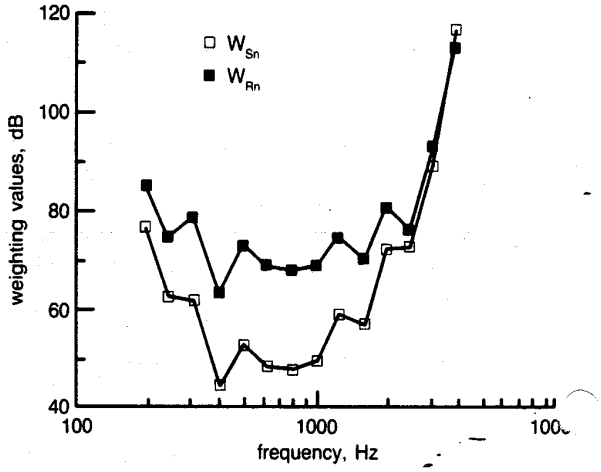


Fig 3 Loudness rating weightings.

The weighting characteristics are applied to discrete frequencies rather than frequency bands.

It is necessary to rework the loudness rating algorithms of P.79 in order to obtain a single figure for signal loudness of zero when the absolute available signal is zero, and to account for the application of the weighting characteristics to BPLs (band pressure levels) rather than discrete measured frequencies. Reworking the algorithms will also enable the calculated metric to be distinguished from the 'loudness rating' which has assumed a definite meaning in telephonometry.

The sending signal-to-noise weighted acoustic response, *SendSNWAR*, is given by:

$$SendSNWAR = 4/5 \sum_{n=1}^{14} SSig_n * 10^{(-0.0175*W_{Sn})}$$

where:

n are the 14 ⅓rd octave bands centred on the standard ISO test frequencies,

$SSig_n$ is the signal/noise band pressure level centred on the nth frequency,

W_{Sn} is the SLR weighting for the nth frequency.

The receiving signal-to-noise weighted acoustic response, *RecSNWAR*, is given by:

$$RecSNWAR = 2/3 \sum_{n=1}^{14} RSig_n * 10^{(-0.0175*W_{Rn})}$$

where:

n are the 14 ⅓rd octave bands centred on the standard ISO test frequencies,

$RSig_n$ is the signal/noise band pressure level centred on the nth frequency,

W_{Rn} is the RLR weighting for the nth frequency.

Adaptation of the P.79 loudness rating algorithms has proved successful in previous work on noise cancelling handsets [8], when the single figure performance metric obtained was validated using CATNAP modelling [1].

5. Experimental investigation

Four telephones with known field performance were selected. These telephones represent different technologies and a significant range of subjective performance in conditions of high background noise.

- CT2 — this handset is a short folding design with a plastic flap intended to direct the user's speech back toward the microphone which is located close to the cheek. The design also features a slotted earcap.
- Cellphone A — the long, straight handset places the microphone away from the mouth and directed into the ambient noise. The concave earcap is rather shallow and is awkward to locate accurately on to the ear in natural use.
- Cellphone B — this folding pocket size cellphone has its single pressure gradient noise cancelling microphone in the folding portion which brings it close to the mouth during use. The earcap is adequately concave and locates naturally on to the ear during use.
- BT table-top payphone — this payphone uses a conventional handset design with a pressure-gradient, noise cancelling microphone arrangement. The earcap is well designed and provides a good seal to the ear in natural use.

The experimental set-up for the evaluation of the signal-to-noise ratio is shown in Fig 4. The set-up is used to measure the signal sent to line, or in the telephony ear. The signal sent to line is measured at the POI (point of interconnect) or junction. The available signal in the telephony ear is measured via an IEC 711 coupler within the head and torso simulator.

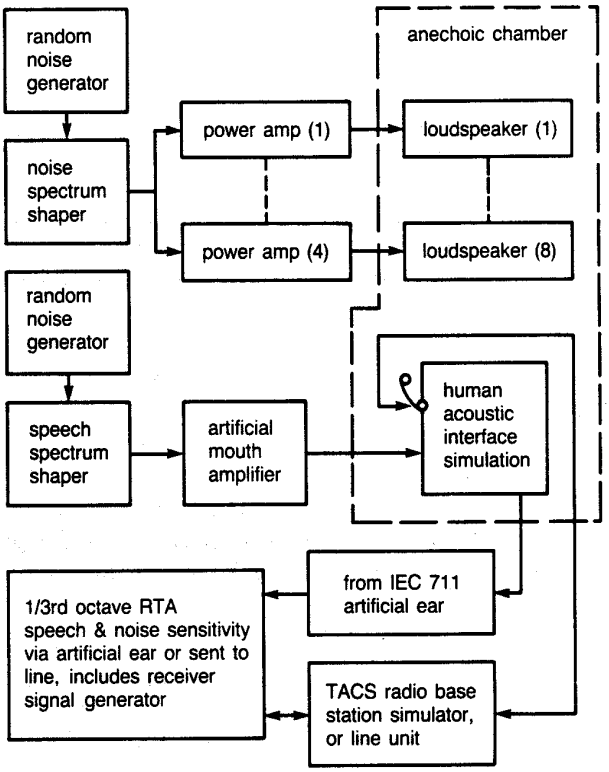


Fig 4 Experimental set-up.

The user's speech and background noise spectra and levels were discussed in section 2. It is also necessary to provide a 'speech' signal to the telephone under test in order to measure the signal in the telephony ear. Since the experimental method is fundamentally comparative, it was deemed satisfactory to set a signal voltage from the network to the telephone which would:

- be the same for all telephones tested to provide the required comparison,
- produce an SPL in the artificial ear representative of received speech in field conditions.

An appropriate junction voltage was set by measuring the SPL of the telephone receiver into an IEC 318

artificial ear. Figure 5 shows the junction and POI for simple telephone and TACS equipment.

In practice, it is possible to validate transmit and receive signal levels to and from the TACS handsets by monitoring the peak carrier frequency deviation of the radio link between the handset and base-station. According to the TACS compatibility specification [12] a peak carrier frequency deviation of ±2.3 kHz shall produce a nominal acoustic output level of −10 dB Pa (84 dB SPL) into a sealed IEC 318 coupler.

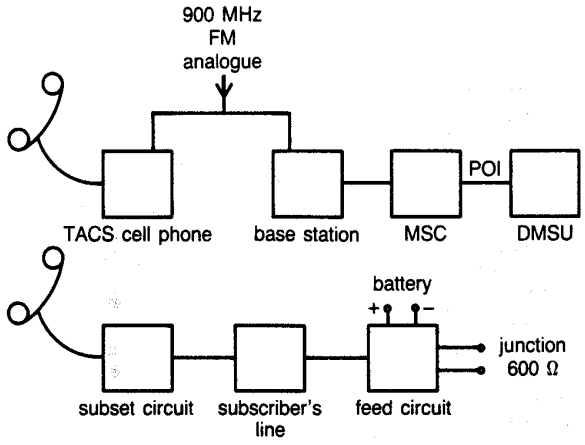


Fig 5 Junction location for simple telephone and TACS.

6. Experimental results

The sample telephones were tested as described above and the resulting spectra processed to give single figure performance ranking estimates for sending and receiving performance, as shown in Table 1.

Table 1 Ranking estimates for sending and receiving performance.

Instrument	SendSNWAR	RecSNWAR
CT2	8.7	−16.2
Cellphone A	7.9	−10.9
Cellphone B	11.3	−5.9
BT table-top payphone	26.2	−2.7

The value of the result increases (become less negative or more positive) with improving performance, hence ranking orders are apparent.

7. Experiment validation

The telephones tested have known performance in noise. They have been the subject of previous studies and experience of their field performance has been gained. The perceptual ranking of the designs in question was thus established. To provide a further check a simple subjective ranking was performed, increasing the level of shaped Gaussian background noise until a receive speech signal was rendered unintelligible. The results are shown in Table 2.

Table 2 Background noise level rendering a speech signal unintelligible.

Instrument	Noise level which prevented communication (dB SPL)
CT2	77.8
Cellphone A	82.4*
Cellphone B	82.5*
BT table-top payphone	83.6

*Very high levels can be set with receiver volume controls. It is likely that customers would tolerate only levels lower than those tested here which would reduce the noise level shown.

The range of noise level which prevented communication is of the order 10 dB. This is a relatively small range and is consistent with modelling results which show a very rapid loss of utility causing the MOS to fall rapidly with increasing background noise once a certain level is exceeded. A typical characteristic, generated by the CATNAP model [1], is shown in Fig 6.

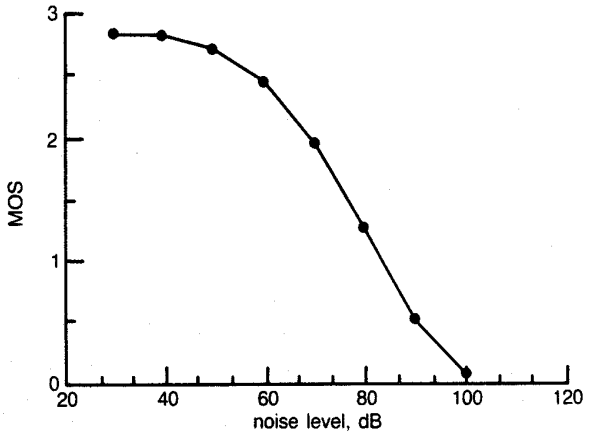


Fig 6 Background noise vs MOS, generated by CATNAP model.

The rapid fall in MOS with increasing background noise indicates the major performance advantage from relatively small gains in noise rejection performance. Such improvements will significantly improve MOS and even enable communication where it would not otherwise be possible when noise is at a critical level. Hence the significance of the spread in reported performance.

8. Conclusions

The objective ranking method described produces a performance ranking in general agreement with experience of field performance and simple subjective ranking.

Objective ranking is consistent with handset design parameters, e.g. long handsets, poor earcap design and noise cancelling microphones. It is interesting to note that the table-top payphone achieved a higher noise performance ranking than the TACS cellphones which produce extremely high receive levels. This is attributed

to the superior earcap design on the table-top payphone which provided a good seal to the ear, while the very loud receive levels from the TACS cellphones caused the user to break the earcap seal to the ear.

The test method described is expected to be improved with a series of refinements, but even in its simplified form provides results which are in broad agreement with subjective results.

As further improvements are introduced the performance of the measurement method will improve. In particular, routine use will be simplified and it will become possible to distinguish more subtle differences in handset design.

References

1 Webb P K: 'The background and philosophy of the telephone network assessment program (CATNAP)', Internal BT memorandum (1979).

2 CCITT Recommendation P.30, Artificial Voices, Melbourne (1988).

3 Richards D L: 'Some aspects of the behaviour of telephone users as affected by the physical properties of the circuit', Communication Theory, Butterworths (1953).

4 BS6317: British Standard for simple extension telephones for connection to the British Telecommunications Public Switched Telephone Network (1982).

5 The Noise at Work Regulations, Health and Safety Executive (1989).

6 IEC Publication 318, An artificial ear of the wideband type for the calibration of earphones used in audiometry.

7 BS6310: (ISO 711:1981), Occluded-ear simulator for the measurement of earphones coupled to the ear by ear inserts (1982).

8 Hollier M P: 'Sound fields around the head/handset system and their exploitation for pressure gradient noise cancellation', Proceedings of the IOA, 12, Part 10, pp 523—530, Windermere (1990).

9 IEC 959:1990, Provisional head and torso simulator for acoustic measurements on air conduction hearing aids (1990).

10 Bell Northern Research, Measurement of carbon microphone using real voice and artificial test signals, CCITT Q11/XII, Contribution No 11 (April 1981).

11 Stuart J R: 'Psycho-acoustic models for evaluating errors in audio systems', Proceedings of the IOA, 13, Part 7, pp 11—34, Windermere (1991).

12 United Kingdom total access communications system mobile station — land station compatibility specification, Issue 4 (August 1989).

Mike Hollier obtained a BEng(Hons) degree in mechanical engineering from Plymouth Polytechnic in 1987 and joined BT Laboratories on graduation. He obtained an IOA diploma in Acoustics in 1989 and is a member of the IOA.

With a background in audio engineering he has worked on a number of telephony acoustics projects including the design of a novel noise cancelling payphone handset. He is currently an external PhD student to the University of Essex developing advanced speech performance measurement methods.

David Guard graduated in 1971 with a BSc in physics, followed by a PhD in medical acoustics in 1975. He then joined BT Laboratories to investigate loudspeaking telephones, culminating in the development of the Orator executive audio conferencing system. He moved into the marketing area to sell Orator nationwide. He returned to BT Laboratories in 1983 to develop and support telephony terminals. Since 1987 he has worked on speech transmission performance of the complete network.

Malcolm Hawksford is a Reader in the Department of Electronic Systems Engineering at the University of Essex, where his principal interests are in the fields of audio engineering and electronic circuit design. He studied at the University of Aston in Birmingham and gained both a first class honours BSc and PhD.

Since his employment at Essex, he has established the Audio Research Group, where research on amplifier studies, digital signal processing and loudspeaker systems has been undertaken. Since 1982 research into digital crossover systems had begun within the group and, more recently, oversampling and noise shaping investigated as a means of analogue-to-digital/digital-to-analogue conversion.

He is a member of the IEE, a chartered engineer, a fellow of The AES and of the Institute of Acoustics.

Characterization of Communications Systems Using a Speechlike Test Stimulus*

M. P. HOLLIER, MALCOLM O. J. HAWKSFORD***,** *AES Fellow*, **AND D. R. GUARD****

***BT Labs, Martlesham Heath, Ipswich, IP5 7RE, UK*
****Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, CO4 3SQ, UK*

Conventional objective means of characterizing communications systems are often inadequate. In particular when a system contains nonlinear elements, such as low-bit-rate encoders, some types of speech transducer, active gain control, and echo cancelers, simple steady-state measurements are not sufficient to predict subjective performance. The use of speechlike test stimuli to characterize nonlinear systems is not new, but a new speechlike composite signal is presented which works in conjunction with a new perception-based objective measure to provide a versatile measurement technique. Rigorous subjective validation of the new method has still to be completed, but early results are given which indicate the usefulness of the method to provide a perceptual ranking for communications systems and also as a diagnostic tool.

0 INTRODUCTION

Ever increasing demands on available bandwidth for transmission, more demanding ambient noise environments, and increasing user expectation, all contribute to the need for reliable characterization of speech systems which accord with customer perceptions. By optimizing system design for human auditory perception it is possible to maximize subjective performance while ignoring perceptually insignificant errors.

Established measurement methods employ a variety of test stimuli: discrete-frequency sine waves [1], swept sine waves, chirp [2], fast Fourier transform (FFT) analysis with random noise excitation [3], maximum-length-sequence analysis (MLSA) with pseudorandom noise [4], and true impulse testing [5]. With the exception of chirp and impulse testing, all these test stimuli result in a pseudo-steady-state system excitation. The adequate characterization of a system with such test stimuli is thus dependent on 1) the signal that the system is intended to reproduce in service, and 2) whether the system can be reasonably approximated as linear.

The communications systems of interest here must reproduce a speech signal and can certainly contain nonlinear elements. Hence it is apparent that the pseudo-steady-state excitation produced by these stimuli may not lead to a satisfactory performance measure. True impulse testing and chirp testing are not without difficulty and cannot be said to be speechlike. Accepting that it is the speech performance of the system that is of interest, two key considerations are apparent:

1) The test stimulus must contain the salient features of speech in terms of physical excitation but not necessarily intelligence. The properties of natural and artificial speech are introduced in Section 1.

2) The analysis of the system output in response to the speechlike signal must reflect the perceptual significance of any errors or distortion produced. In this way the performance of the system under test is related to the subjective performance that will be experienced by a user.

Human auditory perceptual modeling is a rapidly developing field in which objective evaluation is used to predict subjective performance. Perception-based analysis is introduced in Section 4.

The evaluation of relatively gross distortions which are certainly audible is required. This can be compared with other models examining the performance of hi-fi codecs, such as PERCEVAL [6], which examine the probability of detection and are not suitable for evaluating audible distortions. Further, since the characteristics of real systems may cause loss of the original

* Presented at the 93rd Convention of the Audio Engineering Society, San Francisco, CA, 1992 October 1–4; revised 1993 February 12 and September 7.

signal, an indication of missing or attenuated signal is also required. A practical measure producing error loudness estimates is presented in Section 5.

Finally in Section 6 the measurement method, which has yet to be rigorously validated using subjective test data, was assessed using a set of precisely reproducible nonlinear distortions generated by digital signal processing. The results obtained are described and illustrated in Section 7.

1 SPEECH PROPERTIES

To understand the construction of speech sounds better it is useful to consider its physiological origins [7]. The human speech production organs are shown in the sagittal section of the head in Fig. 1. Air can be expelled through the oral or nasal cavities, or both, depending on the position of the velum and lips. The nasal cavity is fixed and gives the distinctive characteristic to an individual's speech while the oral tract is of highly variable volume.

The different voiced speech sounds (such as M and L) are formed by bringing the vocal folds together and varying the positions of the articulators to create a continuously varying set of resonators. The spectrum of a short segment of a vowel is illustrated in Fig. 2. It shows a number of spectral peaks, called formants, which is the expected characteristic of this combination of resonant cavities.

In an unvoiced sound (such as SH and T) the vocal folds are not brought together, but air is passed between two articulators, resulting in turbulent flow. The rapid release of two articulators also produces an unvoiced sound such as P. Fig. 3 shows the spectral characteristics of a voiceless sound, which has very little formant structure. Some sounds such as V and Z involve simultaneous vocal fold vibration and turbulent airflow through the oral cavity.

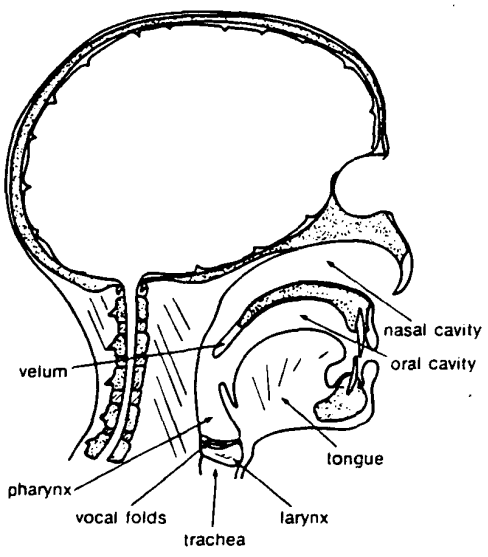


Fig. 1. Sagittal section of human head showing speech production organs.

The continuously varying speech waveform can be represented discreetly as a series of segments known as phonemes. The phonological taxonomy of English contains 40 to 50 phonemes. This number of discrete segments is needed if the intelligence of the signal is to be represented. If, however, the representative physical excitation of a system is required, then the intelligence of the signal is not important and energetically similar segments can be grouped. It is therefore proposed that an artificial speech signal intended to be used for the characterization of physical systems could be composed of a relatively small number of segments.

The generation of an artificial voice signal without intelligence but suitable for the excitation of a physical system is not new and the CCITT P.50 recommendation [8] describes such a signal. The signal is generated by applying two different types of excitation source signals to a time-variant spectrum-shaping filter. The two excitations are glottal and random noise, corresponding, respectively, to voiced and unvoiced sounds. The frequency response of the spectrum-shaping filter simulates the transmission characteristics of the vocal tract.

The P.50 signal was originally continuous, but a modification to allow the simulation of conversational speech has been subsequently incorporated into the recommendation. The speech characteristics reproduced by the artificial voice signal are:

1) Long-term average spectrum
2) Short-term average spectrum
3) Instantaneous amplitude distribution
4) Voiced and unvoiced structure of the speech waveform
5) Syllabic envelope.

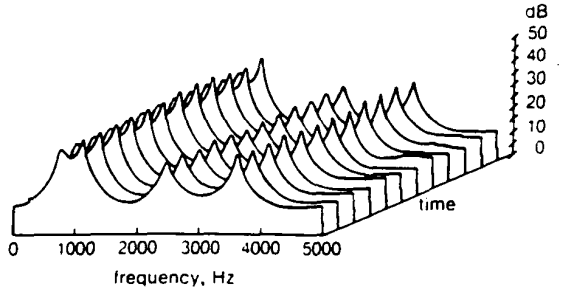


Fig. 2. Spectral characteristics of voiced vowel sound.

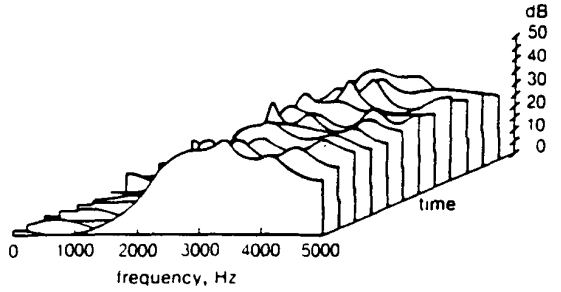


Fig. 3. Spectral characteristics of voiceless sound.

Further development of artificial speech signals is planned. In this paper artificial speech events arranged into a test composite are introduced and were used in the experimental investigation presented in Section 7.

The artificial voice can exist as an electrical signal or an acoustic waveform via an artificial mouth [9].

The long-term spectrum and amplitude distribution characteristics of P.50 are developed over 10 s. It is thus anticipated that a conventional FFT transfer function or time-averaged one-third-octave analysis will be performed over a comparable time. Temporally varying system characteristics and distortions important to perceptual performance will be similarly averaged over this period, "smearing" their effect.

2 SPEECH EVENT CONTEXT

Accepting that the systems to be characterized may contain perceptually significant nonlinearities, a measurement method is therefore required that will reflect the characteristics of 1) the excitation signal and 2) the perceptual significance of the distortions produced.

The way in which a nonlinear system will respond to a given speech event will be dependent on the condition of the system, that is, what has occurred previously. It is therefore proposed that in order to determine the system response to any given speech event, the "event context" must be considered. It follows that any distortion produced by the system will also depend on the event context. Speech event sequences are determined by physiological and linguistic considerations.

In order to test anything other than the simplest systems the conditioning signal must also be speechlike. An example stimulus construction is shown in Fig. 4.

The length of the conditioning portion of the signal must be commensurate with the time constants of the system under test. Example time constants include codec adaption and active gain control, with time constants on the order of a few seconds, and speech transducer transient response, which is on the order of a few milliseconds.

A set of composite signals which create the required range of event contexts and then exercise the required system parameters with a speech characterized transient event will allow a speech system to be fully characterized. Example composites are used in the experimental investigation reported in Section 6.

The system response to the speech transient must be analyzed to determine the perceptual significance of any response errors.

3 NONLINEAR SYSTEMS

Examples of nonlinearities in the telecommunications system are numerous. Examples of strong and weak nonlinearities in the frequency and time domains may be cited. Examples include speech transducer transient response, low-bit-rate coders (including adaptive forms), radio fading, echo cancelers, burst errors, voice switches and voice activity detectors (VADs), and automatic gain controls (AGCs).

It is significant that these examples include temporal "distortions," including characteristics that make the system response time variant. Further, even a simple voice-switched telephone will recognize that steady-state noise and pure tones are not speech and revert to "idle," underlining the inadequacy of many established test stimuli. CCIR TG10/2 [10] and others have previously assessed perceptual encoding methods, including for use in high-quality audio transmission over telephone lines [11].

4 HUMAN AUDITORY PERCEPTUAL MODELING

There are two main approaches to the prediction of subjective performance from objective measurements:

1) The empirical (or statistical mapping) approach uses one or more objective measures, such as cepstral distance and segmental signal-to-noise ratio (SNR), and then employs advanced statistical methods such as clustering theory to map these objective measures onto a field of known subjective data. It is then possible to predict the subjective rating of distortions within the known data field.

2) The second approach is the subject of this section and is generally referred to as auditory modeling. In this approach the perceptual significance of an audio event or "error" is assessed by modeling the stimulation of the auditory system either in terms of overall psychoacoustic quantities such as masking and loudness, or by modeling the individual processes in the peripheral auditory system. Such an approach is robust and able to evaluate distortion types that were not known during the design of the model.

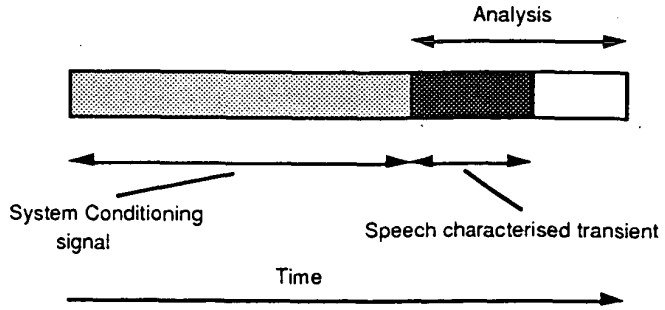


Fig. 4. Example construction for speech characterized composite test signal.

4.1 Frequency Discrimination

We do not hear frequency, but we perceive pitch. There is a nonlinear relationship between frequency and pitch, sometimes referred to as frequency warping. Auditory frequency selectivity is commonly represented as a set of bandpass filters. Examples are Zwicker and Zwicker's critical bandwidth [12] (Figs. 5 and 6) and Patterson and Moore's equivalent rectangular noise bandwidth [13].

Parallel filters are physically simple to construct or synthesize, and many models have been devised. There are two main methods for implementation:

1) *FFT Spectrogram.* The time-domain signal is converted to frequency using the Fourier transform and postprocessed to simulate bandpass filtering [14].

2) *Digital Filters.* The time-domain signal is digitally filtered, with filter characteristics representing the human auditory response [15].

The basic concept for such a model is shown in Fig. 5.

4.2 Auditory Masking

The presence of an auditory stimulus will limit the audibility of other stimuli. The modified auditory threshold which thus occurs when an audio event is present is known as the masked threshold.

The masked threshold due to a signal is dependent on the frequency and level of that signal, that is, it is nonlinear with respect to frequency and level (intensity). Fig. 6 shows the psychoacoustical masking of test tones due to narrow-band noise at different levels, after Zwicker and Zwicker.

Fig. 7 shows the auditory masking patterns for narrow-band noise at 60-dB SPL masking pure tones on a critical-band rate rather than frequency scale, after Zwicker and Zwicker [12]. (The broken line is the threshold in quiet conditions.) The critical-band rate, or bark, scale is derived by considering a set of narrow-band excitations where the adjacent masking curves cross at the −3-dB point. The characteristic shape of masking patterns due to narrow-band noise is directly related to the characteristic of the auditory filter shape for frequency discrimination.

When an auditory stimulus is removed, the masking effect of this stimulus does not cease immediately, but decays with a characteristic that it dependent on the duration of the masker. This phenomenon is known as forward masking. Fig. 8 shows the relationship between the duration of a masking tone burst and the duration of the masking effect after the tone burst has finished, after Zwicker and Zwicker.

4.3 Auditory Stimulation

We do not hear level (intensity), but we perceive loudness and other subjective quantities. The relationship between intensity and perceived sensation is nonlinear with respect to frequency and level. In particular small changes in intensity close to threshold will give a far greater change in perceived loudness than the same change in intensity at a higher level. This is easily explained in evolutional terms since we must be able to hear high-energy sounds while on other occasions we must be aware of low-energy sounds such as distant prey or the approach of a large carnivore. The evolution of our hearing has also been central to the evolution of our speech.

4.4 Anatomical Models

A further class of auditory model attempts to estimate the perceptual effect of a given signal by applying this signal to a model of the human auditory anatomy. Models tend to concentrate on the inner ear, the cochlea. Models of greatly varying complexity have been reported in the literature (such as [16]), ranging from full three-dimensional models of the complete cochlea to more practical one-dimensional models of the basilar membrane (the cochlea partition). The auditory stimulation due to a given input can then be described in terms of its effect on the hearing sensation detection

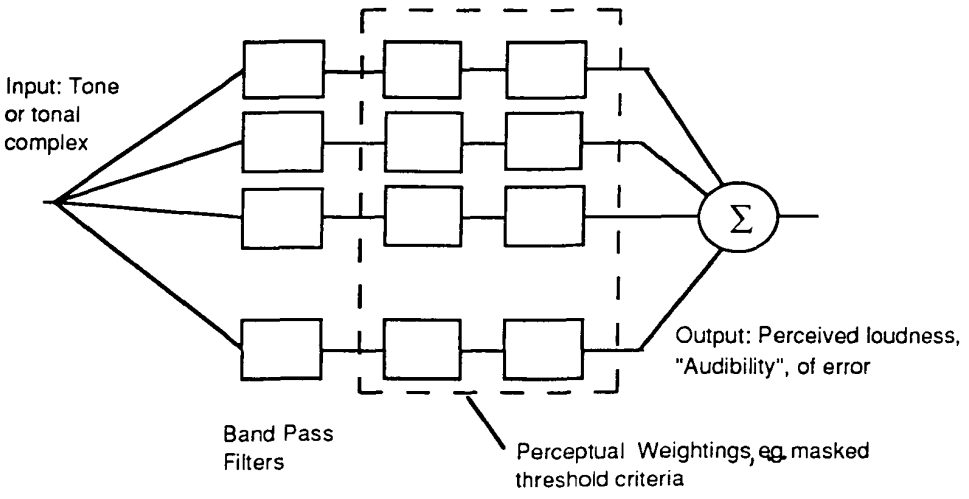


Fig. 5. Parallel filter concept for auditory modeling.

structures. Further levels of processing are then required to interpret the perceptual significance of such sensation.

5 PRACTICAL PERCEPTUAL MEASURE

A practical measurement process is required to objectively assess the perception of distortion of the test stimulus when passed through a system.

5.1 Segmentation in Time and Frequency

Zwicker's critical bands [10] are similar in shape below 500 Hz when represented on a linear frequency scale and above 500 Hz when viewed on a logarithmic frequency scale. The telephony bandwidth is typically 300–3150 Hz, and so similar filter shapes viewed on a logarithmic frequency axis may be used with little compromise.

The ISO one-third-octave band filters are similar in shape when viewed on a log frequency axis and are similar to the critical bands discussed in the preceding.

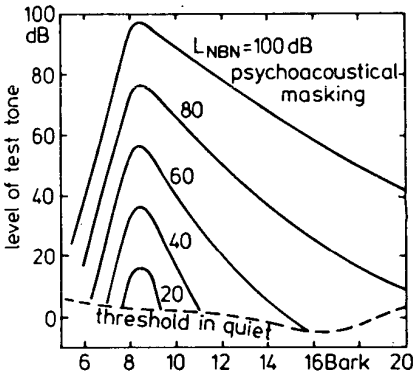


Fig. 6. Psychoacoustic masking of test tones due to narrow-band noise at different levels (after Zwicker).

The ISO bands were developed to be representative of human perception of sound in general acoustics and form the basis of many perceptual analyses such as ISO 532B, the calculation of the subjective loudness of complex spectra. Because the use of one-third-octave analysis is highly developed, proprietary equipment can perform the required acquisition efficiently, yielding results in a variety of formats.

Time and frequency segmentation is achieved by analyzing the signal as a one-third-octave multispectrum with $T_{av} \geq 4$ ms (Fig. 9).

5.2 Excitation Estimation

As described in Section 4.3, the auditory sensation resulting from an auditory stimulus depends on the frequency and level of the stimulus. The auditory threshold and equal-loudness contours are given in ISO 226. The equal-loudness contours are used to obtain the nonlinear mapping from intensity to auditory excitation.

5.3 Threshold and Temporal Masking

The acquired time segments are postprocessed to apply the threshold of sensation, and calculation products are carried forward from segment to segment to produce a forward masking estimate.

The forward masking calculation used in the perceptual model is based on a simplification of the data for Fig. 10. The approximated decay characteristics are shown in Fig. 10 for 5- and 200-ms-duration masking signals. These durations are the likely extremes encountered in speech.

The frequency discrimination of the ear and the temporal extension of masking, known as forward masking, are combined with the auditory thresholds given in ISO 226 to produce an excitation estimate "surface"

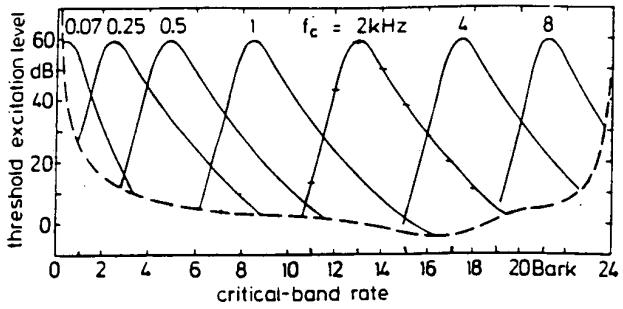


Fig. 7. Excitation level versus critical-band rate for narrow-band noise of given center frequency and 60-dB SPL (after Zwicker).

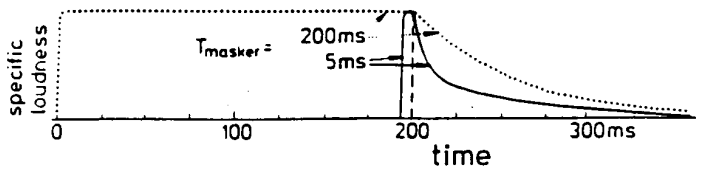


Fig. 8. Specific loudness of masker bursts of 200 and 5 ms versus time (after Zwicker).

due to the input signal. The excitation estimate for a typical speech characterized transient is shown in Fig. 11.

The speech event shown comprises a voiced sound followed by an unvoiced sound. The formant structure of the first sound and the broadband character of the second element can readily be distinguished. The excitation estimate for two pulses is shown in Fig. 12, in which the forward masking characteristic is clearly visible.

5.4 Error Surface

For a unity-gain system the audible error is the amplitude transfer function of the original and distorted excitation estimate surfaces. If the measured system is not unity gain, it is necessary to normalize the amplitude of the two speech signals. This can be achieved by amplitude scaling (speech signal/rms power) or by normalizing each psychoacoustically transformed segment with the average psychoacoustically weighted energy of the corresponding original segment.

The error surface will be flat for a perfectly reproduced output. When errors occur, additive distortions are represented as peaks and signal loss as troughs. Examples of typical error surfaces are given in Section 6.

5.5 Error Loudness

In order to simplify interpretation of the results, a "loudness of error" is calculated for the error spectrum of each time segment. There is a nonlinear mapping between the error surface segments and subjective quantities such as intelligibility and loudness.

The calculation of the perceptual loudness of a complex spectrum is given in ISO 532B. However, this calculation assumes that the sound was binaural. A useful simplification is therefore possible by utilizing the established monaural telephony perceptual weightings for loudness [17]. The perceptual weightings for monaural telephony loudness are shown in Fig. 13.

The single-figure loudness is obtained for a narrowband telephony model using the following expression:

$$\mathrm{ErrLoud}_t = 0.8 \sum_{n=1}^{14} \mathrm{Er}_n * 10^{-0.0175 * W_{Sn}}$$

where

$\mathrm{ErrLoud}_t$ = error loudness at time t (positive and negative parts calculated separately)

n = nth one-third-octave band from 200 Hz to 4 kHz

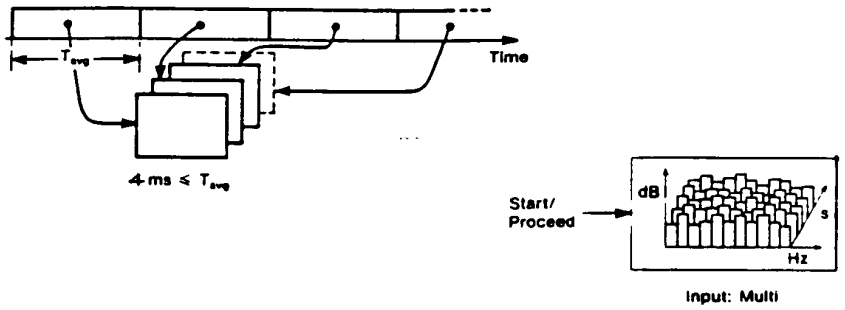


Fig. 9. Time and frequency segmentation of signal.

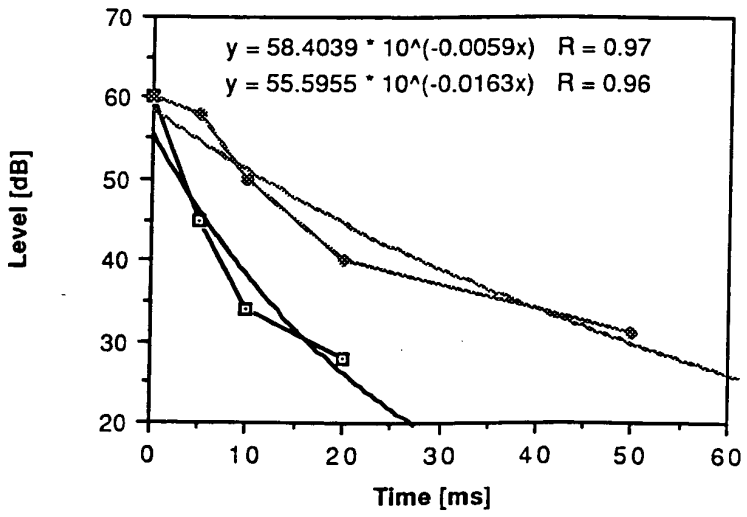


Fig. 10. Forward masking approximation for 5- and 200-ms duration masking extremes.

Er_n = error amplitude, dB
W_{Sn} = SLR weighting for nth frequency.

For a broad-band telephony model it is obtained from

$$\text{ErrLoud}_t = 1.28 \sum_{n=1}^{21} Er_n * 10^{(-0.0175 * W_{Sn})}$$

where the terms are as before except for n being the nth one-third-octave band from 100 Hz to 8 kHz.

The additive error (+ve) and the signal shortfall (−ve) are calculated separately for each error spectrum, yielding an error loudness versus time result of the form shown in Fig. 14.

Simple acceptance criteria can be devised, such as a maximum acceptable error loudness coupled with a maximum acceptable running average error loudness. Subjective tests will be required to set appropriate acceptance criteria for a fully developed test.

Alternatively the phon-to-sone mapping, originally proposed by Stevens [18], can be considered for the estimation of error loudness.

6 NONLINEAR TEST SYSTEMS

In order to investigate and refine the measurement method described, a series of experiments were undertaken to illustrate the method's key features using a set of nonlinear test systems. To maximize the accuracy of these experiments, the distortions and system

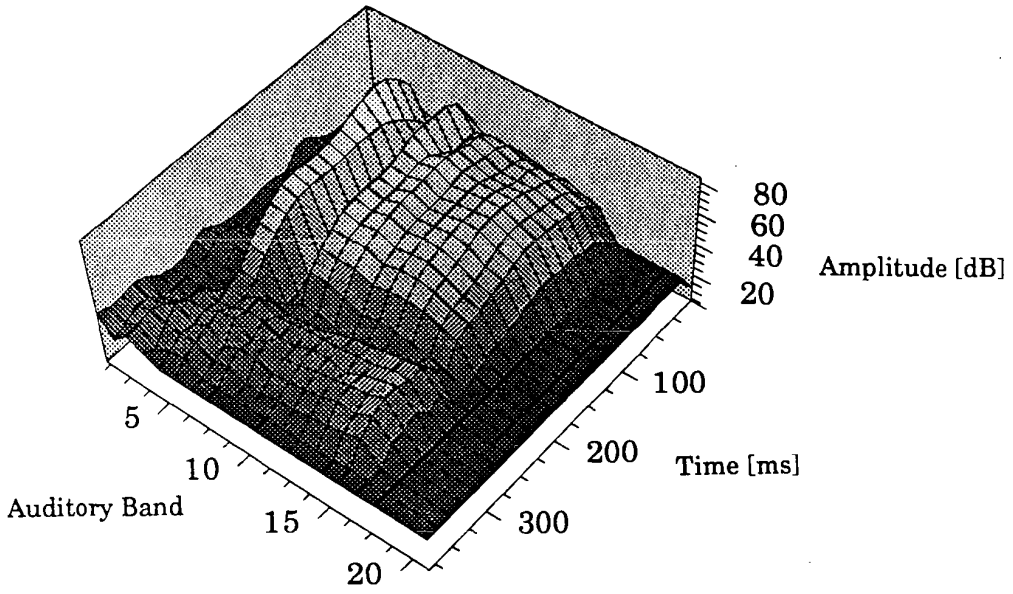


Fig. 11. Auditory excitation estimate for speech characterized transient event.

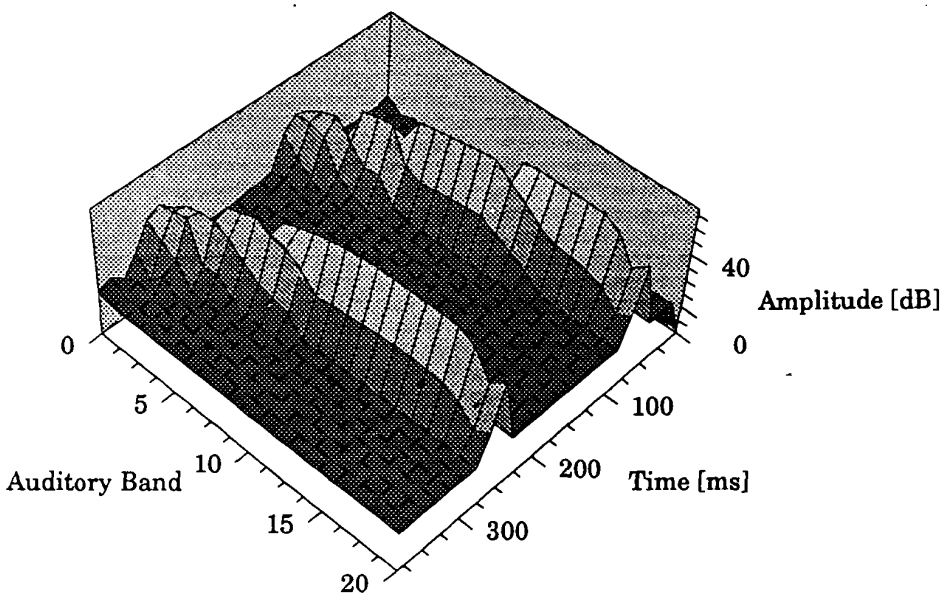


Fig. 12. Auditory excitation estimate for clicks, showing calculated forward masking characteristic.

parameters of the nonlinear systems under test were generated algorithmically using the DSP network emulator [19] developed at BTL. The network emulator is based on cards of AT&T DSP32C 32-bit floating-point digital signal processors which can emulate a wide range of generic distortions, speech coders, network types, echo paths, and VADs.

Four error types were investigated to illustrate the diagnostic and performance ranking properties of the measurement method. In general two levels of each distortion type were chosen to give just audible distortion and fully audible distortion. These levels were not rigorously established, but simply subjectively selected by the experimenter. Characterization of gross temporal effects was of particular interest since this is generally difficult with conventional objective metrics.

1) *Instantaneous Amplitude Nonlinearity.* Measurements of the error surface for "just" audible and "fully" audible degrees of distortion were made, although the subjective impact of the distortions was not rigorously established for this initial investigation. The additive distortion is governed by a polynomial where the output y is related to the input x,

$$y = x + 2 \times 10^{-5}x^2 + 1 \times 10^{-8}x^3$$

for the just audible case

$$y = x + 2 \times 10^{-5}x^2 + 1 \times 10^{-7}x^3$$

for the fully audible case .

2) *Modulated Noise Reference Unit (MNRU)* [20]. 1 MNRU is theoretically equivalent to the distortion introduced by one A-law PCM stage. Error surfaces for just audible and fully audible distortion were generated:

$a = 0.2$ to generate just audible distortion

$a = 0.8$ for fully audible distortion .

where MNRU is given by

$$n = \frac{\log(a/1.0789)}{-4.9424 \times 10^{-2}} \quad \text{and} \quad Q = 37 - 15 \log n$$

as per Annex D to Supplement 14, CCITT Blue Book V1.V.

3) *Crossover Distortion.* Crossover distortion is amplitude nonlinear and governed by the expressions

$$y = mx + c , \quad x > 0$$

$$y = mx - c , \quad x < 0 .$$

The cross over intersection c can be set to any required value (maximum voltage = 5 V). Again two values were chosen:

$c = 45.8$ mV for just audible distortion

$c = 91.5$ mV for fully audible distortion .

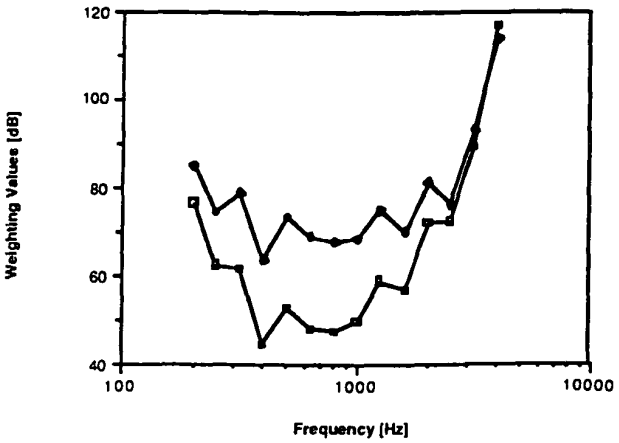


Fig. 13. P.79 monaural loudness rating weightings for telephony. Upper curve—receiving loudness weightings W_{Rn}; lower curve—sending loudness weightings W_{Sn}.

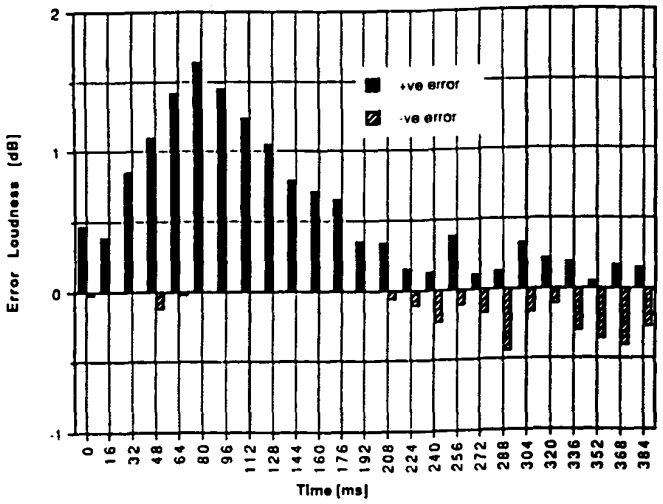


Fig. 14. Example error loudness versus time result.

4) *VAD—Gross Temporal Nonlinearity*. The VAD emulation used allowed the onset, or clipping, time to be controlled. A range of onset times was investigated, including 50 and 100 ms.

7 EXPERIMENTAL RESULTS

Sample error surface and error loudness versus time results are shown for the four example measurements.

7.1 Instantaneous Amplitude Nonlinear Distortion

The error surface and error loudness results for the barely audible and clearly audible cases of this distortion are shown in Figs. 15–18. The results accord well with intuitive prediction. The increased distortion level leads to increasing error loudness. Most of the error loudness is positive due to the additive distortion. The majority of the distortion loudness coincides with the voiced part of the speechlike test event since this contains low-frequency formant tones whose harmonics will be perpetually significant.

7.2 MNRU

Figs. 19 and 20 show the error surface and error loudness for the audible MNRU distortion type. The result shows a number of important features; distortion ridges appear in the error surface which correspond to the frequency and time of the strong voiced formants. This should be expected since the distortion increases

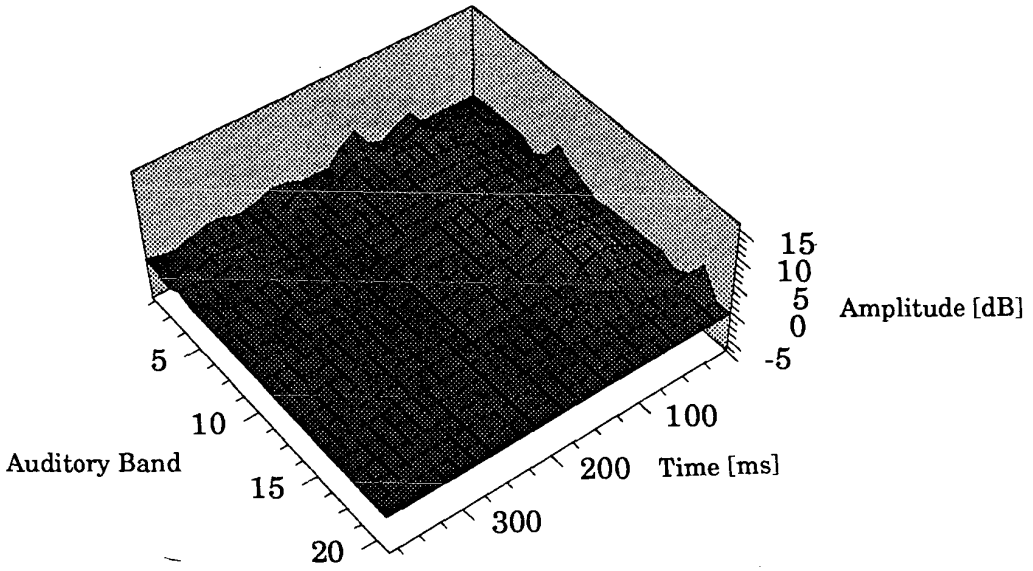


Fig. 15. Error surface for barely audible instantaneous amplitude nonlinear distortion.

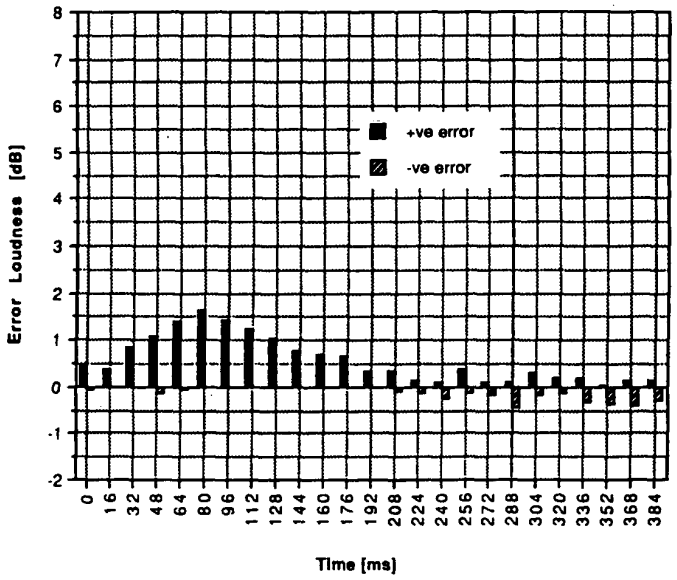


Fig. 16. Error loudness versus time for barely audible instantaneous amplitude nonlinear distortion.

with signal amplitude. Conversely, little distortion coincides with the lower energy unvoiced sound.

7.3 Crossover Distortion

The error surface and error loudness results for audible crossover distortion are shown in Figs. 21 and 22. Low-amplitude signals are not transmitted and so the lower energy unvoiced sound is drastically attenuated. A very significant subjective impact is appropriately predicted.

7.4 VAD

The error surface and error loudness for a 50-ms (audible) VAD onset time are shown in Figs. 23 and 24. The perpetual impact of the "clipped" signal is immediately apparent. Any transient overshoot or settling will also be shown.

8 DISCUSSION AND CONCLUSIONS

The principle of using a composite speechlike test stimulus and perception-based analysis to characterize nonlinear and temporally varying systems has been introduced. The approach reported provides a system characterization which reflects perceived speech performance better than other conventional measurement techniques.

Early results indicate that the initial implementation is sufficiently sensitive to predict the onset of audibility for the distortion types tested, and subjective perform-

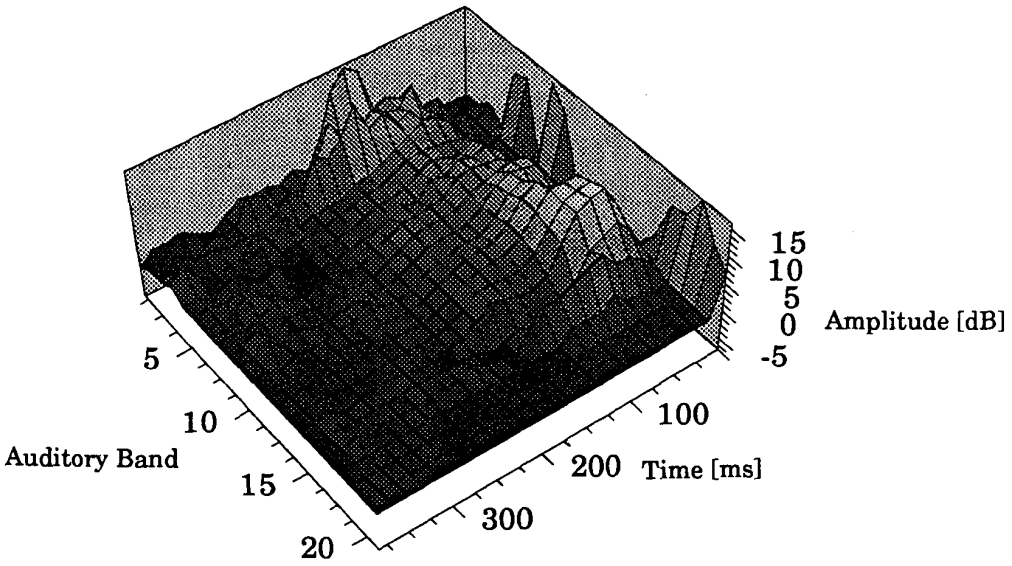


Fig. 17. Error surface for clearly audible instantaneous amplitude nonlinear distortion.

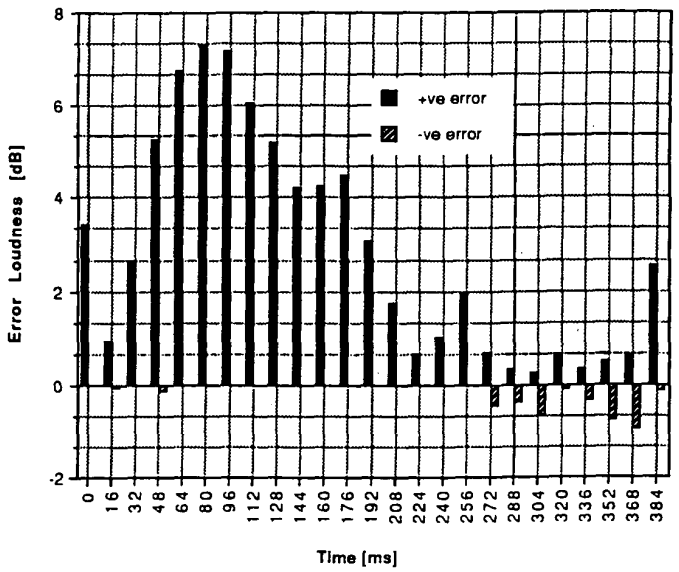


Fig. 18. Error loudness versus time for clearly audible instantaneous amplitude nonlinear distortion.

ance ranking can be predicted. Further work will refine the prediction capabilities of the method using rigorous subjective validation.

Further work is under way to employ an alternative approximation to the auditory filter shape, such as that described by Patterson and Moore [13]. Fine tuning of the temporal resolution and forward masking elements of the model will be addressed by examining accurate prediction of VAD onset time detection and the audibility of burst errors.

The diagnostic value of the technique is illustrated by results consistent with intuitive prediction, in particular, the MNRU distortion peaks relating to the voiced sound formant frequencies, the addition of harmonic distortion to the energetic voiced sound formants, and the loss of the unvoiced sound due to crossover distortion.

The evaluation of time-variant system elements such as VADs, burst errors, and AGCs is particularly well addressed by this technique. In particular the evaluation of practical AGCs, intended to improve performance in conditions of high ambient noise, must deal with temporally varying gain, different expansion and compression rates, and soft clipping. The need for event context and perceptual analysis approach is thus immediately apparent.

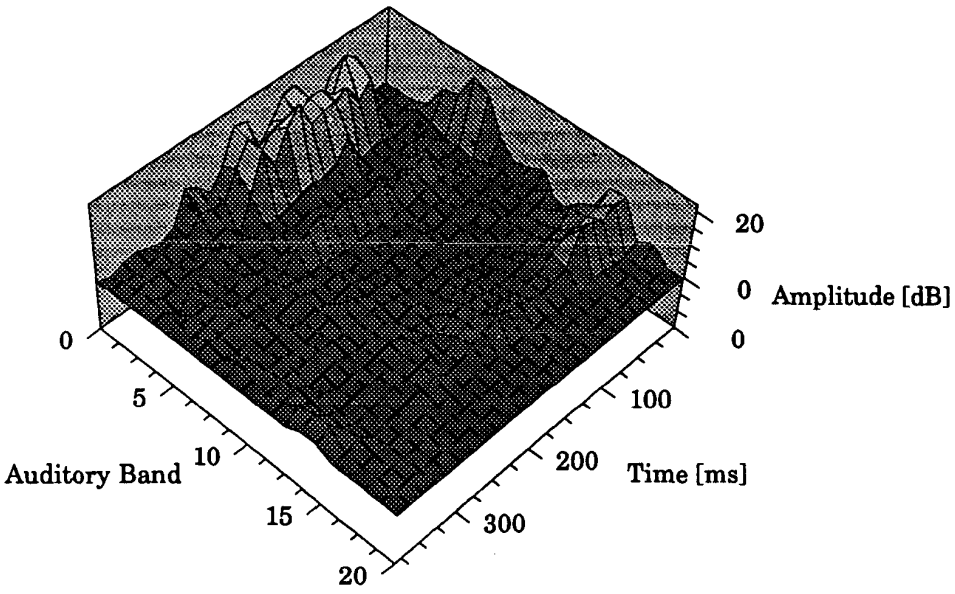


Fig. 19. Error surface for MNRU distortion.

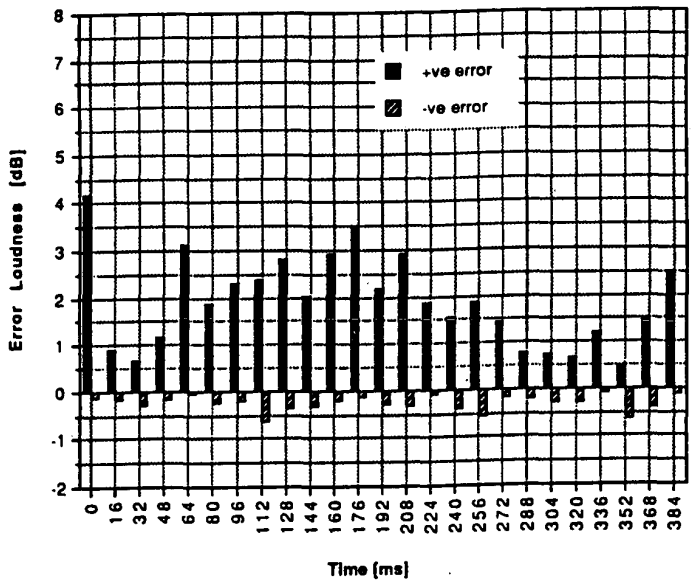


Fig. 20. Error loudness versus time for MNRU distortion.

9 ACKNOWLEDGMENT

The authors would like to thank Bob Stuart, of Meridian Audio Ltd., for supplying information on perceptual modeling, and their colleagues at BT Labs for their encouragement and support.

10 REFERENCES

[1] BS 6317, "British Standard for Simple Extension Telephones for Connection to the British Telecommunications Public Switched Telephone Network" (1982).

[2] P. M. Djuric and S. M. Kay, "Parameter Estimation of Chirp Signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38 (1990 Dec.).

[3] R. B. Randall, *Frequency Analysis*, 3rd ed. (Brüel and Kjær, Denmark, 1987 Sept.).

[4] D. D. Rife and J. Vanderkooy, "Transfer-Function Measurement with Maximum-Length Sequences," *J. Audio Eng. Soc.*, vol. 37, pp. 419–444 (1989 June).

[5] A. Duncan, "The Analytic Impulse," *J. Audio Eng. Soc.*, vol. 36, pp. 315–327 (1988 May).

[6] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Audio Eng. Soc.*, vol. 40, pp. 21–31 (1992 Jan./Feb.).

[7] C. Wheddon and R. Linggard, Eds., *Speech and Language Processing* (Chapman and Hall, London, 1990).

[8] CCITT P.50, "Recommendation on Artificial Voices," vol. V, Rec. P.50, Melbourne (1988).

[9] CCITT P.51, "Recommendation on Artificial Ear and Artificial Mouth," vol. V, Rec. P.51, Melbourne (1988).

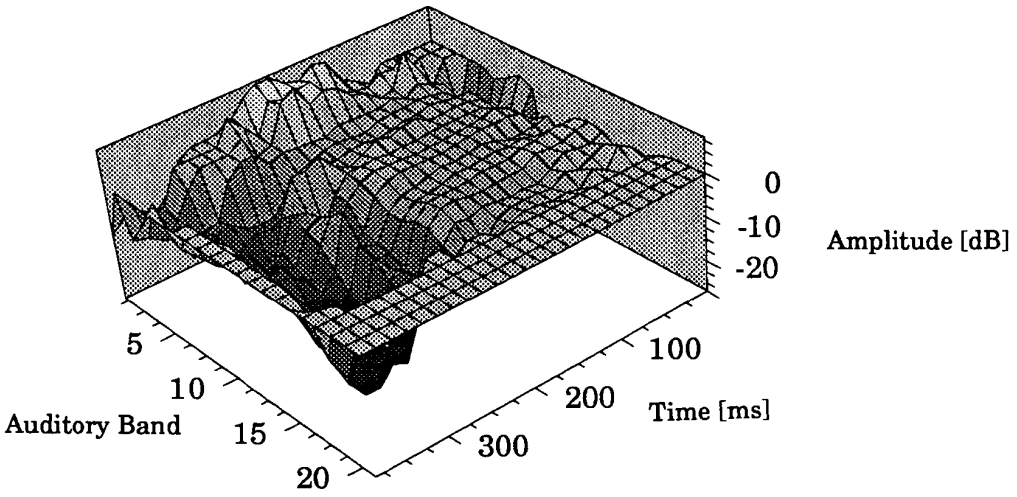


Fig. 21. Error surface for audible crossover distortion.

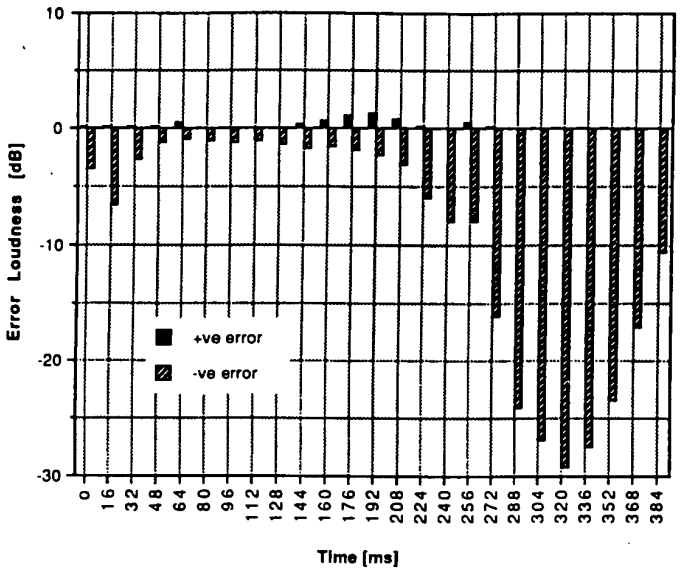


Fig. 22. Error loudness versus time for audible crossover distortion.

[10] CCIR Doc. 10/2-11-E, "Chairman Report of the First Meeting of the Task Group 10/2 Geneva 7–20 November 1991" (1992 Jan.).

[11] T. Scott and M. Bosi, "Use of Low Bit-Rate Coding for High Quality Audio over Telephone Lines," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1040 (1992 Dec.), preprint 3362.

[12] E. Zwicker and U. T. Zwicker, "Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory Systems," *J. Audio Eng. Soc.*, vol. 39, pp. 115–126 (1991 Mar.).

[13] R. D. Patterson and B. J. C. Moore, "Auditory Filter and Excitation Patterns as Representations of Frequency Resolution," in B. J. C. Moore, Ed., *Frequency Selectivity in Hearing* (Academic Press, New York, 1986).

[14] J. G. Beerends and J. A. Stemerdink, "Measuring the Quality of Audio Devices," presented at the 90th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 388 (1991 May), preprint 3070.

[15] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Selected Areas in Comm.*, vol. 10 (1992 June).

[16] E. Ambikairajah, N. D. Black, and R. Linggard, "Digital Filter Simulation of the Basilar Membrane," *Comput. Speech Lang.*, no. 3 (1989).

[17] CCITT Rec. P.79 (1988).

[18] S. S. Steven and H. Davis, *Hearing, Its Psychology and Physiology*, 4th Edition (Wiley and Sons, Inc., 1954).

[19] D. R. Guard and I. Goetz, "The DSP Network Emulator for Subjective Assessment," *BT Technol. J.*, vol. 10 (1992 Jan.).

[20] MNRU (Modulated Noise Reference Unit), in CCITT Rec. P.81, Annex A (1988).

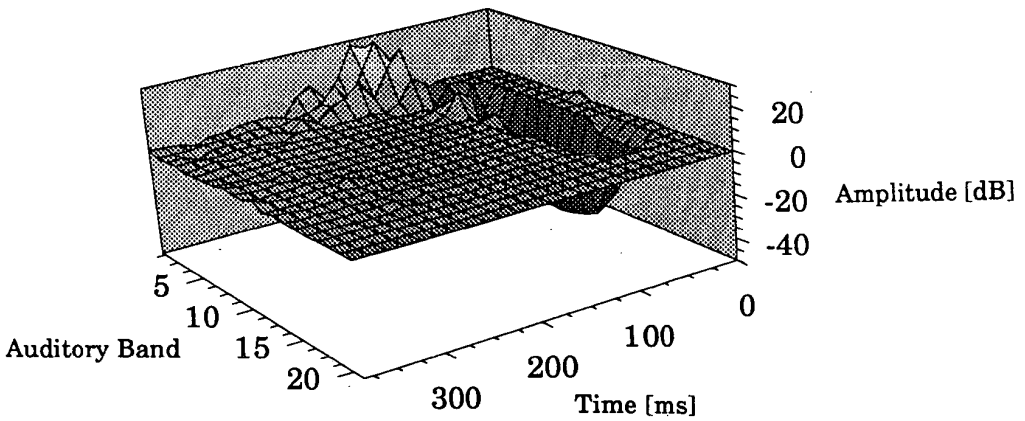


Fig. 23. Error surface for 50-ms VAD onset.

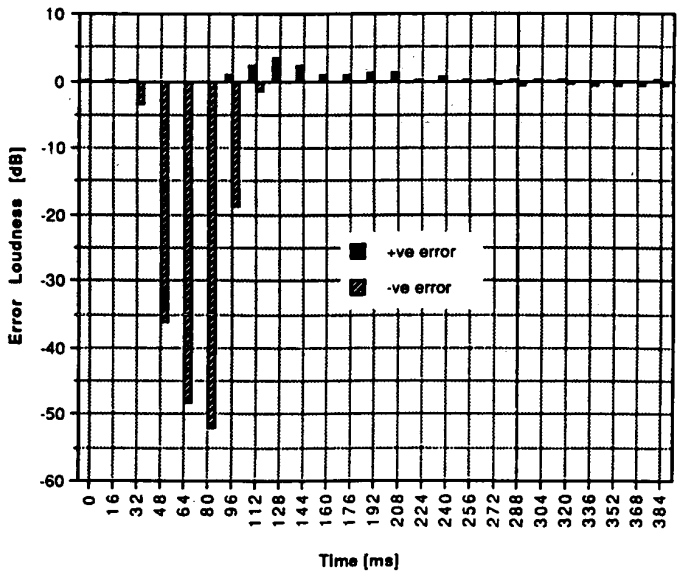
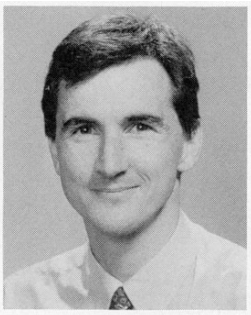


Fig. 24. Error loudness versus time for 50-ms VAD onset.

THE AUTHORS

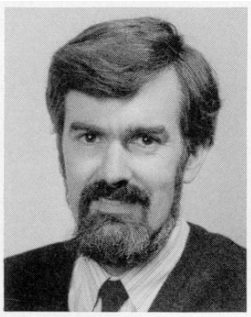


M. Hollier

Mike Hollier obtained a B.Eng.(Hons) degree in mechanical engineering from Plymouth Polytechnic in 1987, and joined BT Laboratories after graduation. He obtained an IOA diploma in acoustics in 1989, receiving an industry award for research into the vibrational behaviour of light structures. He is a member of the IOA.

With a background in audio engineering, he has worked on a number of telephony acoustics projects including the development of a novel noise canceling handset. His current work concerns the development of objective measurement methods capable of predicting the subjective performance of non-linear audio systems.

Mr. Hollier is currently employed at BT Laboratories, and is an external Ph.D. student to the University of Essex in Colchester, Essex, U.K.



D. Guard

David Guard graduated in 1971 with a B.Sc. in physics, followed by a Ph.D. in medical acoustics in 1975. He then joined BT Laboratories to investigate loudspeaking telephones, culminating in the development of the Orator executive audio conferencing system. He then moved into the marketing area to sell Orator nationwide.

Dr. Guard returned to BT Laboratories in 1983 to develop and support telephony terminals. From 1987 he worked on transmission performance of the complete network. Since 1992 he has been a BT Laboratories systems designer concerned with particular speech technology services.

The biography for Malcolm Omar Hawksford was published in the 1993 March issue of the *Journal*.

Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain

M.P. Hollier
M.O. Hawksford
D.R. Guard

Indexing terms: Audible error surfaces, Auditory model, Cell entropy, Speech codes

Abstract: Several models have been described in the literature which seek to represent audio stimuli in the perceptual domain to best predict the audibility of errors and distortions. By modelling the principal nonlinear processes of human hearing it is possible to calculate a perceptual-domain error surface that represents the audible difference between distorted and original audio signals. A further stage of analysis is required to maximise the usefulness of the auditory model output. The audible error surface must be interpreted to produce an estimate of the overall subjective judgement which would result from the particular distortion. Ideally, the interpretation of the error surface should be broadly analogous to human perceptual mechanisms, and equally, it would be desirable to avoid the complex and cumbersome statistical mapping and clustering techniques proposed by some authors. A technique employed in adaptive transform coding of images, namely cell entropy, offered several desired properties. The paper reports the extension and application of such a technique to the interpretation of perceptual-domain error surfaces produced by an auditory model. Speech data were subjected to an example, algorithmically generated, nonlinear distortion and then processed by the auditory model. The usefulness of the error-activity and error-entropy quantities are illustrated, without optimisation, by comparison of model predictions and experimentally determined opinion scores.

1 Introduction

It is increasingly difficult to predict the subjective performance of speech systems with conventional objective measurement techniques. This is due to the use of nonlinear and time-variant processes for echo cancellation, low-bit-rate coding, DCME (digital circuit multiplication equipment), etc. As a result, new measurement methods are under development that seek to predict the subjective performance of audio systems by analysing the audio signal in a way that is analogous to the human hearing process, i.e. auditory models.

Several recent publications have described models which seek to process audio signals to represent perceived error in the perceptual domain, [1–4]. Others have attempted to predict subjective opinion by relating a variety of objective measurements to known subjective test data using statistical mapping and clustering techniques [5–7].

In general, the auditory model seeks to process the audio signal to include the peripheral auditory mechanisms and certain central nervous system behaviour. For example, by modelling the frequency discrimination and auditory threshold, which would be described by a psychoacoustic experiment, the peripheral auditory system and certain cortex behaviour have been modelled. By representing 'original' and 'distorted' versions of an auditory stimulus in perceptual space it is possible to predict the 'audible error' resulting from the distortion in the form of a surface, i.e. an error surface.

A significant challenge, in this rapidly emerging field, is the development of relationships between the audible error surface and an overall subjective opinion. Several possible approaches are given in Section 2, and a new technique based on the distribution of the error is developed and tested in Sections 3 and 4, respectively.

2 Relating audible error to subjective scale

To readily compare the performance of two systems, or to iteratively develop a prototype system, a single figure of merit is required in place of the audible-error surface. Further, it is desirable that this single figure of merit be directly related to existing subjective opinion scales, such that measured performance can be directly related to exiting subjective test data (which is employed in transmission performance plans, etc.).

2.1 Weighted sum and curve fit

The simplest way to develop a relationship between the error surface and a target subjective opinion scale is to operate a weighted summation along the pitch and time axis. This type of weighted summation is proposed in Reference 8, where it is applied to the speech signals themselves rather than an audible error surface predicted by an auditory model. Applying the method to the audible error surface, the relative significance of pitch

Paper 1248K (E5), first received 25th November 1993 and in revised form 29th March 1994

M.P. Hollier and D.R. Guard are at BT Laboratories, Martlesham Heath, Ipswich IP5 7RE, United Kingdom

M.O. Hawksford is with the Department of Electronic Sytems Engineering, University of Essex, Colchester CO4 5SQ, United Kingdom

and time features can be adjusted to fit predictions to known average subjective judgements.

Imagine an error-surface fragment relating to a time segment of the required duration, Fig. 1. Perform a

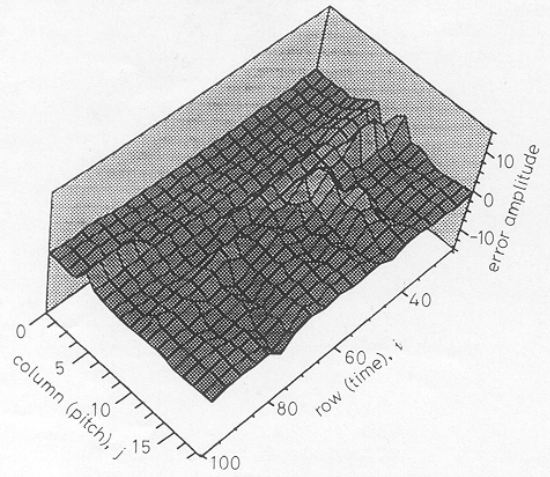


Fig. 1 *Error-surface fragment*

weighted sum of the error-surface values to yield a single figure metric, SF_{ws}

$$SF_{ws} = \sum_{i=1}^{n} Wt_i \sum_{j=1}^{m} e_{ij} Wf_j \quad (1)$$

where Wt and Wf are weightings for time and pitch dimensions, respectively, n and m are the discrete dimensions of the error-surface fragment, and e_{ij} is the value in the ith row and jth column of the error-surface fragment. The Wf weightings emphasise those regions of the pitch scale most significant in determining the target subjective opinion. The Wt weightings emphasise those parts of the time history most critical to the subjective judgement, for example emphasising the significance of the beginning of a certain sounds. (The Wt weighting could be set, for a particular test stimulus, to correspond with the timing of critical events such as second formant transitions.)

Once the weighted sum has been calculated for a range of distortion values it is possible to fit the calculated distortion to the target opinion scale Y. Typically, this is achieved using a sigmoid function such as the logistic [9], defined as follows

$$\text{logit}\ (y) = \ln\ (P/(1 - P)) \quad (2)$$

where P is the normalised variate $(Y - Y_{min})/(Y_{max} - Y_{min})$, Y is the experimentally determined opinion, and y the distortion variable. The logit is the deviate corresponding to a given (cumulative) probability for the logistic distribution in the same way as the normit is the deviate corresponding to a given (cumulative) probability for the normal distribution. The cumulative probability can therefore be obtained by inverting eqn. 2

$$P = 1/(1 + \exp\ (-y)) \quad (3)$$

whence by differentiation the probability density is

$$z = \exp\ (-y)/(1 + \exp\ (-y))^2 \quad (4)$$

Other forms of eqns. 2 and 3 are possible in terms of hyperbolic tangents and their inverses. The logistic distribution has the important practical advantage, compared with the normal distribution, that the logistic and its inverse can both be represented by a simple explicit formula rather than tedious numerical integration.

The advantage of the weighted sum and logistic mapping method is its simplicity. The principal disadvantage is that it is only possible to optimise the mapping for one distortion type at a time, i.e. predictions of subjective performance could only be made for the distortion type for which the mapping was developed. A practical assessment method would thus have to determine the class of the distortion and alert the analysis algorithm which mapping to employ for the subjective performance prediction. This renders the apparently simple approach rather complicated and would lead to serious error if the wrong class of distortion was predicted, perhaps owing to an unknown distortion type or a combination of distortions.

2.2 Empirical methods

Empirical, or statistical, mapping methods employ advanced statistical techniques such as clustering theory to develop a relationship between the audible-error surface for particular distortions, and a corpus of known subjective responses for the same distortions. By training the clustering relationship for the corpus of known subjective test data it is possible to predict the subjective opinion which will result for a distortion value within the trained range.

This form of statistical process can also be used to predict subjective performance using a series of more conventional objective measurement techniques, such as cepstral distance, segmental signal-to-noise ratio (SNR), coherence, and correlation. In this case, the results of several objective measurements are used to develop the mapping between distortion and known subjective opinion. This is the basis of the approach proposed by several authors [5–7].

There is a hybrid solution between using objective measures (SNR, etc.) plus statistical mapping and using the audible error (predicted by the auditory model) plus statistical mapping. The hybrid solution is to employ psychoacoustically sensitised objective measurements such as perceptually weighted segmental SNR.

Methods based on clustering are very powerful and can be trained to provide accurate predictions of subjective opinion within the range of the training data. The disadvantage of such methods is that opinion predictions for distortions outside the range of the training data are less good and rapidly become unreliable if the distortion type is significantly different from the training-data distortion. The technique is more robust in assessing different distortion types than the simple weighted sum approach, but still cannot provide reliable predictions of opinion on an arbitrary distortion type.

2.3 Error significance

Assuming a reliable prediction of the audible error can be produced, say by an auditory model, then it would be highly desirable to interpret the subjective significance of this error in terms which are broadly analogous to the average perceptual processes which actually occur. In this way it would be possible to produce reliable subjective opinion estimates for all types of distortion including nonlinear and time-variant behaviours.

To interpret the significance of the audible error, the features of the error effecting its subjective impact must be considered. In particular, it is not simply the total error activity that will determine its subjective significance but also the distribution of this error, i.e. whether it

permeates the entire error surface or is concentrated into a particular peak or ridge. For example, a time-varying gain or frequency response change may have little subjective impact and yet create as much audible error as a highly significant nonlinear distortion.

Adaptive transform coding of video images [10] dynamically allocates the number of bits used to code each frame, dependent on the information contained in the frame. This determination of frame information requires that the amount and distribution of activity within the frame is assessed. A set of equations to produce numerical indicators for 'cell activity' and 'cell entropy' are given in Reference 10. Such numerical descriptions would appear to exhibit useful properties for describing the subjective significance of audible error.

The remaining feature of the error which determines its subjective impact is the degree to which the error is correlated to the desired (original) signal. This of course may be calculated directly since the error surface and original signal are available. The cross correlation R_{xy} is calculated between each original and degraded auditory band with a discrete algorithm implementation, for which the elements of h are obtained using

$$h_j = \sum_{k=0}^{q-1} x_k y_{j+k} \tag{5}$$

for $j = -(q-1), -(q-2), \ldots, -2, -1, 0, 1, 2, \ldots, (r-2), (r-1)$, and where q and r are the time dimensions of the error surface and original signal, respectively, i.e. they will typically be equal to n in eqn. 1. The elements of the output sequence Rxy are related to the elements in the sequence h by

$$Rxy_i = h_{i-(n-1)} \quad \text{for } i = 0, 1, 2, \ldots, \text{size} - 1$$

size $= n + m - 1$, where n and m are as eqn. 1. Since the error correlation is calculated for each auditory band the result is in the form of a surface. The error-correlation quantity is not explicitly developed in the following analysis since all the distortions discussed have a similar high degree of correlation. The error correlation requires further development and must be included to account for the subjective effect of delay and echo. This will be the subject of separate study.

3 Error activity and error entropy

The total audible error may be referred to as the error activity. Based on the expression in Reference 10 for block activity, we define error activity as

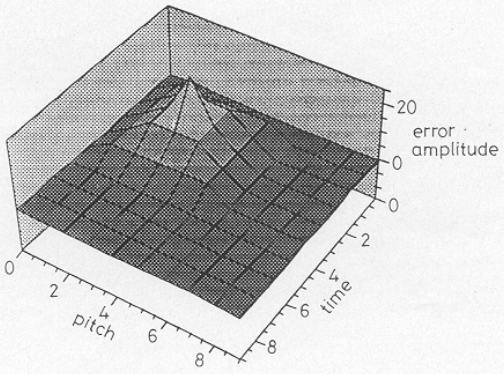
$$E_a = \sum_{i=1}^{n} \sum_{j=1}^{m} |e(i, j)| \tag{6}$$

where n and m are the dimensions of the error surface or error-surface fragment. The distribution of the audible error may be referred to as the error entropy. Again, we base an expression on that given in Reference 10 to define a suitable quantity

$$E_e = -\sum_{i=1}^{n} \sum_{j=1}^{m} a(i, j) * \ln (a(i, j)) \tag{7}$$

where $a(i, j) = |e(i, j)|/E_a$. The behaviour of the error activity and error-entropy quantities are illustrated in Figs. 2 and 3 using hypothetical error-surface fragments which show the same error activity distributed in different ways. A flat distribution of error activity produces an error entropy of 4.605. Error entropy is independent of scaling.

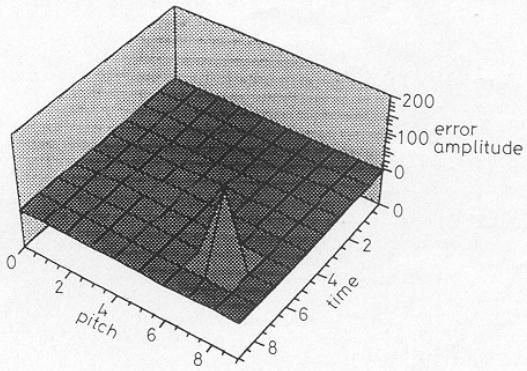
The variation of error entropy with different distributions of the error activity is apparent. The subjective significance of the error will thus be related to the inverse error entropy, since a highly concentrated error, which will yield a low error entropy is found to have a greater subjective impact than a distributed error. The concentration of error is described in the perceptual domain and so the temporal discrimination of the ear has already been included. This means that the concentration of error can be related directly to the subjective impact, since the rates of change of error reflect the psychoacoustic responses of hearing.



0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	5	5	5	5	5	0.1	0.1	0.1	0.1
0.1	5	11	11	11	5	0.1	0.1	0.1	0.1
0.1	5	11	24.5	11	5	0.1	0.1	0.1	0.1
0.1	5	11	11	11	5	0.1	0.1	0.1	0.1
0.1	5	5	5	5	5	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

error activity = 200 error entropy = 3.294

Fig. 2 *Error present as single feature*



0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	190	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

error activity = 200 error entropy = 0.425

Fig. 3 *Error concentrated into single peak*

4 Example analysis and error calculation

Example speech files were degraded with a nonlinear distortion MNRU (modulated noise reference unit) [11] and processed by an auditory model [4] to produce an audible error surface. The use of MNRU distortion has the advantage that, for a given listening level, the average subjective opinion for different degrees of distortion is known. By calculating the error activity, error entropy, and correlation for a corpus of speech it is possible to investigate the relationship between the error descriptors and the average subjective opinion that would have resulted.

The error descriptor quantities are calculated and averaged and compared with opinion scores for speech degraded by the same distortion. The performance of the error descriptors can be assessed by examining the relationship between them and the corresponding subjective opinion.

The error descriptors are computed for commercially available low-bit-rate coders; and the resulting subjective opinion score predictions compared with the experimentally determined MOS.

4.1 Choice of nonlinear distortion

The MNRU is representative of an important general class of distortion which occurs in digital systems. It can be conveniently generated algorithmically, and has been included in a large number of international subjective experiments. The subjective opinion scale referred to here is listening effort Y_{LE}, see Reference 12 for definition. The relationship between Y_{LE} and MNRU distortion, determined experimentally [13], is shown in Fig. 4. The data can be approximated by

$$(Y_q - 1)/(Y_{q\,max} - 1) = 1/(1 + e^{4.5(M - Q)}) \tag{8}$$

where Y_q is the opinion score, $S = 0.0551$, $M = 11.449$, $Y_{max} = 4.31$, and Q is the equivalent quantisation distortion in dB.

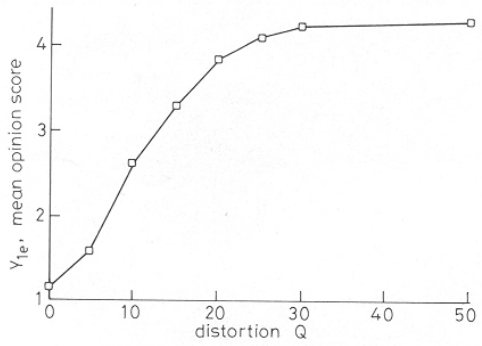


Fig. 4 *Experimentally determined relationship between MNRU and Y_{LE}*

4.2 Speech material

The corpus of speech data consisted of four sentences each from two male and two female talkers. A small speech corpus is deemed to be acceptable to illustrate the proposed error interpretation technique. Each sentence was about 2 s long and contained normal conversational English.

4.3 Generation of error surfaces and calculation of error descriptors

The error surface for the 16 speech files were generated by an auditory model and then postprocessed to produce the activity and entropy error descriptors. The stages of speech processing were

(i) Speech files preprocessed with different amounts of MNRU distortion, for which average subjective opinion is known.

(ii) Original and distorted speech files processed by the auditory model [4] to predict the audible error due to the different distortion settings.

(iii) Postprocess audible error surfaces to produce error descriptor quantities.

Fig. 5 shows example error-surface fragments for the first talker at the highest distortion value. It is necessary to

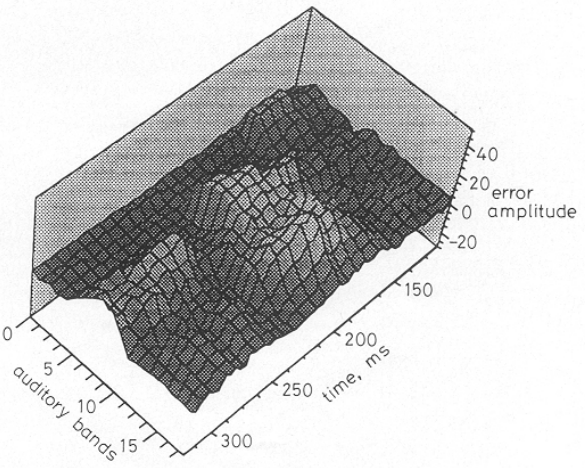


Fig. 5 *Part of audible error surface for first talker $Q = 0$*

rework the E_a and E_e terms to account for the logarithmic units of the audible error amplitude scale

$$E'_a = \sum_{i=1}^{n} \sum_{j=1}^{m} 10^{|e(i,\,j)|} \tag{9}$$

and

$$E'_e = -\sum_{i=1}^{n} \sum_{j=1}^{m} 10^{e(i,\,j)}/E'_a * \ln\{10^{(e(i,\,j))}\}/E'_a \tag{10}$$

Figs. 6 and 7 show the relationship between the averaged $\log_{10} E'_a$ and E'_e values and the distortion Q. A combination of the error-activity and error-entropy quantities was developed by fitting $\log_{10} E'_a$ and E'_e values to

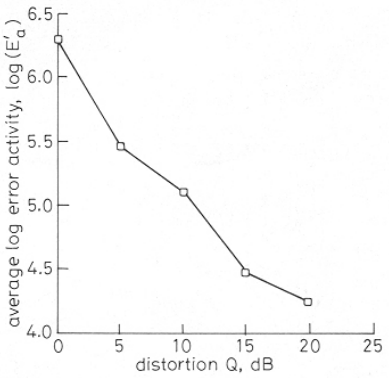


Fig. 6 *Average $\log_{10} E'_a$ against distortion Q*

the distortion Q by linear regression. The resulting fit is given in eqn. 11

$$\text{distortion } Q = -55.09 - 0.5556 \log_{10} E'_a + 2.624E'_e \quad (11)$$

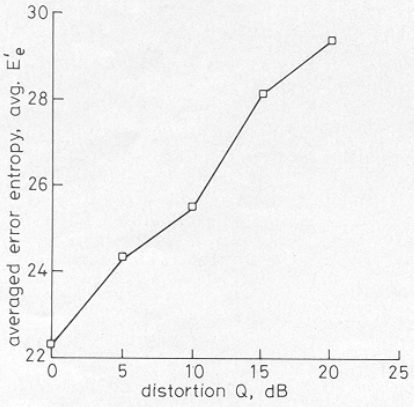


Fig. 7 *Average E'_e against distortion Q*

4.4 Relationship between error descriptors and Y_{LE}

The relationship between Y_{LE} and MNRU shown in Fig. , and eqn. 8 can be used to relate the error descriptors to MOS. The resulting relationship is given in eqn. 12

$$Y_{LE} = -8.373 + 0.05388 \log_{10} E'_a + 0.4090E'_e \quad (12)$$

Fig. 8 shows a graph of the calculated opinion score against experimental Y_{LE} approximated by eqn. 8.

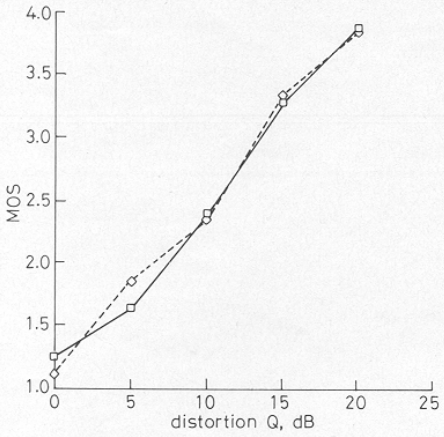


Fig. 8 *Calculated and experimental MOS against distortion Q*
—□— MOS fit to experimental data
···◇··· MOS calculated

Alternatively, the error descriptors can be fitted to the known opinion score values using a logistic and double quadratic expansion

$$Y_{LE} = \text{fn}\ \{E_a, E_e\} \quad (13)$$

One may choose to use some further function of the error descriptors in the following expansion, for example, the reciprocal of E_e and the log of E_a. The modified forms of E_a and E_e are shown as E''_a and E''_e

$$\text{logit}\ (Y_{LE}) = b_0 + b_1 E''_a + b_2 E''^2_a + b_3 E''_a E''_e + b_4 E''_e + b_5 E''^2_e \quad (14)$$

$$= \ln\ (Y_{LE}/(4 - Y_{LE})) \quad (15)$$

$$= w$$

Then the estimated value of opinion score Y' is

$$Y' = 4/(1 + e^{-w}) \quad (16)$$

Fitting is achieved via iterative weighted least squares, which is accomplished using proprietary software, such as GLIM.

4.5 Application to practical nonlinear processes

The sample speech corpus was processed by three commercial coding algorithms, for which the MOS has been experimentally determined. The resulting degraded speech files were processed as described in Section 4.3 and the error descriptors used to predict the subjective opinion of an average user. The predicted and experimentally determined MOSs are shown in Table 1.

Table 1: Experimental and predicted MOS for three codecs

Coding algorithm	MOS (experimental)	MOS (prediction)
Commercial low-rate codec A	3.39	2.90
Commercial low-rate codec B	3.16	2.67
Commercial low-rate codec C	2.65	2.94

The experimental MOS is predicted to within 0.5 of an opinion score for all three examples. The combination of error descriptors has not been optimised across a range of distortions. The results are not therefore intended to demonstrated the superior performance of the technique but rather the potential for a significant saving in complexity compared with the empirical methods introduced in Section 2.2.

To further illustrate the behaviour of the proposed technique, the error-surface descriptors were calculated for the same speech corpus processed by four different types of degradation. The degradations were algorithmically generated and deliberately chosen to span a range of subjective opinions. The distortion types ranked by a panel of six subjects in rank order were

(i) DCME (including VAD and data reduction)
(ii) Speech codec A (with error pattern EPX)
(iii) Nonlinear amplifier distortion
(iv) Speech codec C (with error pattern EPY)

The error patterns EPX and EPY were nonstandard and chosen to give the desired subjective ranking. The types of distortion include different codec types, a nonlinear amplifier characteristic, and a temporal distortion owing to the VAD (voice activity detector) in the DCME. As such the distribution of error is very different for the four distortion types and a simple average of the error surface fails to provide the correct subjective ranking. Table 2 shows the average error and MOS prediction for the four degradations.

Table 2: Average error and predicted MOS for different degradations

Degradation	Subjective ranking		Average $\log_{10}$(error)	MOS prediction
DCME	best	1	7.288	4.04
Speech codec A (EPX)		2	7.232	3.12
Nonlinear amplifier		3	7.810	2.97
Speech codec C (EPY)	worst	4	7.555	2.36

It is apparent that the average error does not indicate the correct subjective ranking while the MOS prediction, which takes account of the distribution of the error, does predict the correct subjective ranking.

5 Discussion and conclusions

The development of a novel objective measurement technique capable of predicting the subjective performance of real systems has been introduced. The requirement to relate audible error surfaces to overall subjective opinion has been described, and the advantage of a process that is broadly analogous to psychoacoustic mechanisms as compared with empirical methods highlighted. In particular, the use of psychoacoustic rules represents a substantial reduction in complexity compared with empirical mapping methods.

The useful properties of quantities used in adaptive transform coding have been noted and applied in modified form to the interpretation of audible error surfaces. A parallel is drawn between the factors effecting the overall psychoacoustic consequences of an audible error, and calculated quantities describing the total, distribution, and correlation of the error.

The behaviour of the error-activity and error-entropy quantities were illustrated with hypothetical error-surface fragments. To further illustrate the technique, a relationship between the error descriptors and MNRU distortion was developed by linear regression and used to predict the MOS for three commercial speech codecs. A more sophisticated optimisation of how the error descriptors are combined is clearly possible. A further set of four different degradations are used to illustrate that average error is insufficient to predict subjective ranking, while the overall measure, which includes the distribution of error, can predict the subjective ranking.

The error descriptors developed are much simpler than the alternative empirical mapping techniques, and are not related to any particular distortion type. As such the error-surface analysis is potentially independent of the distortion type and capable, with the inclusion of error correlation, of assessing the full range of linear, nonlinear and time-variant distortions.

6 References

1 WANG, S., SEKEY, A., and GERSHO, A.: 'An objective measure for predicting subjective quality of speech coders', *IEEE J. Sel. Areas Commun.*, June 1992, **10**, (5), pp. 819–829

2 BEERENDS, J.G., and STEMERDINK, J.A.: 'A perceptual audio quality measure based on a psychoacoustic sound representation', *J. Audio Eng. Soc.*, 1992, **40**, (12), pp. 963–978

3 STUART, J.R.: 'Psychoacoustic models for evaluating errors in audio systems'. Proceedings of the Institute of Acoustics conference, vol. 13, part 7, November 1991

4 HOLLIER, M.P., HAWKSFORD, M.O., and GUARD, D.R.: 'Characterisation of communications systems using a speech-like test stimulus', *J. Audio Eng. Soc.*, 1993, **41**, (12), pp. 1008–1021

5 HALKA, U., and HEUTER, U.: 'A new Approach to objective quality-measures based on attribute matching', *Speech Commun.*, 1992

6 'Effects of speech amplitude normalization on NTIA objective voice quality assessment method'. CCITT SG XII NTIA, contribution Doc. SQ-74.91, December 1991

7 IRII, H., KOZONO, J., and KURASHIMA, K.: 'PROMOTE — a system for estimating speech transmission quality in telephone networks', *NTT Review*, September 1991, **3**

8 QUAKENBUSH, S.R., BARNWELL, T.P., and CLEMENTS, M.A.: 'Objective measures of speech quality' (Prentice Hall, New Jersey, 1988)

9 FINNEY, D.J.: 'Probit analysis' (Cambridge University Press, 1971, 3rd edn.)

10 MESTER, R., and FRANKE, U.: 'Spectral entropy-activity classification in adaptive transform coding', *IEEE J. Sel. Areas Commun.*, June 1992, **10**, (5) pp. 913–917

11 MNRU (Modulated Noise Reference Unit). Annex A of CCITT Recommendation P. 81

12 CCITT P Series Recommendations, Volume V, supplement 3, sections 2.3 to 2.6

13 'Selection tests — basic data', TD92/39, ETSI/TM/TM5/TCH-HS expert group traffic channel half-rate speech, December 1992

ENGINEERING REPORTS

Algorithms for Assessing the Subjectivity of Perceptually Weighted Audible Errors*

M. P. HOLLIER, M. O. HAWKSFORD***,** *AES Fellow,* **AND D. R. GUARD****

***BT Labs, Martlesham Heath, Ipswich IP5 7RE, UK*
****Department of Electronic Systems Engineering, University of Essex, Colchester, Essex CO4 3SQ, UK*

Auditory modeling is used increasingly to provide an objective prediction of the subjective performance of audio systems. Such techniques typically compare the predicted auditory stimulation of original and processed audio signals to produce an estimate of audible error that can be presented as an error surface. The use of audible error surfaces as a diagnostic tool is investigated and an algorithmic interpretation provided to predict subjective opinion. The applicability of the technique across a range of industries is discussed.

0 INTRODUCTION

There is an increasing use of auditory models to evaluate audio systems. This approach is advantageous since it gives a prediction of the errors that will be perceptible to a listener, and necessary since many complex coding and reproduction processes cannot be adequately characterized with conventional engineering performance metrics. The combination of a number of nonlinear processes occurs routinely in telecommunications, and increasingly in other industries. The concatenation of nonlinear processes is a key issue for performance, since an inaudible distortion produced by one nonlinear process becomes part of the input to subsequent processes where it may yield an audible error. Auditory models are introduced in Section 2.

The use of audible error surfaces as a diagnostic tool is considered in Section 3 by examining the relationship between particular signal artifacts and the resulting audible errors. Signal artifacts and the corresponding distortions are presented for a low-bit-rate speech codec, illustrating the benefit of a perceptually weighted measurement. The potential benefits of such a technique both for performance evaluation and as a design tool are highlighted.

Many commercial applications for perceptual analysis do not require a diagnostic representation of an audible error surface, but require a specific single-figure performance metric that indicates the subjective audio quality experienced by the user. Such a metric is relevant for performance assessment across a number of industries such as the following.

1) Communications: Performance assessment during design and commissioning, for use in nonlinear network planning tools, and for the assessment of new products and services.

2) Professional and domestic audio equipment: Performance evaluation of data-reduction schemes.

3) Broadcasting: Codecs and perceptual coding schemes.

In order to predict a listener's opinion of audio quality it is necessary to interpret the audible error surface in a way that is analogous to human psychoacoustic perception of audible errors. The algorithmic interpretation of the audible error surface, in order to predict the subjectivity of the errors present, is discussed in Section 4. The opinion prediction required is often against a particular established response scale such as the listening effort scale (see [1] for definition).

1 AUDITORY MODELS

The authors, together with several others [2]–[5], have previously presented the use of auditory models to objectively assess nonlinear audio processes. Such models are intended to be analogous to the main functions of human hearing and are thus used to predict the auditory sensation that will result due to any given audio stimulus.

The limitations of the alternative empirical methods for predicting subjective audio quality are discussed briefly in [5]. This engineering report is exclusively concerned with the diagnostic and algorithmic interpretation

* Presented at the 97th Convention of the Audio Engineering Society, San Francisco, CA, 1994 November 10–13; revised 1995 October 26.

of audible errors predicted by an auditory modeling approach which has the advantage of being potentially independent of the distortion type. The auditory model transforms an audio stimulus into perceptual space, mapping frequency to pitch, level to sensation, and representing magnitudes over perceptually relevant intervals of time. In this way the auditory stimulation resulting from an audio event can be predicted together with simultaneous and temporal masking [5], [6].

It is important that the temporal resolution of hearing be modeled since the performance of real systems at encoding and reproducing transient information is critical to the perceived quality of speech and music. Indeed it has been previously illustrated [5] that it is during rapidly changing portions of the signal that certain data-reduction schemes are most prone to errors. A perceptual analysis measurement system must therefore employ a test stimulus which has the time-varying properties of the in-service signal, such as speech, and an analysis which includes an estimate of the temporal behavior of hearing.

A comparison of the auditory stimulation which occurs due to an original and degraded signal provides a prediction of the audible error—which can be presented in the form of an error surface. The error surface is a valuable diagnostic tool and can also be interpreted algorithmically to predict a listener's opinion of subjective performance.

2 DIAGNOSTIC INTERPRETATION

The audible error surface produced by comparing the auditory stimulation of an original and the degraded version of an audio stimulus can provide a valuable insight into subjective performance. In particular, it is possible to visualize the audible error which results from a particular signal artifact. The audible error surface has perceptual dimensions, and therefore features which appear on it have exceeded the masked threshold of the surrounding signal, and existed for a perceptually relevant time. It follows that a visible feature on the error surface is predicted to be audible. Indeed, it is possible to show that just audible errors do appear on the error surface. With experience it is sometimes possible to interpret both possible causes of a visible error feature and its subjective consequences.

Fig. 1 shows the auditory sensation predicted for the 0.10–0.85-s portion of the sentence "He retired quickly to his seat," and Fig. 2 the 1.15–1.85-s portion. Figs. 3 and 4 shows the error-surface fragments corresponding to the intervals in Figs. 1 and 2 when the speech has been processed by a low-bit-rate speech codec. When the processed sentence is replayed in a listening session, two dominant errors are perceived. These two dominant errors are clearly visible on the predicted error surface. First there is a click near the start of the file. This can be seen in Fig. 3 (≈300 ms), and second after about 1.5 s there is a "whoop" noise, shown in Fig. 4 (≈1350 ms).

The first error, Fig. 3, is a burst error and not attributable to a particular signal artifact. The second error, Fig. 4, is caused by an inappropriate selection of voiced/unvoiced parameters for the vocal tract model by the codec. This is a familiar error from a prototype linear predictive codec and occurs in this example with a typical signal artifact, that is, a consonant onset. A speech codec designer, who would be familiar with the detailed operation of the codec algorithm, would be able both to observe the audibility of the resulting error and to hypothesize which parameters within the codec should be modified to improve performance. The perceptual analysis, used to give a diagnostic visualization of the error, would provide a useful tool for iterative development.

3 ALGORITHMIC INTERPRETATION OF AUDIBLE ERRORS

In many instances it is not sufficient to present the error surface, but necessary to interpret the audible error automatically in order to produce an overall estimate of a listener's opinion. It is shown, with reference to specific examples, that the average error is not sufficient to predict subjective opinion. However, if the distribution of the error is also considered, a general prediction of opinion can then be made.

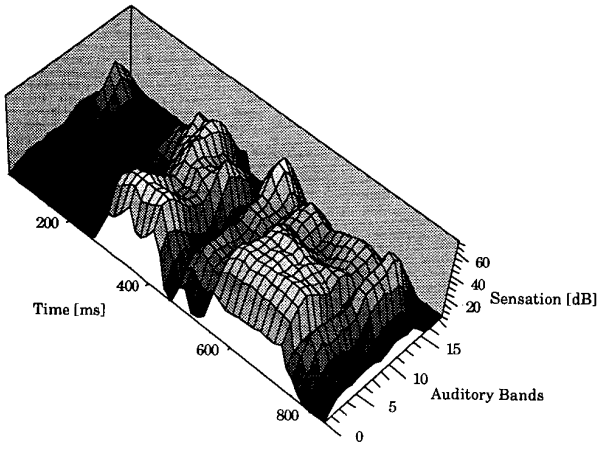


Fig. 1. Sensation surface for speech fragment.

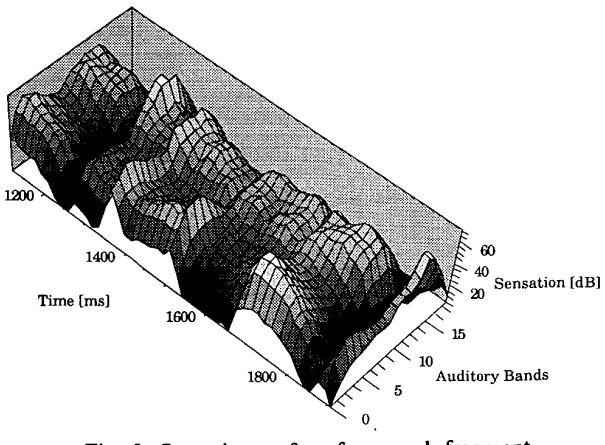


Fig. 2. Sensation surface for speech fragment.

3.1 Error Subjectivity

In order to describe the subjectivity of the features on an audible error surface algorithmically it is necessary to devise "rules" which are broadly analogous to the psychoacoustics that determine subjective significance. It is possible to determine such rules by starting with functions which accord with experimentally based knowledge on what determines the subjectivity of error features. Such an approach was followed by the authors in [7], and three error-descriptor parameters were proposed, which are found to provide an analogy to psychoacoustic significance. These parameters describe quantities related to the total (sum), distribution, and correlation of the audible error.

It is also possible to generalize about certain parts of cognitive signals, such as speech, being more important than others, such as transient onset sounds (consonant plus formant transition, for example), being critical for intelligibility. Because the test signal (original signal) is known in advance it is possible to label parts of the signal known to be cognitively important and increase their contribution to the subjective score. Such an approach is limited to cases where the cognitive significance of different parts of the signal can be reasonably predicted, that is, it is feasible to select and weight certain parts of speech signals but difficult to allocate appropriate weightings for music signals.

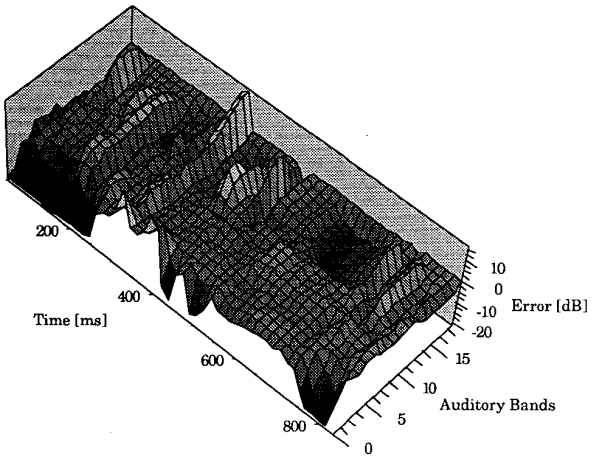


Fig. 3. Error surface for speech fragment in Fig. 1.

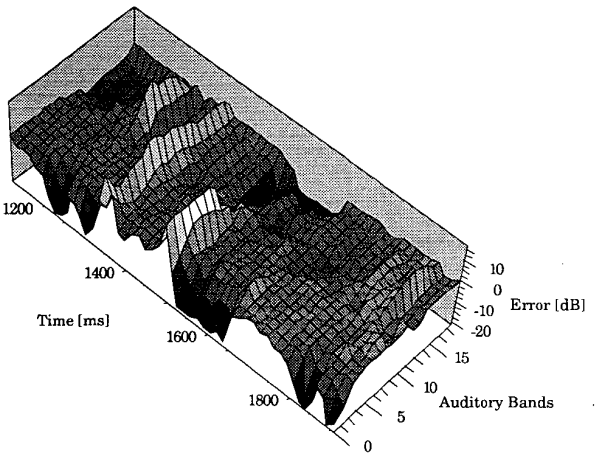


Fig. 4. Error surface for speech fragment in Fig. 2.

3.2 Error Descriptors

The error descriptors proposed by the authors in [7] reflect the total, distribution, and correlation of the audible error. It was observed that a useful measure of distribution can be taken from adaptive transform coding of images [8], where the number of bits used to encode a frame is dynamically allocated depending on the information in the frame.

The total audible error may be referred to as the error activity. Based on the expression in [8], for block activity we can define error activity as

$$E_a = \sum_{i=1}^{n} \sum_{j=1}^{m} |e(i,j)| \tag{1}$$

where n and m are the dimensions of the error surface or error-surface fragment, and $e(i,j)$ is the error surface value in the ith time segment of the jth auditory band.

The distribution of the audible error may be referred to as the error entropy. Again we may base an expression on that given in [8] to define a suitable quantity:

$$E_e = -\sum_{i=1}^{n} \sum_{j=1}^{m} a(i,j) * \ln[a(i,j)] \tag{2}$$

where

$$a(i,j) = \frac{|e(i,j)|}{E_a}.$$

To illustrate the concept, consider the hypothetical error fragments in Figs. 5, 6, and 7, as presented in Table 1.

The overall error activity in each fragment is equal, 250, whereas the error entropy varies between 4.63 and 0.87, depending on the distribution of the error. As the error varies from a uniform response shaping (Fig. 5) to discrete error features (Fig. 6) and finally a single peak (Fig. 7), the degree of disorder, and hence the entropy, decreases.

The error descriptor expressions must be modified to suit the dimensions of the perceptual domain error surface and combined to yield an algorithm which will automatically interpret the subjectivity of audible error. The combination of the error descriptors is achieved using a sigmoid function, such as a logistic [9], to ensure that even at the extremes, error descriptor values are asymptotic with the boundary values of the chosen opinion scale [7].

The error-correlation quantity is required to account for distortion types where the error is significantly delayed, such as listener echo. It is therefore omitted from the example results since the distortion processes chosen do not contain perceptually significant delays. The interval over which the error parameters are calculated is important and 0.5s was chosen for the example results shown.

Further consideration must also be given to higher level psychological behavior if the subjective opinion for an extended audio stimulus is to be predicted. This is because periods of distorted signal are not compensated for by an equal period without distortion, that is, a nonlinear averaging function over time is required.

3.3 Prediction of Subjective Audio Quality

A corpus of known sujective test results, available at BT Labs, has been used to calibrate the combination of the error descriptors so that performance predictions can be made for a range of nonlinear processes. A key issue for both calibrating and testing the error parameters is the inclusion of a wide range of nonlinear process types. Different nonlinear processes cause radically different distributions of audible error which have different degrees of subjective impact. It is therefore crucial that widely differing distortion types be investigated.

The major advantage of an analysis which is analogous to human perception is that it is potentially independent of distortion type—which is not the case with empirical mapping methods. Table 2 illustrates the performance of the analysis with a range of distortion types, and shows that the total distortion alone is not a sufficient predictor of subjective ranking, whereas the ranking prediction, which includes error entropy, does agree with subjective opinion (after [7]).

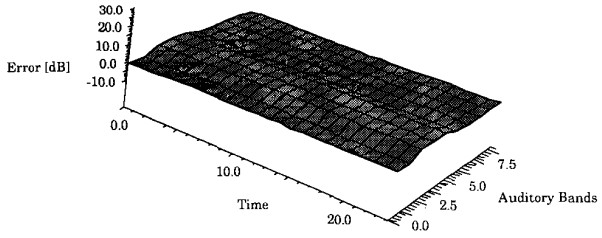


Fig. 5. Error-surface fragment, response shaping.

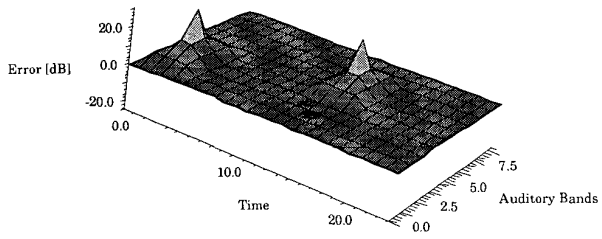


Fig. 6. Error-surface fragment, two peaks.

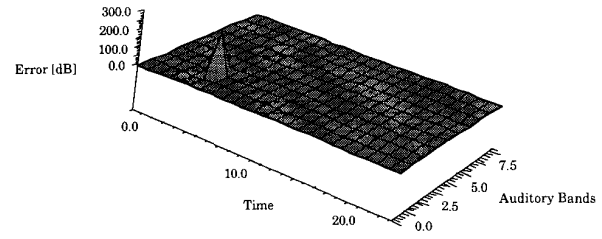


Fig. 7. Error-surface fragment, single peak.

The error patterns EPX and EPY were nonstandard and specifically chosen to give the desired subjective ranking. The types of distortion include different codec types, a nonlinear amplifier characteristic, and a temporal distortion due to the voice activity detector (VAD) in the digital circuit multiplication equipment (DCME).

4 DISCUSSION AND CONCLUSIONS

Perceptual analysis, based on models which are analogous to the human auditory system, can provide objective performance assessment of complex and nonlinear systems, which are not adequately characterized by conventional engineering performance metrics. Auditory-model-based perceptual analysis is potentially independent of distortion type and can provide an objective performance assessment even when a number of nonlinear processes are combined.

Since the analysis techniques are robust across a range of distortions, they provide useful predictions of performance for unknown and unusual distorting processes. This robustness cannot be simply assumed for empirical mapping methods.

The concatenation of a number of nonlinear processes represents a very complex system in which an inaudible artifact generated by one process may yield an audible error artifact when it becomes the input to further processes. It is interesting to consider:

1) That lower than audible thresholds may be appropriate when designing individual data compression stages.

2) That the error artifacts resulting from data reduction are less likely to produce audible errors from subsequent processes if they are "well conditioned," that is, they retain the normal characteristics of the signal. In other words, a nonspeechlike error artifact is more likely to produce an audible error from a subsequent speech cod-

Table 1. Error descriptors for hypothetical error-surface fragments.

Figure	Error Activity (E_a)	Error Entropy (E_e)
5, response shaping	250	4.63
6, two peaks	250	3.78
7, single peak	250	0.87

Table 2. Average error and predicted ranking for different degradations.

Degradation	Subjective Ranking	Average $\log_{10}$(error)	Ranking Prediction
DCME (including VAD and data reduction)	1 (best)	7.288	4.04
Speech codec A (with error pattern EPX)	2	7.232	3.12
Nonlinear amplifier distortion	3	7.810	2.97
Speech codec C (with error pattern EPY)	4 (worst)	7.555	2.36

ing process.

The model described in this engineering report is band-limited to suit telecommunications applications but in principle could be extended to cover the full auditory bandwidth. This would require further subjective testing in order to validate the model's performance and predictions.

Patent applications have been made with regard to both the novel features of the auditory model and the algorithms used to interpret the audible error surfaces.

In conclusion, the value of perceptual analysis techniques across a range of applications is underlined. The diagnostic use and single-figure performance metric aspects of the analysis have been shown, illustrating the success of appropriate psychoacoustic analogies for error-surface interpretation.

5 ACKNOWLEDGMENT

The authors would like to thank their colleagues at BT Labs for their encouragement and support.

6 REFERENCES

[1] "P Series Recommendations," CCITT vol. V, suppl. 3, secs. 2.3–2.6 (1988).

[2] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978 (1992 Dec.).

[3] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. Selected Areas Comm.*, vol. 10 (1992 June).

[4] J. Stuart, "Psychoacoustic Models for Evaluating Error in Audio Systems," *Proc. Inst. Acoust.*, vol. 13 (1991 Nov.).

[5] M. P. Hollier, M. O. Hawksford, and D. R. Guard, "Objective Perceptual Analysis: Comparing the Audible Performance of Data Reduction Schemes," presented at the 96th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 394 (1994 Feb.), preprint 3797.

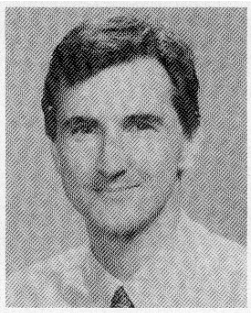
[6] M. P. Hollier, M. O. Hawksford, and D. R. Guard, "Characterization of Communications Systems Using a Speechlike Test Stimulus," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 41, pp. 1008–1021 (1993 Dec.).

[7] M. P. Hollier, M. O. Hawksford, and D. R. Guard, "Error-Entropy and Error-Activity as a Measure of Psychoacoustic Significance in the Perceptual Domain," *IEE Proc. Vision, Image and Signal Process.*, vol. 141, no. 3, p. 203 (1994 June).

[8] R. Mester and U. Franke, "Spectral Entropy-Activity Classification in Adaptive Transform Coding," *IEEE J. Selected Areas Comm.*, vol. 10 (1992 June).

[9] D. J. Finney, *Probit Analysis*, 3rd ed. (Cambridge University Press, London, 1971, reprinted 1980).

THE AUTHORS

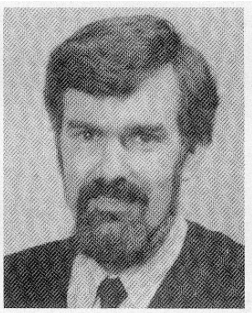


M. Hollier

Mike Hollier obtained a B.Eng. (Hons) degree in mechanical engineering from Plymouth Polytechnic in 1987, and joined BT Laboratories after graduation. He obtained an IOA diploma in acoustics in 1989, receiving an industry award for research into the vibrational behaviour of light structures. He is a member of the IOA.

With a background in audio engineering, he has worked on a number of telephony acoustics projects including the development of a novel noise canceling handset. Recent work has included the development of objective measurement methods capable of predicting the performance of nonlinear audio systems.

Dr. Hollier received a Ph.D. from the University of Essex in Colchester, UK, in 1995 July for his work on objective audio quality assessment. He is currently employed at BT Laboratories, where he is extending his research into multimedia assessment.



D. Guard

David Guard graduate in 1971 with a B.Sc. in physics, followed by a Ph.D. in medical acoustics in 1975. He then joined BT Laboratories to investigate loudspeaking telephones, culminating in the development of the Orator executive audio conferencing system. He then moved into the marketing area to sell Orator nationwide.

Dr. Guard returned to BT Laboratories in 1983 to develop and support telephony terminals. From 1987 he worked on transmission performance of the complete network. Since 1992 he has been a BT Laboratories systems designer concerned with particular speech technology services.

The biography for Malcolm Omar Hawksford was published in the 1993 March issue of the *Journal*.

Scalable Multichannel Coding with HRTF Enhancement for DVD and Virtual Sound Systems*

M. O. J. HAWKSFORD

Centre for Audio Research and Engineering, University of Essex, UK CO4 3SQ

A scalable and reverse compatible multichannel method of spatial audio using transaural coding designed for multiple-loudspeaker feeds is described with a focus on attaining optimum ear signals. A Fourier transform method for computing HRTF matrices is employed, including the generation of a subset of band-limited reproduction channels. Applications considered embrace multichannel audio, DVD, virtual reality, and telepresence.

0 INTRODUCTION

The purpose of this paper is to investigate how transaural processing can enhance conventional multichannel audio both by embedding perceptually relevant information and by improving image stability using additional loudspeakers integrated with supplementary digital processing and coding. The key objective is to achieve scalability in spatial performance while retaining full compatibility with conventional multichannel formats. This enables the system in its most basic form with unprocessed loudspeaker feeds to be used in a conventional multichannel installation. However, by appropriate signal processing additional loudspeaker feeds can be derived, together with the option of exploiting buried data to extract more signals in order to improve spatial resolution. The system is therefore hierarchical in terms of number of loudspeakers, channels, and ultimately spatial resolution, while in its simplest incarnation it remains fully compatible with the system configurations used with multichannel DVD-A and SACD replay equipment.

The multichannel capabilities of DVD[1] technology [1], [2] were designed to enhance stereo[2] sound reproduction by offering surround image and improved envelopment capabilities. Normally multichannel audio encoded onto DVD assumes the ITU standard of a five-loudspeaker configuration driven by five discrete wide-band "loudspeaker feeds." However, a limitation of this system is the lack of a methodology to synthesize virtual images capable of three-dimensional audio (that is, a perception of direction, distance, and height together with acoustic envelopment) rather than just "sound effects" often (although not exclusively) associated with surround sound in a home theater context. The ITU five-channel loudspeaker configuration can also be poor at side image localization, although this deficiency is closely allied to a sensitivity to room acoustics. Nevertheless, DVD formats still offer only six discrete channels, which if mapped directly into loudspeaker feeds remain deficient in terms of image precision, especially if height and depth information is to be encoded.

The techniques described in this paper support scalable spatial audio that can remain compatible with conventional multichannel systems. It is shown that in this class of system, under anechoic conditions, signal processing can be used to match theoretically the ear signals to either a real or an equivalent spatially synthesized sound source. Also, in order to improve image robustness, directional sound-field encoding is retained as exploited in conventional surround sound to match the image synthesized through transaural processing. It may be argued that as the number of channels is increased, there is convergence toward wavefront synthesis [3], where by default optimum ear-signal reconstruction is achieved. However, the proposed system is positioned well into the middle territory[3] and is far removed from the array sizes required for broadband wavefront synthesis. Consequently, from the perspective of wavefront synthesis the transition frequency above which spatial aliasing occurs is located at a relatively low frequency, implying that for the proposed system the core concepts of wavefront synthesis do not apply

* Presented at the 108th Convention of the Audio Engineering Society, Paris, France, 2000 February 19–22; revised 2001 October 10 and 2002 July 15.

[1] Includes both DVD-A and SACD formats.

[2] Of Greek origin, meaning solid, stereo is applicable universally to multichannel audio.

[3] A range of 5 to 32 loudspeakers is suggested.

over much of the audio band.

It is emphasized that an *n*-channel system does not necessarily imply *n* loudspeakers. Indeed, as is well known, it is possible for a two-loudspeaker system to reproduce virtual-sound sources [4], while using more than *n* loudspeakers can help create a more robust and stable illusion. Also, the mature technology of Ambisonics [5]–[7] is scalable and can accommodate both additional loudspeakers and information channels. However, here the encoding is hierarchical in terms of spatial spherical harmonics, although no attempt is made to reconstruct the ear signal directly at the listener. Consequently the approach taken in this paper differs in a number of fundamental aspects from that of Ambisonics, especially since there is no attempt to transform a sound field directly into a spherical harmonic set. Thus it remains for future work to establish the relative merits of these approaches although, because similar loudspeaker arrays are used, there is no fundamental compromise should the system be used either for Ambisonics or for conventional surround sound encoded audio.

The method of spatial audio described in this paper uses a conventional loudspeaker array to surround the listener and to reproduce a directional sound field. In addition, ear signals are simultaneously synthesized using head-related transfer functions (HRTFs) matched to the source image, where it is assumed in all cases that loudspeaker transfer functions have been equalized or otherwise taken into account. A number of examples illustrate the computational methods, which include pairwise transaural image synthesis[4] reported in earlier work, where some preliminary experimental results were also discussed [8]–[10] to establish the efficacy of the method. This technique is especially well matched to multichannel multiloudspeaker installations, where transaural coding can be applied during encoding and recording while processing within the decoder located within the reproduction system can accommodate both additional loudspeakers and loudspeaker positions that differ from those assumed at the encoder. Consequently for an *n*-channel system it is straightforward to employ only *n* loudspeakers, although additional loudspeaker feeds can be derived, while still retaining correct ear signals, either by using matrix techniques or deterministically within the DVD-A format using additional embedded code. However, it is emphasized that in the simplest configuration, using only direct loudspeaker feeds and provided the loudspeakers are correctly located, there are no additional decoding requirements and the system remains fully compatible with all existing recordings.

Alternative technologies such as Ambisonics [11] have used fewer loudspeakers together with sophisticated matrix encoding. Also, there has been substantial research into perceptually based processing to reconstruct a three-dimensional environment using only two channels and two loudspeakers. More recently DOLBY EX[5] has been introduced as a means of synthesizing a center rear channel using nonlinear Prologic[6] processing applied to the rear two channels of a five-channel system. However, this technology is aimed principally at surround sound as conceived for cinema and home theater, with a bias toward sound effects and ambience creation. Nevertheless there exists a grey area between cinema applications, music reproduction, gaming applications, and the synthesis of virtual acoustics, especially as at their core the same multichannel carriers can link all systems. It is therefore not unreasonable to anticipate some degree of convergence as similar theoretical models apply. Also, with conventional multichannel technology it is often the listening environment and the methods used to craft the audio signals that impose the greatest performance limitations.

Multichannel stereo on DVD allows for improved methods of spatial encoding that can transcend the common studio practice of using just pairwise amplitude panning with blending to mono. It is conjectured that by including perceptually motivated processing, three-dimensional "soundscapes" can be rendered rather than just peripheral surround sound. Complex HRTF data by default encapsulate all relevant spatial information [that is, interaural amplitude difference (IAD), interaural time difference (ITD), and directional spectral weighting] and form a generalized approach. However, to reduce signal coloration, a method of HRTF equalization is proposed with an emphasis on characterizing the interear difference signal computed in the lateral plane. The extension to height information in the equalized HRTFs is also discussed briefly.

In Section 5.2.1 a special case is presented for narrow subtended angle, two-channel stereo where it is shown that ear signals derived from a real acoustic source located on the arc of the loudspeakers can be closely synthesized using a mono source with amplitude-only panning. Critically in this example, the HRTFs are defined by the actual locations of the loudspeakers, and so are matched automatically to an individual listener's HRTF characteristics. This is an important aspect of the proposal, which is directly extendable to the method of pairwise association where mismatch sensitivity between listener HRTFs and target HRTFs is reduced. This approach also encapsulates succinctly the principles of two-channel amplitude-only panning stereo while exposing inherent errors as the angle between the loudspeakers is increased.

To summarize, four core elements constitute the proposed scalable and reverse compatible spatial audio system:

- A vector component of the sound field is produced as a loudspeaker array surrounds the listener following conventional multichannel audio practice.
- Pairwise transaural techniques are used to code directional information and to create ear signals matched to the required source signal.
- Matrix processing can increase the number of loudspeakers used in the array while simultaneously preserving the ear signals resulting from transaural processing and nonoptimum loudspeaker placements.
- Embedded digital code[7] [12] is used to create additional channels for enhanced resolution while remaining com-

[4] Subject to a British Telecommunications patent application.
[5] Dolby Laboratories, channel extension technology to AC-3 perceptual coding.
[6] Registered trademark of Dolby Laboratories.
[7] Applicable only to the DVD-A format.

patible with the basic system already enhanced by pairwise transaural processing.

1 HRTF NOTATION

A set of HRTFs is unique to an individual and describes a continuum of acoustic transfer functions linking every point in space to the listener's ears. HRTFs depend on the relative position of the source to the listener and are influenced by distance, reflection, and diffraction around the head, pinna, and torso. In this paper HRTFs were derived from measurements taken at BT Laboratories (BTL) utilizing an artificial head and small microphones mounted at the entrance of each ear canal. Measurements of head-related impulse responses (HRIRs) were performed in an anechoic chamber at 10° intervals using a maximum-length-sequence (MLS) excitation, and the corresponding HRTFs were computed using a time window and the Fourier transform.

To define the nomenclature used for various HRTF subfunctions, consider the arrangement shown in Fig. 1, where the listener's ears are labeled A (left) and B (right) when viewed from above. In a sound reproduction system all sound sources and loudspeakers have associated pairs of HRTFs, uniquely linking them to the listener, whereas in this paper these transfer functions are called the HRTF coordinates for each object given. In Fig. 1 the single sound source X has the HRTF coordinates $\{h_{xa}, h_{xb}\}$, while the three loudspeakers 1, 2, and n, with arbitrary positions, have the coordinates $\{h_a(1), h_b(1)\}$, $\{h_a(2), h_b(2)\}$, and $\{h_a(n), h_b(n)\}$. In specifying the loudspeaker HRTF coordinates, a left–right designation can be included when the loudspeaker array is known to be symmetrical about the centerline. Consequently $h_{la}(r)$ denotes the HRTF between the left-hand loudspeaker r and the left-hand ear A. However, for arrays having only three symmetrically positioned loudspeakers (left, center, and right) a simpler notation is used in Section 3, namely, $\{h_{la}, h_{lb}\}$, $\{h_{ca}, h_{cb}\}$, and $\{h_{ra}, h_{rb}\}$. It should be observed that in typical HRTF calculations, such as the evaluation of the positional transfer functions G_R and G_L used in transaural signal processing (see Section 3.1), ear-canal equalization need not be incorporated, provided the same set of HRTFs is used for both loudspeaker and image locations. Then the ear-canal transfer functions cancel, assuming they are not directionally encoded. For example, in Section 4 Eqs. (23a) and (23b) describe typical transaural processing to derive the positional transfer functions G_R and G_L, where any transfer function components common to all HRTFs cancel.

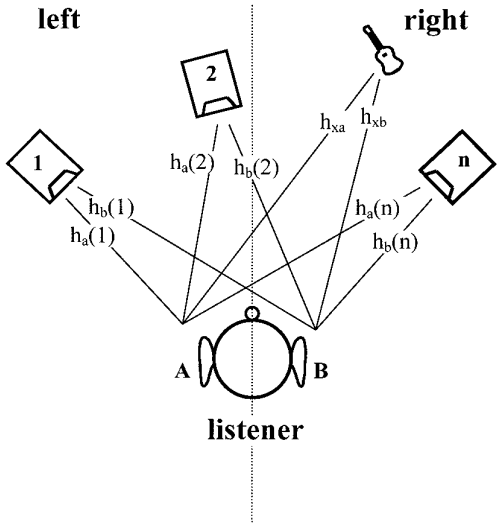


Fig. 1. Definition of HRTF notation for n loudspeakers positioned arbitrarily around the listener.

2 HRTF EQUALIZATION

The HRTFs used in transaural processing reveal frequency response variations that may contribute tonal coloration when sound is reproduced. In this section a strategy for equalization is studied that reduces the overall spectral variation, yet retains the key attributes deemed essential for localization. A simplified form of HRTF is also defined, which can prove useful in multiloudspeaker systems.

2.1 Methods of Equalization

When sound is reproduced over a conventional multiloudspeaker array, where for example a signal is spatialized using pairwise amplitude panning, equalization as a function of direction is not normally employed. In such a system sound is perceived generally as uncolored, even though the ears, head, and torso impose direction-specific spectral weighting. However, although HRTFs used in transaural processing take account of both the source location and the loudspeakers, reducing frequency response variations can ameliorate tonal variance, which may become accentuated as the phantom image moves away from the loudspeaker locations and when the listener turns away from the optimum forward orientation. Also, systems are rarely optimally aligned and exhibit sensitivity to small head motions, both of which map into frequency response errors in the reconstructed ear signals. Consequently the aim is to introduce minimum spectral modifications commensurate with achieving spatialization.

It is proposed that for image localization within the lateral plane the relationship between the complex interaural difference signal and the signal components common to both ears is the critical factor. This conjecture is based on the premise that for lateral images, spectral components common to both ears relate closely to the source spectrum whereas the interaural spectrum is strongly influenced by source direction. Consequently it is argued that modification to the common spectrum causes principally tonal coloration, whereas the relationship between common spectrum and interaural spectrum is more critical to localization, even though spectral cues embedded in the source can induce an illusion of height. This approach may be extendable to include height localization, although it is recognized that additional spectral weighting of the monaural component can be required following, for example, the boosted-band experiments performed by Blauert [13].

2.1.1 Lateral-Plane HRTF Equalization

The proposed method of lateral-plane HRTF equalization first transforms each HRTF pair into sum and difference (M-S) coordinates and then performs equalization on the corresponding pair by dividing by the corresponding sum spectrum. It is proposed that all HRTFs in a set should be equalized using this technique in order to maintain relative group delay and, with appropriate weighting, relative level.

As defined in Section 1, let the HRTFs for a given source location X be h_{xa} and h_{xb}, and let the corresponding complex sum and difference transforms be HSUM_x and HDIFF_x. Thus

$$\mathrm{HSUM}_x = h_{xa} + h_{xb} \tag{1}$$

$$\mathrm{HDIFF}_x = h_{xa} - h_{xb} \, . \tag{2}$$

Four methods of HRTF equalization that match this objective are identified, where $\{h_{xea}, h_{xeb}\}$ are the resulting HRTFs after equalization.

Method 1: Equalization by the modulus of the complex sum spectrum,

$$h_{xea} = \frac{h_{xa}}{|h_{xa} + h_{xb}|} W_{nx} \tag{3a}$$

$$h_{xeb} = \frac{h_{xb}}{|h_{xa} + h_{xb}|} W_{nx} \, . \tag{3b}$$

Method 2: Equalization by complex sum spectrum,

$$h_{xea} = \frac{h_{xa}}{h_{xa} + h_{xb}} W_{nx} \tag{4a}$$

$$h_{xeb} = \frac{h_{xb}}{h_{xa} + h_{xb}} W_{nx} \, . \tag{4b}$$

Method 3: Equalization by the derived minimum-phase spectrum of the complex sum spectrum,

$$h_{xea} = \frac{h_{xa}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa} + h_{xb})))))} W_{nx} \tag{5a}$$

$$h_{xeb} = \frac{h_{xb}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa} + h_{xb})))))} W_{nx} \, . \tag{5b}$$

Here W_{nx} are the normalization coefficients calculated to maintain the relative levels after equalization of all HRTF coordinates in the set. Each form of equalization delivers identical magnitude spectra in the HRTFs, although there are variations in the time-domain waveforms resulting from phase response differences. To illustrate these variations, consider an example HRTF pair corresponding to a nominal 30° off-axis image source. Fig. 2(a) shows the measured HRIRs, whereas Fig. 2(b)–(d) presents the impulse responses resulting from each form of equalization in order that both pre- and postring can be compared. Fig. 3 shows the corresponding amplitude spectra before and after equalization, and Fig. 4 illustrates the sum and difference spectra, again before and after equalization.

In selecting a potential equalization strategy, it is a necessary condition that the relative time difference between HRTF pairs be maintained. Also, the time-domain waveforms should not accentuate or exhibit excessive prering or postring, as this can produce unnatural sound coloration. Although each equalization method meets the principal objective, the technique of forging the denominator from the minimum phase of the sum spectrum yields results with minimum prering. In essence, the minimum-phase information common to both ear signals is removed, leaving mainly excess phase components to carry the essential time-delay information. Equalization using the complex sum spectrum (method 2) also yields results close to the requirements. However, inspection of Fig. 2(c) shows that the right-ear response, which in this case has the greater delay, exhibits prering extending back in time to the commencement of the left HRIR.

However, experience gained with equalization has revealed that certain image locations, particularly toward the center rear, can yield excessive ringing after equalization. Consequently a further equalization variant is proposed. This is similar to method 3, but it differs in the way the sum spectrum is computed and is defined as follows.

Method 4: Equalization by the derived minimum-phase sum of the moduli of each complex spectrum,

$$h_{xea} = \frac{h_{xa}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa}) + \mathrm{abs}(h_{xb})))))} W_{nx} \tag{6a}$$

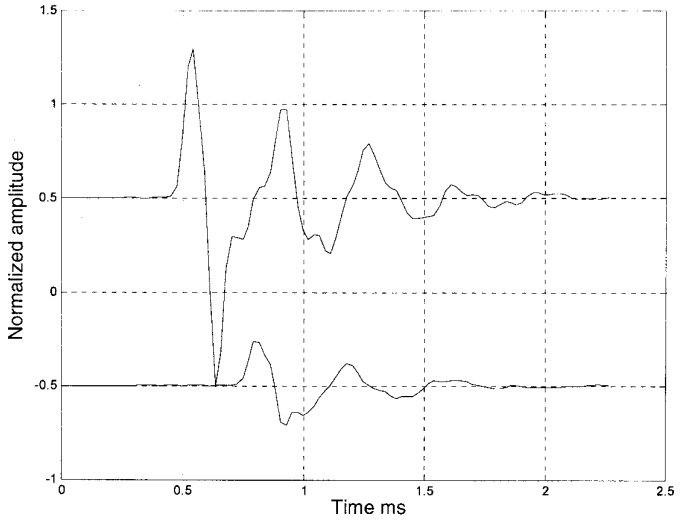
$$h_{xeb} = \frac{h_{xb}}{\exp(\mathrm{conj}(\mathrm{hilbert}(\log(\mathrm{abs}(h_{xa}) + \mathrm{abs}(h_{xb})))))} W_{nx} \, . \tag{6b}$$

In the denominator this algorithm factors out the interaural time difference between left and right signals, which otherwise map into artificial amplitude response variations in the complex sum spectrum. As such this procedure could be argued to be a better estimator of the common spectrum, as human auditory processing does not sum ear signals directly. Overall the effect on HRTFs is minor. Fig. 5 presents results that should be compared directly with those in Fig. 3.

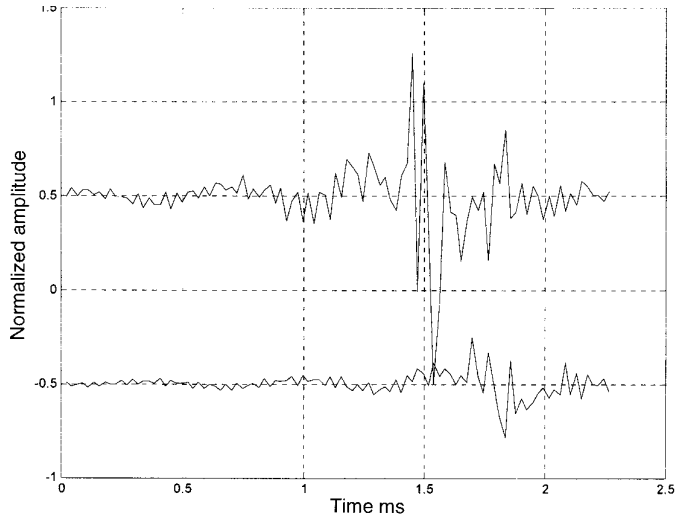
Finally a further variant of equalization is where an average of all sum spectra is formed and the HRTFs are modified following procedures similar to those reported in this section but with particular emphasis on the minimum-phase and sum-of-moduli techniques. However, in this case, since all HRTFs are modified by a common equalization function in a way similar to ear-canal equalization, when positional transfer functions are calculated, their form is unchanged.

2.1.2 Equalization with the Addition of Height Cues

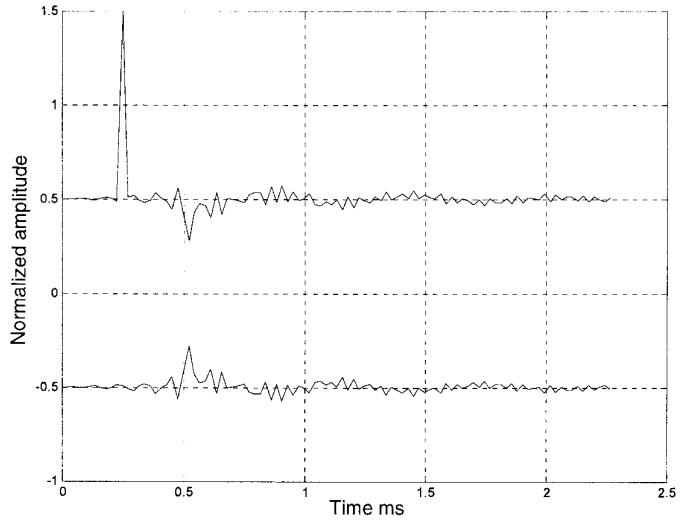
Research by Blauert [13] has shown that by introducing specific frequency-dependent characteristics into the HRTFs a sensation of height is achievable. However, the



(a)

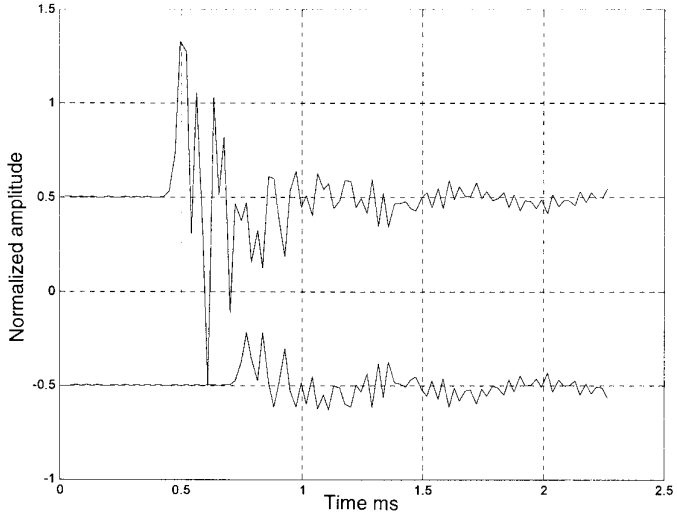


(b)



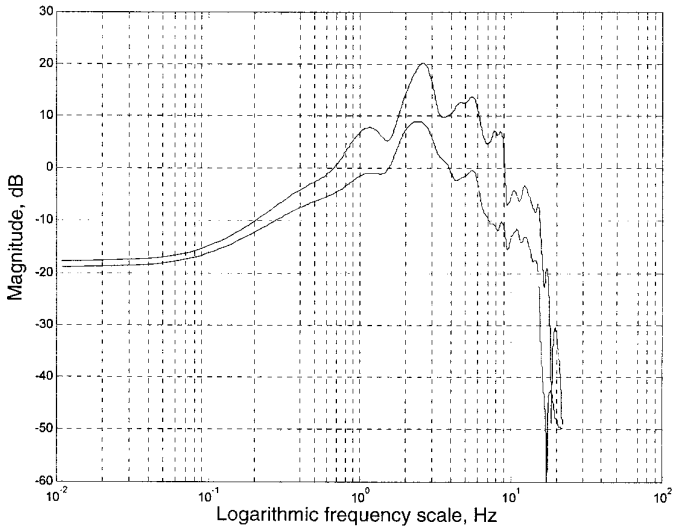
(c)

Fig. 2. Normalized time-domain left–right HRTFs at 30°. Top—left ear; bottom—right ear. (a) Measured. Observe relative time displacement revealing ITD and lack of prering in natural responses. (b) Method 1 equalized. Observe excessive prering that blurs commencement of the two HRIRs. (c) Method 2 equalized. HRIRs exhibit mirror images except for initial impulse. (d) Method 3 equalized. Prering reduced and initial ITD of HRIRs maintained.

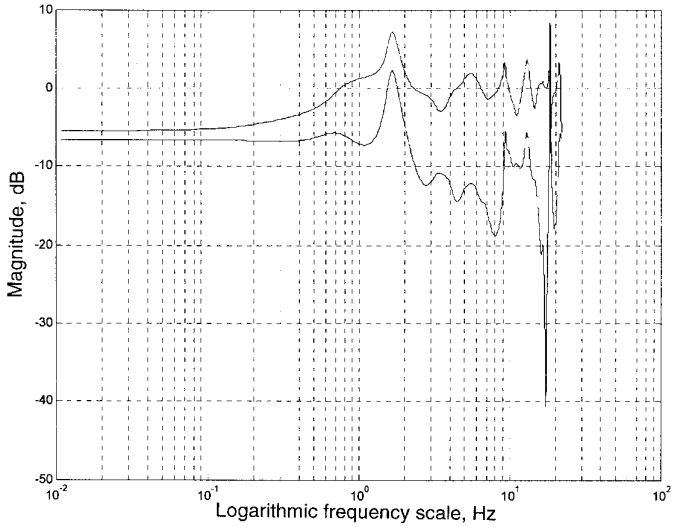


(d)

Fig. 2. *Continued*

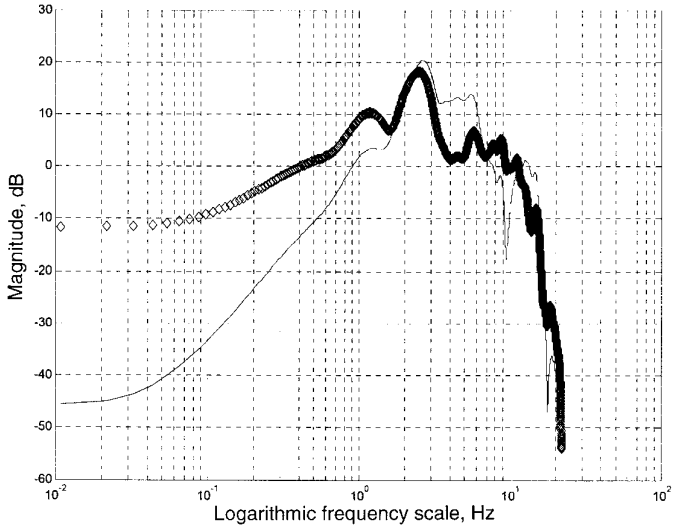


(a)

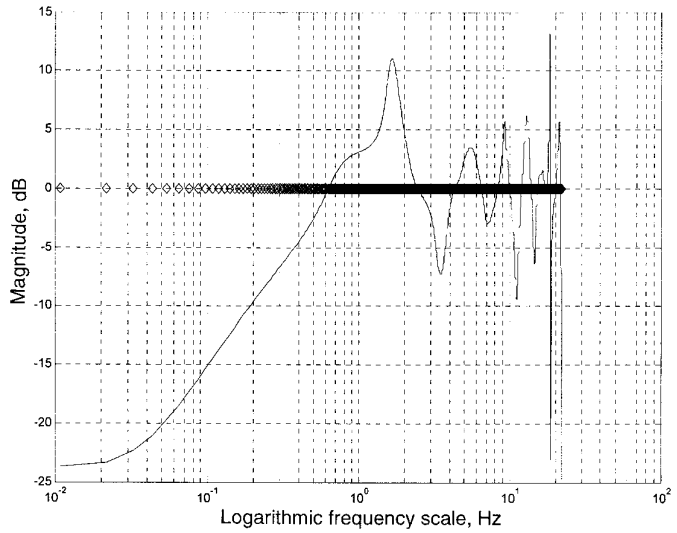


(b)

Fig. 3. Magnitude HRTF pair at 30°. Top—left ear; bottom—right ear. (a) Measured. (b). Equalized. Results are identical for methods 1, 2, and 3.



(a)



(b)

Fig. 4. Magnitude HRTF sum and difference spectra at 30°. Sum—"diamond" line; difference—continuous line. (a) Unequalized. (b) Equalized, applicable to methods 1, 2, and 3. (Note constant-level sum spectrum following equalization.)

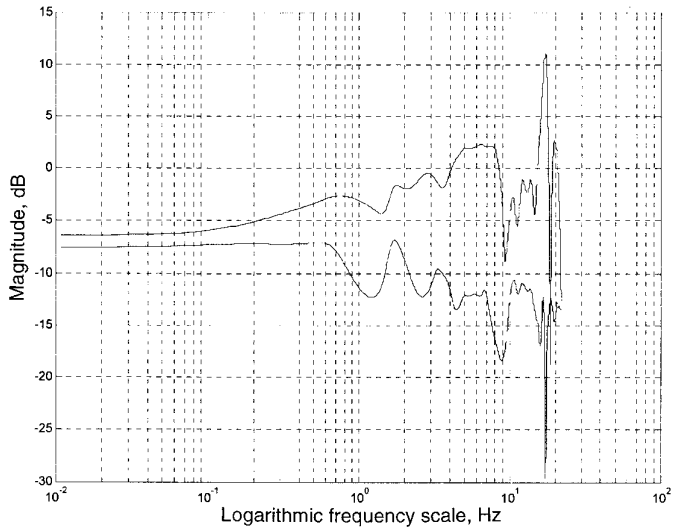


Fig. 5. Equalized HRTF pair at 30°, applicable to method 4 only. Top—left ear; bottom—right ear.

question arises as to whether this modification is compatible with the equalization strategies presented in Section 2.1.1. For example, the following questions need to be considered:

- Is it sufficient to measure the HRTF coordinates only at the required location above the lateral plane and then apply equalization, and will then sufficient information remain buried in the interear difference signal with a unique characterization to discriminate against lateral images with equivalent interaural time differences?
- Does the absolute amplitude response variation, rather than just the difference response variation inherent in the HRTFs, represent a major factor in producing height cues?
- Are there secondary factors, such as ground reflections, which introduce additional cues to aid height localization? Effectively this would require at least two interfering sets of HRTFs to be summed.

A full investigation of these points relating to height is beyond the scope of the present study. However, if the ground reflection model were responsible, then the equalization methods could be applied individually to the direct source and to the ground reflection, with the results combined by taking the path difference into account.

2.2 Simplified HRTF Models

In applications where phantom images are positioned close to the locality of the loudspeakers, it may be sufficient to use a simple form of HRTF. This is particularly applicable with multiloudspeaker arrays where a vector component already forms a strong localization clue. Fig. 6 shows a source image at angle θ defined by h_{xa} and h_{xb} with respect to a human head of diameter d meters.

2.2.1 Simple HRTF Model 1

The model ignores head shadowing and assumes that only the interaural time difference is significant. Hence for a source at angle θ from the forward position, the respective HRTF coordinates are approximately

$$h_{xa} = \exp\left\{-j2\pi f\left[T - \frac{d}{2c}\sin(\theta)\right]\right\} \quad (7a)$$

$$h_{xb} = \exp\left\{-j2\pi f\left[T + \frac{d}{2c}\sin(\theta)\right]\right\} \quad (7b)$$

where the velocity of sound is c m/s and T s is the time delay from the source to the center of the head. In this model the advantage of equalization method 3 is evident as no equalization need be applied.

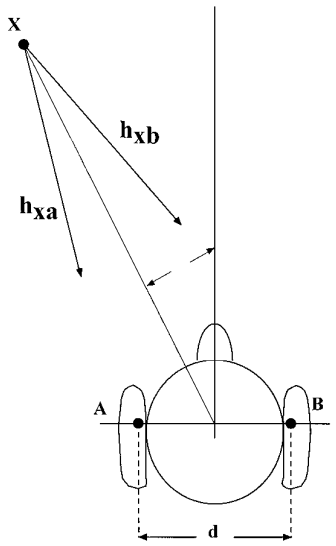


Fig. 6. Sound source X and simplified head model used to derive approximate HRTFs.

2.2.2 Simple HRTF Model 2

In this second model both the front and the back waves are considered, where the back wave results from head defraction. In this representation a wave incident on ear A produces, by head defraction, a secondary signal at ear B. The head diffraction transfer function from ears A to B, $DH_{A-B}(r, \theta, \phi)$, is a function of the direction of the incident wave defined by the spherical coordinates $\{r, \theta, \phi\}$. A similar function linking ears B to A is defined, $DH_{B-A}(r, \theta, \phi)$. Hence for the source HRTF coordinates $\{h_{xa}, h_{xb}\}$,

$$h_{xa} = \exp\left\{-j2\pi f\left[T - \frac{d}{2c}\sin(\theta)\right]\right\} + DH_{B-A}(r,\theta,\phi)\exp\left\{-j2\pi f\left[T + \frac{d}{2c}\sin(\theta)\right]\right\} \quad (8a)$$

$$h_{xb} = \exp\left\{-j2\pi f\left[T + \frac{d}{2c}\sin(\theta)\right]\right\} + DH_{A-B}(r,\theta,\phi)\exp\left\{-j2\pi f\left[T - \frac{d}{2c}\sin(\theta)\right]\right\}. \quad (8b)$$

In a simple model the diffraction transfer functions could be represented as attenuation DH_k with a time delay of approximately the interaural time delay ΔT_{A-B},

$$DH_{A-B}(r,\theta,\phi) = DH_{B-A}(r,\theta,\phi) \approx DH_k \exp(-j2\pi f \Delta T_{A-B}). \quad (9)$$

3 MULTILOUDSPEAKER ARRAYS IN TWO-CHANNEL STEREO

This section introduces variants to two-channel, two-loudspeaker transaural processing to demonstrate how a two-channel signal format can be mapped into n feeds to drive a multiloudspeaker array [14]. It is assumed that more than two loudspeakers are driven simultaneously by signals derived from a single-point sound source, while formal methods show that the correct ear signals can be retained. Besides supporting stand-alone applications, these transformations are relevant in the development of multichannel transaural stereo, as described in Section 5.

The outputs of an *n*-array of loudspeakers combine by acoustical superposition at the entrance to each ear canal. The principal condition for accurate sound localization is that these signals match the signals that would have been generated by a real sound source, both in the static case and in the case for small head rotations. Also, by using several loudspeakers placed to surround the listener, sound-field direction can make the system more tolerant to head motion. Consequently changes in ear signals with head motion match more closely those of a real image.

A static sound source, whatever its size and physical location, produces two ear signals that fully define the event, provided the relative head position to source is fixed. In practice it is possible to generate the correct ear signal from two or more noncoincident loudspeakers that can take any arbitrary position around the head. However, if the position of the head moves, then a change in the ear signals results, which no longer match the phantom image correctly, and a localization error is perceived. In a system of wavefront reconstruction this distortion is minimized, although the penalty is a large number of loudspeakers and channels. However, if a limited number *n* of loudspeakers is used (for example, $n = 12$), then although image position distortion still occurs, the effect is reduced as there is a robust directional component. Also, when a phantom image coincides with a loudspeaker position, the positional distortion as a function of head position tends to zero, although it is debatable whether this is a desirable situation as images from other locations are represented differently in terms of their radiated cones of sound.

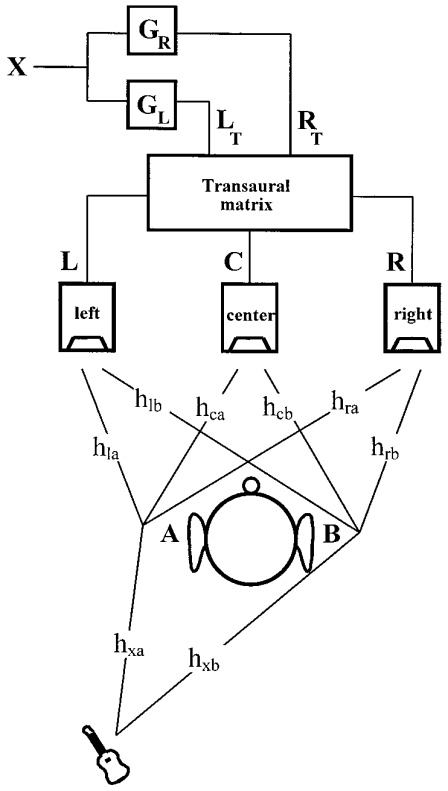


Fig. 7. Transaural processing using symmetrical three-loudspeaker array.

3.1 Three-Loudspeaker Transaural Processing

To illustrate how to accommodate more than two loudspeakers in an array while retaining the requirements for precise HRTF formulation, consider a three-loudspeaker array as illustrated in Fig. 7. In this system a mono source signal X is filtered by the positional transfer functions G_R and G_L to form L_T and R_T, which in turn form inputs to the matrix $[a]$. By way of example a Trifield[8] matrix (after Gerzon [15]) is selected, which is defined as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = \begin{bmatrix} 0.8850 & -0.1150 \\ 0.4511 & 0.4511 \\ -0.1150 & 0.8850 \end{bmatrix}. \tag{10}$$

By applying the coefficients defined in matrix $[a]$, the three loudspeaker signals L, C, and R (left, center, right) can be derived. However, to reproduce optimum localization, the system requires that the ear signals produced by the three loudspeakers match the ear signals that would be produced by the real source.

3.1.1 Analysis

The positional transfer function matrix $[G]$ converts the mono signal X to L_T and R_T as

$$\begin{bmatrix} L_T \\ R_T \end{bmatrix} = \begin{bmatrix} G_R \\ G_L \end{bmatrix} X \ . \tag{11}$$

Using matrix $[a]$, the loudspeaker feeds L, C, and R are then derived from $[G]X$,

$$\begin{bmatrix} L \\ C \\ R \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} G_L \\ G_R \end{bmatrix} X \ . \tag{12}$$

However, recalling the HRTFs as defined in Fig. 7, where h_{xa} and h_{xb} are the HRTF coordinates of source image X, then

$$\begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} X = \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} L \\ C \\ R \end{bmatrix} \tag{13}$$

where, substituting for L, C, and R,

$$\begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} = \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} G_L \\ G_R \end{bmatrix}. \tag{14}$$

The positional transfer functions G_R and G_L then follow by matrix inversion,

$$\begin{bmatrix} G_L \\ G_R \end{bmatrix} = \left\{ \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right\}^{-1} \begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} \tag{15}$$

[8] Registered trademark describing a two-channel to three-loudspeaker mapping proposed by Gerzon [15].

from which the loudspeaker feeds *L*, *C*, and *R* are calculated,

$$\begin{bmatrix} L \\ C \\ R \end{bmatrix} = \left\{ \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right\}^{-1} X \; . \tag{16}$$

In practice *L*, *C*, and *R* are calculated directly using matrix inversion. However, because the transfer functions can have several thousand elements, to avoid large-dimension matrices the solution can be decomposed as follows. Define

$$\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} = \begin{bmatrix} h_{la} & h_{ca} & h_{ra} \\ h_{lb} & h_{cb} & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \tag{17}$$

giving

$$\begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} G_L \\ G_R \end{bmatrix} \tag{18}$$

where, using matrix inversion, the positional transfer functions are

$$G_R = \frac{h_{xa} h_{21} - h_{xr} h_{11}}{h_{12} h_{21} - h_{11} h_{22}} \tag{19a}$$

$$G_L = \frac{h_{xa} h_{12} - h_{xb} h_{22}}{h_{12} h_{21} - h_{11} h_{22}} \tag{19b}$$

which enable *L*, *C*, and *R* to be calculated. Fig. 8 shows example transfer functions linking the system input to the three loudspeaker inputs (*L*, *C*, and *R*) located at −45°, 0°, and 45°, with a source location at 30°. Simulations confirm that the correct ear signals are produced as shown in Fig. 9, while Figs. 10 and 11 present the magnitudes of the positional transfer functions G_L and G_R and their differential phase response, respectively.

3.2 Three-Loudspeaker Matrix with Band-Limited Center Channel

This section extends multiloudspeaker transaural processing by considering a case where the center channel is band-limited by a low-pass filter with a transfer function $\lambda(f)$. For example, $\lambda(f)$ could constrain the center channel to operate only in the band where the ear and brain employ interaural time differences for localization. Alternatively the center channel may be used mainly for low-frequency reproduction.

The inclusion of $\lambda(f)$ yields effective HRTF center channel coordinates $\{h_{ca.}*\lambda(f),\ h_{cb.}*\lambda(f)\}$, where .* implies element-by-element vector multiplication. Hence, from the equations derived in Section 3.1, *L*, *C*, and *R* follow,

To illustrate this system with band-limited center channel, Fig. 12 shows again the system input to the loudspeaker transfer functions for the three loudspeakers located at −45°, 0°, and 45°, with a phantom source location at 30°. The low-pass filter $\lambda(f)$ in the center channel has a cutoff frequency of 100 Hz with an asymptotic attenuation slope of 40 dB per octave. The ear signals are formed correctly and are identical to those presented in Fig. 9.

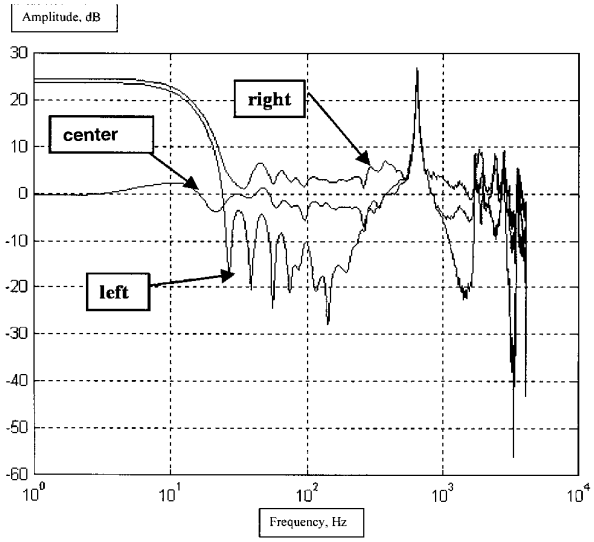


Fig. 8. Loudspeaker feed transfer functions for Trifield matrix linking input to three loudspeakers located at −45°, 0°, and 45°, image at 30°.

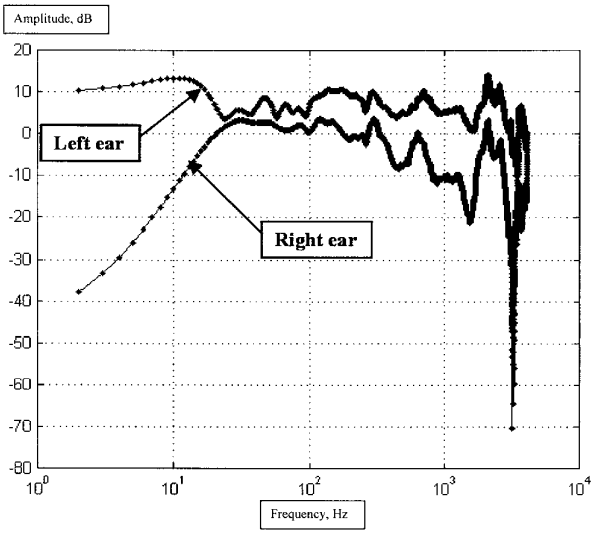


Fig. 9. Input-to-ear transfer functions for three-loudspeaker system with Trifield matrix, corresponding to functions shown in Fig. 8.

$$\begin{bmatrix} L \\ C \\ R \end{bmatrix} = \left\{ \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} h_{xa} \\ h_{xb} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_{la} & h_{ca.}*\lambda(f) & h_{ra} \\ h_{lb} & h_{cb.}*\lambda(f) & h_{rb} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \right\}^{-1} \begin{bmatrix} 1 \\ \lambda(f) \\ 1 \end{bmatrix} X \; . \tag{20}$$

An attraction of this configuration is that the center channel can have a limited bandwidth while offering improvements in bass quality both in terms of power handling and by improving modal dispersion in the listening room. Alternatively, if a loss of spatial resolution at low frequency is permitted, then the center channel could function as a subwoofer with an upper response that extends only into the lower midband frequency range. The left- and right-hand loudspeakers would extend to high frequencies, although with restricted low-frequency performance.

3.3 *n*-Loudspeaker Array with Two-Channel Transaural Processing

The method of using more than two loudspeakers can be generalized to *n* loudspeakers while retaining only two information channels, where for example the loudspeakers surround the listener in a symmetrical array. The left- and right-hand loudspeakers in the array are fed by one of two information signals, and each loudspeaker has individual weighting defined by a coefficient matrix [*a*]

$$\begin{bmatrix} \mathrm{LS}(1) \\ \mathrm{LS}(2) \\ \vdots \\ \mathrm{LS}(n) \end{bmatrix} = \left\{ \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \begin{bmatrix} h_{\mathrm{xa}} \\ h_{\mathrm{xb}} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_{\mathrm{a}}(1) & h_{\mathrm{a}}(2) & \ldots & h_{\mathrm{a}}(n) \\ h_{\mathrm{b}}(1) & h_{\mathrm{b}}(2) & \ldots & h_{\mathrm{b}}(n) \end{bmatrix} \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \right\}^{-1} X \, . \tag{21}$$

Although this matrix equation cannot be solved in general as there are too many independent variables, solutions can be achieved when the matrix [*a*] is specified. For general two-channel stereophonic reproduction this system offers little advantage. However, in a telepresence and teleconference environment the coefficient matrix [*a*] may be transmitted for a given talker alongside the two information channels. A transaural reproduction system can then be conceived, where the coefficients are updated dynamically to enhance directional coding. This becomes particularly attractive where there are a number of talkers, as the coefficient matrix could be adjusted dynamically to enhance localization.

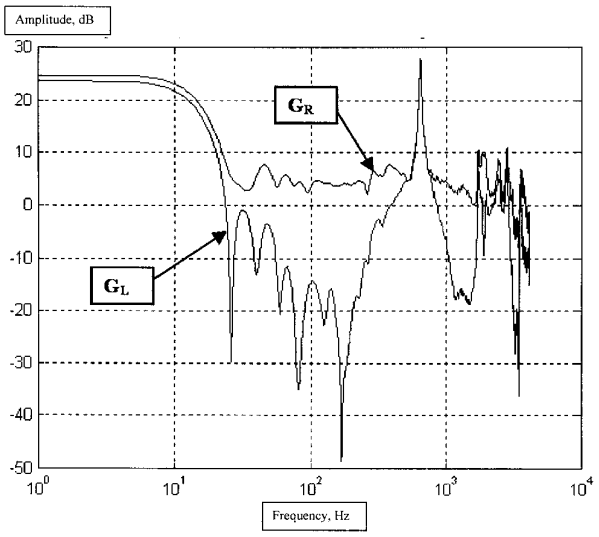


Fig. 10. Positional transfer functions G_L and G_R, corresponding functions shown in Fig. 8.

4 MULTICHANNEL PARADIGM EXPLOITING PAIRWISE TRANSAURAL STEREO

This section reviews a spatial audio paradigm that links multichannel audio and transaural processing. The technique augments the directional clues inherent in multichannel stereo reproduction by embedding HRTF data such that the ear signals are matched more accurately to those produced by the source image. By default, such processing includes both frequency (interaural amplitude response) and time (interaural time response) information and therefore forms an elegant method of virtual image manipulation. Also, because HRTFs vary with both angular position and distance, sound sources can be synthesized and manipulated in three-dimensional space, together with reflections spatialized using their HRTF coordinates, which further enhances this process. In a multichannel system this processing is performed during source coding and is therefore compatible with all DVD formats.

The proposal operates at six principal levels:

- Selecting a pair of loudspeakers whose subtended angle includes the position of the phantom image helps reinforce the sound direction and matches conventional mixing practice for localization in multichannel systems.
- Encoded amplitude differences in signals above about 2 kHz support localization using interaural amplitude differences.
- Encoded time differences in signals below about 2 kHz support localization using interaural time differences where an extended bass performance is desirable.
- The addition of transaural processing based on HRTF data enables the construction of ear signals that match the original event and aids localization.
- Closer spacing of loudspeakers in a multiloudspeaker array reduces sensitivity to the precise form of HRTF characteristics, thus making an averaged HRTF set more applicable to a wide range of listeners.
- The effect of moderate head motion, which is a desirable attribute for improving localization, is supported. For relatively small loudspeaker subtended angles the error in ear-signal reconstruction is reduced when the head is moved by a small angle such that the vector component reenforces localization.

As an example, consider a circular array of n loudspeakers, as shown in Fig. 13. In this system pairwise coding (PWC) selects the two closest loudspeakers such that an image X falls within the subtended angle at the listening position. Two-channel HRTF synthesis is then used to form the optimum ear signals. For example, if loudspeakers r and $r + 1$ are selected from the n-array of loudspeakers, then loudspeaker feeds LS(r) and LS($r + 1$) are computed,

$$\begin{bmatrix} \mathrm{LS}(r) \\ \mathrm{LS}(r+1) \end{bmatrix} = \begin{bmatrix} h_\mathrm{a}(r) & h_\mathrm{a}(r+1) \\ h_\mathrm{b}(r) & h_\mathrm{b}(r+1) \end{bmatrix}^{-1} \begin{bmatrix} h_\mathrm{xa} \\ h_\mathrm{xb} \end{bmatrix} X = \begin{bmatrix} G_r \\ G_{(r+1)} \end{bmatrix} X \ . \quad (22)$$

Matrix [G] defines a set of positional transfer functions, which are effectively filters located between source and loudspeaker feed, where the HRTF notation was defined in Section 1 with reference to Fig. 1.

Solving for the positional transfer function matrix [G],

$$G_r = \frac{h_\mathrm{xa} h_\mathrm{b}(r) - h_\mathrm{xb} h_\mathrm{a}(r)}{h_\mathrm{a}(r+1) h_\mathrm{b}(r) - h_\mathrm{b}(r+1) h_\mathrm{a}(r)} \quad (23a)$$

$$G_{r+1} = \frac{h_\mathrm{xb} h_\mathrm{a}(r+1) - h_\mathrm{xa} h_\mathrm{b}(r+1)}{h_\mathrm{a}(r+1) h_\mathrm{b}(r) - h_\mathrm{b}(r+1) h_\mathrm{a}(r)} \ . \quad (23b)$$

This result can be generalized for an n-array of loudspeakers as

$$\begin{bmatrix} \mathrm{LS}(1) \\ \mathrm{LS}(2) \\ \vdots \\ \mathrm{LS}(n) \end{bmatrix} = \left\{ \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \begin{bmatrix} h_\mathrm{xa} \\ h_\mathrm{xb} \end{bmatrix} \right\} \left\{ \begin{bmatrix} h_\mathrm{a}(1) & h_\mathrm{a}(2) & \ldots & h_\mathrm{a}(n) \\ h_\mathrm{b}(1) & h_\mathrm{b}(2) & \ldots & h_\mathrm{b}(n) \end{bmatrix} \begin{bmatrix} a_1 & a_1 \\ a_2 & a_2 \\ \vdots & \vdots \\ a_n & a_n \end{bmatrix} \right\}^{-1} X = \begin{bmatrix} G_1 \\ G_2 \\ \vdots \\ G_n \end{bmatrix} X \ . \quad (24)$$

If a sound source falls between loudspeakers r and $r + 1$, then $a_r = a_{r+1} = 1$; otherwise all remaining coefficients in matrix [a] are set to zero. Consequently for a sound source to circumnavigate the head, the HRTF coordinates h_xa and h_xb must change dynamically, whereas as the source moves between loudspeaker pairs, the coefficient matrix is switched to redirect the sound.

A four-channel, four-loudspeaker PWC scheme is shown in Fig. 14. The positional transfer functions {G_1, G_2, G_3, G_4} are calculated for each source location, which then filters the source signal X to form the loudspeaker feeds. Because transaural processing is performed at the encoder, a simple replay system is supported. Consequently complete compatibility with conventional multichannel audio is retained.

5 INCREASED NUMBERS OF LOUDSPEAKERS

An increase in the number of loudspeakers can achieve more even sound distribution, distribute power handling, and possibly lower the sensitivity to room acoustics. This section considers methods by which the number of loudspeakers in an array can be increased.

In the basis system, where loudspeakers are linked directly to information channels located at coordinates compatible with the encoding HRTF coordinates, the loudspeakers are designated nodal loudspeakers. An n-

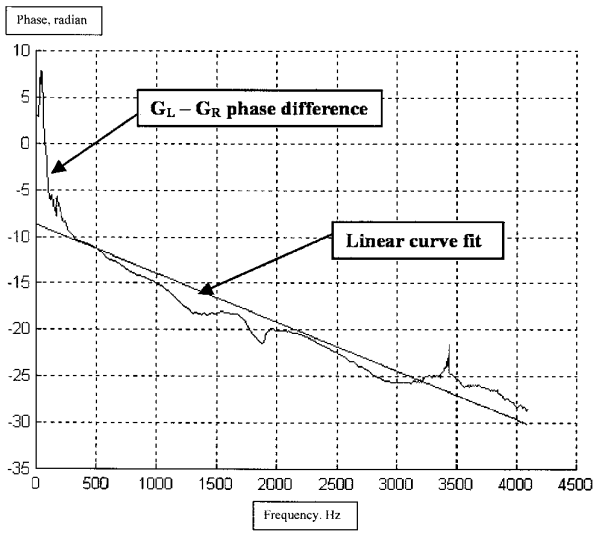


Fig. 11. Differential phase response between positional transfer functions G_L and G_R corresponding to functions shown in Fig 8.

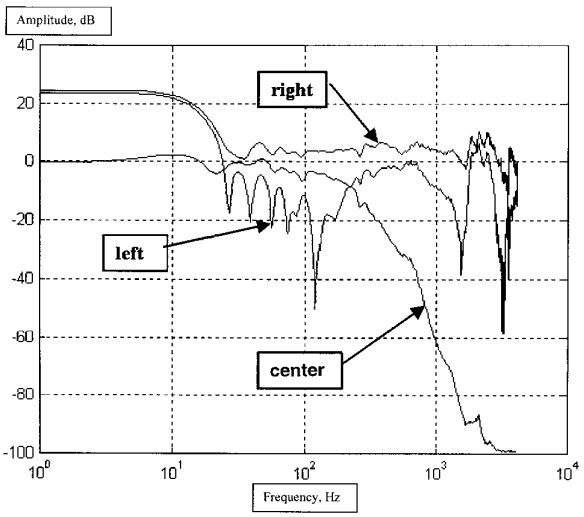


Fig. 12. Transfer functions linking input to three-loudspeaker feeds for Trifield matrix, but with center channel band-limited to 100 Hz, corresponding to functions shown in Fig. 8.

channel system therefore has *n* nodal loudspeakers, which constitute the basis array, with the corresponding drive signals, or primary signals, collectively forming the primary signal set. Loudspeakers in addition to the nodal loudspeakers are termed secondary loudspeakers.

5.1 Compensation for Inclusion of Secondary Loudspeakers

The objective is to derive additional signals within the decoder to drive secondary loudspeakers located between the nodal loudspeakers. However, the ear signals must be conserved and theoretically remain identical to the case where only the nodal loudspeakers are present. It is assumed here that all loudspeakers in the array have identical transfer functions and therefore do not affect the decoder process. If this is not the case, then they require individual correction. A sector of such an array is shown in Fig. 15. In this array nodal loudspeakers r and $r + 1$ have the respective HRTF coordinates $[h_a(r), h_b(r)]$ and $[h_a(r + 1), h_b(r + 1)]$ at the listener, whereas the single secondary loudspeaker p has the HRTF coordinates $[h_a(p), h_b(p)]$.

The synthesis of the drive signal for secondary loudspeaker p employs two weighting functions λ_{p1} and λ_{p2} applied to the respective primary signals LS_r and LS_{r+1} such that LS_p, the drive signal to the secondary loudspeaker p, is

$$LS_p = \lambda_{p1}\, LS_r + \lambda_{p2}\, LS_{r+1}\,. \tag{25}$$

However, when the secondary loudspeaker enters the array, it is necessary to compensate the output from the two adjacent nodal loudspeakers in order that the ear signals remain unchanged. Ideally this needs to be achieved without knowledge of the encoding parameters. Otherwise the modified array cannot be used universally in a multichannel audio system. A scheme capable of meeting this objective is shown in Fig. 16, where the compensation transfer functions γ_{p1} and γ_{p2} filter the signal LS_p to yield signals that are added to LS_r and LS_{r+1}.

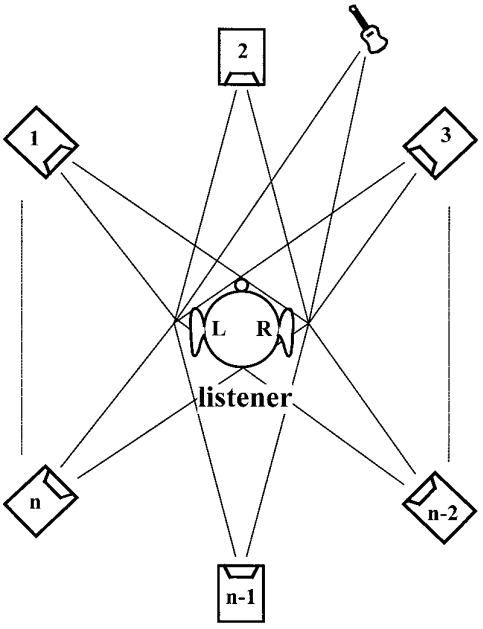


Fig. 13. *n*-loudspeaker array, suitable for transaural pairwise stereo.

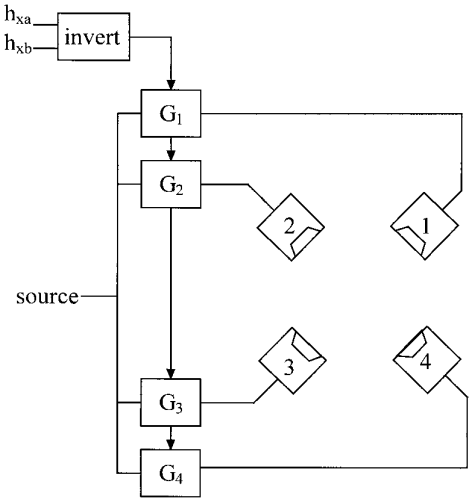


Fig. 14. Four-loudspeaker array with pairwise transaural synthesis.

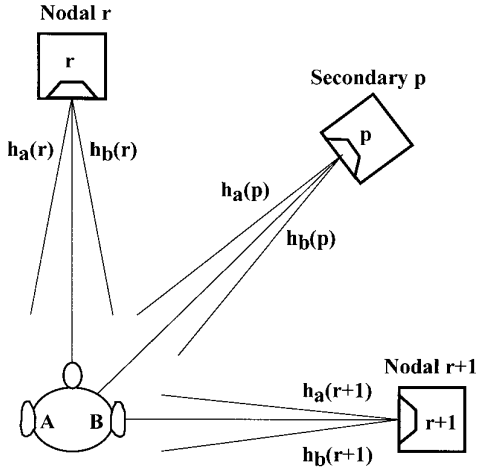


Fig. 15. Nodal loudspeakers with additional secondary loudspeaker.

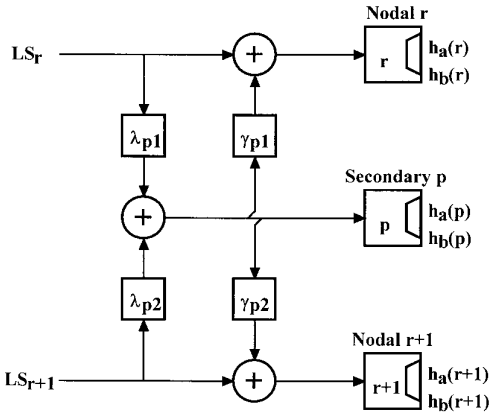


Fig. 16. Decoder processing to compensate for secondary loudspeaker *p*.

5.1.1 Analysis

Consider initially the case when the secondary loudspeaker p is absent from the array. The left- and right-ear signals e_a and e_b are expressed in terms of the HRTFs corresponding to the two adjacent nodal loudspeakers r and $r + 1$,

$$e_a = LS_r h_a(r) + LS_{r+1} h_a(r+1) \tag{26a}$$

$$e_b = LS_r h_b(r) + LS_{r+1} h_b(r+1)\,. \tag{26b}$$

When the secondary loudspeaker p is introduced into the array, the modified ear signals e'_a and e'_b become

$$\begin{aligned} e'_a = {} & LS_r h_a(r) + LS_{r+1} h_a(r+1) + \left\{LS_r \lambda_{p1} + LS_{r+1}\lambda_{p2}\right\}\gamma_{p1} h_a(r) \\ & + \left\{LS_r \lambda_{p1} + LS_{r+1}\lambda_{p2}\right\}\gamma_{p2} h_a(r+1) + \left\{LS_r \lambda_{p1} + LS_{r+1}\lambda_{p2}\right\} h_a(p) \end{aligned} \tag{27a}$$

$$\begin{aligned} e'_b = {} & LS_r h_b(r) + LS_{r+1} h_b(r+1) + \left\{LS_r \lambda_{p1} + LS_{r+1}\lambda_{p2}\right\}\gamma_{p1} h_b(r) \\ & + \left\{LS_r \lambda_{p1} + LS_{r+1}\lambda_{p2}\right\}\gamma_{p2} h_b(r+1) + \left\{LS_r \lambda_{p1} + LS_{r+1}\lambda_{p2}\right\} h_b(p)\,. \end{aligned} \tag{27b}$$

Forcing $e'_a = e_a$ and $e'_b = e_b$, then

$$\begin{bmatrix} h_a(r) & h_a(r+1) \\ h_b(r) & h_b(r+1) \end{bmatrix} \begin{bmatrix} \gamma_{p1} \\ \gamma_{p2} \end{bmatrix} = -\begin{bmatrix} h_a(p) \\ h_b(p) \end{bmatrix} \tag{28}$$

from which the correction functions γ_{p1} and γ_{p2} follow,

$$\gamma_{p1} = -\left[\frac{h_a(p)h_b(r+1) - h_b(p)h_a(r+1)}{h_a(r)h_b(r+1) - h_b(r)h_a(r+1)}\right] \tag{29a}$$

$$\gamma_{p2} = -\left[\frac{h_b(p)h_a(r) - h_a(p)h_b(r)}{h_a(r)h_b(r+1) - h_b(r)h_a(r+1)}\right]. \tag{29b}$$

These equations reveal that the correction functions γ_{p1} and γ_{p2} depend only on the HRTFs corresponding to the loudspeaker array. Consequently they are purely a function of the replay system, and its loudspeaker layout thus can be computed within the replay decoder as part of the installation procedure. However, the weighting functions γ_{p1} and γ_{p2} can be selected independently, provided the system is stable.

5.2 Relationship of System Parameters to Loudspeaker HRTFs

Section 5.1 presented an analysis of decoder parameters where precise HRTF measurement data are known for each loudspeaker position. However, there are some interesting observations and simplifications that can be made, which are considered in this section.

5.2.1 HRTFs Derived Using Linear Interpolation

Consider a pair of nodal loudspeakers within an n-array PWC system where the proximity of loudspeakers r and $r + 1$ is such that the image source HRTFs can be approximated by linear interpolation, such that

$$h_{xa} = \beta(m_r)\left\{m_r h_a(r) + (1 - m_r)\, h_a(r+1)\right\} \tag{30a}$$

$$h_{xb} = \beta(m_r)\left\{m_r h_b(r) + (1 - m_r)\, h_b(r+1)\right\} \tag{30b}$$

where m_r is the panning variable with a range of 1 to 0, which corresponds to an image pan from loudspeaker r to $r + 1$, and the function $\beta(m_r)$ is a moderator chosen to achieve a constant subjective loudness with variations in m_r. By substitution, the positional transfer functions G_r and G_{r+1} then simplify to

$$G_r = m_r \beta(m_r) \tag{31a}$$

$$G_{r+1} = (1 - m_r)\beta(m_r)\,. \tag{31b}$$

Consequently, for an image source that lies on a radial arc between the two nodal loudspeakers, simple amplitude panning yields the optimum panning algorithm. Effectively, HRTF coding information is derived directly from the loudspeaker locations and therefore is matched precisely to the listener. This assumes that intermediate HRTFs are derived by linear interpolation. It should also be noted that as the image source moves away from the radial arc containing the loudspeaker array, changes in HRTFs occur, making the positional transfer functions complex. Nevertheless, even when more exact HRTF data are available, there remains a strong desensitization to the exact form of the HRTFs when the positional transfer functions are calculated because of the relatively close proximity of loudspeakers in an n-array.

Consider next a secondary loudspeaker p that is added to the array, again assuming a linear interpolation model. It is assumed that the secondary loudspeaker is located midway along the same radial arc as the nodal loudspeakers and that its HRTFs $h_a(p)$ and $h_b(p)$ with respect to the listener can be determined by linear interpolation. Thus for the midpoint location,

$$h_a(p) = 0.5h_a(r) + 0.5h_a(r+1) \tag{32a}$$

$$h_b(p) = 0.5h_b(r) + 0.5h_b(r+1)\,. \tag{32b}$$

From these data the compensation gamma functions γ_{p1}

and γ_{p2} follow,

$$\gamma_{p1} = \gamma_{p2} = 0.5 \tag{33}$$

revealing once more a simple form for this special case. The modified positional transfer functions GM_r GM_{r+1} and GM_p for the respective nodal and secondary loudspeakers r, $r + 1$, and p are calculated as

$$GM_p = \lambda_{p1} G_r + \lambda_{p2} G_{r+1} \tag{34a}$$

$$GM_r = \left(1 + \gamma_{p1}\lambda_{p1}\right) G_r + \gamma_{p1}\lambda_{p2} G_{r+1} \tag{34b}$$

$$GM_{r+1} = \gamma_{p2}\lambda_{p1} G_r + G_{r+1}\left(1 + \gamma_{p2}\lambda_{p2}\right). \tag{34c}$$

Assuming symmetry, let $\lambda_{p1} = \lambda_{p2} = 0.5$ and consider the following examples:

1) $G_r = 1$ and $G_{r+1} = 0$, yielding $GM_r = 0.75$, $GM_{r+1} = -0.25$, and $GM_p = 0.5$.
2) $G_r = 0.5$ and $G_{r+1} = 0.5$, yielding $GM_r = 0.25$, $GM_{r+1} = 0.25$, and $GM_p = 0.5$.

For an image located coincident with the secondary loudspeaker there is a 6-dB level difference between secondary and nodal loudspeaker input signals. Also signal processing is simple and uses only real coefficients in the matrices.

5.2.2 Compensation and Incorporation of Exact HRTF Data

It is instructive to compare three other options for incorporating secondary loudspeaker and image HRTF coordinates. In each case correction functions λ_{p1} and λ_{p2} are calculated to match the selected HRTFs and the corresponding transfer functions for system input-to-loudspeaker inputs evaluated for loudspeakers r, p, and $r + 1$ located at 20°, 30°, and 40°, respectively, with an image at 30°. The four cases are as follows.

Case 1:

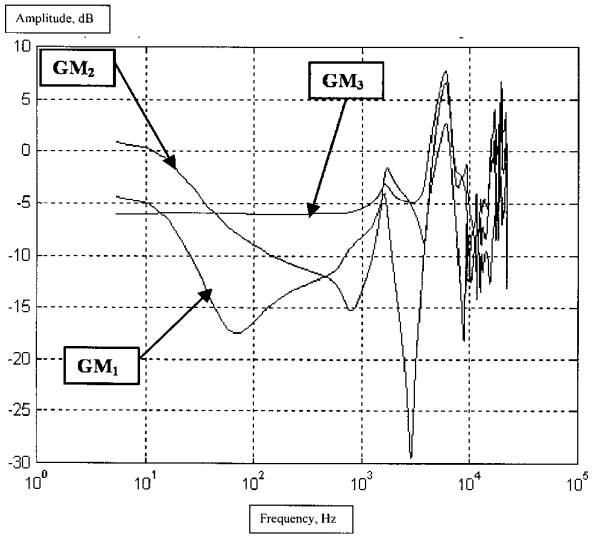
Image	Measured HRTF data
Secondary loudspeaker	Measured HRTF data
Results	See Fig. 17(a)

Case 2:

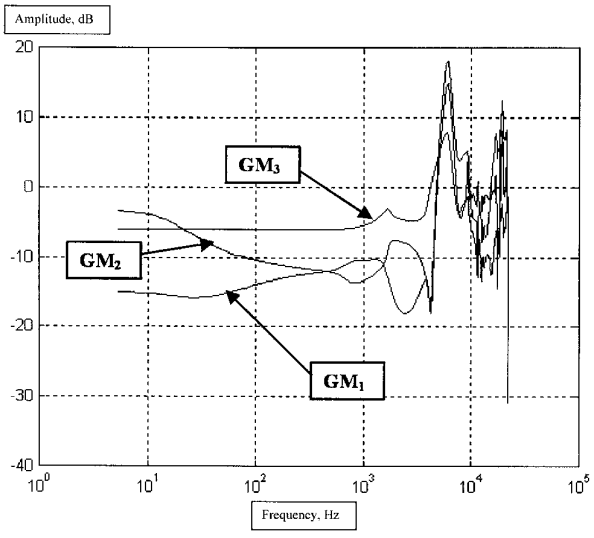
Image	Measured HRTF data
Secondary loudspeaker	HRTF derived by linear interpolation
Results	See Fig. 17(b)

Case 3:

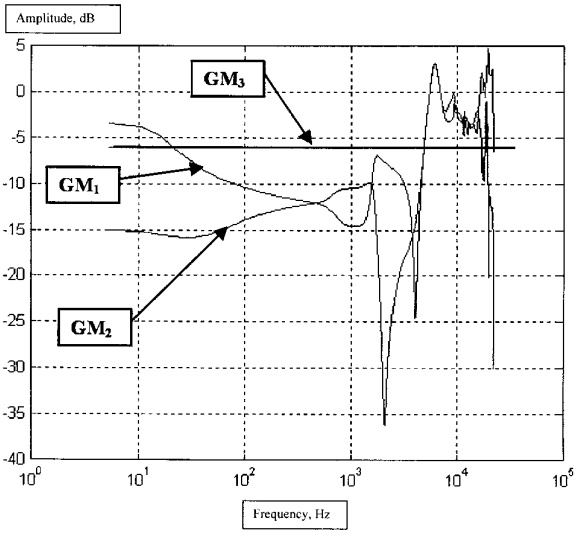
Image	HRTF derived by linear interpolation
Secondary loudspeaker	Measured HRTF data
Results	See Fig. 17(c)



(b)



(a)



(c)

Fig. 17. Transfer functions linking input to three-loudspeaker feeds at 20°, 30°, and 40°, image at 20°. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4.

Case 4:

Image	HRTF derived by linear interpolation
Secondary loudspeaker	HRTF derived by linear interpolation
Results	See Fig. 17(d) and discussion in Section 5.2.1

Fig. 17(c) shows that when the HRTF coordinates for the image are derived by linear interpolation, the center channel response is constant with frequency, even though the secondary loudspeaker HRTF coordinates are derived by measurement. Also, when both sets of coordinates are derived by linear interpolation, all three responses are constant, as demonstrated in Section 5.2.1. However, for cases 1 and 2 all responses vary with frequency, although case 2 maintains a greater high-frequency content in the center channel response, which is desirable for an image located coincident with the secondary loudspeaker.

5.2.3 Compensation for Encoder–Decoder Loudspeaker Displacement Error

The encoder assumes a nominal loudspeaker array location when determining the positional transfer functions. If the nodal loudspeakers are placed in the listening space in equivalent positions, then no additional processing is required at the decoder. However, in circumstances where the loudspeaker locations are displaced, positional compensation is required. It is important that positional compensation be independent of source coding and can be applied at the decoder without knowledge of the encoding algorithm. A pairwise positional correction scheme is shown in Fig. 18, where the compensation functions $GC_{11}(r)$, $GC_{12}(r)$, $GC_{11}(r+1)$, and $GC_{12}(r+2)$ are used to derive modified loudspeaker feeds. The form of this compensation is not unique, as correction signals can be applied to other loudspeakers in the array via appropriate filters. However, it is suggested that the loudspeaker selected to process the correction signal be the one closest to the loudspeaker that has been displaced. Hence by way of example, consider the scheme shown in Fig. 18. When the only active primary fed is LS_r, and equating the ear signals for both optimum and displaced loudspeaker locations, the positional compensation filters are as follows:

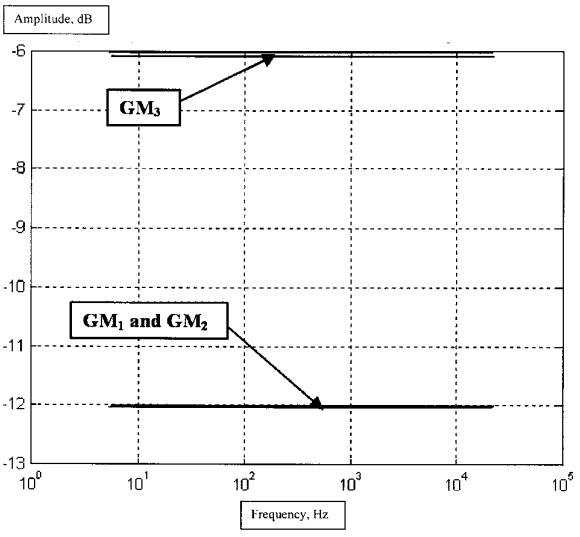
$$\begin{bmatrix} GC_{11}(r) \\ GC_{12}(r) \end{bmatrix} = \begin{bmatrix} h'_a(r) & h'_a(r+1) \\ h'_b(r) & h'_b(r+1) \end{bmatrix}^{-1} \begin{bmatrix} h_a(r) \\ h_b(r) \end{bmatrix}. \quad (35)$$

Similarly, when only LS_{r+1} is active, then

$$\begin{bmatrix} GC_{11}(r+1) \\ GC_{12}(r+1) \end{bmatrix} = \begin{bmatrix} h'_a(r+1) & h'_a(r) \\ h'_b(r+1) & h'_b(r) \end{bmatrix}^{-1} \begin{bmatrix} h_a(r+1) \\ h_b(r+1) \end{bmatrix}. \quad (36)$$

5.3 Standardization of HRTF Grid and Nodal Loudspeaker Locations

The techniques described in the preceding require knowledge of the HRTF coordinates for each nodal loudspeaker. Establishing a standardized encoding grid where each grid point is assigned nominal HRTF coordinates and where nodal loudspeakers are assigned nominal grid locations can satisfy this requirement. All encoder and decoder users then universally know this information. An example grid proposal, as shown in Fig. 19, is based on a 60° subtended angle for nodal loudspeakers with two additional



(d)

Fig. 17. *Continued*

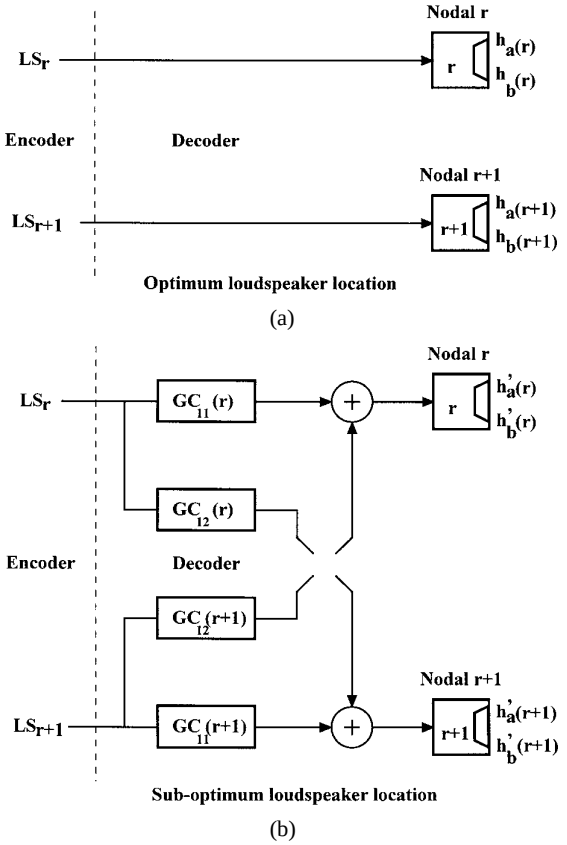


Fig. 18. Compensation process for nodal loudspeaker positional errors. (a) Optimum loudspeaker location. (b) Suboptimum loudspeaker location.

secondary loudspeakers. Three layers of nodes are suggested at radii of 1.5 m, 3 m, and 6 m. It is recognized that HRTFs are not unique, being listener specific, but when a multiloudspeaker array is formed using a set of HRTFs that are shared with image synthesis, errors are reduced.

HRTF coordinates that are noncoincident with the nodal points can be inferred by interpolation. For example, assume that an image X is located at the cylindrical coordinates $\{r, \theta\}$, where the four nearest nodes are $\{r_1, \theta_1\}$, $\{r_1, \theta_2\}$, $\{r_2, \theta_1\}$, and $\{r_2, \theta_2\}$. The interpolated HRTFs $h_a(r, \theta)$ and $h_b(r, \theta)$ are then

$$h_a(r,\theta) = m_r\left[m_\theta h_a(r_1, \theta_1) + (1 - m_\theta)h_a(r_1, \theta_2)\right] + (1 - m_r)\left[m_\theta h_a(r_2, \theta_1) + (1 - m_\theta)h_a(r_2, \theta_2)\right] \quad (37a)$$

$$h_b(r,\theta) = m_r\left[m_\theta h_b(r_1, \theta_1) + (1 - m_\theta)h_b(r_1, \theta_2)\right] + (1 - m_r)\left[m_\theta h_b(r_2, \theta_1) + (1 - m_\theta)h_b(r_2, \theta_2)\right] \quad (37b)$$

where m_θ and m_r are the angular and radial linear interpolation parameters defining the image X. For images that lie either within the inner radius or beyond the outer radius, angular interpolation is performed first, followed by an appropriate adjustment to the amplitude and time delays based on the radial distance from the head.

6 PERCEPTUALLY BASED CODING EXPLOITING EMBEDDED CODE IN PRIMARY SIGNALS TO ENHANCE SPATIAL RESOLUTION

The techniques described in Section 5 can be extended to a system with any number of nodal and secondary loudspeakers and thus can be matched to a wide variety of multichannel system configurations. However, inevitably there is a limit to spatial resolution arising from the use of matrixing only, which imposes crosstalk between nodal and secondary loudspeaker feeds. Some advantage may be gained by using nonlinear decoding with dynamic parameterization, although for high-resolution music reproduction linear decoding should be retained.

Because DVD-A is capable of six channels at 24 bit 96 kHz, some of the lower bits in the LPCM stream can be sacrificed [12] while still retaining an exemplary dynamic range by using standard methods of psychoacoustically motivated noise shaping and equalization [16]. The least significant bits in the LPCM streams together with a randomization function can then be used to encode additional audio channels using perceptual coders such as AC-3,[9] DTS,[10] or MPEG.[11] For example, 4 bit per sample per LPCM channel

[9] Proprietary perceptual coding developed by Dolby Laboratories.
[10] Digital Theatre Systems.

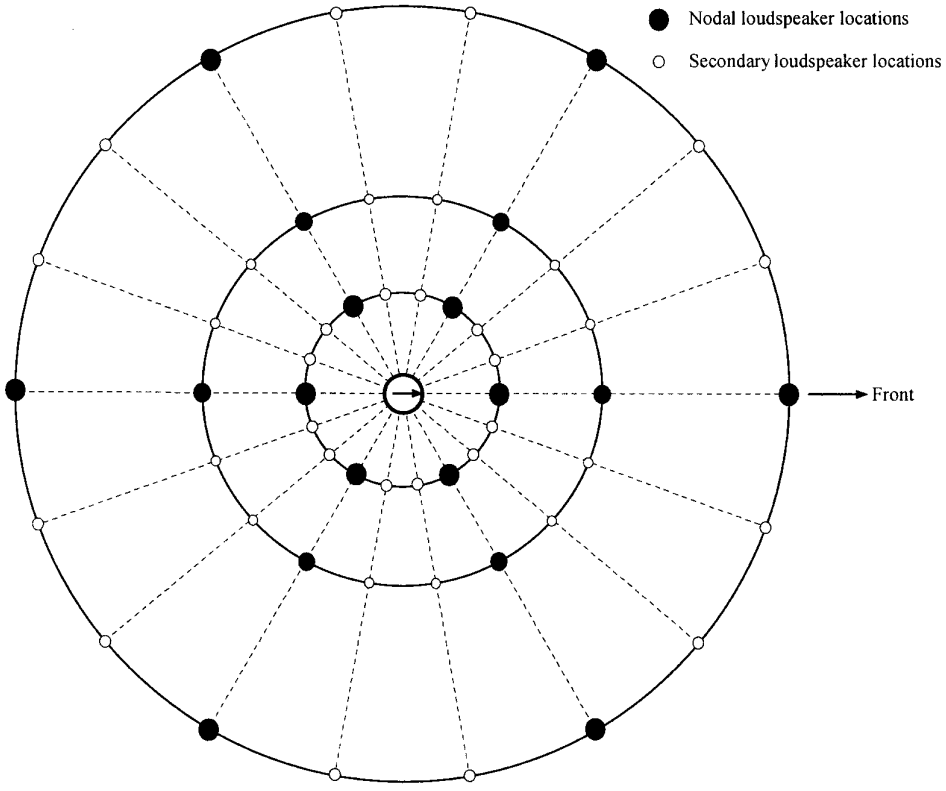


Fig. 19. Proposed 18-segment constellation map to define a standard set of HRTFs.

at 96 kHz yields a serial bit rate of 384 kbit/s.

The proposal retains the primary signals in high-resolution LPCM and uses the matrix methods described in Section 5 to estimate the secondary loudspeaker feeds. Spatially related difference signals are then calculated from the discrete secondary loudspeaker signals available at the encoder and the matrix-derived signals. Also, because of close spatial clustering of the additional channels there is a high degree of interchannel correlation with both the primary and the secondary loudspeaker signals, a factor that bodes well for accurate perceptual coding. Close clustering also implies that perceptual coding errors are not widely dispersed in space, yielding an improved masking performance. Given that DVD-A already supports six LPCM channels, it is suggested that an extra two encoded signals per primary signal is a realistic compromise, yielding a total of 18 channels, as proposed in the standardized HRTF constellation illustrated in Fig. 18. In the grand plan there would be *n* perceptual coders in operation, one per nodal loudspeaker feed. In such a scheme further gains are possible by integrating dynamic bit allocation across all coders as well as using a perceptual model designed specifically for multichannel stereo encoding. In difficult encoding situations dynamic spatial blending can be used to reduce the difference signals prior to perceptual encryption. Fig. 20 shows the basic encoder architecture where the error signals D_1 and D_2 are indicated. In Fig. 21 a decoder is shown where identical estimates are made of the secondary loudspeaker input signals, but with the addition of the difference signals to yield discrete loudspeaker feeds. Of course, if the embedded perceptually coded difference signals are not used, estimates can still be made for the secondary loudspeakers as shown in Fig. 16. Alternatively, for a basic scheme an array of nodal loudspeakers only can be used.

[11] Perceptual based audio coding proposed by the Motion Picture Expert Group.

[12] Registered trademark of Company name, New Transducers plc, UK.

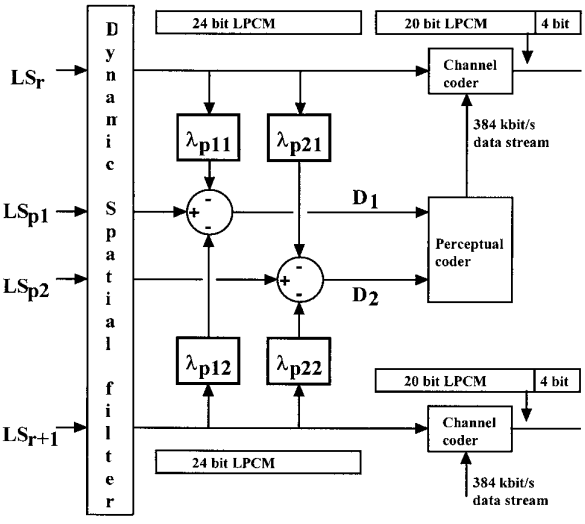


Fig. 20. High-level processor architecture for encoding discrete secondary loudspeaker feeds.

7 CONCLUSIONS

This paper has presented a method for multichannel audio that is fully compatible with DVD (DVD-A and SACD) multichannel formats, which have the capability of six high-resolution signals. The key to this technology is the exploitation of spatial coding using HRTF data to enhance the positional representation of sound sources. As such it forms a link between two-channel transaural techniques and conventional multichannel audio using many loudspeakers. This technique has already been demonstrated in telepresence and teleconferencing applications [9], [10] to be effective in representing spatial audio. However, the methods described here show how a particular loudspeaker array can be configured where issues of positional calibration were discussed for loudspeakers displaced from those locations assumed during coding.

The method is scalable and fully backward compatible. In a simple system there is no additional processing at the decoder where, for example, the outputs of a DVD player are routed directly to an array of loudspeakers. However, if additional loudspeakers are used, as might be envisaged with tiled walls of flat-panel NXT[12] loudspeakers (see, for example, Fig. 22), then formal methods exist, enabling the correct ear signals at the listening position to be maintained. Also for DVD-A, a method was suggested where perceptually coded information is embedded within the LPCM code to enable discrete loudspeaker signals to be derived. It was proposed that an upper limit of 18 channels should be accommodated, although full compatibility with systems down to the basic array is maintained.

An interesting observation for systems using a large number of closely spaced loudspeakers is that image positioning on the arc of the array can use simple linear amplitude panning applied between pairs of adjacent loud-

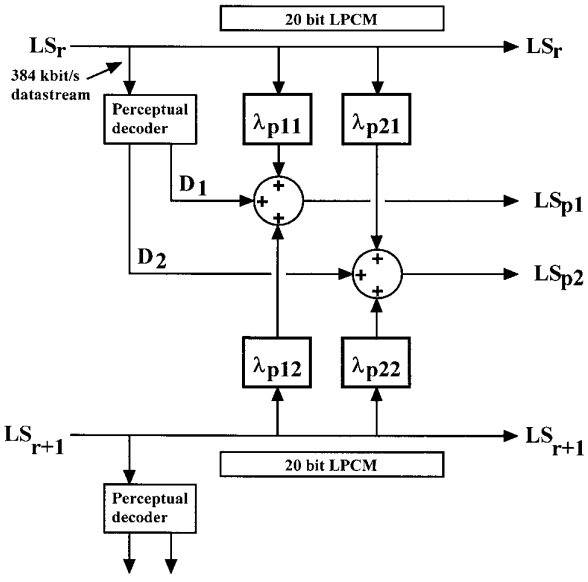


Fig. 21. High-level decoder architecture to derive discrete secondary loudspeaker feeds.

speakers. This approximation assumes that the image HRTF coordinates can be estimated to sufficient accuracy using linear interpolation between adjacent loudspeaker HRTF coordinates. This expedience effectively embeds HRTF data matched exactly to the listener simply because of the physical location of the loudspeakers. However, as the loudspeaker spacing increases, this approximation fails, requiring then the use of more accurate HRTF image coordinates together with transaural PWC, as described. This is particularly important where an image is located away from the arc of the loudspeaker array and where reflections are to be rendered to craft a more accurate virtual acoustic.

This work is also targeted at new communication formats for virtual reality, telepresence, and video conferencing [17], [18], where future research should investigate its application. Such schemes are not constrained by the normal paradigms of multichannel stereophonic reproduction, nor is compatibility necessarily sought. The approach is to form an optimum methodology for constructing phantom images and to consider coding paradigms appropriate for communication. For example, one possible communication format assigns a discrete channel to each phatnom image. The channel then conveys the auditory signals together with the spatial coordinates updated at a rate compatible with motion tracking of the sound source. At the receiver, a processor carries a downloaded program with knowledge of the positional data and source acoustics from which the required reflections and reverberation are computed. These data would then be formatted to match the selected loudspeaker array. Such a scheme has great flexibility and can allow many mono sources to contribute to the final soundscape.

In conclusion, the techniques presented describe a means by which spatial resolution and image coding performance can transcend the six-channel limitation of the current DVD formats, yet without requiring additional storage capacity. Also, by basing signal processing on a perceptual model of hearing, it is revealed how sound images can be rendered and, in particular, how interaural amplitude differences and interaural time differences can be accommodated without seeking tradeoffs between time and amplitude clues. Essentially the work has presented a scalable and reverse compatible solution to multichannel audio that is particularly well matched to an LPCM format on DVD-A.

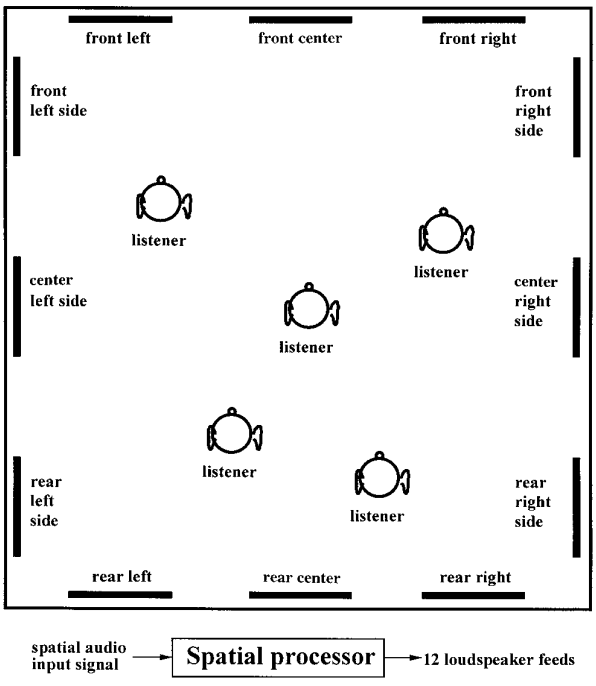


Fig. 22. Multichannel configuration using array of wall-mounted flat-panel diffuse loudspeakers.

8 REFERENCES

[1] *HFN/RR*, "Digital Frontiers," vol. 40, pp. 58–59, 106 (1995 Feb.).

[2] M. O. J. Hawksford, "High-Definition Digital Audio in 3-Dimensional Sound Reproduction," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 1016 (1997 Nov.), preprint 4560.

[3] A. J. Berkout, D. de Vries, and P. Vogel, "Acoustic Control by Wavefield Synthesis," *J. Acoust. Soc. Am.*, vol. 93, pp. 2764–2778 (1993).

[4] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (AP Professional, 1994).

[5] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10 (1973 Jan./Feb.).

[6] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J. Audio Eng. Soc.*, vol. 33, pp. 859–871 (1985 Nov.).

[7] M. A. Gerzon, "Hierarchical Transmission Systems for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 692–705 (1992 Sept.).

[8] K. C. K. Foo and M. O. J. Hawksford, "HRTF Sensitivity Analysis for Three-Dimensional Spatial Audio Using the Pairwise Loudspeaker Association Paradigm," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 1018 (1997 Nov.), preprint 4572.

[9] K. C. K. Foo, M. O. J. Hawksford, and M. P. Hollier, "Three-Dimensional Sound Localization with Multiple Loudspeakers Using a Pairwise Association Paradigm and Embedded HRTFs," presented at the 104th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 46, p. 572 (1998 June), preprint 4745.

[10] K. C. K. Foo, M. O. J. Hawksford, and M. P. Hollier, "Pairwise Loudspeaker Paradigms for Multichannel Audio in Home Theatre and Virtual Reality," presented at the 105th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 46, p. 1035 (1998 Nov.), preprint 4796.

[11] M. Gerzon, "Practical Periphony: The Reproduction of Full-Sphere Sound," presented at the 65th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 28, p. 364 (1980 May), preprint 1571.

[12] M. A. Gerzon and P. G. Craven, "A High-Rate

Buried Data Channel for Audio CD," *J. Audio Eng. Soc.*, vol. 43, pp. 3–22 (1995 Jan./Feb.).

[13] J. Blauert, *Spatial Hearing*, rev. ed. (MIT Press, Cambridge, MA, 1997).

[14] J. Bauck and D. H. Cooper, "Generalized Transaural Stereo and Applications," *J. Audio Eng. Soc.*, vol. 44, pp. 683–705 (1996 Sept.).

[15] M. A. Gerzon, "Optimum Reproduction Matrices for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 571–589 (1992 July/Aug.).

[16] J. R. Stuart and R. J. Wilson, "Dynamic Range Enhancement Using Noise-Shaped Dither Applied to Signals with and without Preemphasis," presented at the 96th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 400 (1994 May), preprint 3871.

[17] "Telepresence Theme," *BT Technol. J.*, vol. 15 (1997 Oct.).

[18] D. M. Burraston, M. P. Hollier, and M. O. J. Hawksford, "Limitations of Dynamically Controlling the Listening Position in a 3-D Ambisonic Environment," presented at the 102nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 45, p. 413 (1997 May), preprint 4460.

THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and investigated delta modulation and sigma–delta modulation (SDM, now known as bitstream coding) for color television and produced a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems with a strong emphasis on audio systems including loudspeaker technology. Since 1982, research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. A first in 1986 was for a prototype system to be demonstrated at the Canon Research Centre in Tokyo, work sponsored by a research contract from Canon. Much of this work has appeared in the *JAES*, together with a substantial number of contributions at AES conventions.

His research has also encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with a special emphasis on SDM. Other research has included the linearization of PWM encoders, diffuse loudspeaker technology, and three-dimensional spatial audio and telepresence including multichannel sound reproduction.

Dr. Hawksford is a recipient of the 1997/1998 AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover." He is a chartered engineer as well as a fellow of the AES, IEE, and IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA). He is also a technical consultant for NXT, UK and LFD Audio, UK.

6 Measurement systems

6-1 MLS systems

6-1 DISTORTION IMMUNITY OF MLS-DERIVED IMPULSE RESPONSE MEASUREMENTS, Dunn, C. and Hawksford, M.O.J., *JAES*, vol. 41, no.5, pp 314-335, May 1993

6-23 DISTORTION ANALYSIS OF NON-LINEAR SYSTEMS WITH MEMORY USING MAXIMUM LENGTH SEQUENCES, Greest, M. and Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 142, no 5, Oct 1995, pp 345-352

6-2 Volterra modeling

6-29 IDENTIFICATION OF DISCRETE VOLTERRA SERIES USING MAXIMUM LENGTH SEQUENCES, Reed, M.J. and Hawksford, M.O.J., IEE Proceedings on Circuits, Devices and Systems, vol. 143, no 5, Oct. 1996, pp 241-248

6-37 EFFICIENT IMPLEMENTATION OF THE VOLTERRA FILTER, Reed, M.J. and Hawksford, M.O.J., IEE Proc.-VIS. Image Signal Processing, Vol. 147, No. 2, pp 109-114, April 2000

6-43 SYSTEM MEASUREMENT AND IDENTIFICATION USING PSEUDORANDOM FILTERED NOISE AND MUSIC SEQUENCES, Hawksford, M.O.J. *JAES*, vol. 52, no. 4, pp. 275-296, April 2005

Distortion Immunity of MLS-Derived Impulse Response Measurements*

CHRIS DUNN, *Student Member,* **AND MALCOLM OMAR HAWKSFORD,** *AES Fellow*

Audio Research Group, Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, UK

Maximum-length-sequence (MLS) measurement of system impulse responses offers a potential enhancement in error immunity over periodic impulse testing, although care must be exercised in setting the MLS excitation amplitude in order to realize this potential. The effects of nonlinearity in MLS measurement are studied, in particular the way in which impulse response errors due to nonlinearity are distributed across the measurement period. The consequences of such errors in cumulative spectral display plots are also investigated. Finally inverse-repeat sequences (IRS) are shown to have complete immunity to even-order nonlinearity while maintaining many of the advantages of MLS.

0 INTRODUCTION

Perhaps the most fundamental evaluation of an audio system is the determination of the linear transfer function, defined by the impulse response (IR) in the time domain from which the frequency response can be calculated. Some applications require highly accurate linear transfer function measurement, for example, equalization of loudspeakers in the digital domain where measurement accuracy must match equalization performance (that is, better than ± 0.5 dB across wide regions of the audio spectrum; see, for example, [1]). Another application that requires highly accurate linear transfer function measurement is a technique proposed by the authors to measure low-level errors within audio systems [2].

There are three established methods of linear transfer function measurement—periodic impulse excitation (PIE), maximum-length sequences (MLS), and time-delay spectrometry (TDS). PIE reveals the periodic impulse response (PIR) of the device under test (DUT) directly by applying a periodic short-duration impulse to the DUT and measuring the output [3]. The main problem encountered in PIE is poor noise immunity due to low excitation signal energy; this drawback can be overcome to some degree by averaging several measurements. Alternatively an MLS can be used, which, compared to a periodic impulse of similar repetition rate, has a much higher excitation energy for the same peak output (that is, a lower crest factor). An MLS is a pseudorandom binary sequence which yields a unit impulse upon circular autocorrelation, and this property allows the PIR of a test system to be obtained by applying an MLS to the DUT and cross-correlating the system output with the input. An excellent introduction to MLS techniques is provided by Rife and Vanderkooy [4]. PIE and MLS initially reveal the transfer function in the time domain while TDS methods yield transfer function information as a complex frequency response (which can of course be converted to the time domain by using the inverse Fourier transform). TDS techniques utilize swept sine waves, or "chirps," which are input to the DUT and recover the complex frequency response of the test system after hardware or software processing [5]. Both TDS and MLS offer an increase in noise and distortion immunity over PIE. However, it can be shown that in achieving a similar frequency resolution to that available from MLS, a typical TDS implementation will take considerably longer to execute [4]–[6]. The additional disadvantage that TDS suffers in terms of hardware

* Manuscript received 1992 January 3; revised 1992 November 9 and 1993 March 1.

and software complexity [7] also helps to explain the growing popularity of MLS [8], [9].

Noise and distortion present in any practical measurement environment reduce the accuracy of linear transfer function measurement. A simulated example of the effects that nonlinearity can have on impulse measurement is shown in Fig. 1. The true magnitude response in the frequency domain [Fig. 1 (a)] of a 1-kHz low-pass finite-impulse-response (FIR) filter with less than 0.001 dB passband ripple can be compared against the MLS-derived magnitude response of the same filter in Fig. 1(b), where the measurement has been corrupted by gross second-order nonlinearity. Clearly the distortion has caused the recovered transfer function to appear much more ragged over the filter's passband than its linear specification would suggest.

This paper investigates by simulation the effects that nonlinearity can have on MLS measurement. In particular we examine the way in which errors due to nonlinearity are spread across the period of the recovered impulse response, and the consequences of such distributions for increasing distortion immunity by truncating the impulse response. A comparison of overall (noise *and* distortion) error immunity between MLS and PIE techniques is followed by an assessment of the suitability of inverse repeat sequences (IRS) for linear transfer function measurement.

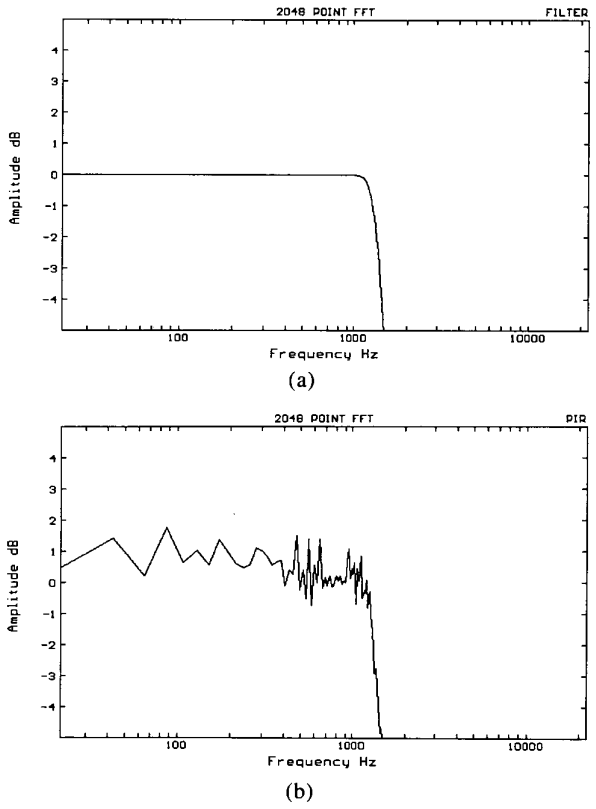


Fig. 1. (a) Magnitude response of 1-kHz low-pass FIR filter. (b) MLS-derived magnitude response of same filter where measurement has been corrupted by gross second-order nonlinearity.

1 DETERMINING DISTORTION IMMUNITY BY SIMULATION

Any system with weak (that is, non-overloading) nonlinearity can be modeled in the frequency domain by the nonlinear transfer function shown in Fig. 2(a) [4]. This includes a linear stage $H(f)$ and nonlinear stages $H(f_1, f_2)$, $H(f_1, f_2, f_3)$, etc., which represent different distortion orders. When a multitone signal is input to the model, harmonic and intermodulation error products will corrupt the output signal. Nonlinearity can also be represented in the time domain as a distributed model [Fig. 2(b)], where distortion polynomials *dm* appear in parallel with linear filters *hm*. Although the distributed time-domain model is capable of modeling complex nonlinearities, we require a simpler model for simulation. The nonlinear model used throughout this paper is similar to that used by Rife and Vanderkooy in a previous study of distortion in MLS measurements [4], consisting of a filter $h(n)$ followed by a nonlinearity $d\{\ \}$ [Fig. 2(c)]. For most of the simulations presented in this paper the nonlinearity is memoryless, that is, the error sequence output from the nonlinear stage depends solely on its instantaneous input, and the char-

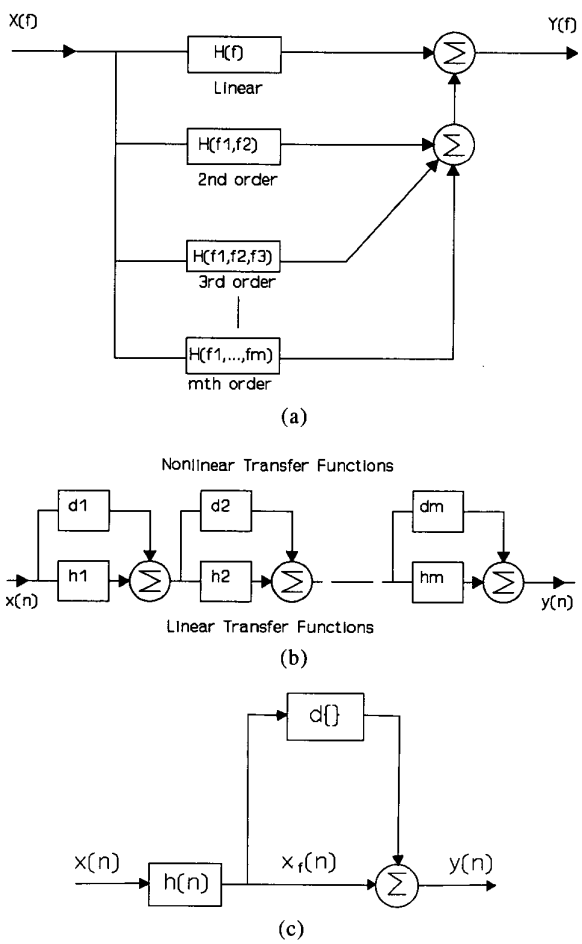


Fig. 2. Nonlinear system modeling. (a) Frequency-domain model. (b) Distributed time-domain model. (c) Lumped time-domain model.

acteristics of the nonlinearity are thus independent of frequency. Despite the simplicity of this model, it is a fairly good representation of many physical nonlinear processes. For example, a signal applied to a loudspeaker will typically pass through an essentially linear crossover filter before reaching the weakly nonlinear drive unit; moreover, the drive unit itself will tend to limit at its "output" as excursion limits are approached. Furthermore a practical MLS measurement system will use some low-pass filtering before the MLS is applied to the DUT in order to minimize slew-related artifacts in the recovered impulse response [10], [11]. In fact, prefiltering the MLS or PIE excitation is a necessary condition that must be met if the nonlinearity is to change the shape of the recovered impulse response. If unfiltered MLS or PIE signals are input to a memoryless nonlinearity, then the binary excitation will merely result in a dc offset error in the recovered impulse response for even-order nonlinearity or a pure gain change error for odd-order nonlinearity.

We will now describe the general simulation process used to determine distortion immunity in impulse response measurements. A periodic driving signal $x(n)$ (either a PIE or an MLS) is convolved with a known linear impulse response $h(n)$ over a measurement period L of 2047 samples. Unless otherwise indicated, $h(n)$ is a 1-kHz low-pass FIR filter, the first 256 samples of which are plotted in the time domain in Fig. 3(a). Noting that the sampling frequency is set to 44.1 kHz, the frequency-domain magnitude response of the filter is illustrated in Fig. 3(b). If $h(n)$ is nonzero for a time less than the period of the driving signal, then time aliasing in the convolution operation is avoided and just one period is required in the simulations to describe the periodic system behavior accurately. The filtered driving signal $x_f(n)$ is distorted by a known polynomial $d\{\ \}$, and after appropriate postprocessing has extracted the distorted impulse response $h_0(n)$, the error component $e(n)$ can be calculated by subtracting the known $h(n)$ from $h_0(n)$. The simulation process therefore allows us to examine the impulse error sequence $e(n)$ associated with a particular nonlinearity $d\{\ \}$. Summarizing the general simulation procedure,

$$\begin{aligned} x_f(n) &= x(n) \otimes h(n) \\ y(n) &= x_f(n) + d\{x_f(n)\} \\ h_0(n) &= P[y(n)] \\ e(n) &= h_0(n) - h(n) \end{aligned} \tag{1}$$

where

$\otimes$ = convolution
$y(n)$ = distorted output driving signal
$P[\ ]$ = postprocessing operation required to yield impulse response from output driving signal

In general a memoryless rth-order nonlinearity $d\{\ \}$ can be written

$$d\{x_f(n)\} = A_d \left[\frac{x_f(n)}{x_{ref}}\right]^r \tag{2}$$

where A_d sets the amplitude of the nonlinearity and x_{ref} is a reference scaling level. We must set x_{ref} such that a valid comparison can be made between PIE and MLS error immunity. The obvious choice is to adopt an "absolute scaling" where $x_{ref} = 1$, and make the peak levels of the unfiltered PIE and MLS signals equal (so that the noise immunity advantage of MLS is known [4]). Thus a third-order nonlinearity at -20 dB would be written

$$d\{x_f(n)\} = 0.1\ [x_f(n)]^3\ . \tag{3}$$

All of the variables used in the simulations are represented by double-precision floating-point numbers which offer a relative error due to mantissa quantization of 2^{-53} (-319 dB).

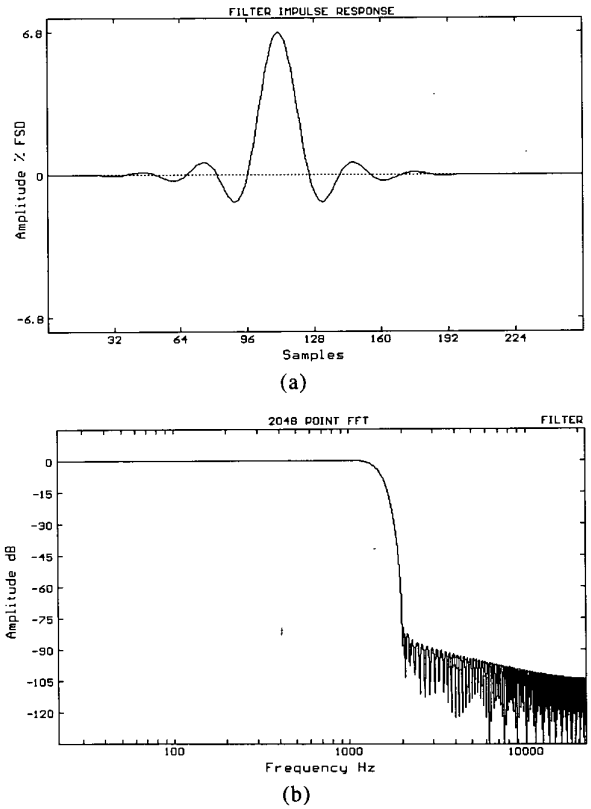


Fig. 3. 1-kHz FIR low-pass filter. (a) Time-domain impulse response $h(n)$. (b) Frequency-domain magnitude response.

2 PIE DISTORTION IMMUNITY

Periodic impulse testing reveals the periodic impulse response of the system under investigation directly by applying a periodic impulse to the DUT and sampling the output signal. Physically this excitation can be ob-

tained using a pulse generator or one of the many test compact discs with a periodic impulse signal track. No postprocessing upon the measured system output signal is required (although averaging several impulse periods can improve random noise immunity; see Section 4), and the PIE simulation process can be summarized as

$$\begin{aligned} x(n) &= \delta(n) \\ x_f(n) &= h(n) \\ P[y(n)] &= y(n) \\ e(n) &= d\{x_f(n)\} \ . \end{aligned} \tag{4}$$

Referring to Eq. (2), we now examine the consequences of second-order nonlinearity by setting $r = 2$ and the distortion level $A_d = 0.1$ (−20 dB). The first 256 samples of the error sequence $e(n)$ are plotted in the time domain in Fig. 4(a), and can be seen to contain a large peak coincident with the linear impulse response. (The amplitude scaling in the time-domain plots is relative to the peak level of the unfiltered driving signal.) In general the error due to nonlinearity will contain a "linear" component $e_l(n)$ identical in shape to the linear impulse response of the system, and also a nonlinear part $e_{nl}(n)$. It is the nonlinear component $e_{nl}(n)$ of the impulse error which causes the raggedness seen in the magnitude response of the example presented in Fig. 1. Conversely the linear component $e_l(n)$ represents a gain change in the measurement, that is, the observed gain of the system under examination has changed due to nonlinearity. With some applications this gain change is important (see, for example, [2]), but in many situations such as loudspeaker testing we are interested in the *relative* linear transfer function rather than the absolute gain of the DUT. Hence in our study of distortion immunity it is useful to be able to distinguish between linear and nonlinear errors in the impulse response measurement. We can perform such an "error normalization" by subtracting a scaled version of the linear impulse response $h(n)$ from the overall error sequence $e(n)$,

$$e_{nl}(n) = e(n) - gh(n) \ . \tag{5}$$

It is easy to show that $e_{nl}(n)$ is minimized in an rms sense by setting the gain error g to

$$g = \frac{\sum_{k=0}^{L-1} e(n)h(n)}{\sum_{k=0}^{L-1} h(n)^2} \ . \tag{6}$$

This analysis of gain change due to nonlinearity is similar to the study undertaken by Vanderkooy [11]. Fig. 5 shows the nonlinear component of the overall error given in Fig. 4.

The distortion immunity I_d of the impulse measurement is calculated as the ratio of linear impulse response energy to nonlinear error energy,

$$I_d = 10 \log_{10} \left[\frac{\sum_{k=0}^{L-1} h(n)^2}{\sum_{k=0}^{L-1} e_{nl}(n)^2} \right] \ . \tag{7}$$

Simulations were performed for second- to seventh-order nonlinearity with the 1-kHz low-pass FIR filter and distortion immunity tabulated in Table 1. The accuracy of the PIE simulations is extremely high because:

1) The error sequence can be calculated directly [Eq. (4)], rather than by subtracting the $h_0(n)$ sequence from $h(n)$, as described in Section 1, Eq. (1), and

2) No postprocessing operation is required.

Direct calculation of $e_{nl}(n)$ allows simulated distortion immunity to exceed the limits that would otherwise be present due to mantissa quantization in the floating-point variables. Theoretically the maximum distortion immunity that can be recorded using this simulation technique is bounded by the *range* of the simulation variables, which is on the order of 6000 dB for double-precision floating-point numbers. The PIE distortion immunity results collated in Table 1 are well below this limit. The relative error of $e_{nl}(n)$ (that is, the error of the impulse error sequence) is limited by quantization effects in the floating-point variables at approximately −300 dB.

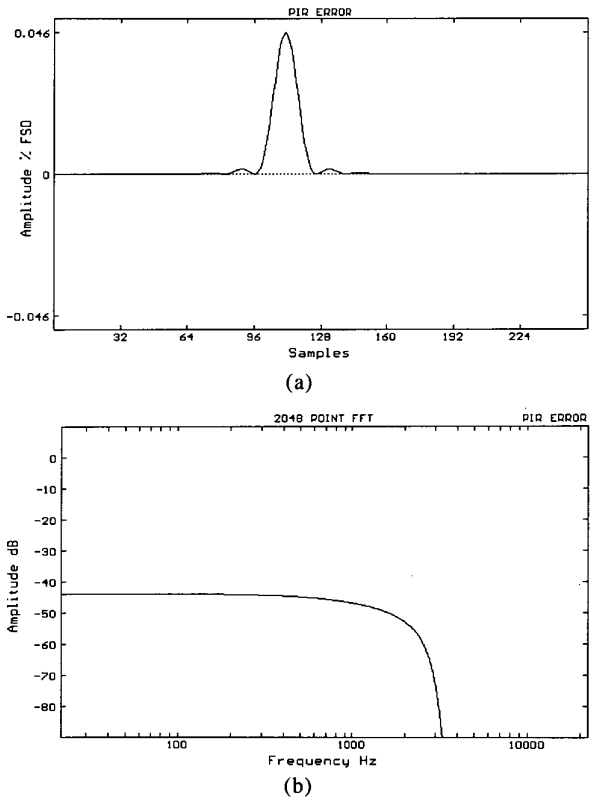


Fig. 4. PIE-derived error due to second-order nonlinearity. (a) Time domain $e(n)$. (b) Frequency domain.

3 MLS DISTORTION IMMUNITY

3.1 Review of MLS Measurement Techniques

MLS are pseudorandom binary signals that can be generated from digital shift registers with appropriate EXCLUSIVE-OR feedback structures. If an MLS $s(n)$ is generated from an mth-order shift register (that is, one with m stages), then all shift register states bar one (all 0s) are included in each MLS period of length L. Thus $L = 2^m - 1$ samples. Fig. 6(a) shows a fifth-order MLS where the (1, 0) shift register logic output has been transformed to (−1, 1) scaled voltages and a zero-order hold is used between samples. If we follow the convention adopted by Rife and Vanderkooy [4] of scaling autocorrelation and cross-correlation operations by $1/(L + 1)$ rather than the usual $1/L$, then the first-order circular autocorrelation Ω_1 of an MLS is a unit impulse with a dc offset,

$$\Omega_1(n) = s(n) \,\Phi\, s(n) = \frac{1}{L+1}\sum_{k=0}^{L-1} s(k)s(k+n) = \delta(n) - \frac{1}{L+1}, \quad 0 \leq n < L \tag{8}$$

where Φ represents circular cross correlation and all indices within the summation are calculated mod L. Fig. 6(b) shows the autocorrelation of the fifth-order sequence shown in Fig. 6(a). When an unfiltered MLS is applied to a linear system with impulse response $h(n)$ of length less than the period of the MLS, then cross-correlating the input and output of the system recovers the ac component of $h(n)$ together with an attenuated dc component [4],

$$s(n) \,\Phi\, y(n) = \frac{1}{L+1}\sum_{k=0}^{L-1} s(k)y(k+n) = \left[h(n) - \frac{1}{L}\sum_{k=0}^{L-1} h(k)\right] + \left[\frac{1}{L(L+1)}\sum_{k=0}^{L-1} h(k)\right] = h_{AC}(n) + \frac{1}{L+1}h_{DC}(n)\,. \tag{9}$$

In many applications such as loudspeaker testing the DUT will be ac coupled, and so the dc component will essentially be equal to zero. A practical MLS mea-

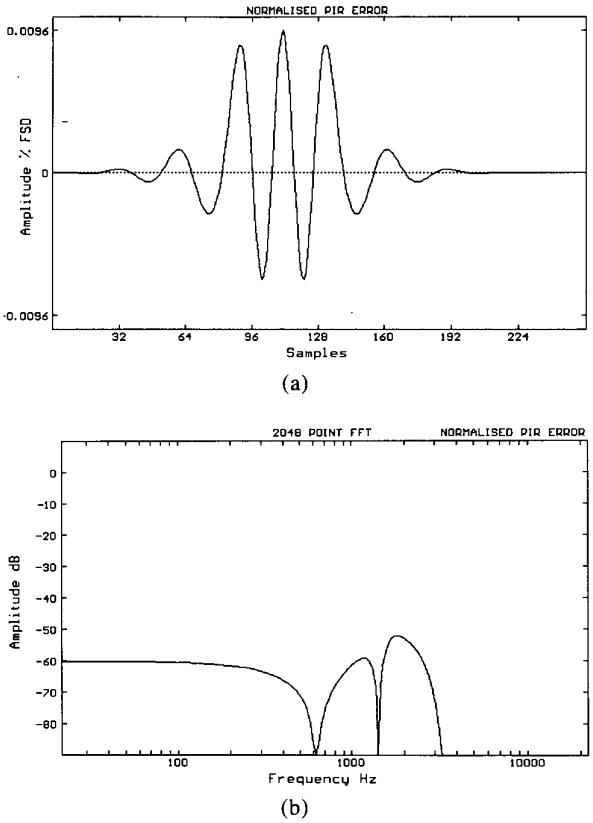


Fig. 5. PIE-derived normalized nonlinear error due to second-order nonlinearity. (a) Time domain $e_{nl}(n)$. (b) Frequency domain.

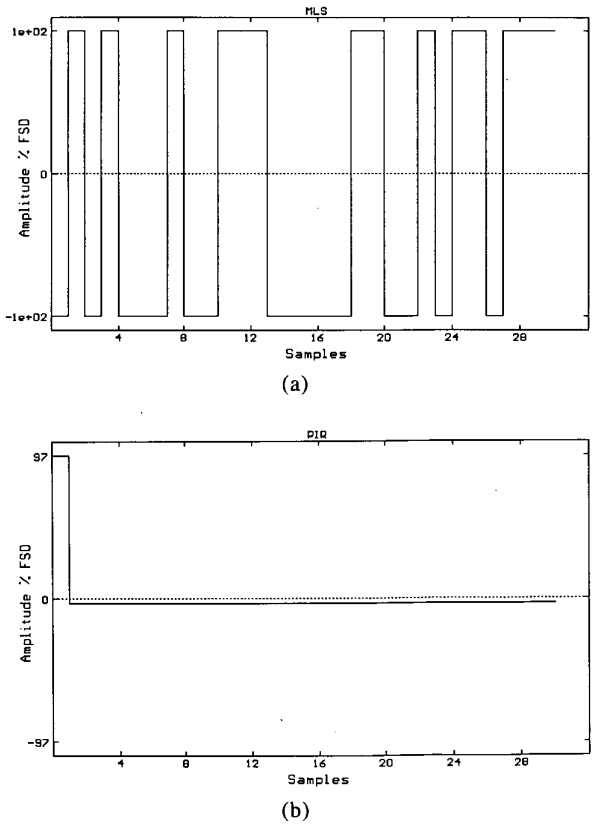


Fig. 6. (a) Unfiltered 31-point MLS $s(n)$ generated from fifth-order shift register with zero-order hold. (b) Autocorrelation $\Omega_1(n)$ of (a).

surement system based around a personal computer would generate the MLS excitation using a shift register in either software or hardware. The MLS is applied to the input of the DUT and the output signal sampled using an analog-to-digital converter over at least one measurement period after the system has settled to steady-state operation. The measured output sequence is then cross-correlated with the known input sequence to reveal the impulse response of the system. There are several methods of performing the cross correlation, the most efficient of which is by fast Hadamard transform (FHT), where an L-point cross correlation can be performed with only $2.5L \log_2 L$ floating-point additions; see Borish and Angell [12] and Borish [13] for details.

3.2 MLS Distortion Immunity

We analyze MLS distortion immunity using an eleventh-order sequence with EXCLUSIVE-OR feedback taps from the second and eleventh shift register stages [13]. This results in a commonly used sequence period of 2047 samples. The MLS signal is convolved with the linear impulse response of the simulated system, again the 1-kHz low-pass FIR filter shown in Fig. 3. The MLS convolution, like an MLS cross correlation, is most efficiently performed by FHT [14]. Distorting the convolved driving signal will corrupt the impulse response obtained from the measurement, but the nature of the distortion is different from that of a PIE-derived measurement because of the postprocessing cross-correlation operation,

$$P[y(n)] = s(n) \,\Phi\, y(n)$$

$$= \frac{1}{L+1} \sum_{k=0}^{L-1} s(k)y(n+k) \,. \qquad (10)$$

The effect that nonlinear distortion has on MLS-derived impulse responses is shown in Fig. 7. The 1-kHz FIR low-pass filter whose full impulse response $h(n)$ is shown in Fig. 7(a) is convolved with the 2047-point MLS, resulting in the filtered MLS signal in Fig. 7(b). The filtered MLS signal is then distorted by second-order nonlinearity and the result cross-correlated with the unfiltered MLS to recover the corrupted impulse response $h_0(n)$ in Fig. 7(c). Although for most of the simulations in this paper we have adopted a distortion amplitude $A_d = -20$ dB, in this example we have set $A_d = -10$ dB in order to clearly show the nature of the artifacts in the recovered impulse response. (The amplitude of the nonlinearity does not change the shape of the impulse error, just its amplitude relative to the linear impulse response.) Immediately evident is the "spiky" or "lumpy" nature of the error in the tail of the impulse response, an observation which has been noted many times [10], [15], [16], [4], [17], [11].

Because cross correlation is a distributive process, the impulse error $e(n)$ can be calculated by cross-correlating $s(n)$ with the driving sequence distortion $d\{x_f(n)\}$. Again, this has the benefit of increasing simulation accuracy since calculating $e(n)$ does not now involve a subtraction from $h(n)$. The MLS simulation summary is thus

$$x(n) = s(n)$$

$$x_f(n) = s(n) \otimes h(n) \qquad (11)$$

$$e(n) = s(n) \,\Phi\, d\{x_f(n)\}$$

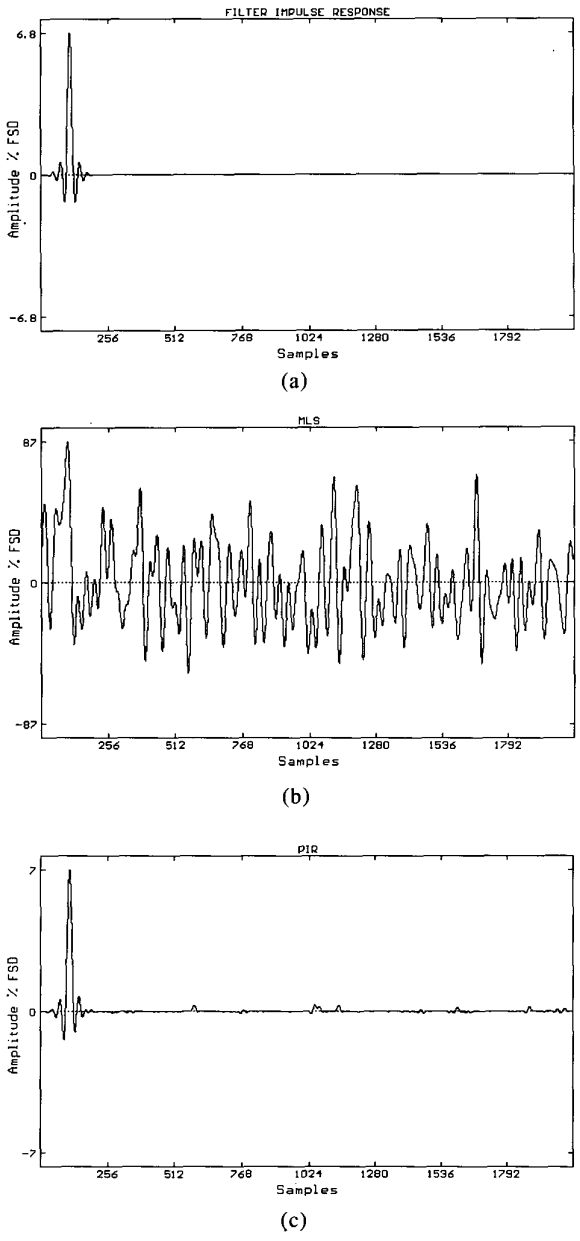


Fig. 7. (a) Impulse response $h(n)$ of 1-kHz low-pass FIR filter. (b) MLS signal after filtering with 1-kHz low-pass filter. (c) MLS-derived impulse response of 1-kHz filter when filtered MLS has been distorted with second-order nonlinearity at -10 dB. Spiky error can clearly be seen in tail of impulse response, and can be compared with uncorrupted impulse response $h(n)$ of low-pass filter in (a).

The nonlinear component $e_{nl}(n)$ of $e(n)$ is calculated using the methodology described in Section 2. Although the error is calculated directly [Eq. (11)], the accuracy of the MLS simulations is not as high as that obtained from the PIE experiments. The cross-correlation postprocessing operation required for MLS involves $\log_2(L/2)$ floating-point additions or subtractions for each error sample when the FHT is used, hence the theoretical maximum MLS distortion immunity that can be recorded using this simulation technique is bounded by floating-point mantissa quantization at approximately 300 dB. The results recorded in Tables 1–4 are well within this limit.

Fig. 8 shows the impulse distortion $e(n)$ due to a second-order nonlinearity for $A_d = -20$ dB. Because a filtered MLS possesses an approximately symmetrical amplitude distribution, then even-order nonlinearity results in very low gain error, and the normalized error sequence is very similar to the unnormalized error [11]. However, odd-order memoryless nonlinearity results in a large error sequence component that is coincident with the linear impulse response, and error normalization thus results in a significant fall in error level. For example, the third-order impulse error with the 1-kHz filter falls by 4 dB after normalization (compare Figs. 9 and 10). Table 1 presents the simulation results for (untruncated) MLS distortion immunity with the 1-kHz filter and nonlinearities ranging from second- to seventh order.

3.3 Enhancing MLS Distortion Immunity by Truncation

We will now examine the effect that truncating a recovered impulse response has on MLS distortion immunity. Rife and Vanderkooy [4] conjecture that distorting an MLS results in an MLS error sequence that can be viewed as a phase-randomized signal in the frequency domain which will, after cross correlation, result in a recovered nonlinear impulse error $e_{nl}(n)$ that is evenly spread over the measurement period. This error distribution model will henceforth be referred to as the constant error density model. Since the linear impulse response $h(n)$ will typically be contained in the first few samples of the measurement, then truncating a measurement of period L at t samples should result in an increase in distortion immunity of T dB

Table 1. Distortion immunity of impulse response measurements for 1-kHz FIR low-pass filter with $A_d = -20$ dB.

Distortion Order	PIE Distortion Immunity (dB)	MLS Distortion Immunity (dB)	IRS Distortion Immunity (dB)
2	54.7	29.4	>262
3	77.2	35.4	36.6
4	99.7	35.9	>265
5	123	38.4	41.4
6	146	39.7	>267
7	169	41.4	46.2

Table 2. Noise immunity advantage of MLS over PIE for 1-kHz FIR low-pass filter, when distortion immunity has been normalized. Results extrapolated from Table 1 data and confirmed by additional simulations.

Distortion Order	Distortion Immunity (dB)	Relative MLS Excitation Amplitude (dB)	MLS Noise Immunity Advantage (dB)
2	54.7	−25.3	7.8
3	77.2	−20.9	12.2
4	99.7	−21.3	11.8
5	123	−21.2	11.9
6	146	−21.3	11.8
7	169	−21.3	11.8

Table 4. Noise immunity advantage of MLS over PIE for 10-kHz FIR low-pass filter, when distortion immunity has been normalized. Results extrapolated from Table 3 data and confirmed by additional simulations.

Distortion Order	Distortion Immunity (dB)	Relative MLS Excitation Amplitude (dB)	MLS Noise Immunity Advantage (dB)
2	36.4	−14.5	18.6
3	42.4	−9.3	23.8
4	47.6	−10.3	22.8
5	53.3	−9.3	23.8
6	59.2	−9.7	23.4
7	65.3	−9.4	23.7

Table 3. Distortion immunity of impulse response measurements for 10-kHz FIR low-pass filter with $A_d = -20$ dB.

Distortion Order	PIE Distortion Immunity (dB)	MLS Distortion Immunity (dB)	IRS Distortion Immunity (dB)
2	36.4	21.9	>263
3	42.4	23.9	23.9
4	47.6	16.6	>254
5	53.3	16.1	16.1
6	59.2	10.6	>246
7	65.3	8.9	8.9

given by

$$T = 10 \log_{10} \left[\frac{L}{t} \right] \text{ dB} . \quad (12)$$

To investigate the accuracy of the constant error density model some tests were performed on the normalized error sequences $e_{nl}(n)$ obtained from the simulations. Fig. 11 shows the normalized error sequences and error distributions for distortion orders 2 through 5. The error distributions $P(n)$ are obtained by calculating error energy accumulation across the sequence as a proportion of total error energy,

$$P(n) = \frac{\sum_{k=0}^{n} e_{nl}(k)^2}{\sum_{k=0}^{L-1} e_{nl}(k)^2} . \quad (13)$$

Hence for a constant error density, $P(n)$ should plot as a straight line from point 0 (0%) to point $(L - 1)$ (100%).

An examination of Fig. 11 indicates two trends:

1) For similar orders of nonlinearity, even-order error distributions exhibit a greater degree of lumpiness than

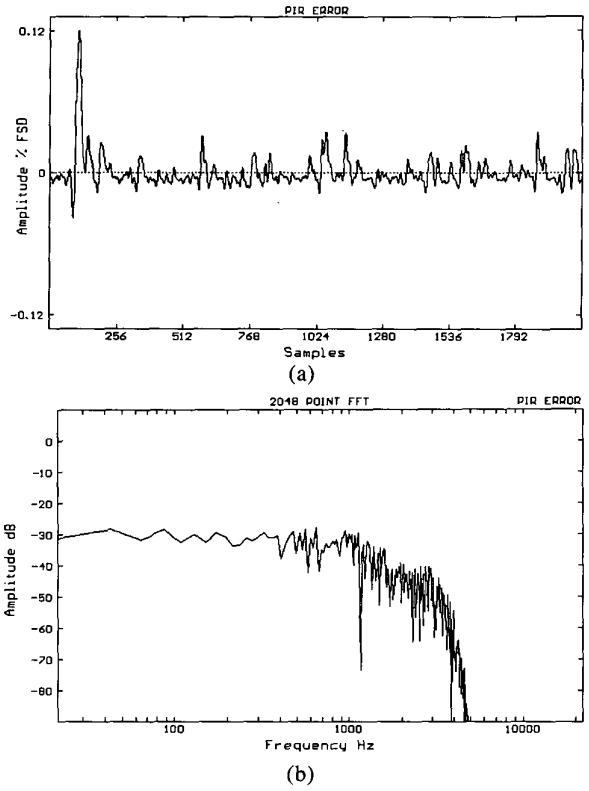


Fig. 9. MLS-derived impulse response error for 1-kHz filter and third-order nonlinearity. (a) Time domain $e(n)$. (b) Frequency domain.

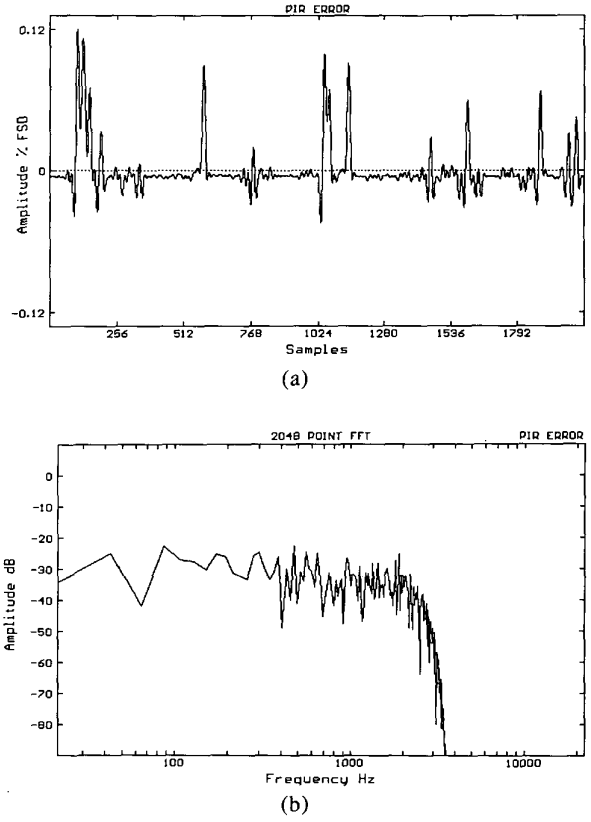


Fig. 8. MLS-derived impulse response error for 1-kHz FIR filter with second-order nonlinearity at −20 dB. (a) Time domain $e(n)$. (b) Frequency domain.

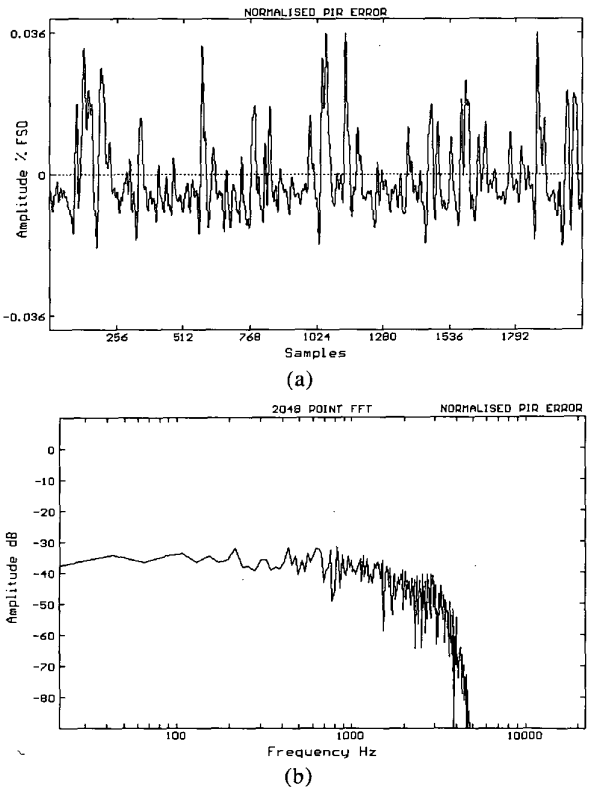


Fig. 10. MLS-derived nonlinear impulse response error (following error normalization) for 1-kHz filter and third-order nonlinearity. (a) Time domain $e_{nl}(n)$. (b) Frequency domain.

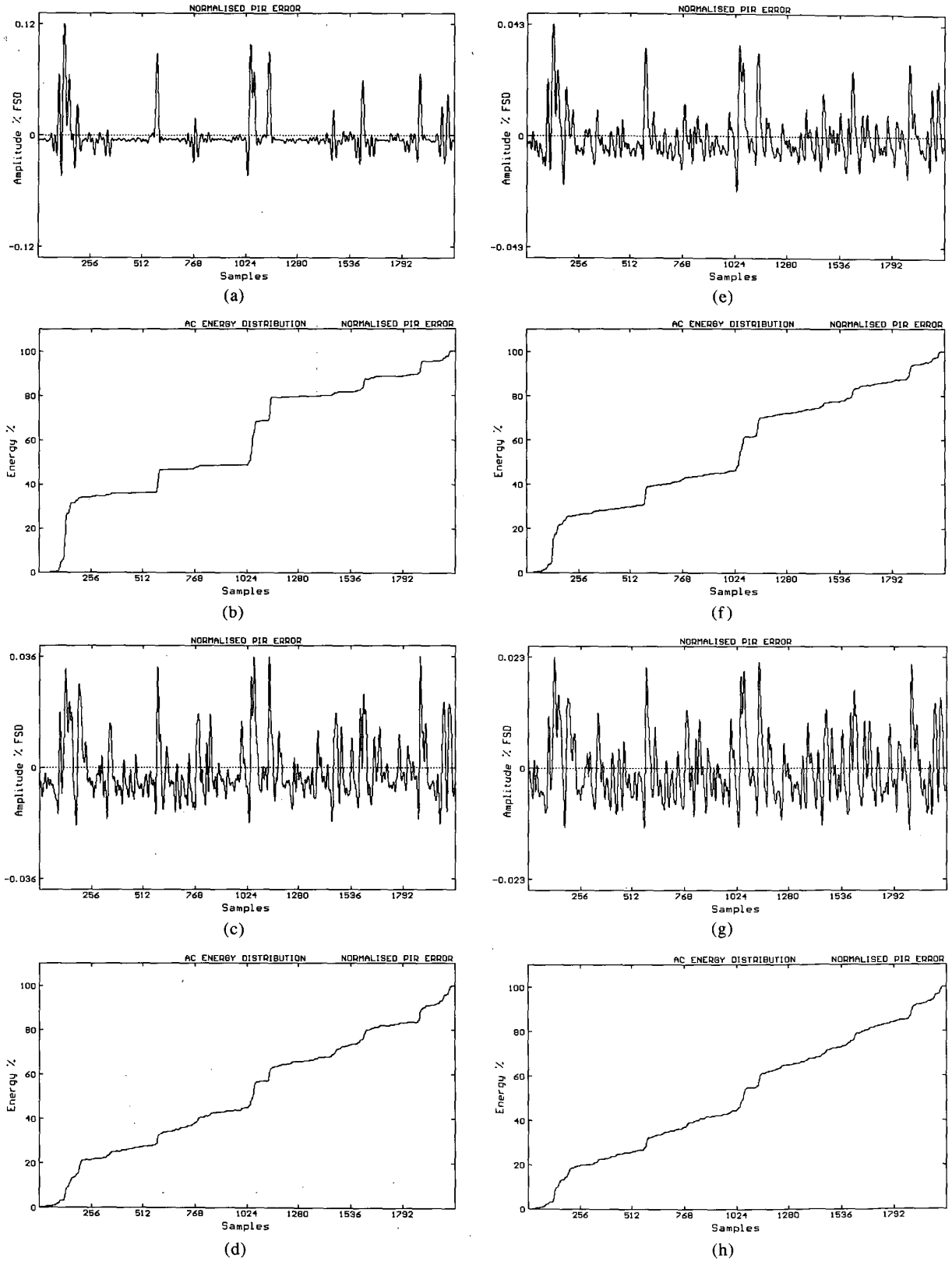


Fig. 11. Normalized nonlinear impulse response errors $e_{nl}(n)$ and error distributions $P(n)$ for MLS measurement of 1-kHz low-pass filter with (a), (b) second-, (c), (d) third-, (e), (f) fourth-, (g), (h) fifth-order nonlinearity.

the odd-order distributions.

2) The distributions become smoother as the order of nonlinearity increases.

Further simulations with different MLS periods L (not shown here) have confirmed that these results are general. Hence the second-order error sequence will usually exhibit the lumpiest distribution; for the examples shown in Fig. 11, truncating the second-order measurement at 256 samples would increase distortion immunity by 4.5 dB rather than the 9 dB predicted by Eq. (12). However, a truncation anywhere in the fifth-order error sequence will result in an increase in distortion immunity close to the predicted improvement, a result that is due to the smoother error distribution of the higher order nonlinearity. The smoothing of error distribution with increasing order of nonlinearity is not surprising; the amplitude peaks in the filtered MLS will be accentuated by nonlineariy and, for high-order distortion, tend to result in impulselike transient errors in the distorted MLS. These transients result in an error sequence evenly spread across the recovered impulse response since cross correlation with the unfiltered MLS $s(n)$ is equivalent to a time-reversed convolution with $s(n)$ (that is, the error transients are convolved with $s(-n)$; see [4] for a more detailed discussion of transient noise immunity in MLS measurements).

It is important to note that, for low and particularly even-order nonlinearity, the increase in distortion immunity due to truncation is not necessarily lower than predicted by Eq. (12). Depending on the particular MLS period used and the distortion order, it is also possible that the increase in distortion immunity will be higher than predicted. In general, the uneven error distributions introduce a degree of uncertainty to the improvement in distortion immunity gained from truncation. Since the degree of uncertainty in MLS error distribution is generally larger for even-order nonlinearity than for odd-order nonlinearity, a significant reduction in uncertainty is obtained by using inverse repeat sequences (IRS), which exhibit complete immunity to even-order nonlinearity (see Section 6).

Rife and Vanderkooy also suggest that the error will tend to spread more evenly as the stimulus applied to the nonlinearity approaches a Gaussian amplitude distribution. The simulations so far have used MLS signals filtered using a 1-kHz low-pass filter, resulting in a quasi-Gaussian amplitude distribution (Fig. 12). This can be compared to the MLS amplitude distribution shown in Fig. 13, obtained by low-pass filtering with a higher cutoff frequency (20 kHz), which is evidently *not* Gaussian. However, the impulse error distribution for the 20-kHz filtered MLS distorted with second-order nonlinearity [Fig. 14(b)] is very similar to the 1-kHz result shown in Fig. 11(b). Conversely the third-order error distribution for the 20-kHz filtered MLS [Fig. 14(d)] is less smooth than the 1-kHz result presented in Fig. 11(d). These results suggest an ambiguous relationship between the filtered MLS amplitude distribution and error distribution.

Although we have so far used only memoryless distortion in our simulations, many physical nonlinear error mechanisms exhibit memory, that is, where the current value of the error sequence depends on previous values of the signal applied to the nonlinearity. Distortion mechanisms with memory usually exhibit frequency-dependent characteristics. In a private communication to the authors, Vanderkooy conjectures that MLS error sequence distributions due to nonlinearity with memory exhibit a degree of smoothing compared to the equivalent memoryless cases. In order to test this hypothesis we performed further simulations using a simple model for distortion with memory, where one argument to the nonlinearity is delayed by b samples. Hence for an rth-order nonlinearity,

$$d\{x_f(n)\} = A_d \left[\frac{x_f(n)}{x_{ref}}\right]^{r-1} \left[\frac{x_f(n-b)}{x_{ref}}\right]. \qquad (14)$$

The results of our simulations confirm that for *even*-order nonlinearity, memory in the distortion mechanism does tend to smooth error distributions. For example, Fig. 15(a) and (b) shows the error sequence and distribution for second-order nonlinearity when $b = 60$; the plots can be compared to the memoryless error sequences in Fig. 11(a) and (b). However, for *odd*-order nonlinearity with memory, the linear gain error

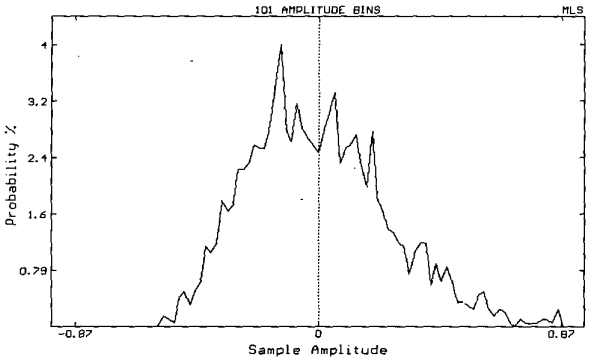


Fig. 12. Amplitude distribution for 1-kHz low-pass filtered MLS, showing almost Gaussian distribution.

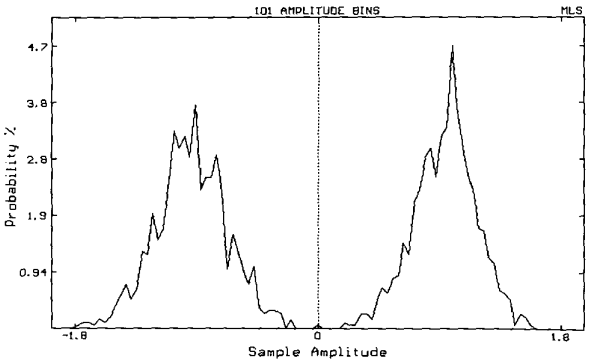


Fig. 13. Amplitude distribution for 20-kHz low-pass filtered MLS.

is delayed relative to $h(n)$ and is not removed during normalization. This behavior results in an error spike close to the linear impulse response, which cannot be removed by truncating the impulse response. Our simulations have indicated that the magnitude of the error spike is generally largest for third-order nonlinearity with memory, and for this case truncation will not yield large increases in distortion immunity. Fig. 15(c) and (d) shows the error sequence and distribution for third-order nonlinearity with memory ($b = 60$), where 45% of the total impulse response error energy is concentrated in the first 256 samples (12.5%) of the measurement period.

Generally then the error distributions are evenly spread across the measurement period, a result which validates the constant error density model. Since noise artifacts will also be spread evenly across the measurement period, then MLS allows a basic separation of linear and error impulse response components. As we shall see in the next section, this behavior is extremely useful in achieving optimal excitation amplitude.

4 OPTIMAL EXCITATION AMPLITUDE AND PERIOD IN MLS MEASUREMENTS

So far we have considered distortion immunity of MLS and PIE measurements, but real measurement environments will tend to suffer both distortion and noise corruption. In this section we ask: what is the total error immunity advantage of MLS over PIE, if any? We will also examine methods of maximizing total error immunity in practical MLS measurement systems, including the selection of optimal excitation amplitude and measurement period.

4.1 Total Error Immunity Advantage of MLS over PIE

Rife and Vanderkooy [4] show that an unfiltered MLS has $L + 1$ times the signal power of a PIE at all signal frequencies bar dc when the peak signal voltage and sequence period are the same for both cases. This result is a direct consequence of the uniform spread of MLS excitation energy across the measurement period, compared to the localized (unit impulse) PIE signal. The noise power in a system is usually fixed in level (for example, room noise in a loudspeaker measurement). Thus the excitation signal-to-noise ratio in an MLS measurement is $10 \log_{10}(L + 1)$ dB higher than PIE. The energy conservation property of MLS cross correlation [18] preserves the signal-to-noise ratio advantage of MLS through to the recovered impulse response. Thus MLS techniques possess a noise immunity advantage of approximately $10 \log_{10}(L + 1)$ dB over PIE, a result that is well known [4], [12], [19]. However, again compared to a PIE with equal (unfiltered) peak excitation amplitude, MLS has a distortion immunity *disadvantage* when measuring systems with bandwidths significantly lower than half the sampling frequency of the test system. In the following discussion we outline

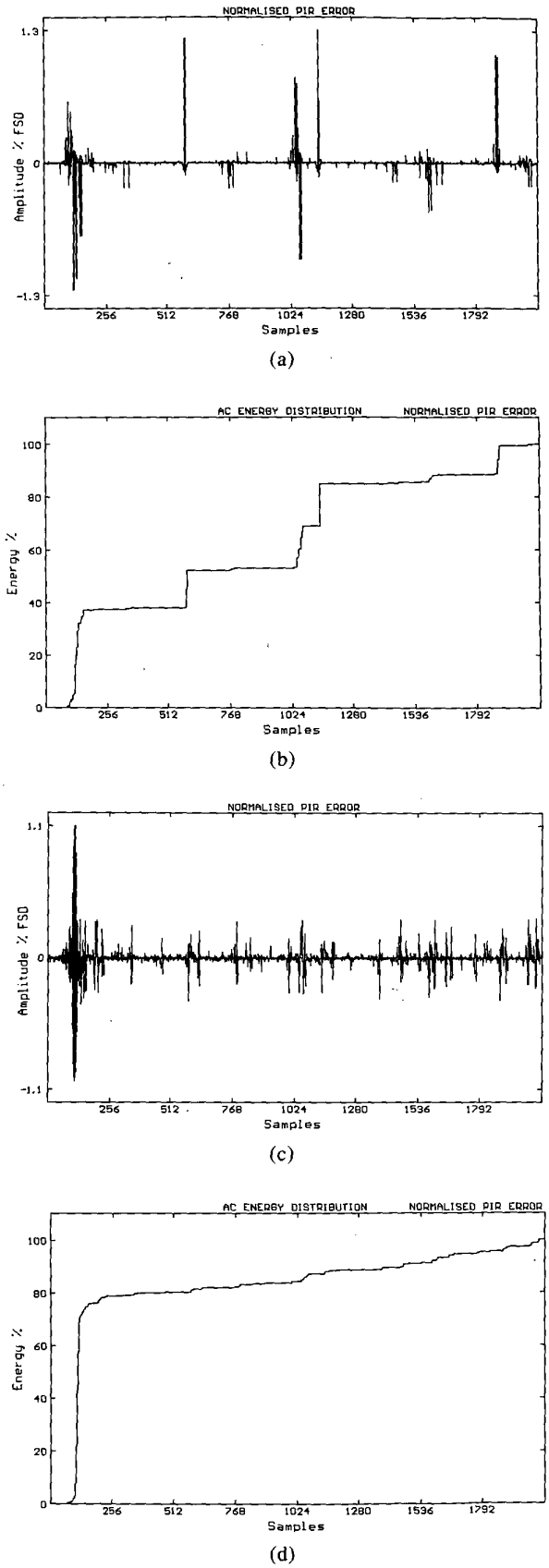


Fig. 14. Normalized error sequence $e_{nl}(n)$ and error distribution $P(n)$ for MLS-derived impulse response fo 20-kHz low-pass FIR filter with (a), (b) second- and (c), (d) third-order memoryless nonlinearity.

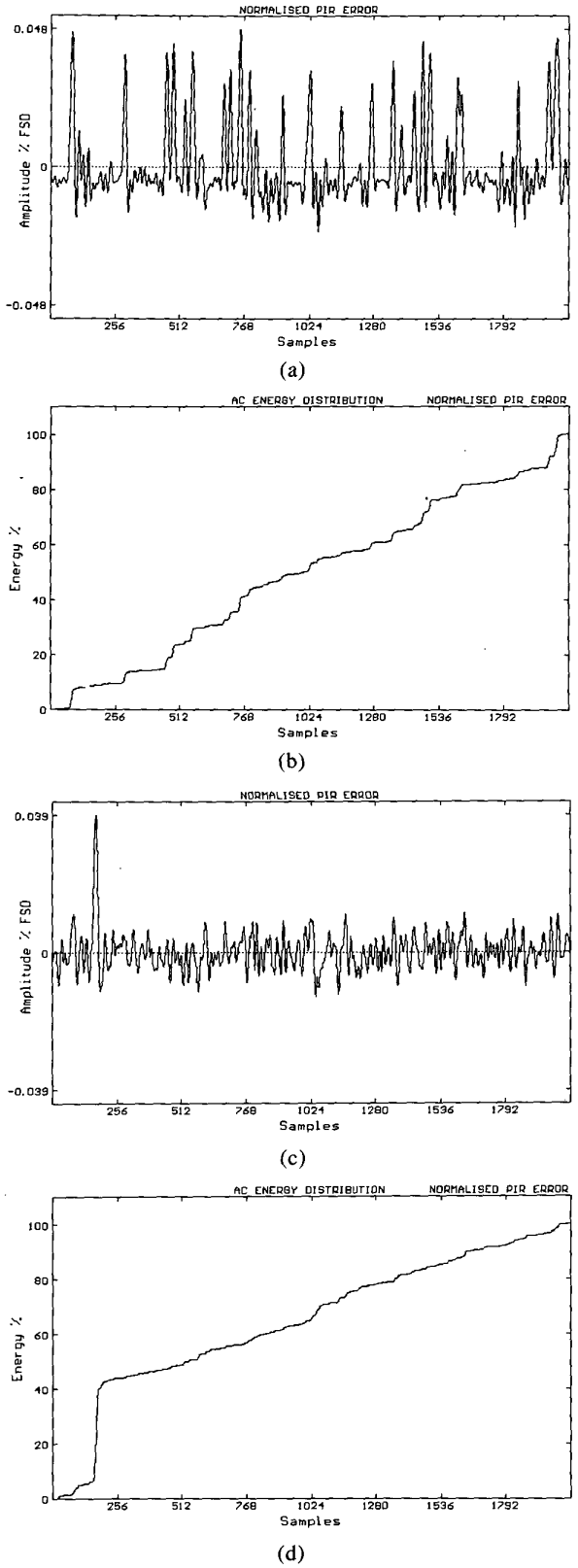


Fig. 15. Normalized error sequence $e_{nl}(n)$ and error distribution $P(n)$ for MLS-derived impulse measurement of 1-kHz FIR filter with (a), (b) second-order distortion with memory (delay $b = 60$ samples) showing smoothing of MLS error distribution compared to memoryless case, while (c), (d) third-order distortion with memory (delay $b = 60$ samples) indicates delayed gain error.

the reasons for this behavior and argue that, broadly speaking, the total error immunity advantage shown by MLS is somewhat less than $10 \log_{10}(L + 1)$ dB.

Consider a hypothetical system with a linear transfer function characterized by a unit impulse response, that is, $h(n) = \delta(n)$. Applying either an MLS or a PIE to such a system will result in an output signal unchanged from the input, that is, for both cases the output signal has a binary amplitude distribution. Thus for equal peak input amplitudes and measurement periods, the "filtered" MLS and PIE signals possess identical peak voltages, although the MLS has $L + 1$ times the ac signal power of the PIE. If the filtered signals are now subject to memoryless nonlinearity, then the resultant MLS error signal will also have $L + 1$ times the power of the PIE error signal (ignoring for the moment the characteristics of the error signals). Because both the MLS signal and nonlinear error powers are $L + 1$ times the respective powers for PIE, the distortion immunities for the two stimuli are equal. Again this is a direct consequence of the energy conservation property of MLS cross correlation. Now consider what happens when the main lobe of the system impulse response broadens over at least a few samples, for example, when the system bandwidth is lower than half the measurement system sampling frequency. If the peak input excitations and the measurement periods are again equal, the filtered excitation will tend to have a higher peak voltage for MLS compared to PIE. This occurs because "runs" of successive 1s and -1s in the unfiltered MLS stimulus [20] effectively integrate the system impulse response over regions of the measurement period during convolution, whereas for PIE the system is stimulated for one sample only. This behavior is illustrated by comparing the peak voltage for the 1-kHz filtered PIE [Fig. 3(a)] to that of the 1-kHz filtered MLS [Fig. 7(b)]. Note the change in vertical axis scaling. The higher peak voltage of the filtered MLS stimulates any nonlinearity more vigorously than for PIE, hence the error energy within the filtered and distorted MLS is greater than $L + 1$ times the error energy encountered in PIE. Consequently MLS has a lower excitation signal-to-distortion ratio than PIE, and hence a distortion immunity disadvantage for the lower system bandwidth. As the system bandwidth increases and the system impulse response approaches a unit impulse, then the MLS distortion immunity disadvantage decreases, eventually becoming zero when the test system impulse response is equal to a unit impulse.

Two examples that help illustrate this behavior are presented in Tables 1 and 3, where the distortion immunities of (untruncated) MLS-derived impulse responses are compared to PIE measurements for memoryless nonlinearities ranging from second to seventh order. Both the unfiltered peak amplitude prior to filtering and the measurement period ($L = 2047$) are constant throughout the MLS and PIE simulations. Table 1 compares the distortion immunities for the two methods with 1-kHz FIR filtering, indicating a clear disadvantage for MLS under these conditions with all dis-

tortion orders. MLS distortion immunity disadvantage remains evident when the filter cutoff frequency is increased to 10 kHz (Table 3), although the disadvantage is now reduced somewhat (as predicted). MLS distortion immunity disadvantage under these excitation conditions must be weighed against the improved noise performance compared to PIE which, for $L = 2047$, will equal 33.1 dB. The question now posed is, what is the effect of varying the excitation amplitude upon the relative noise and distortion immunities of the two methods?

For any transfer function measurement strategy and simple distortion mechanisms described by Eq. (2), the error due to nonlinearity increases as the peak level of the excitation increases, although the rate at which the error increases depends on the order of nonlinearity that the test system is subject to. A 6-dB increase in driving level will decrease distortion immunity by 6 dB for second-order nonlinearity, 12 dB for third-order, and so on. If the increase in peak unfiltered excitation amplitude is ΔA dB, then we can write

$$\Delta I_d = -(r - 1)\Delta A \ . \tag{15}$$

For system noise that is fixed in level, the noise immunity will increase by 6 dB for every 6-dB increase in excitation amplitude. If I_n represents noise immunity in decibels, then

$$\Delta I_n = \Delta A \ . \tag{16}$$

We can use Eqs. (15) and (16) to predict the noise immunity advantage of MLS over PIE following distortion immunity normalization. Consider the third-order result from Table 1 where, for equal unfiltered peak excitation amplitudes, PIE has a distortion immunity of 77.2 dB while MLS offers 35.4 dB. To make the distortion immunities for both techniques equal to 77.2 dB, the MLS excitation amplitude must be reduced. Using Eq. (15), the required reduction in amplitude is $(77.2 - 35.4)/2 = 20.9$ dB. Reducing the MLS excitation amplitude by this amount yields a reduced noise immunity advantage for MLS. Using Eq. (16), the new MLS noise immunity advantage is $33.1 - 20.9 = 12.2$ dB. This procedure was followed for all of the examples listed in Tables 1 and 3, and the new noise immunity results were recorded in Tables 2 and 4, respectively. These predicted results were found to agree identically with a further set of MLS simulations performed at the lower excitation amplitudes. Generally it is clear that some degree of MLS noise immunity advantage remains following distortion immunity normalization, although the noise immunity advantage is then not as high as 10 $\log_{10}(L + 1)$ dB.

How can we use this information to predict the minimum total error immunity advantage of MLS over PIE? The answer to this question evidently depends on the relative levels of noise and distortion present in the test environment. For a system with low levels of weak nonlinearity the noise error will dominate and the total error immunity will be limited by the peak amplitude allowed in the system before overload. In this case MLS will show a 10 $\log_{10}(L + 1)$-dB total error immunity advantage over PIE. Most systems, however, will exhibit some degree of noise *and* distortion error. For these situations there will be some optimum driving level where the error contributions due to noise and distortion are equal and together result in the highest possible overall error immunity. When the driving signal is increased above this optimum level, the distortion error will dominate, while noise will be dominant for lower signal levels. If MLS still offers a noise immunity advantage over PIE when distortion immunities for both techniques are equal, then a little thought reveals that MLS must also possess some degree of total error immunity advantage over PIE when driving levels are optimized individually for both cases. Generally we can state that given the same measurement period L but optimal excitation amplitudes, MLS offers a total error immunity advantage over PIE of between 0 dB and 10 $\log_{10}(L + 1)$ dB, the exact advantage depending on (1) the relative levels of noise and distortion in the test system and (2) the sampling frequency used for the measurement. We have seen that with normalized distortion immunities, MLS noise immunity advantage tends to increase as the bandwidth of the DUT increases relative to the sampling frequency of the test system. Hence the greatest overall error immunity advantage offered by MLS over PIE for a given DUT occurs when the MLS/PIE sampling rate is as low as possible.

The arguments presented so far are based on measurements where the recovered impulse response remains untruncated. As we have seen in Section 3, truncation will yield enhancements in distortion immunity for MLS where impulse error due to nonlinearity is generally evenly spread across the measurement period. Noise immunity will also be improved by truncation for both MLS and PIE, since the noise error is spread evenly across the measurement period for both techniques. However, nonlinear errors in PIE measurements are coincident with the linear impulse (Figs. 3–5), and thus PIE distortion immunity will not be reduced by truncation. Thus for a system prone to both noise and distortion, truncation improves the minimum total error immunity advantage of MLS over PIE. We should also note that we have characterized nonlinearity in the measurement system by simple power laws [see Eq. (2)], although many practical systems also suffer from other types of nonlinearity. For example, analog-to-digital converter quantization distortion can corrupt the data acquisition stage of measurement, while crossover distortion and slew limiting can occur in power amplifiers used to drive loudspeakers under test. To a first-degree approximation, both quantization distortion and crossover nonlinearity remain fixed in level as the excitation amplitude is varied. These errors can therefore be treated as system noise and do not change the basic arguments presented earlier. For a full discussion of the effects of slew limiting in MLS measurements

we refer the reader to Godfrey and Murgatroyd [10] and Vanderkooy [11].

4.2 Determining Optimal Excitation Amplitude

We have seen that, given optimal excitation amplitudes for both techniques, MLS measurements theoretically exhibit at least some degree of overall error immunity advantage over PIE. However, practical PIE measurements have a further disadvantage since it is often difficult to achieve optimal excitation amplitude. First, the driving signal energy is relatively low so that the effects of distortion are often negligible in comparison to the noise error. The driving system (for example, the power amplifier in a loudspeaker PIE measurement) is often overloaded before the optimal driving amplitude is achieved. Second there is the problem of effectively monitoring the error due to nonlinearity in PIE measurements, because the linear impulse and nonlinear error signals are coincident in the time domain. Conversely, MLS allows the overall error level to be easily monitored by windowing the tail of the recovered impulse response, which given a sufficiently long measurement period will contain only noise and distortion error components (see Section 3). Optimal MLS excitation amplitude is achieved when there is minimal energy in the tail of the impulse response with respect to the energy of the linear (initial) part of the impulse response. This is equivalent to maximizing the MLS coherence function proposed by Rife and Vanderkooy [4]. Consider Fig. 16, which shows the last 1023 points of the recovered impulse responses from simulated MLS measurements which have been corrupted by nonlinearity (third order) and Gaussian noise. Remembering that the amplitude scaling in the time-domain plots is relative to the peak level of the driving signal, Fig. 16(a) shows the impulse tail when the driving level is 6 dB too high and distortion artifacts dominate the error signal. Optimal error immunity is achieved in Fig. 16(b) by reducing the MLS amplitude by 6 dB, which results in noise and distortion errors of equal energy. The overall error immunity is increased by about 9 dB over Fig. 16(a). When the driving level is further reduced by 6 dB in Fig. 16(c), noise now dominates and overall error immunity is reduced by approximately 3 dB from the optimal arrangement of Fig. 16(b).

4.3 Selecting Optimal Measurement Period

Given that we can determine the optimal excitation amplitude in an MLS measurement by examining the tail of the recovered impulse response, what is the best MLS period L to use? Although we have only presented results for $L = 2047$, additional simulations with L ranging from 255 to 8191 samples have indicated that MLS noise and distortion immunity for a given DUT both remain fairly constant as L changes. If the MLS period is much longer than the length of the impulse response being measured, then noise and distortion immunity can both be increased substantially by discarding the tail of the recovered PIR, because as we have seen in Section 3, both the error due to noise and that due to distortion are evenly distributed across the measurement period. Of course the increase in noise immunity obtained from such a truncation could also be effected by averaging several shorter measurements, but averaging has no effect on distortion immunity, and the single long MLS measurement ultimately takes less time to execute. Thus we conclude that for maximum overall error immunity the period of MLS measurements should be made as large as possible, and the PIR recovered from cross correlation should be truncated to as short a length as possible.

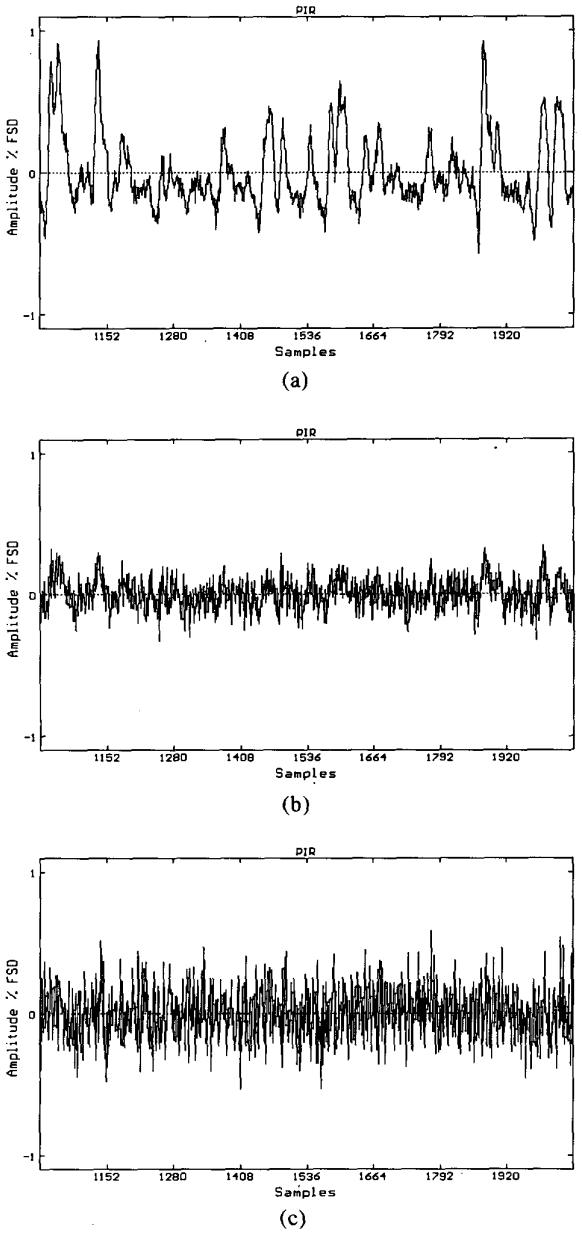


Fig. 16. Tail of impulse response from noisy and nonlinear MLS measurements. (a) MLS amplitude 6 dB too high; third-order distortion artifacts dominate. (b) Optimal excitation amplitude; noise and distortion powers are equal. (c) MLS amplitude 6 dB too low; noise dominates.

5 CUMULATIVE SPECTRAL DECAY PLOTS OF WEAKLY NONLINEAR SYSTEMS

In general, impulse response corruption due to nonlinearity will also corrupt information derived from the basic impulse response data. For example, Vanderkooy [11] has studied the effect of nonlinearity upon reverberation plots obtained from MLS measurements. In this section we examine the effect that weak nonlinearity has upon the accuracy of cumulative spectral decay (CSD) plots generated from impulse response measurements. Examples illustrate the increases in decay plot resolution to be gained from optimizing the stimulus amplitude in the impulse response measurement.

CSD or "waterfall" plots indicate how linear systems respond to tone bursts. Their use is becoming more widespread in loudspeaker evaluation [8], [9] because they allow a simple analysis of delayed resonances. An appropriately apodized CSD plot $C_a(\tau, \nu)$ is generated using the following function:

$$C_a(\tau, \nu) = \mathrm{F}\{h(t)\, w(t, \tau)\} \quad (17)$$

where

τ = time variable of plot
ν = frequency variable of plot
F = Fourier operator
$h(t)$ = impulse response of DUT
$w(t, \tau)$ = window function which influences time and frequency resolution of plot generated

For a full discussion of CSD plots we refer the reader to [21] and [22]. Following the conventions established in these works, the simulations presented here were generated using a raised-cosine window function with a rise time of 0.5 ms, fixed window time of 3.5 ms, and a fall time of 2 ms [Fig. 17(a)]. The 0-ms graduation on the decay plot time axes corresponds to the peak of the impulse response under evaluation. Fig. 17(b) indicates the spectral decay of a 20-kHz low-pass FIR filter. Apart from the characteristic FIR structure ringing around 20 kHz, this waterfall diagram shows a rapid energy decay across the audio band. Now consider Fig. 17(c), which shows the cumulative spectral decay of the same filter using a PIE-derived impulse response corrupted by second-order nonlinearity. Other than some additional high-frequency energy above 20 kHz, the plot is very similar to that of the linear filter plotted in Fig. 17(b). This behavior is expected since nonlinear artifacts in a PIE measurement are concentrated in the vicinity around the linear impulse response (see Section 2), and the error component therefore falls outside the

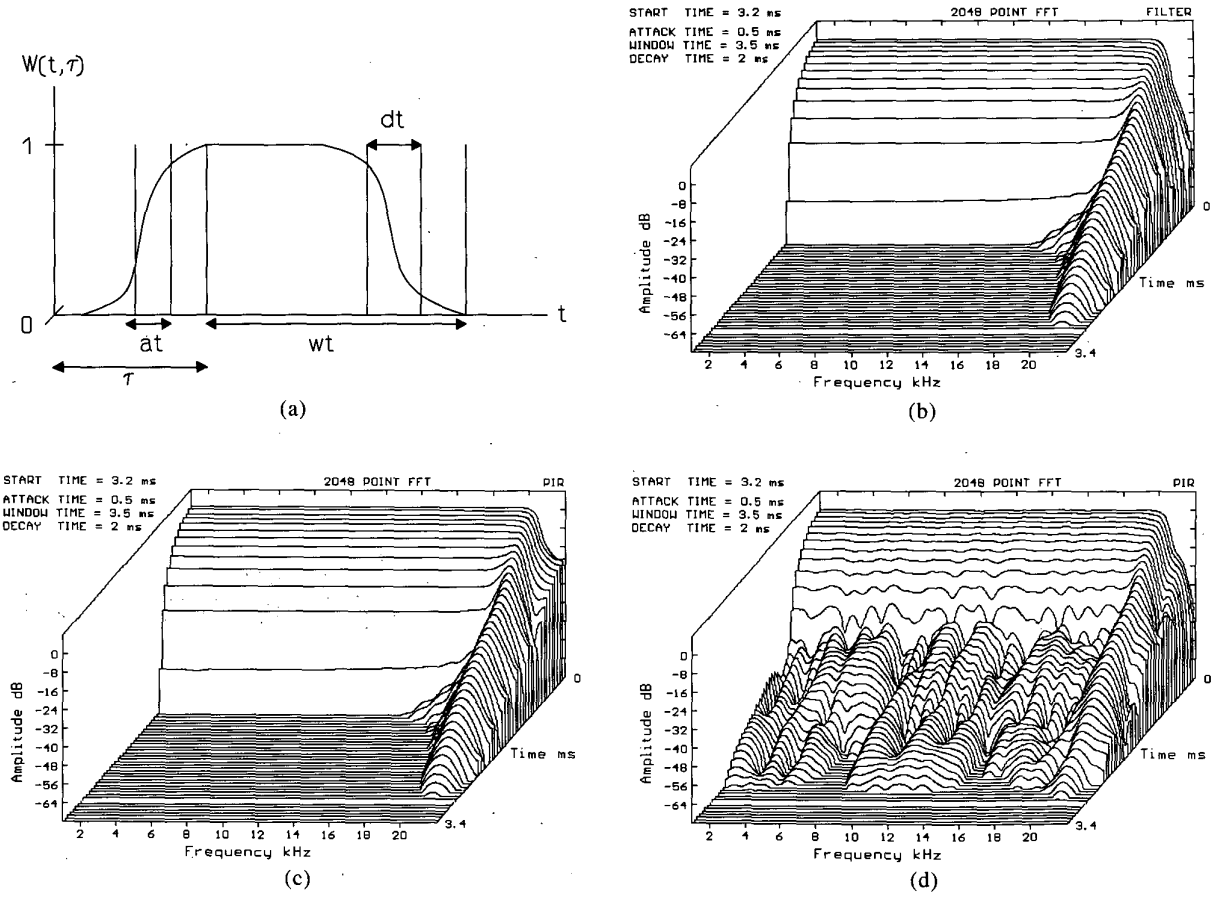


Fig. 17. Effects of nonlinearity on cumulative spectral decay plots. (a) Window function $w(t, \tau)$. (b) Decay plot of 20-kHz FIR filter. (c) Decay plot of impulse response from distorted PIE measurement. (d) Decay plot from distorted MLS measurement.

analysis window after the first few CSD sections. However, this is not true of an MLS-based measurement, where we have seen in Section 3 that error due to nonlinearity is spread evenly over the entire measurement period. Here the error will affect the cumulative spectral decay right across the plot and will be especially influential toward the front of the diagram where the main impulse energy has fallen outside the analysis window. This behavior is illustrated in Fig. 17(d), where the waterfall has been generated from an MLS measurement with second-order nonlinearity. After the initial energy decay, nonlinear artifacts within the analysis window cause ripples in the decay plot, which could be mistaken for delayed resonances attributable to the linear characteristics of the DUT. Similar errors occur in both MLS and PIE measurements that are corrupted with noise (which is, of course, evenly spread across the measurement period for both techniques). Indeed, poor noise immunity in PIE measurements can severely limit the resolution obtained from PIE-derived decay plots, and the advantages that MLS measurements possess in terms of total error immunity once optimal excitation amplitude is established are of real benefit here. An example used to illustrate the benefits of optimizing excitation levels in an MLS-derived decay plot is presented in Figs. 18, where white Gaussian noise and fourth-order nonlinearity corrupt impulse response measurements of the 20-kHz FIR filter used for the simulations of Fig. 17. The sequence of excitation conditions is similar to that of Fig. 16, that is, we examine the plots obtained from measurements where the driving amplitude is 10 dB too high [Fig. 18(a)], followed by optimal excitation amplitude [Fig. 18(b)], and finally an amplitude which is 10 dB below optimum [Fig. 18(c)]. In Fig. 18(a) the error due to nonlinearity severely corrupts the decay plot, while in Fig. 18(c) noise is largely responsible for the decay-plot error. The decay plot shown in Fig. 18(b) clearly possesses the lowest error component, a condition which is coincidental with optimal excitation amplitude being achieved. Although this is a slightly contrived example in that the use of high-order nonlinearity accentuates the increase in total impulse response error when the excitation amplitude is too high, it does illustrate the gains to be made from optimizing excitation conditions. It is interesting to compare the decay plot obtained from the optimal MLS measurement [Fig. 18(b)] to that obtained from an optimal PIE stimulus (Fig. 19). The total error immunity advantage of MLS is clearly evident.

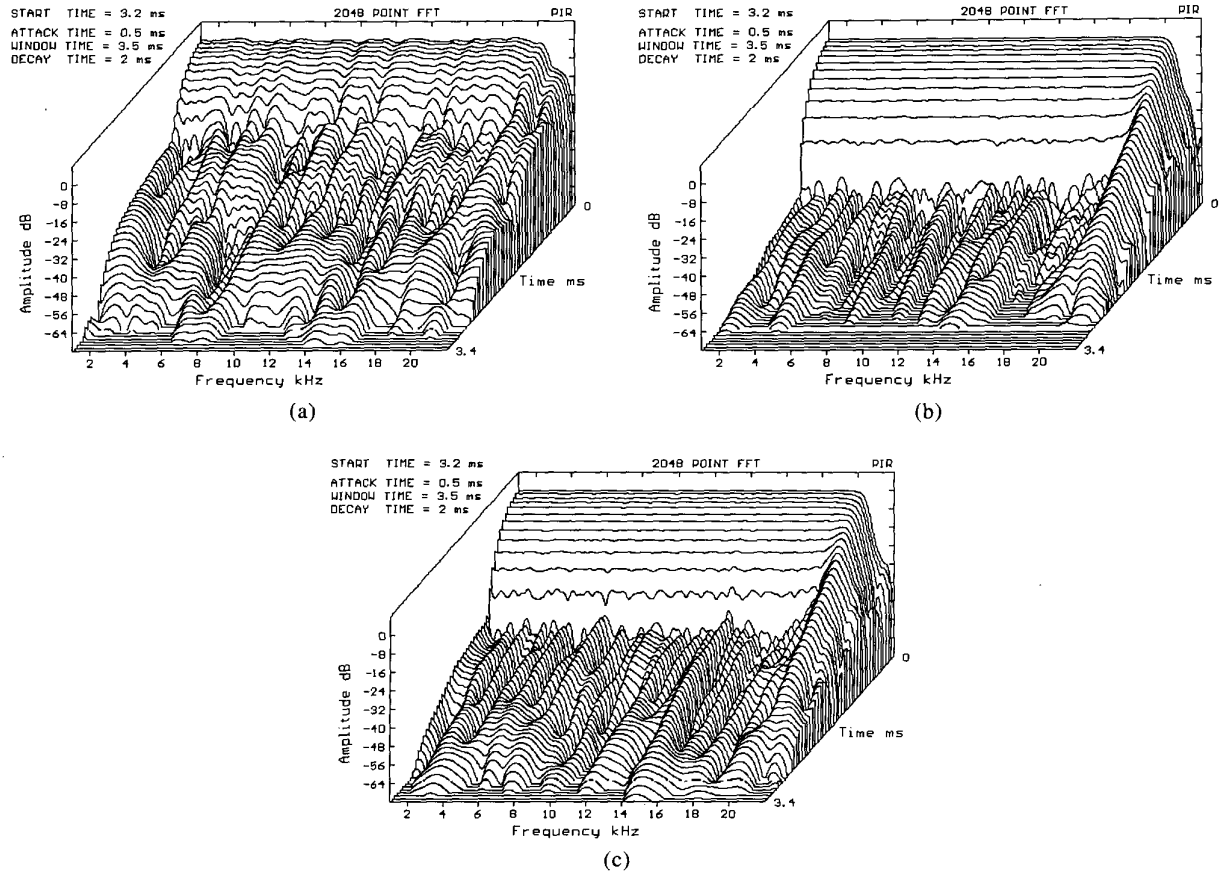


Fig. 18. Effect of optimizing excitation amplitude in cumulative decay plots derived from MLS impulse response measurements of 20-kHz FIR filter corrupted with white Gaussian noise and fourth-order nonlinearity. (a) Excitation amplitude 10 dB too high; nonlinear error dominates. (b) Optimal excitation amplitude, resulting in minimal plot corruption. (c) Excitation 10 dB too low, where noise error dominates.

6 INVERSE REPEAT SEQUENCES

Referring to our nonlinear system model [Fig. 2(c)], consider the error signal $d\{x_f(n)\}$ due to a second-order nonlinearity,

$$d\{x_f(n)\} = A_d\,[x_f(n)]^2 \ . \tag{18}$$

Now

$$\begin{aligned} x_f(n) &= x(n) \otimes h(n) \\ &= \sum_{i=0}^{L-1} h(i)x(n - i) \ . \end{aligned} \tag{19}$$

Combining Eqs. (18) and (19) yields

$$d\{x_f(n)\} = A_d \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} h(i)h(j)x(n - j) \ . \tag{20}$$

When the input signal is an MLS, then $x(n) = s(n)$ and the impulse error $e(n)$ is obtained by cross-correlating $d\{x_f(n)\}$ with $x(n)$ [as in Eq. (11)],

$$\begin{aligned} e(n) &= x(n)\ \Phi\ d\{x_f(n)\} \\ &= \frac{1}{L+1}\sum_{k=0}^{L-1} x(k)d\{x_f(n + k)\} \\ &= \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} A_d h(i)h(j)\left[\frac{1}{L+1}\sum_{k=0}^{L-1} x(k)x(k + n - i)x(k + n - j)\right] \\ &= \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} A_d h(i)h(j)\Omega_2(n - i, n - j) \ . \end{aligned} \tag{21}$$

Here Ω_2 is the second-order autocorrelation of the unfiltered MLS. We can write Eq. (21) in more general terms as a function of the second-order kernel of the system $h_2(i, j)$, of which our memoryless second-order nonlinearity is a specific case, that is,

$$h_2(i, j) = A_d h(i)h(j) \ . \tag{22}$$

Generalizing Eq. (21) to include a number of nonlinear kernels yields

$$\begin{aligned} e(n) = &\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} h_2(i, j)\ \Omega_2(n - i, n - j) \\ &+ \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\sum_{k=0}^{L-1} h_3(i, j, k)\ \Omega_3(n - i, n - j, n - k) \\ &+ \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\sum_{k=0}^{L-1}\sum_{l=0}^{L-1} h_4(i, j, k, l)\ \Omega_4(n - i, n - j, n - k, n - l) \\ &+ \cdots \ . \end{aligned} \tag{23}$$

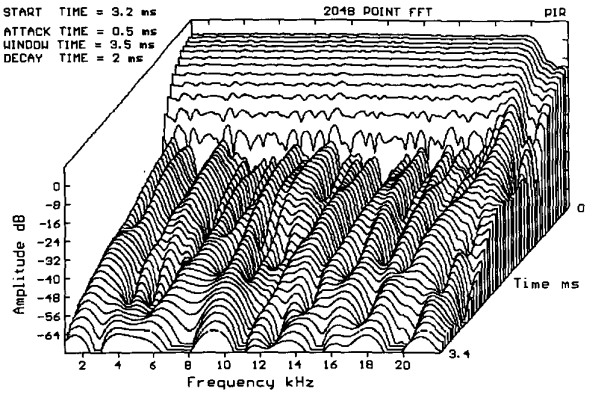


Fig. 19. Decay plot obtained from PIE-derived impulse response measurement of 20-kHz FIR filter. Measurement is corrupted with white Gaussian noise and fourth-order nonlinearity, while PIE excitation amplitude is optimal.

Each term in Eq. (23) is an r-dimensional convolution of a system kernel $h_r(n_1, n_2, \ldots, n_r)$ with the appropriate autocorrelation function $\Omega_r(n_1, n_2, \ldots, n_r)$ of the input sequence, where

$$\begin{aligned} &\Omega_r(n_1, n_2, \ldots, n_r) \\ &= \frac{1}{L+1}\sum_{k=0}^{L-1} x(k)x(k + n_1)x(k + n_2)\cdots x(k + n_r) \ . \end{aligned} \tag{24}$$

Using the shift and add property of MLS signals [23] it is easy to show that Ω_r can only take on two values, $L/(L + 1)$ (a "spike") and $-1/(L + 1)$. It is the spikes in Ω_r that cause the spikiness in the impulse error sequence $e(n)$ that we noted in Section 3.2. Following arguments similar to those presented in [24] it can be shown that there are approximately L^{r-1} spikes in Ω_r, evenly distributed across the L^r coordinates of the r-dimensional autocorrelation function. It is the even distribution of spikes in Ω_r which causes $e(n)$ to be evenly distributed across L, as we have seen in Section 3.3.

Now consider a periodic binary signal $x(n)$ suitable for impulse response measurement, where the second half of the sequence is the exact inverse of the first half, that is,

$$x(n + L) = -x(n) . \tag{25}$$

Note that the period of $2L$ of such a sequence will always contain an even number of samples. Referring to Eq. (24) and extending the limits of the summation to $2L - 1$, all even-order autocorrelations (r even) will be exactly zero simply because for all $n_1, n_2, \ldots, n_r$, each $x(k)x(k + n_1) \cdots x(k + n_r)$ term within the summation will exactly cancel with the corresponding $x(k + L)x(k + L + n_1) \cdots x(k + L + n_r)$ term. Such a sequence would therefore also possess complete immunity to even-order nonlinearity after cross correlation. Due to the antisymmetry in $x(n)$ the first-order autocorrelation will also possess antisymmetry about L, that is, $\Omega_1(n) = -\Omega_1(n + L)$. It is desirable that $x(n)$ be chosen such that the first half of Ω_1 is as close to a unit impulse as possible so that the linear impulse response of the DUT can be easily measured by cross-correlating the system output with the input (the second inverted half of the cross correlation can simply be discarded). A signal that satisfies these conditions is the so-called inverse-repeat sequence (IRS), obtained from two periods of an MLS, where every other sample of the MLS is inverted [24], that is,

$$x(n) = \begin{cases} s(n) , & n \text{ even} , 0 \leq n < 2L \\ -s(n) , & n \text{ odd} , 0 < n < 2L \end{cases} \tag{26}$$

where L is the period of the generating MLS. (Note that the IRS period is $2L$.) A 62-point IRS generated from a 31-point MLS (five-stage shift register) is shown in Fig. 20(a). The first-order autocorrelation of an IRS Ω_{IRS1} is related to the corresponding signal for the generating MLS by the following expression:

$$\Omega_{IRS1}(n) = \frac{1}{2(L + 1)} \sum_{k=0}^{2L-1} x(n)x(n + k)$$

$$= \begin{cases} \Omega_{MLS1}(n), & n \text{ even} \\ -\Omega_{MLS1}(n), & n \text{ odd} \end{cases}$$

$$= \delta(n) - \frac{(-1)^n}{L + 1} - \delta(n - L), \quad 0 \leq n < 2L . \tag{27}$$

The first-order autocorrelation for the 62-point IRS is presented in Fig. 20(b), clearly showing antisymmetry about L. There is also a small term oscillating at a rate of half the sampling frequency due to the $(-1)^n/(L + 1)$ factor in Eq. (27). The power spectrum of a periodic sampled signal is defined as the discrete Fourier transform (DFT) of its autocorrelation. Thus the IRS is spectrally flat at all frequencies except for dc and half the sampling frequency where the power is exactly zero. In a practical measurement system this would not be of concern because many systems such as loudspeakers do not possess a magnitude response that extends to dc, while the analog-to-digital converter used to digitize the system output signal will usually employ an antialiasing filter that will reject all information at half the sampling frequency. Thus by exciting a linear system with an IRS, sampling the output of the system, and cross-correlating that output with the (known) unfiltered IRS we obtain the impulse response of the system in much the same way that we would if using an MLS excitation. Of course the antisymmetry in the IRS autocorrelation results in an inverted copy of the system impulse response beginning at L samples. For example, Fig. 21 shows the recovered impulse response from a simulated 4094-point IRS measurement of the 1-kHz FIR filter. Since the second half of the cross correlation contains no additional information, it can simply be discarded.

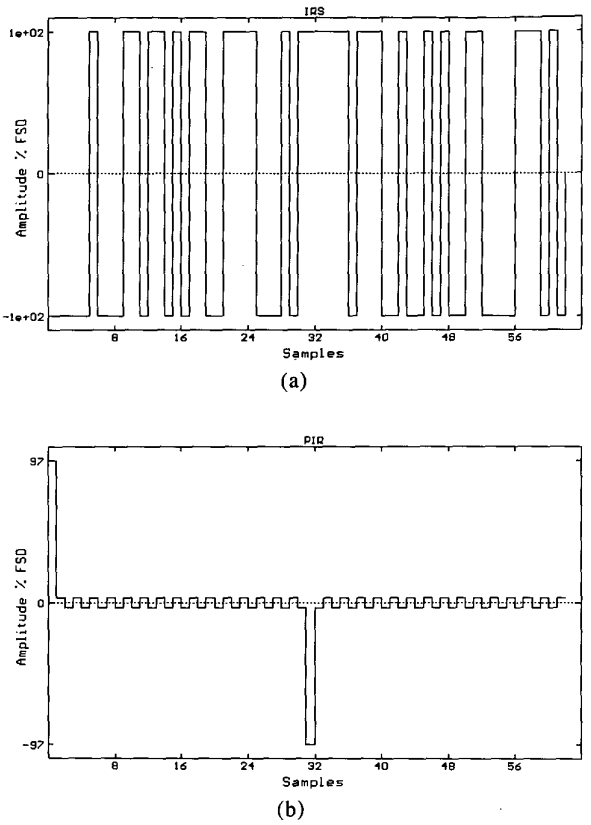


Fig. 20. (a) IRS formed from fifth-order MLS. (b) Autocorrelation sequence.

We have shown that impulse response measurements from IRS signals offer complete immunity to even-order nonlinearity. However, for odd-order distortion the recovered impulse response will still contain an error component, although, like MLS, this error component will tend to be spread evenly across the measurement period after error normalization. Fig. 22(a) shows the nonlinear error component for third-order nonlinearity at −20 dB with the 1-kHz low-pass FIR filter. This figure can be compared directly with Fig. 11(c), which shows the equivalent MLS error sequence.

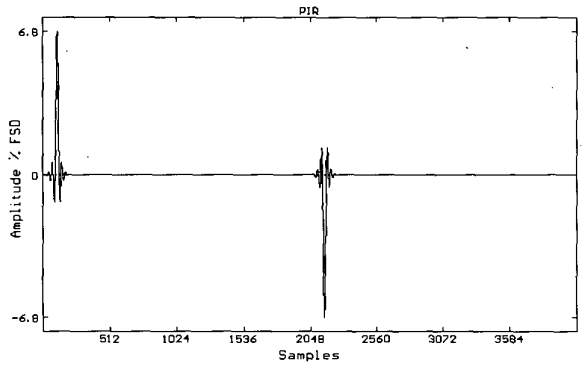
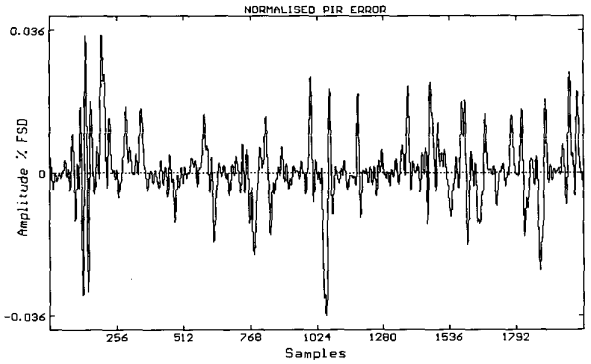
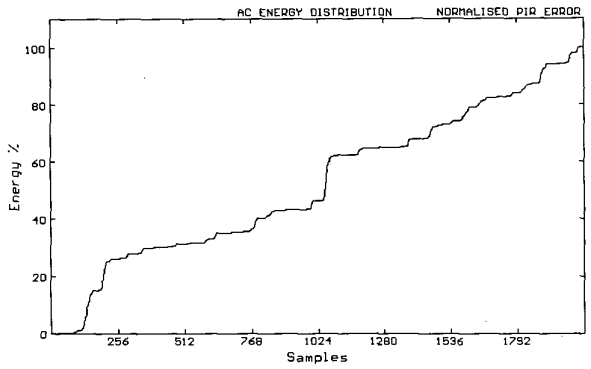


Fig. 21. Output of IRS cross correlation indicating antisymmetry about L samples.



(a)



(b)

Fig. 22. (a) Normalized nonlinear impulse error $e_{nl}(n)$. (b) Error distribution $P(n)$ from IRS-derived impulse response measurement of 1-kHz FIR filter with third-order nonlinearity.

The most obvious difference is that the IRS error sequence has bipolar spikes. This is because the odd-order autocorrelation functions for an IRS can assume four levels, ± 1 and $\pm 1/(L + 1)$, whereas we have seen that MLS autocorrelation functions are two valued. For an IRS the spikes appear in ± 1 pairs, and so for low-bandwidth systems, where the system kernels $h_r(n_1, n_2, \ldots, n_r)$ change slowly across r-dimensional space, some of the spike pairs tend to partially cancel out in the cross-correlation operation [Eq. (23)]. This behavior results in an increase in IRS distortion immunity for low-bandwidth systems (compared with the equivalent MLS case).

In order to determine IRS distortion immunity, simulations were performed with various nonlinearities using a process similar to that used for the MLS simulations [Eq. (11)]. MLS convolution and cross-correlation routines take advantage of the FHT, as discussed in Section 3, and in principle there is no reason why IRS convolution and cross correlation cannot also be performed using the FHT. However, permutation routines that make the most efficient use of the FHT (as developed by Borish for MLS) are not yet available for IRS. Hence for the IRS simulations all convolutions and cross correlations were performed by fast Fourier transform (FFT) in the frequency domain. Because the number of samples in an IRS period is not an exact power of 2, double-length FFTs must be used with zero padding (see [25, chap. 12]). The results of IRS distortion immunity simulations for second- to seventh-order nonlinearity with the 1-kHz FIR filter are tabulated in Table 1. Distortion immunity for even-order nonlinearity is extremely high and is only limited by the accuracy of the calculations in the simulations at approximately −250 dB (lower than MLS simulation accuracy because of the sizable increase in the number of computational operations required for the IRS simulations). For odd-order nonlinearity the IRS excitation shows a small increase in distortion immunity over MLS. This is not true for higher bandwidth simulations tabulated in Table 3. Here the FIR low-pass filter now has a cutoff frequency of 10 kHz and IRS odd-order distortion immunity can be seen to be identical to MLS. Finally we should note that, like MLS, IRS impulse response error due to nonlinearity is generally spread evenly across the measurement period [Fig. 22(b)]. Thus an IRS impulse measurement will also show an increase in distortion immunity after truncation. A point worth noting is that the second half of the autocorrelation output is the *exact* inverse of the first half, including any artifacts due to noise or distortion. Hence IRS error immunity is in no way affected by subtracting the second half of the recovered antisymmetric impulse response from the first half.

Inverse repeat sequences thus possess an impressive distortion immunity advantage over MLS, and given that the basic theory has been known for some time [24],[26],[27], it is surprising that IRS techniques are not in more widespread use. Are there any disadvantages suffered by IRS measurements? Obviously for an L-

point impulse response measurement, $2L$ samples must be generated, stored in memory, and then cross-correlated. There is no difficulty in generating the inverse repeat sequence since this can be formed using a shift register in a similar fashion to MLS. However, given a memory limit in a practical measurement arrangement, MLS will recover an impulse response that is twice as long as that available from IRS. For impulse lengths found in typical test devices such as loudspeakers, this does not represent a problem. Since an IRS of period $2L$ will have the same noise immunity as an MLS of length $2L + 1$, there is also no penalty to pay in terms of maximum noise immunity. There is, however, a disadvantage in cross-correlation time for IRS when using an FFT-based cross-correlation algorithm. If we assume that the execution time of an $(L + 1)$-point FFT is the same as an $(L + 1)$-point FHT, and that the IRS cross correlation comprises a forward FFT followed by an inverse FFT, both of length $4(L + 1)$, then the $2L$-point IRS cross correlation will take approximately eight times as long to execute as an L-point MLS cross correlation by FHT. In a 386 PC based system this does not present a serious problem since a 2046-point IRS cross correlation, coded in C language, can be performed in about 10 s. Although this is quick enough for on-site measurements, the cross correlation would be faster if assembly-level routines were used or FHT routines were developed for IRS.

Finally we should note that inverse repeat sequences are not the only signals that possess complete immunity to even-order nonlinearity while also displaying favorable first-order autocorrelation characteristics. Ternary sequences, that is, periodic signals with three levels generated from a shift register using modulo-3 arithmetic [28],[29], possess both of these characteristics and in fact show third-order distortion immunity advantages over both MLS and IRS due to their superior third-order autocorrelation characteristics. Nevertheless ternary sequences suffer from two disadvantages compared to IRS. First, three-level ternary sequences are more difficult to generate than binary IRS signals which only require a simple switch circuit for digital-to-analog conversion (although this hardly represents much of an obstacle given the wide availability of low-cost, high-performance multibit digital-to-analog converters). Second, cross-correlation routines for use with three-level ternary sequences cannot utilize the efficient FHT, because the FHT will only perform cross correlations for driving sequences with binary coefficients. More research is required upon the use of ternary and higher order sequences in linear transfer function measurement.

7 CONCLUSIONS

A simulated comparison of PIE and MLS impulse measurement techniques has shown that, given optimal excitation amplitudes, MLS methods possess superior overall error immunity. For excitations of equal peak voltage and period L, MLS offers a $10 \log_{10}(L + 1)$-dB noise immunity advantage over PIE, but suffers a distortion immunity disadvantage when the test device bandwidth is significantly lower than half the system sampling frequency. Once optimal excitation amplitudes have been established, the exact overall MLS error immunity advantage depends on the characteristics of the system under test, but will for all cases be between 0 dB and $10 \log_{10}(L + 1)$ dB.

An investigation into nonlinear error distribution in MLS-derived impulse measurements has confirmed that, in general, the error is evenly distributed across the period of the recovered impulse response. Even-order nonlinearity tends to result in less evenly spread error distributions compared to odd-order error distributions, while as the order of nonlinearity increases, the error distributions become smoother and more evenly spread. Memory in the nonlinearity also tends to smooth the even-order error distributions. Further simulations have shown that the amplitude distribution of the filtered MLS is not necessarily the major factor in determining the error distribution.

The even spread of error across the measurement period for both noise and nonlinearity has several important implications for MLS measurements. First the tendency toward the separation of linear and error components of the recovered impulse response can be used to monitor the relative error level in an MLS measurement and adjust excitation amplitude for optimal error immunity, a feature that is not available with PIE. Second, MLS noise and distortion immunity can be enhanced by truncating the recovered impulse response after the linear part of the impulse has decayed to zero. In fact maximum overall error immunity from a given MLS measurement system is obtained by choosing the longest period available and truncating the recovered impulse response as early as possible. Finally it is important to remember that the evenly spread MLS distortion error can also corrupt certain measurements such as cumulative spectral decay plots, although care taken in setting the stimulus amplitude should ensure that in most circumstances MLS offers superior performance compared to PIE.

The nonlinear impulse artifacts in MLS can be described in terms of the nonlinear kernels of the system under test and the higher order autocorrelation functions of the unfiltered MLS signal. Inverse repeat sequences can be formed by inverting every other sample of an MLS, and they possess even-order autocorrelation functions equal to zero. This feature endows IRS measurements with complete immunity to even-order nonlinearity in the test device. Furthermore, for test devices where the bandwidth is significantly lower than the sampling frequency of the measurement system, IRS will also show some error immunity advantage over MLS for odd-order nonlinearity. An IRS shows no disadvantages compared to the MLS other than a halving in measurable impulse length given a hardware memory limit, and an increase in cross-correlation time, which is of small consequence for typical measurement periods.

8 ACKNOWLEDGMENT

This work was supported by the Science and Engineering Research Council, UK. The authors would like to thank John Vanderkooy of the University of Waterloo for many fruitful discussions during a 4-month stay at the University of Essex, and also Richard Greenfield of Essex Electronic Consultants. They would also like to acknowledge the many helpful comments and suggestions made by the reviewers upon previous versions of the manuscript.

9 REFERENCES

[1] R. G. Greenfield and M. O. J. Hawksford, "Efficient Filter Design for Loudspeaker Equalization," *J. Audio Eng. Soc.*, vol. 39, pp. 739–751 (1991 Oct.).

[2] C. Dunn and M. O. J. Hawksford, "Towards a Definitive Analysis of Audio System Errors," presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, p. 994 (1991 Dec.), preprint 3137.

[3] J. M. Berman and L. R. Fincham, "The Application of Digital Techniques to the Measurement of Loudspeakers," *J. Audio Eng. Soc.*, vol. 25, pp. 370–384 (1977 June).

[4] D. D. Rife and J. Vanderkooy, "Transfer-Function Measurement with Maximum-Length Sequences," *J. Audio Eng. Soc.*, vol. 37, pp. 419–444 (1989 June).

[5] J. Vanderkooy, "Another Approach to Time-Delay Spectrometry," *J. Audio Eng. Soc.*, vol. 34, pp. 523–538 (1986 July/Aug.).

[6] H. Biering, O. Z. Pedersen, and J. Vanderkooy, "Comments on 'Another Approach to Time-Delay Spectrometry,'" *J. Audio Eng. Soc. (Letters to the Editor)*, vol. 35, pp. 145–146 (1987 Mar.).

[7] R. Greiner, J. Wania, and G. Noejovich, "A Digital Approach to Time-Delay Spectrometry," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 37, pp. 593–602 (1989 July/Aug.).

[8] J. Atkinson, "Measuring with MELISSA," *Stereophile*, vol. 13, pp. 118–119 (1990 Feb.).

[9] M. Colloms, Loudspeaker Reviews, *Hi-Fi News Rec. Rev.*, vol. 35, pp. 51–63 (1990 Dec.).

[10] K. R. Godfrey and W. Murgatroyd, "Input-Transducer Errors in Binary Crosscorrelation Experiments, pts. 1 and 2," *Proc. IEE (London)*, vol. 112, pp. 565–573 (1965 Mar.), vol. 113, pp. 185–189 (1966 Jan.).

[11] J. Vanderkooy, "Aspects of MLS Measuring Systems," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 1048 (1992 Dec.), preprint 3398.

[12] J. Borish and J. B. Angell, "An Efficient Algorithm for Measuring the Impulse Response Using Pseudorandom Noise," *J. Audio Eng. Soc.*, vol. 31, pp. 478–488 (1983 July/Aug.).

[13] J. Borish, "Self-Contained Crosscorrelation Program for Maximum-Length Sequences," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 33, pp. 888–891 (1985 Nov.).

[14] J. Borish, "An Efficient Algorithm for Generating Colored Noise Using a Pseudorandom Sequence," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 33, pp. 141–144 (1985 Mar.).

[15] C. Swerup, "On the Choice of Noise for the Analysis of the Peripheral Auditory System," *Biol. Cybern.*, vol. 29, pp. 97–104 (1978).

[16] H. Alrutz and M. R. Schroeder, "A Fast Hadamard Transform Method for the Evaluation of Measurements Using Pseudorandom Test Signals," *Proc. 11th Int. Conf. on Acoust.* (Paris, 1983), pp. 235–238.

[17] C. Dunn and M. Hawksford, "Distortion Immunity of MLS-Derived Impulse Response Measurements," in *Proc. IOA*, vol. 13, pt. 7, pp. 191–206 (1991).

[18] D. R. Rife, "Modulation Transfer Function Measurement with Maximum-Length Sequences," *J. Audio Eng. Soc.*, vol. 40, pp. 779–790 (1992 Oct.).

[19] M. R. Schroeder, "Integrated-Impulse Method for Measuring Sound Decay without Using Impulses," *J. Acoust. Soc. Am.*, vol. 66, pp. 497–500 (1979 Aug.).

[20] S. W. Golomb, *Shift Register Sequences* (Holden-Day, San Francisco, 1967).

[21] J. D. Bunton and R. H. Small, "Cumulative Spectra, Tone Bursts, and Apodization," *J. Audio Eng. Soc.*, vol. 30, pp. 386–395 (1982 June).

[22] S. P. Lipshitz, T. C. Scott, and J. Vanderkooy, "Increasing the Audio Measurement Capability of FFT Analyzers by Microcomputer Postprocessing," *J. Audio Eng. Soc.*, vol. 33, pp. 626–648 (1985 Sept.).

[23] S. H. Tsao, "Generation of Delayed Replicas of Maximal-Length Sequences," *Proc. IEE (London)*, vol. 111, pp. 1803–1806 (1964 Nov.).

[24] H. R. Simpson, "Statistical Properties of a Class of Pseudorandom Sequences," *Proc. IEE (London)*, vol. 113, pp. 2075–2080 (1966 Dec.).

[25] W. H. Press et al., *Numerical Recipes in C: The Art of Scientific Programming* (Cambridge University Press, Cambridge, UK, 1988).

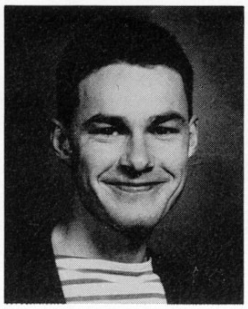
[26] N. Ream, "Nonlinear Identification Using Inverse Repeat Sequences," *Proc. IEE (London)*, vol. 117, pp. 213–218 (1970 Jan.).

[27] P. A. N. Briggs and K. R. Godfrey, "Pseudorandom Signals for the Dynamic Analysis of Multivariable Systems," *Proc. IEE*, vol. 113, pp. 1259–1267 (1966 July).

[28] E. P. Gyftopoulos and R. J. Hooper, "Signals for Transfer-Function Measurement in Nonlinear Systems," in *Proc. Noise Analysis in Nuclear Systems*, USAEC Div. of Tech. Inf., Symp. ser. 4 (TID-7679), pp. 335–345 (1964).

[29] E. P. Gyftopoulos and R. J. Hooper, "On the Measurement of Characteristic Kernels of a Class of Nonlinear Systems," in *Neuron Noise, Waves and Pulse Propagation*, USAEC Div. of Tech. Inf., Rep. 660206, pp. 343–356 (1967).

THE AUTHOR



Chris Dunn was born in Oxford, England, in 1966, and educated at Emmanuel College, Cambridge University, where he gained a degree in electrical and information sciences in 1988. Following graduation from Cambridge, he spent a year with Cambridge Systems Technology, which manufactures the Audiolab brand of audio hardware, where his duties involved a study of audio amplifier circuit topologies.

Mr. Dunn is currently lecturing in the Department of Electronic Systems Engineering at the University of Essex, while also completing a Ph.D. program of study within the Audio Research Group at the university under the supervision of Malcolm Hawksford. His current research interests include the effects of timing jitter in digital audio systems, psychoacoustic models for nonlinear error evaluation, and MLS measurement techniques.

The biography for Malcolm Omar Hawksford was published in the 1993 March issue of the *Journal*.

Distortion analysis of nonlinear systems with memory using maximum-length sequences

M.C. Greest
M.O. Hawksford

Indexing terms: Nonlinear systems, Maximum-length sequences

Abstract: Crosscorrelation using a binary MLS excitation yields a linear system response, whereas system nonlinearity produces a residue dispersed over the MLS period. By incorporating precomputed templates, coefficients of a nonlinear polynomial descriptor can be decoded from this distortion residue, and, by incorporating complex coefficients, nonlinearity with memory can be accommodated. A nonlinear model results from which general THD and IMD can be derived from a single MLS measurement.

1 Introduction

The use of maximum-length sequence (MLS) methods in linear system analysis is rapidly gaining in popularity. Rife and Vanderkooy [1] give a comprehensive account of current MLS methods and show that MLS is a viable and practical measurement technique that is superior to periodic impulse testing and time delay spectrometry in both noise and distortion immunity. Vanderkooy [2] explores distortion effects that can falsify a reverberation plot or reduce noise immunity, and Dunn and Hawksford [3] examine methods of increasing distortion immunity or eliminating distortion artefacts from the result altogether.

The MLS measurement process is a simple and efficient procedure. An MLS is a pseudorandom binary signal that yields an impulse upon circular autocorrelation. This signal is applied to a DUT in an analogue form, and the output is sampled and crosscorrelated with the original MLS, often using an efficient fast Hadamard transform (FHT) algorithm [4] to perform the crosscorrelation. The result is the periodic impulse response (PIR) of the system.

The papers above make no attempt to characterise the nonlinear behaviour of the DUT from the crosscorrelation result. This has been regarded as unfeasible, but is essential to obtain a comprehensive system characterisation. This paper is an extension of a previously published letter [5] and shows that it is technically possible to analyse aspects of the DUT using the MLS system, including its nonlinear components, and construct a more complete model of the DUT.

Paper 2025G (E6), firsted received 25th August 1994 and in final revised form 6th April 1995

The authors are with the Centre for Audio Research & Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom

One of the more interesting and important aspects of the MLS technique occurs during the crosscorrelation operation, which effectively separates the linear and nonlinear components of the DUT. The linear elements of the output signal are mapped into the impulse response, which is concentrated in the initial period of the sequence, whereas elements caused by nonlinearities are mostly spread across the entire crosscorrelation period. Thus, if the period is significantly longer than the impulse response, only a fraction of the nonlinear energy will corrupt the linear impulse trace [3]. This paper shows that further processing of the distortion artefacts can then characterise the nonlinear aspects of the DUT.

All examples and processes described are modelled by computer software.

2 Review of MLS theory

2.1 MLS generation and properties

A maximum-length sequence is a pseudorandom binary sequence (PRBS) that is generated recursively using a shift register with feedback (Fig. 1).

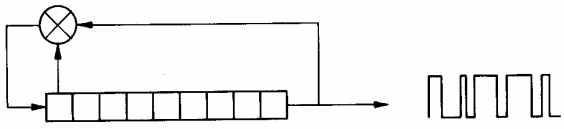


Fig. 1 *Generation of PRBS using shift register with modulo-2 addition feedback*

The PRBS has a period L determined by the position of the feedback taps along the register. An MLS is the maximum sequence period obtainable for a given length of register (N bits) and is given by

$$L = 2^N - 1 \tag{1}$$

Over a signal period, the MLS has similar properties to those of white noise; Figs. 2 and 3 show the frequency magnitude and phase spectrum of the MLS.

Thus an MLS contains discrete frequencies of equal magnitudes, up to the MLS clocking frequency, but spread in phase.

2.2 Extraction of PIR

The autocorrelation Ω_{ss} of a discrete sequence $s(n)$ is defined as

$$\Omega_{ss}(n) = \frac{1}{L+1} \sum_{k=0}^{L-1} s(k)s(k+n) \tag{2}$$

When the binary MLS is mapped so that a 1 is represented by -1 and a 0 is represented by $+1$, the autocorrelation of the symmetrical MLS is a periodic unit

impulse sequence [1]

$$\Omega_{ss}(0) = \frac{L}{L+1} \tag{3}$$

$$\Omega_{ss}(n) = -\frac{1}{L+1} \quad 0 < n < L \tag{4}$$

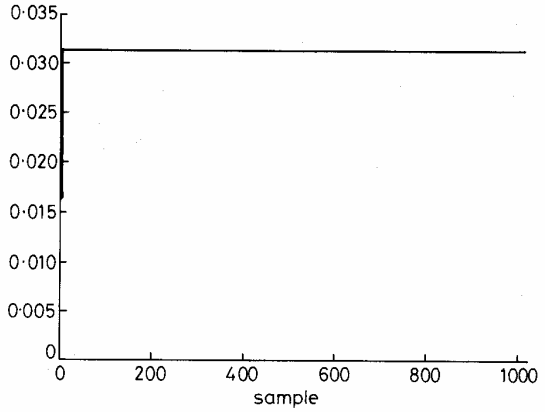


Fig. 2 *MLS frequency magnitude spectrum*

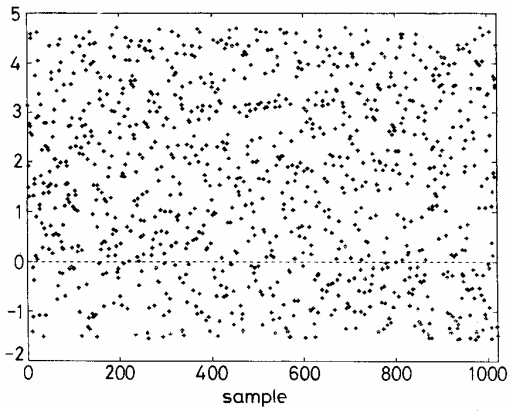


Fig. 3 *MLS frequency phase spectrum*

It can be shown [1] that if the MLS is applied to a system with the PIR of $h(n)$, the crosscorrelation Ω_{sy} of the input $s(n)$ and the output $y(n)$ is

$$\Omega_{sy}(n) = \sum_{k=0}^{L-1} \Omega_{ss}(n)h(n-k) \tag{5}$$

By comparing eqns. 3, 4 and 5, it can be seen that the PIR $h(n)$ can be extracted. A discrete Fourier transform (DFT) on $h(n)$ will then give the linear transfer function.

2.3 Phase randomisation

The nonlinear effects evident after the crosscorrelation are due to the spread of the phases in the frequency mix of the MLS. This mechanism is called phase randomisation and occurs because the crosscorrelation operation is equivalent to the coherent demodulation of the output signal of each particular input frequency component. The nonlinearities will generate frequency components not in the original MLS mix and not following the original phase pattern. For example, if the DUT exhibits some degree of second-harmonic distortion and two frequencies f_a and f_b are input, then the output would contain extra intermodulation frequencies of $(f_a + f_a)$, $(f_b + f_b)$, $(-f_a - f_a)$, $(-f_b - f_b)$, $(f_a - f_b)$ and $(f_b - f_a)$ plus DC. Thus the components of the output signal that are due to nonlinearities will have no correlation with the original MLS and will appear as almost random noise spread over the whole crosscorrelation period.

Distortion artefacts in the PIR do not show a totally random noise-like trace because the phase distribution of the MLS is not totally random. However, through appropriate windowing techniques, the initial PIR portion of the correlation results can be selected, rejecting most of the distortion products.

Phase randomisation will not occur in certain circumstances. For example, if a system consists of a third-order (x^3) nonlinearity, the binary MLS signal (with levels 1 and -1) will pass through the nonlinearity unchanged. Therefore a grossly nonlinear system will be measured as being totally linear. If the MLS is band-limited, however, before application to the DUT, it will be converted to a multilevel signal that will excite the nonlinearity over a large range of values, thus alleviating the above problem. This expedient is followed in this paper.

2.4 Comparison of MLS with other measurement techniques

Two other commonly used impulse response measurement systems are periodic impulse excitation (PIE) and time delay spectrometry (TDS).

PIE is the direct application of a periodic impulse signal to the DUT and measurement of the output. This, in particular, suffers from poor noise immunity because of the low signal excitation energy. To achieve a high signal-to-noise ratio during measurement of the DUT, it is necessary to employ an excitation signal that contains as much energy as possible but does not saturate the DUT. Thus the crest factor (ratio of peak to RMS level) of the excitation needs to be low. Clearly an MLS signal has a much lower crest factor than PIE. Averaging several measurement periods can improve noise immunity (as with MLS), but PIE, unlike MLS, is unable to separate linear and nonlinear components of the DUT in the result, and so it is inappropriate for distortion analysis.

TDS uses a swept frequency sine wave, or 'chirp', excitation, where the response is the complex frequency response of the DUT. TDS has similar noise immunity to MLS but requires longer measurement times and more complex processing for a given resolution [1].

Alternatively, it is possible to achieve the same result as MLS using a white-noise source and a dual channel FFT analyser. A noise generator would be truly-random and would spread nonlinear effects over the whole crosscorrelation period, but requires much longer measurement times and complex window processing.

2.5 Length of MLS

The minimum requirement of MLS length is that the sequence period must be longer than the time it takes for the impulse response of the DUT to decay to zero, thus avoiding aliasing effects that will occur in the crosscorrelation. Ideally, the length should be several times longer than the PIR, increasing the distortion immunity of the PIR as the distortion artefacts are spread more thinly across the correlation period. Also, if any impulsive noise occurs during the measurement, its energy will be spread across the entire time period of the crosscorrelation; thus, the longer the sequence, the greater the transient noise immunity. These benefits must, however, be weighed against increased measurement times and increased processing complexity.

3 MLS techniques

3.1 Linear system

To demonstrate the MLS measuring technique, a digital FIR lowpass filter was modelled on computer. A 1023-point MLS was generated and fed into the filter. The output was crosscorrelated with the MLS using an FHT algorithm. A standard crosscorrelation of two sequences of length N requires N^2 multiplication operations; the FHT calculation reduces this to $2.5N \log_2 N$ additions, a substantial saving of computation time [4]. The result is the impulse response of the filter $h_f(n)$ (see Fig. 4) and, hence, the linear transfer function (Fig. 5). All vertical scales are linear.

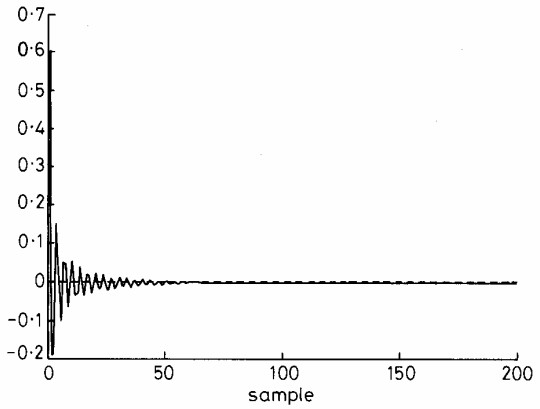


Fig. 4 *Impulse response $h_f(n)$ of lowpass digital filter*

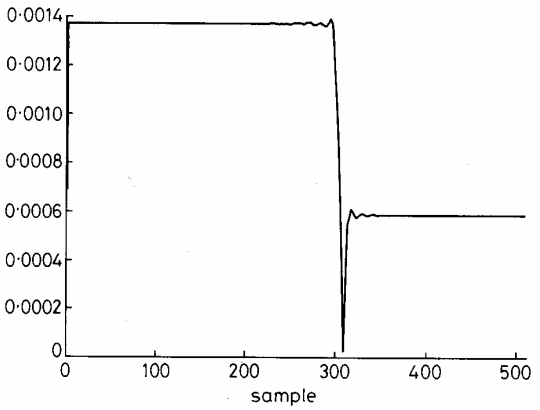


Fig. 5 *Linear transfer function of filter*

3.2 Single nonlinearity

For simple simulation, a weakly nonlinear memoryless system can be modelled in the time domain by a polynomial, thus

$$y(n) = C_1 x(n) + C_2 x^2(n) + C_3 x^3(n) + \cdots \quad (6)$$

where C_1 is a linear gain term, and C_2, C_3 etc. are the nonlinear distortion coefficients, which, for a memoryless system, are constants independent of time or frequency.

This description is based on the Volterra approach to nonlinear system modelling, where all nonlinearities are described as nonlinear kernels, as opposed to the Wiener perspective, which forms the best-fit linear approximation to a nonlinear system [6].

A system with a single second-order nonlinearity can be modelled as

$$y(n) = x(n) + C_2 x^2(n) \quad (7)$$

Here, the circular crosscorrelation Φ between input and output is

$$\begin{aligned}\Omega_{xy}(n) &= x(n) \Phi y(n) \\ &= \frac{1}{L-1} \sum_{k=0}^{L-1} x(k) y(n+k) \\ &= \frac{1}{L-1} \sum_{k=0}^{L-1} x(k)[x(n+k) + C_2 x^2(n+k)] \\ &= \frac{1}{L-1} \sum_{k=0}^{L-1} x(k) x(k+n) \\ &\quad + \frac{C_2}{L-1} \sum_{k=0}^{L-1} x(k) x^2(k+n) \end{aligned} \quad (8)$$

where the first term is the linear impulse response, and the second term is the distortion artefact. Observe that the magnitude of the distortion trace varies linearly with C_2.

For example, Fig. 6 shows a system with a single second-order nonlinearity, and thus $C_1 = 1$, and $C_2 = 0.1$.

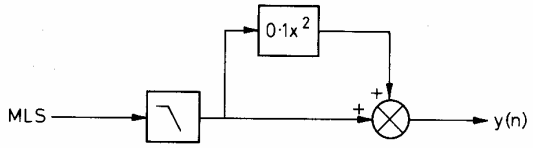


Fig. 6 *System with second-order nonlinearity*

The lowpass filter, described in Section 3.1, is used to band-limit the MLS, to transform it into a multilevel signal to ensure phase randomisation. Fig. 7 shows the

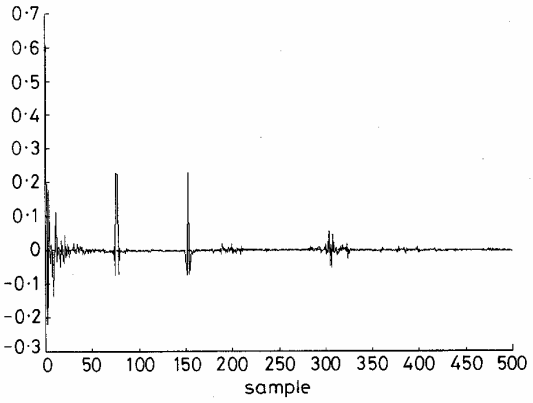


Fig. 7 *Crosscorrelation, with artefacts due to nonlinearity*

crosscorrelation between the input MLS and the output. This is the original impulse response of the filter $h_f(n)$ (Fig. 4), together with noise-like artefacts due to the second-order distortion, described by eqn. 8.

Hence, the crosscorrelation can be decomposed into the summation of the filter impulse response $h_f(n)$ and an error signal $e(n)$ due to distortion, as follows

$$\Omega_{sy}(n) = h_f(n) + e(n) \quad (9)$$

If the result in Fig. 7 is normalised by subtracting the original $h_f(n)$, the impulse error $e(n)$ is revealed (Fig. 8) and is given by

$$e(n) = C_2[s(n) \Phi s_f^2(n)] \quad (10)$$

where $s(n)$ = MLS, and $s_f(n)$ = filtered MLS.

By setting C_2 to unity, the normalised error $e(n)$ can be saved as a template $e_2(n)$, where C_2 is defined as

$$C_2 = \frac{e(n)}{e_2(n)} \quad 0 \leqslant n \leqslant L \tag{11}$$

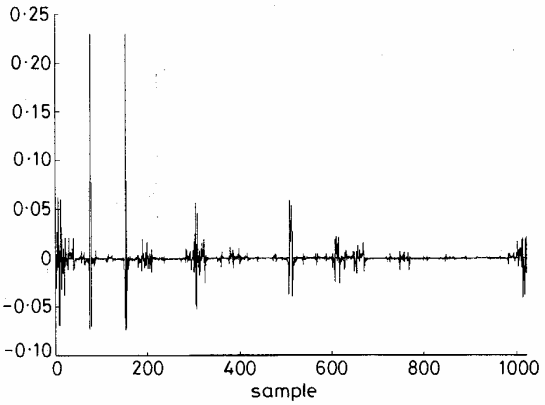


Fig. 8 *Normalised impulse error plot e(n)*

This is only one way of finding C_2 from a set of crosscorrelation results. For a 1023-point MLS, $e(n)$ would also have 1023 points, yet only one point is required to determine C_2. To take into account the possible inaccuracies due to the division by small numbers or random noise, a more robust method of calculating C_2, which takes all points into consideration, is

$$C_2 = \frac{\text{first term of } e(n)\Phi e_2(n)}{\text{first term of } e_2(n)\Phi e_2(n)} = \frac{\Omega_{ee_2}(0)}{\Omega_{e_2e_2}(0)} \tag{12}$$

This method can be generalised for higher orders of nonlinearity, where each order has a distinct template or 'fingerprint', for a particular MLS, that can be compiled into a library of templates. Hence, for a single nonlinearity of order m, where the template e_m is available, the coefficient of an mth-order nonlinearity is

$$C_m = \frac{\Omega_{ee_m}(0)}{\Omega_{e_m e_m}(0)} \tag{13}$$

3.3 Multiple nonlinearities

Consider a more general system

$$y(n) = x(n) + C_2 x^2(n) + C_3 x^3(n) \tag{14}$$

where, by inspection of eqns. 8 and 10, $e(n)$ follows as

$$\begin{aligned} e(n) &= C_2[s(n)\Phi s_f^2(n)] + C_3[s(n)\Phi s_f^3(n)] \\ &= C_2 e_2(n) + C_3 e_3(n) \end{aligned} \tag{15}$$

Thus the resulting $e(n)$ is the summation of the respective amounts of template for each of the nonlinear orders present. For example, Fig. 9 shows a system with second- and third-order nonlinearities ($C_1 = 1$, $C_2 = 1$, $C_3 = 1$). The crosscorrelation was performed, and the normalised impulse error plot $e(n)$ is shown in Fig. 10.

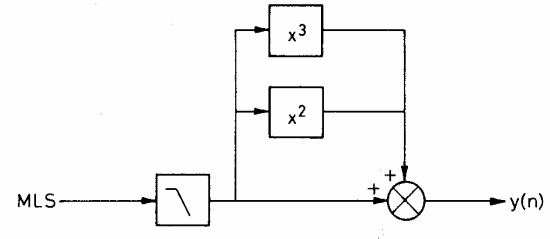


Fig. 9 *System with 2nd- and 3rd-order nonlinearities*

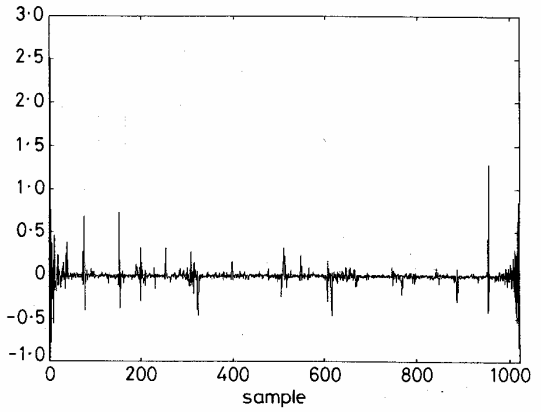


Fig. 10 *e(n) for 2nd- and 3rd-order summation*

To decode the two values of C_2 and C_3, it is necessary to crosscorrelate with the two templates $e_2(n)$ and $e_3(n)$, as follows

$$\begin{aligned} &e(n)\Phi e_2(n) = C_2 e_2(n)\Phi e_2(n) + C_3 e_3(n)\Phi e_2(n) \\ &\quad \rightarrow \Omega_{ee_2}(n) = C_2 \Omega_{e_2e_2}(n) + C_3 \Omega_{e_2e_3}(n) \end{aligned} \tag{16}$$

and

$$\begin{aligned} &e(n)\Phi e_3(n) = C_2 e_2(n)\Phi e_3(n) + C_3 e_3(n)\Phi e_3(n) \\ &\quad \rightarrow \Omega_{ee_3}(n) = C_2 \Omega_{e_2e_3}(n) + C_3 \Omega_{e_3e_3}(n) \end{aligned} \tag{17}$$

Hence, the coefficients C_2 and C_3 can be found by solving the following simultaneous equations

$$C_2 = \frac{\Omega_{e_2e}(0) - C_3 \Omega_{e_2e_3}(0)}{\Omega_{e_2e_2}(0)} \tag{18}$$

$$C_3 = \frac{\Omega_{e_3e}(0) - C_2 \Omega_{e_2e_3}(0)}{\Omega_{e_3e_3}(0)} \tag{19}$$

This method can be extended to cover higher-order polynomials and, using matrix techniques to solve the simultaneous equations, there is no limit to the number of coefficients that can be determined, providing the template library is appropriate.

3.4 Alternative distortion types

The MLS technique can recover the coefficients of the memoryless polynomial models used for illustration. However, most real devices show more varied types of distortion, such as exponential or crossover distortion. A simple exponential model is simulated using the following relationship

$$y(n) = \alpha(e^{\beta x(n)} - 1) \tag{20}$$

Setting $\alpha = 1$ and $\beta = 1$ gives the impulse error plot $e(n)$ shown in Fig. 11.

Varying α will change the magnitude of $e(n)$, but a change in β will alter the shape of $e(n)$, so that it is impossible to have a single template for exponential distortion, as β is continuously variable. It is still possible to measure, however, the distortion introduced by the exponential term.

If the multiple nonlinearity measurement process described in Section 3.3 is used on the $e(n)$ of Fig. 11, the result is to give the coefficients of the power series equivalent of eqn. 20. The process is unable to differentiate

between simple types of nonlinearity, i.e. second- or third-order, and more complex exponential type nonlinearities, but it is still able to give an accurate polynomial model of DUT.

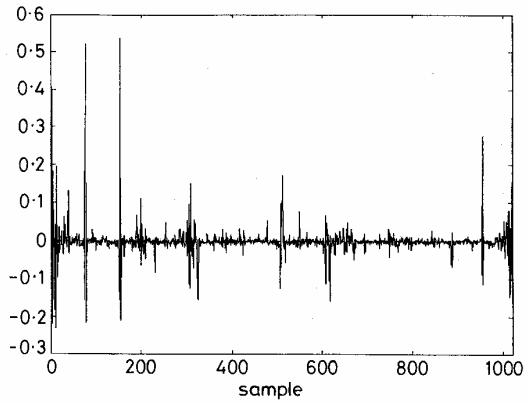


Fig. 11 *e(n) for exponential nonlinearity*

Thus a complete model can be calculated regardless of the type of distortion exhibited by the DUT, providing that the distortion type has an equivalent power series expansion.

3.5 Complex coefficients

We now consider a system with memory: for example, where a nonlinearity is within a feedback loop, the output contains extended distortion harmonics that are phase-shifted to varying degrees. The MLS process described is a method that can recover an overall open-loop model (as in eqn. 6) of the closed-loop system, but the model needs to be extended to encompass both the magnitudes and phases of the distortion products. Complex coefficients can be used in the polynomial model to describe the required parameters, as follows

$$y(n) = (C_1 + c_1 j)x(n) + (C_2 + c_2 j)x^2(n) + (C_3 + c_3 j)x^3(n) + \cdots \quad (21)$$

Consider the following system

$$y(n) = x(n) + C_2 x^2(n) + c_2 j x^2(n) \quad (22)$$

The resulting $e(n)$ after crosscorrelation for such a system is therefore

$$e(n) = C_2 e_2(n) + c_2 j e_2(n) \quad (23)$$

Comparing with eqn. 15, there are two coefficients and two templates required: $e_2(n)$ and an orthogonal version of $e_2(n)$. The Hilbert transform of a template for the order m gives the orthogonal template $e_{Hm}(n)$ and is given by

$$e_{Hm}(n) = \frac{1}{\pi} \sum_{\substack{k=0 \\ k \neq n}}^{L-1} \frac{e_m(n)[1 - e^{j\pi(n-k)}]}{(n-k)} \quad (24)$$

The calculation of the two coefficients is an identical operation to that of Section 3.3, i.e., for an unknown distortion component given by $(C_m + c_m j)x^m(n)$, we must have the template $e_m(n)$ and its Hilbert transform $e_{Hm}(n)$ and solve the following simultaneous equations

$$C_m = \frac{\Omega_{e_m e}(0) - c_m \Omega_{e_m e_{Hm}}(0)}{\Omega_{e_m e_m}(0)} \quad (25)$$

and

$$c_m = \frac{\Omega_{e_{Hm} e}(0) - C_m \Omega_{e_m e_{Hm}}(0)}{\Omega_{e_{Hm} e_{Hm}}(0)} \quad (26)$$

Again, using matrix techniques, many simultaneous equations can be solved for $e(n)$s containing distortions of many orders with varying complex coefficients.

The technique was tested on a simulated amplifier with feedback, where the gain stage within the closed loop exhibited a degree of nonlinearity. An MLS was fed into the amplifier and crosscorrelated with the output. From the measured PIR, coefficients were calculated to form the open-loop model of the closed-loop system. To verify the result, a sine wave signal was then put through both the amplifier and the open-loop model, and the two outputs were found to be comparable. Thus the MLS measurement produced an accurate model of the amplifier system.

3.6 Noise immunity

Noise present in the output signal of the DUT is spread across the result of the crosscorrelation along with the distortion artefacts, and this affects the ability of the template technique to resolve accurate coefficient values for the DUT model. To study the effects of noise on the measurement system, the following is used

$$y(n) = C_2(\alpha N(n) + x_f(n))^2 \quad (27)$$

where $x_f(n)$ = filtered MLS; and $N(n)$ = random noise of approximately the same RMS value as $x(n)$.

The objective is to recover the coefficient C_2 under varying amounts of additive noise. Using $C_2 = 0.5$, for each value of α the simulation was run, and the second-order coefficient was calculated 500 times using eqn. 12. Fig. 12 shows the probability density function (PDF)

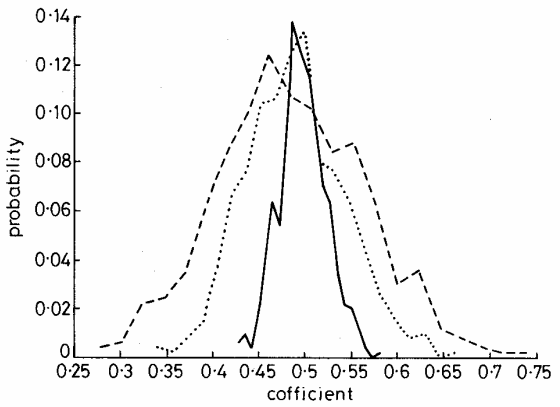


Fig. 12 *PDF of coefficient C_2 for varied amounts of added noise*
——— $\alpha = 0.5$
····· $\alpha = 1.0$
– – – – $\alpha = 1.5$

plots of the coefficient calculated for three values of α. This result demonstrates that, even for large amounts of additive noise ($\alpha = 1.5$), with a sufficient number of repeated measurements the average of the coefficient values will converge to the correct result of 0.5.

Standard pre-averaging practice involves repeating the test several times and averaging the outputs as the test is repeated. For example, the recovery of the PIR when using the MLS techique would be described by

$$\text{PIR} = x(n) \Phi \overline{y(n)} \quad (28)$$

where $x(n)$ is the MLS. The coefficient(s) are then calculated from the PIR.

Fig. 13 shows the convergence of the calculated coefficient using the model of eqn. 27, with $C_2 = 0.5$ and

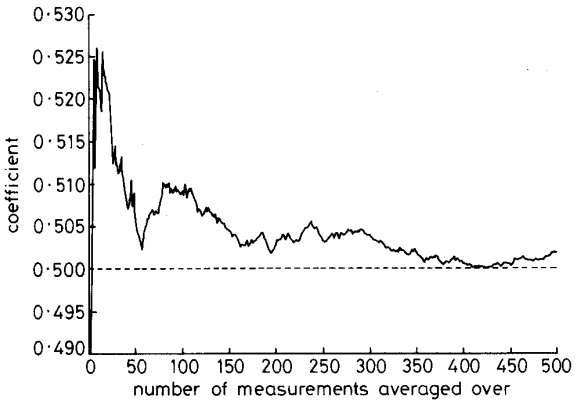


Fig. 13 *C_2 convergence to 0.5 when averaging before crosscorrelation*

$\alpha = 1.0$ as the number of measurements used for the average $y(n)$ increases. In this case, the coefficient value does converge to the required value of 0.5, but with further averaging it diverges. To achieve full convergence, it is necessary to move the point of averaging to after the crosscorrelation operation. Thus eqn. 28 becomes

$$\text{PIR} = \overline{x(n)\Phi y(n)} \tag{29}$$

A repetition of the above experiment (see Fig. 14) shows that convergence results.

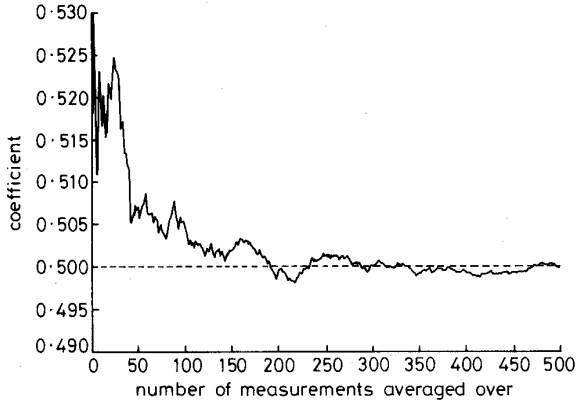


Fig. 14 *C_2 convergence to 0.5 when averaging after crosscorrelation*

Normal pre-averaging fails in this case because of the nonlinearities present that modify the noise, preventing a zero mean. Following crosscorrelation, the resulting noise is effectively re-randomised, enabling averaging to be effective.

The MLS distortion analysis technique is therefore viable in the presence of noise when appropriate averaging practice is employed.

4 Conclusion

Results from an MLS measurement analysis can predict both linear and nonlinear behaviour of the DUT. The technique was tested on a simulated amplifier with feedback, where the gain stage within the closed loop exhibited a degree of nonlinearity.

From the MLS result, coefficients were calculated for the open-loop model of the closed-loop system. This model could then accurately calculate the distortion harmonic structure exhibited by the amplifier when excited by a sine wave of any amplitude or frequency.

The MLS test is an efficient process that uses a repeatable and deterministic driving signal and has substantial noise and distortion immunity advantages over other techniques. For a simple, weakly nonlinear system a complete model can be generated from a single analysis, and its response to standard tests such as total harmonic and intermodulation distortion can then readily be evaluated. However, the results presented are based on simulations, and further work on real systems is a priority to understand fully the MLS measurement process.

5 References

1 RIFE, D.D., and VANDERKOOY, J.: 'Transfer-function measurement with maximum-length sequences', *J. Audio Eng. Soc.*, 1989, **37,** pp. 419–444
2 VANDERKOOY, J.: 'Aspects of MLS measuring systems', *J. Audio Eng. Soc.*, 1994, **42,** pp. 219–231
3 DUNN, C., and HAWKSFORD, M.O.: 'Distortion immunity of MLS-derived impulse response measurements', *J. Audio Eng. Soc.*, 1993, **41,** pp. 314–334
4 BORISH, J., and ANGELL, J.B.: 'An efficent algorithm for measuring the impulse response using pseudorandom noise', *J. Audio Eng. Soc.*, 1983, **31,** pp. 478–488
5 GREEST, M.C., and HAWKSFORD, M.O.: 'Nonlinear distortion analysis using maximum length sequences', *Electron. Lett.*, 1994, **30,** (13), pp. 1033–1035
6 DUNN, C., and RIFE, D.: 'Comments on "Distortion immunity of MLS-derived impulse response measurements"', *J. Audio Eng. Soc.*, 1994, **42,** pp. 490–497

Identification of discrete Volterra series using maximum length sequences

M.J.Reed
M.O.J.Hawksford

Indexing terms: Loudspeaker modelling, Nonlinear system modelling, Volterra algorithm

Abstract: An efficient method is described for the determination of the Volterra kernels of a discrete nonlinear system. It makes use of the Wiener general model for a nonlinear system to achieve a change of basis. The orthonormal basis required by the model is constructed from a modified binary maximum sequence (MLS). A multilevel test sequence is generated by time reversing the MLS used to form the model and suitably summing delayed forms of the sequence. This allows a sparse matrix solution of the Wiener model coefficients to be performed. The Volterra kernels are then obtained from the Wiener model by a change of basis.

1 Introduction

The Volterra series [1] has been used to model nonlinear systems such as electrodynamic loudspeakers [2] and the human auditory system [3]. It expands the impulse response model of a linear system by representing nonlinearity as a set of higher-order impulse responses termed kernels. In its discrete form the Volterra series of a causal system may be expressed as [4]

$$y(n)=\sum_{r=1}^{N}\sum_{i_1=0}^{M-1}\cdots\sum_{i_r=0}^{M-1}h_r(i_1,i_2,\ldots,i_r)u(n-i_1)u(n-i_2)\ldots u(n-i_r) \quad (1)$$

where u and y are the input and output, respectively, and h_r is the rth-order kernel. It has been shown that a wide class of nonlinear systems may be represented as such a Volterra filter of finite order N and memory M for a finite range of excitation [4].

Existing methods for the identification of Volterra kernels have proved computationally burdensome. Many use the crosscorrelation of gaussian random variables [5] requiring long sequence lengths as the result converges for an infinite sequence. Binary maximum length sequences (MLS) have been used as excitation [3] but, as they do not fully excite even a simple nonlinear system, cannot fully identify a Volterra filter [6]. Nowak and Van Veen use a multilevel MLS to fully excite a Volterra filter however, the method requires very long test sequences owing to the extended filter approach [6].

IEE Proceedings online no. 19960726

Paper first received 20th November 1995 and in revised form 1st July 1996

The authors are with the Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

This paper describes a method that efficiently determines the kernels of a nonlinear system that can be represented as a Volterra filter. It makes use of the general Wiener model of a nonlinear system as shown in Fig. 1 [7]. This decomposes the nonlinearity into a parallel set of linear filters which comprise all of the time-varying part of the model, termed the memory of the system. The nonlinear part of the model comprises every possible rth order product of the filter outputs for each order r. The outputs are then multiplied by a coefficient and summed to give the overall model. The impulse responses of the linear filters form an orthonormal set for which Wiener used Laguerre functions to aid identification when gaussian distributed noise is used to excite the system. The model topology and filter impulse responses are predefined so that only the coefficient values need be determined to model a system.

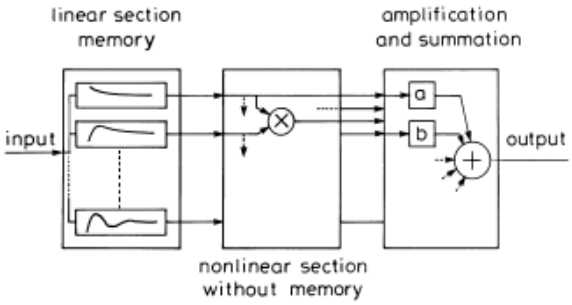


Fig.1 *General Wiener model of nonlinear system*

This paper demonstrates that the discrete Volterra filter can be represented as a Wiener model using a finite set of discrete functions. It then shows how these functions can be made from a single binary MLS. The MLS that forms the linear section of the model is used to generate a multilevel sequence that allows the coefficients of the Wiener model to be efficiently determined. The kernels of the Volterra model can then be obtained by a change of basis.

2 Expression of Volterra series using orthonormal basis

Define a column vector $\mathbf{h}_r$

$$\mathbf{h}_r=\begin{bmatrix} h_r(0,0,\ldots,0) \\ h_r(0,0,\ldots,1) \\ \vdots \\ h_r(M-1,M-1,\ldots,M-2) \\ h_r(M-1,M-1,\ldots,M-1) \end{bmatrix}$$

and a row vector u_n

$$\mathbf{u}_n = [u(n), u(n-1), u(n-2), \ldots, u(n-M+1)]$$

All the *r*th-order products of the input from *u*(*n*) to *u*(*n* – *M* + 1) may be expressed as the *r*th-order Kronecker product (denoted as the superscript $^{[r]}$) of u_n with itself

$$\underbrace{\mathbf{u}_n \otimes \mathbf{u}_n \otimes \ldots \otimes \mathbf{u}_n}_{r\,\text{times}} = \mathbf{u}_n^{[r]}$$

where ⊗ is the Kronecker product between two matrices, a review of which is given in [8, 9]. Then eqn. 1 can be represented in vector notation as

$$y(n) = \sum_{r=1}^{N} \mathbf{u}_n^{[r]} \mathbf{h}_r \tag{2}$$

Define a matrix M of dimension $M \times M$ with columns that form an orthonormal basis which spans the space $\underline{R}^M$. Consider the product of two matrices $A^{[r]}$ and $B^{[r]}$, by the mixed product rule [8]

$$\begin{aligned} \mathbf{A}^{[r]}\mathbf{B}^{[r]} &= (\mathbf{A}^{[r-1]} \otimes \mathbf{A})(\mathbf{B}^{[r-1]} \otimes \mathbf{B}) \\ &= (\mathbf{A}^{[r-1]}\mathbf{B}^{[r-1]}) \otimes (\mathbf{AB}) \end{aligned} \tag{3}$$

Repeated application of eqn. 3 obtains the following:

$$\mathbf{A}^{[r]}\mathbf{B}^{[r]} = (\mathbf{AB})^{[r]} \tag{4}$$

As M is orthonormal $MM^T = I_M$ where I_M is the $M \times M$ identity matrix. Then applying eqn. 4 gives

$$\mathbf{M}^{[r]}(\mathbf{M}^{\mathbf{T}})^{[r]} = \mathbf{I}_M^{[r]} = \mathbf{I}_{M^r}$$

Thus the *r*th-order Kronecker product of M forms an orthonormal basis spanning the space $\underline{R}^{M^r}$. This allows the *r*th-order Volterra kernel h_r which is of dimension M^r to be represented as g_r, a column vector of the same dimension, by

$$\mathbf{h}_r = (1/\alpha)^r \mathbf{M}^{[r]} \mathbf{g}_r \tag{5}$$

where $1/\alpha$ is a constant that is defined later.

Using the symmetry of the kernels of a Volterra series [10] it is possible to reduce the size of the vectors for $r > 1$. If each row of h_r is labelled by an ordered set $<i_1, i_2, \ldots, i_r>$ representing the point $h_r(i_1, \ldots, i_r)$ the rows that contain the same elements in the set irrespective of order are all points of symmetry in h_r. Thus there are only *l* unique rows in h_r given by the binomial [11]

$$l = \binom{M+r-1}{r} \tag{6}$$

Define a matrix Q_r which sums the symmetrical points in h_r removing any duplicate rows such that we are left with rows $<i_1, i_2, \ldots, i_r>$ with the condition $i_1 \le i_2 \le \ldots \le i_r$. Q_r is a {1, 0} matrix, that is it only contains 1 or 0 as its elements. This is demonstrated for a simple case shown below:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} h_2(0,0) \\ h_2(0,1) \\ h_2(1,0) \\ h_2(1,1) \end{bmatrix} = \begin{bmatrix} h_2(0,0) \\ h_2(0,1) + h_2(1,0) \\ h_2(1,1) \end{bmatrix}$$

$$\mathbf{Q}_2\mathbf{h}_2 = \mathbf{h}_{2(reduced)}$$

The formation of $u_n^{[r]}$ by the Kronecker product introduces the same symmetries that allowed the removal of the rows in h_r. Thus the corresponding columns of $u_n^{[r]}$ can be removed by the matrix we define as P_r again a {1, 0} matrix. This operation does not need to add the symmetrical columns just remove them. The following example illustrates the operation:

$$\begin{aligned} &[u(1)u(1), u(1)u(0), u(0)u(1), u(0)u(0)] \begin{bmatrix} 100 \\ 010 \\ 000 \\ 001 \end{bmatrix} \\ &\quad = [u(1)u(1), u(1)u(0), u(0)u(0)] \\ &\mathbf{u}_1^{[2]}\mathbf{P}_2 = (\mathbf{u}_1^{[2]})_{(reduced)} \end{aligned}$$

Finally we incorporate the change of basis and the reduction due to symmetry to give an alternative form of the Volterra filter that was defined by eqn. 2 as

$$\begin{aligned} y(n) &= \sum_{r=1}^{N} (1/\alpha)^r \mathbf{u}_n^{[r]} \mathbf{M}^{[r]} \mathbf{P}_r \mathbf{Q}_r \mathbf{g}_r \\ &= \sum_{r=1}^{N} (1/\alpha)^r (\mathbf{u}_n \mathbf{M})^{[r]} \mathbf{P}_r \mathbf{Q}_r \mathbf{g}_r \end{aligned} \tag{7}$$

This is in fact equivalent to the Wiener general model described by Schetzen [7] and shown in Fig. 1. The columns of matrix M are equivalent to the set of parallel filters in the model which are convolved with the input. The Kronecker product of the resulting matrix is equivalent to all the possible *r*th-order products, each of which is multiplied by a weight given as elements of g_r. The result is then summed to give the scaler output at sample *n*.

Define a partitioned matrix A

$$\mathbf{A} = \left[\begin{array}{c|c|c|c} 1/\alpha \mathbf{u}_0 \mathbf{M} \mathbf{P}_1 & (1/\alpha \mathbf{u}_0 \mathbf{M})^{[2]} \mathbf{P}_2 & \ldots & (1/\alpha \mathbf{u}_0 \mathbf{M})^{[N]} \mathbf{P}_N \\ 1/\alpha \mathbf{u}_1 \mathbf{M} \mathbf{P}_1 & (1/\alpha \mathbf{u}_1 \mathbf{M})^{[2]} \mathbf{P}_2 & & (1/\alpha \mathbf{u}_1 \mathbf{M})^{[N]} \mathbf{P}_N \\ \vdots & \vdots & & \vdots \\ 1/\alpha \mathbf{u}_{L-1} \mathbf{M} \mathbf{P}_1 & (1/\alpha \mathbf{u}_{L-1} \mathbf{M})^{[2]} \mathbf{P}_2 & & (1/\alpha \mathbf{u}_{L-1} \mathbf{M})^{[N]} \mathbf{P}_N \end{array}\right] \tag{8}$$

and a partitioned column vector g

$$\mathbf{g} = \begin{bmatrix} \mathbf{Q}_1 \mathbf{g}_1 \\ \hline \mathbf{Q}_2 \mathbf{g}_2 \\ \hline \vdots \\ \hline \mathbf{Q}_N \mathbf{g}_N \end{bmatrix}$$

From eqn. 6 the dimension of g is given by the function $f(M, N)$ defined as

$$f(M,N) = \sum_{s=1}^{N} \binom{M+s-1}{s} \tag{9}$$

Then the output vector representing the output from sample 0 to $L - 1$

$$\mathbf{y} = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(L-1) \end{bmatrix}$$

is given by

$$\mathbf{y} = \mathbf{A}\mathbf{g} \tag{10}$$

If $L = f(M, N)$ then A is square, and assuming it is nonsingular, the LMS solution of eqn. 10 giving the required coefficients is

$$\mathbf{A}^{-1}\mathbf{y} = \mathbf{g} \tag{11}$$

The inversion of A poses a problem as it can be of very large dimension for a Volterra filter of any but trivial systems. The input vectors u_n and the basis M are defined such that A is sparse and also nonsingular thus, allowing the coefficients to be determined efficiently.

3 Forming an orthonormal basis from an MLS

Binary maximum length sequences are a class of pseudorandom sequences that have an autocorrelation function that approaches a delta function and can be simply generated using a recursive shift register [12]. These properties have been widely used for the measurement of linear systems using MLS excitation [13]. Unfortunately the higher-order autocorrelation function is not as precise so preventing the identification of nonlinear systems using crosscorrelation methods with MLS excitation [14].

We make use of the first-order autocorrelation function of a binary MLS to construct an orthonormal basis for use with the Wiener model discussed previously. Let $w_0(n)$ be a binary MLS which repeats every M samples where $M = 2^q - 1$ for q a positive integer. Then let $w_i(n)$ be $w_0(n)$ shifted by i places, $w_i(n) = w_0(n - i)$. The crosscorrelation between any two of the functions $w_i(n)$ and $w_j(n)$ is [12]

$$R(w_i, w_j) = \sum_{n=0}^{M-1} w_i(n)w_j(n) = \begin{cases} M & i = j \\ -1 & i \neq j \end{cases} \qquad (12)$$

An orthonormal set of functions is needed for which there is a requirement that $R(w_i, w_j) = 0$ for $i \neq j$. For linear measurements this problem of true orthonormality is usually ignored as for large M the result is only a small DC offset in an impulse response measurement. In fact Vanderkooy [15] states that, for linear measurements, not correcting the DC offset is an advantage for systems which possess some even-order nonlinear errors. However, the nonlinear performance is taken into account and strict orthonormality is required for the construction of a basis so the DC offset is corrected using the procedure discussed subsequently.

Introducing a scale factor a and a constant c obtains the function $m_i(n) = a(w_i(n) + c)$ with the corresponding crosscorrelation function

$$R(m_i, m_j) = \sum_{n=0}^{M-1} a^2(w_i(n) + c)(w_j(n) + c) = \begin{cases} b & i = j \\ d & i \neq j \end{cases}$$
$$= a^2 \sum_{n=0}^{M-1} w_i(n)w_j(n) + a^2 \sum_{n=0}^{M-1} w_i(n)c$$
$$+ a^2 \sum_{n=0}^{M-1} w_j(n)c + a^2 \sum_{n=0}^{M-1} c^2$$

If the most frequent symbol of the MLS is cast to +1 and the least to –1 then $\Sigma_{n=0}^{M-1} w_i(n) = 1$. For the functions to be orthonormal requires $b = 1$ and $d = 0$ giving the following simultaneous equations:

$$b = 1 \text{ implies } 1 = a^2M + 2a^2c + Ma^2c^2$$
$$d = 0 \text{ implies } 0 = -a^2 + 2a^2c + Ma^2c^2$$

Solving gives

$$a = \frac{1}{\sqrt{M+1}} \qquad c = \frac{-1 \pm \sqrt{1+M}}{M} \qquad (13)$$

We choose the positive solution of c as it produces the smallest offset but there is no reason against using the negative solution, essentially the choice being arbitrary.

There are M possible functions $m_i(n)$ which form an orthonormal set, each is used as the ith column of the matrix M so constructing the orthonormal basis required by the Wiener model. Introducing an MLS restricts the memory of the model to the length of an MLS given by $M = 2^q - 1$ for q a positive integer.

4 Solution of Volterra filter coefficients

To solve for the coefficients in the model given by eqn. 11 the matrix A needs to be as sparse as possible. To demonstrate how this is achieved consider a model using an MLS of length seven samples. This has a linear section made up of the seven possible orthonormal functions $m_i(n)$, i = 0... 6 which make up the columns of the matrix M. Fig. 2 shows the result of applying an input sequence which is a time reversed form of $m_0(n)$.

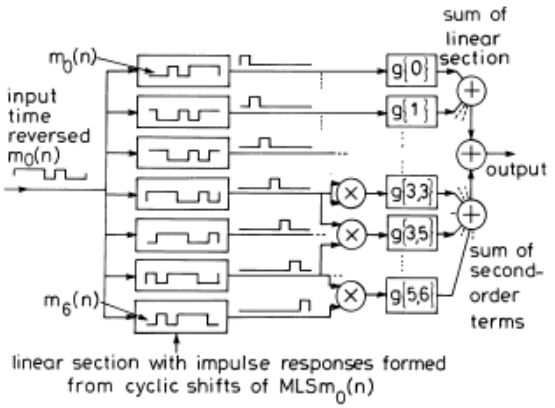


Fig. 2 *Example of Wiener model using MLS to form linear filters*

The outputs of the linear section of the example are a set of delta functions. Thus, the output of the overall model at one sampling instant is dependent on only a limited set of the coefficients. By applying the correct combination of the reversed functions $m_i(n)$ summed appropriately it is possible to determine all of the coefficients. Furthermore, we show that the derived set of excitations form a minimal set which allows the L unique Volterra filter coefficients to be determined with only L observations.

4.1 Input sequence definition to make matrix A sparse

To refer to the rows (or columns) of the partitioned matrices we define an unordered set $I = \{i_1, ..., i_n\} = \{i: 0 \leq i < M\}$. An n-element set is required for each row of the nth partition with all possible n sets in the nth partition. If the elements of each set are arranged such that $i_1 \leq i_2 \leq ... i_n$ the sets increase in natural order with the first row in each partition labelled by the all zero set. A row in the nth partition of a matrix B is then defined as $b\{i_1, ..., i_n\}$. The relationship $I \subset J$ denotes I a proper subset of J whereas $I \subseteq J$ denotes I a subset of J.

Define a matrix X that is partitioned into N parts and forms the excitation matrix which consists of L rows, with $L = f(M, N)$.

$$\mathbf{X} = \begin{bmatrix} \alpha\mathbf{m}_0^T \\ \alpha\mathbf{m}_1^T \\ \vdots \\ \alpha\mathbf{m}_{M-1}^T \\ \hline \alpha(\mathbf{m}_0^T + \mathbf{m}_0^T) \\ \alpha(\mathbf{m}_0^T + \mathbf{m}_1^T) \\ \vdots \\ \alpha(\mathbf{m}_{M-1}^T + \mathbf{m}_{M-1}^T) \\ \hline \vdots \\ \hline \alpha(\mathbf{m}_0^T + \mathbf{m}_0^T + \ldots + \mathbf{m}_0^T) \\ \vdots \\ \alpha(\mathbf{m}_{M-1}^T + \mathbf{m}_{M-1}^T + \ldots + \mathbf{m}_{M-1}^T) \end{bmatrix}$$

Ith row of partition s of $\mathbf{X}$ given by

$$\mathbf{x}\{i_1, \ldots, i_s\} = \alpha \sum_{k=1}^{s} \mathbf{m}_{i_k}^T$$

where m_j is the jth column of the matrix M. Each row is the sum of various combinations of the sequences $m_j(n)$, all such sums up to the Nth order are included in X. The constant α allows the input sequence to have an arbitrary amplitude as is required with measuring practical systems. Manipulating X into the required Toeplitz form is discussed later.

The rows of X form the vectors u_i in the matrix A defined in eqn. 8. If the ith row of A is defined by the set $I = \{i_1, i_2, ..., i_s\}$ and the jth column by the set $J = \{j_1, j_2, ..., j_r\}$ using our defined notation, the element a_{ij} of A is given by

$$a_{ij} = \mathbf{x}_i^{[r]}(1/\alpha)^r(\mathbf{m}_{j_1} \otimes \mathbf{m}_{j_2} \otimes \ldots \otimes \mathbf{m}_{j_r})$$

From the definition of X and by applying the mixed product rule [8]

$$\begin{aligned} a_{ij} &= \sum_{k=1}^{s} \mathbf{m}_{i_k}^T \mathbf{m}_{j_1} \otimes \sum_{k=1}^{s} \mathbf{m}_{i_k}^T \mathbf{m}_{j_2} \otimes \ldots \otimes \sum_{k=1}^{s} \mathbf{m}_{i_k}^T \mathbf{m}_{j_r} \\ &= \sum_{k=1}^{s} \delta(i_k - j_1) \sum_{k=1}^{s} \delta(i_k - j_2) \ldots \sum_{k=1}^{s} \delta(i_k - j_r) \end{aligned} \tag{14}$$

where $\delta(n)$ is the Kronecker delta. Thus $a_{ij} = 0$ unless

$$\{j_1, \ldots, j_r\} \subseteq \{i_1, \ldots, i_s\} \tag{15}$$

By including the scale factor α in eqn. 5 A has been made independent of the input amplitude. As an example, Fig. 3 shows the matrix A for $N = 3$ and $M = 3$.

4.2 Efficient method of inverting matrix A

As we have an excitation matrix with L rows with $L = f(M, N)$ the LMS solution of eqn. 11 is met however, the matrix A cannot be easily inverted using existing routines. Thus the solution of eqn. 11 is achieved by utilising the sparseness of A, breaking up the problem into smaller stages which are carried out in many blocks by back substituting the solution of earlier stages. The solution is carried out in N stages, each stage p will be broken up into many blocks n. The rows of the example matrix shown in Fig. 3 are labelled with the appropriate solution stage and block.

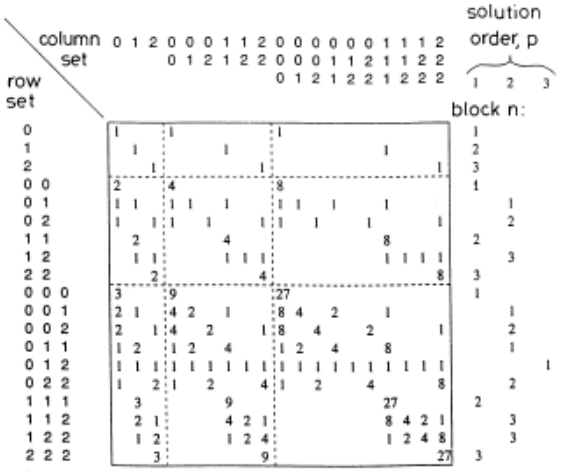


Fig.3 *Example of matrix A for N = 3 and M = 3*

We now give the solution for the general case of stage p and block n. The rows and columns of A are tagged by sets I and J as defined previously. Consider a set: $K_{pn} = \{k_1, ..., k_p\} = \{k: 0 \le k < M$, no repetition of elements$\}$. As there are

$$\binom{M}{p}$$

possible combinations [11],

$$n = 1 \ldots \binom{M}{p}$$

Select a set of rows R_{pn} of the matrix A for which $R_{pn} = \{I : I = K_{pn}\}$. Define a vector y_{pn} which contains the rows of y that correspond to these selected rows. The number of rows is given by

$$\sum_{r=p}^{N} \binom{p+(r-p)-1}{r-p} = \sum_{r=0}^{N-p} \binom{p+r-1}{r} = 1 + f(p, N-p)$$

using the function defined by eqn. 9.

The rows of g are each labelled by a set J as they commute with the columns of A. Define a vector $g_{(new)pn}$ consisting of the rows of g given by the set of rows $new_{pn} = \{J:J = K_{pn}\}$. As $new_{pn} = R_{pn}$ the dimension of $g_{(new)pn}$ is $1 + f(p, N - p)$. Similarly define $g_{(old)pn}$ from the set of rows $old_{pn} = \{J:J \subset K_{pn}\}$, of dimension $f(p, N) - (1 + f(p, N - p))$ which is zero for $p = 1$.

From eqn. 15 the selected rows in A are nonzero only in the columns for which $J \subseteq I$. As $I = K_{pn}$ then the nonzero condition for the selected rows is $J \subseteq K_{pn}$ giving a set of nonzero columns: $C_{pn} = \{J:J \subseteq K_{pn}\}$. As $C_{pn} = old_{pn} + new_{pn}$ all the nonzero elements in the selected rows commute with either $g_{(new)pn}$ or $g_{(old)pn}$. Hence, define two submatrices of A using the nonzero elements of the selected rows giving $A_{(old)p}$ which commutes with $g_{(old)pn}$ and $A_{(new)p}$ that commutes with $g_{(new)pn}$. The elements of $A_{(old)p}$ and $A_{(new)p}$ are given by the relationship eqn. 15 which depends on the intersection of set elements and not their actual value. Thus, the values of the elements depend only on the stage p and the maximum order N, dividing the matrix A into N pairs of identical submatrices.

Now a relationship between these defined subblocks can be given by

$$\mathbf{y}_{pn} = \mathbf{A}_{(old)p}\mathbf{g}_{(old)pn} + \mathbf{A}_{(new)p}\mathbf{g}_{(new)pn} \quad (16)$$

for example from Fig. 3 for $p = 2$ and $n = 1$

$$\begin{bmatrix} \mathbf{y}\{0,1\} \\ \mathbf{y}\{0,0,1\} \\ \mathbf{y}\{0,1,1\} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 4 & 1 & 8 & 1 \\ 1 & 2 & 1 & 4 & 1 & 8 \end{bmatrix} \begin{bmatrix} \mathbf{g}\{0\} \\ \mathbf{g}\{1\} \\ \mathbf{g}\{0,0\} \\ \mathbf{g}\{1,1\} \\ \mathbf{g}\{0,0,0\} \\ \mathbf{g}\{1,1,1\} \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix} \begin{bmatrix} \mathbf{g}\{0,1\} \\ \mathbf{g}\{0,0,1\} \\ \mathbf{g}\{0,1,1\} \end{bmatrix}$$

For the general case eqn. 16 can be expressed as

$$\mathbf{g}_{(new)pn} = \mathbf{A}^{-1}_{(new)p}\left(\mathbf{y}_{pn} - \mathbf{A}_{(old)p}\mathbf{g}_{(old)pn}\right) \quad (17)$$

For $p > 1$ the elements of $g_{(old)pn}$ contain only elements from the vectors

$$\mathbf{g}_{(new)rq} \text{ for } r < p, q = 1 \ldots \binom{M}{r}$$

Thus if the operation given in eqn. 17 is carried out in the order $p = 1 \ldots N$, for all the possible values of n for each p, the solution of g will be formed. The total operation count for the process is of the order

$$\sum\nolimits_{p=1}^{N} \binom{M}{p} f(p,N)(1 + f(p, N-p))$$

not including any necessary book keeping.

The set of matrices $A_{(old)P}$ and $A^{-1}_{(new)p}$ can be calculated if the order N is known as they are independent of any other model parameter. The size and condition of $A_{(new)p}$, which is always square, makes it possible obtain the inverse using existing routines. Table 1 gives the condition number for each matrix for up to a sixth-order solution. As the solution for each stage is the result of a back substitution from earlier stages the condition number of the whole process will be the product of the condition numbers of each of the submatrices.

Table 1: Condition of matrices $A_{(new)p}$

Order N	Condition of matrices $\mathbf{A}_{(new)1} \ldots \mathbf{A}_{(new)N}$					
1	1					
2	10.9	1				
3	141	23.4	1			
4	2500	428	40.0	1		
5	5.70×10^4	8700	988	60.5	1	
6	1.59×10^6	2.13×10^5	2.33×10^4	1950	85.0	1

4.3 Improving the condition of the problem

The solution discussed previously suffers from poor conditioning for higher-order systems. This is in common with many solutions of a Volterra-type problem particularly with higher-order models [6, 16]. The conditioning of the problem is determined only by the condition number of the matrix A. The elements of A are fixed by eqn. 14 from the excitation matrix X. Hence, a method is introduced whereby the elements in A can be altered by adjusting the relative levels of the test sequence while the maximum amplitude is maintained at the same level as previously defined. The levels were set arbitrarily in Section 4.1 as equally spaced points through the transfer function by adding delayed combinations of an MLS. The ideal excitation levels are not easily determined using an analytic approach so an alternative method is used using a numerical search algorithm.

Thus we redefine the constant used in the matrix X to be $\alpha\beta_s$ where s is the partition in X. The constant α still defines the global level of the test sequence while β_s changes the relative levels to improve the condition of $A_{(new)p}$. The peak amplitude of each row of X is given by

$$peak(\mathbf{x}_I) = \alpha\beta_s sa(1 + c) \quad (18)$$

where a and c are the MLS multiplier and offset as defined by eqn. 13. One of the constants β_N is set to a fixed value of 1 giving the peak value of the whole input sequence to be

$$peak(\mathbf{X}) = \alpha N a(1 + c) \quad \text{if} \quad \beta_s s \leq N \quad \text{for all} \quad s$$

The change of basis has a new constant introduced so that eqn. 5 becomes

$$\mathbf{h}_r = (\gamma_r/\alpha)^r \mathbf{M}^{[r]}\mathbf{g}_r \quad (19)$$

Now the element a_{ij} in the matrix A is given by eqn. 14 together with the new constants as

$$a_{ij} = (\gamma_r\beta_s)^r \sum_{k=1}^{s} \delta(i_k - j_1) \sum_{k=1}^{s} \delta(i_k - j_2) \ldots \sum_{k=1}^{s} \delta(i_k - j_r)$$

The choice of such a scheme for the constants is to ensure the matrices $A_{(new)p}$ and $A_{(old)p}$ depend only on the order of the system and the newly introduced constants, allowing them to be calculated in advance of a measurement. A simulated annealing algorithm was applied [17] to give the minimised condition numbers shown in Table 2.

Table 2: Condition number of matrices $A_{(new)p}$ with scale factors introduced

Order N	$\gamma_1 \ldots \gamma_N$	$\beta_1, \ldots, \beta_{N-1}$	Condition $1..N$
2	60.321	−2	1
	5.4919		1
3	0.50241	1.6457	7.56
	0.25346	−1.2503	3.38
	− 0.41108		1
4	10.0	−3.9282	10.1
	− 1.9010	−1.3086	77.4
	− 1.0610	0.94716	30.6
	− 0.79136		1
5	7.7882	−4.2521	286
	3.6020	0.70215	319
	− 1.4055	−0.79262	130
	1.2564	1.0357	48.1
	− 0.91231		1
6	− 7.3791	−0.28323	3120
	− 3.5979	2.8747	3120
	− 1.5454	−1.9999	2690
	− 1.0602	1.1499	992
	0.84527	0.66033	220
	− 0.63963		1

4.4 Change of basis

Unlike the Volterra kernel representation, the vector g of the Wiener model coefficients does not have any directly interpretable results. Thus, the Volterra kernels are obtained from the relationship given in eqn. 19. The size of this matrix operation can be greatly reduced from the order of M^{2r} to rM^{r+1} operations by

making use of the algorithm devised by Nowak and Van Veen [6]. This decomposes the matrix operation $M^{[r]}g_r$ into repeated operations of the type Md where d is a vector containing elements from g_r. It uses the Kronecker product theorem T2.13 shown by Brewer [8] applied recursively.

The operation count of this process can be further reduced by making use of the formation of M from a binary MLS. This allows the process Md to be performed using the fast Hadamard transform in the order of $2.5M\log_2 M$ operations [18, 19]. This gives the operation count for the change of basis for N kernels as the order of $\Sigma_{r=1}^{N} 2.5rM^r\log_2 M$ operations not including some additional book keeping.

5 Generation of test sequence

The input matrix must be of the Toeplitz form to represent a test sequence applied to the model. Define a matrix U that can be permuted by S to give the earlier defined excitation matrix X by the relationship X = SU. The matrix U must be of the form

$$\mathbf{U} = \begin{bmatrix} u_0 & u_{-1} & u_{-2} & u_{-3} & u_{-4} & u_{-5} \\ u_1 & u_0 & u_{-1} & u_{-2} & u_{-3} & u_{-4} \\ u_2 & u_1 & u_0 & u_{-1} & u_{-2} & u_{-3} \\ & & \vdots & & & \end{bmatrix}$$

the first column of which forms the input sequence. The rows of X can be organised into blocks of M rows that are of the Toeplitz form, generally given by

$$\begin{bmatrix} \mathbf{x}\{i_1, & i_2, & \dots & i_s\} \\ \mathbf{x}\{i_1+1, & i_2+1, & \dots & i_s+1\} \\ \mathbf{x}\{i_1+2, & i_2+2, & \dots & i_s+2\} \\ & & \vdots & \\ \mathbf{x}\{i_1+M-1, & i_2+M-1, & \dots & i_s+M-1\} \end{bmatrix}$$

with the indices calculated modulus M. The first line of the block, the start set, defines the whole block which can only be M rows long as $m_{i+M} = m_i$.

The matrix U can thus be generated by including all the possible blocks and inserting $M - 1$ rows between them to make the transition between blocks of the Toeplitz form. These additional rows effectively fill up the memory of the system as shown below for the transition between a block starting with the sequence [*a b c*] and one starting with [*x y z*]

$$\begin{bmatrix} & \vdots & \\ a & b & c \\ c & a & b \\ b & c & a \\ z & b & c \\ y & z & b \\ x & y & z \\ z & x & y \\ y & z & x \\ & \vdots & \end{bmatrix} \quad \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{1st block} \quad \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{transition} \quad \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{2nd block}$$

Due to the circularity of the blocks two different start sets may include the same rows of X. To generate U all the possible start sets are searched through, omitting any that will cause repetition, until all the rows from X have been included in the blocks. As 0 will always be an element of one of the sets in a block it is only necessary to search through the possible sets $\{0, i_2, i_3 \ldots i_s\}$ for which $i_2 \le i_3 \le \ldots \le i_s$. The permutation S is constructed at the same time so that it is possible to check that a start set has not been included in an earlier block.

It is possible that some blocks may repeat in less than M rows giving a few repeated rows in the input matrix. This occurrence has been predicted for up to $N = 6$ and $M = 4095$ and has been found to be very small compared to the length of the test sequence. The exact length of the test sequence will depend on the number of blocks that repeat in less than M rows; for no occurrences it is $2L - L/M$ samples. For the predicted occurrences the length never exceeds $2L$ so giving a good estimate of the test sequence length.

5.1 Practical implementation and algorithm summary

The system to be modelled is assumed to have known memory of less than M and maximum nonlinearity order N. The value of M is restricted to $M = 2^q - 1$ for q a positive integer. This gives L coefficients to solve where $L = f(M, N)$ as defined by eqn. 9. In advance of measuring such a system, matrices $A^{-1}_{(new)p}$ and $A_{(old)p}$ for $p = 1 \ldots N$ are generated and stored. The largest matrix to invert for up to a sixth-order system is of dimension 20×20 and is nonsingular. A binary MLS of length M is generated using the appropriate shift register [12]. The test sequence of the desired peak level is generated from this MLS giving the required multilevel sequence which is approximately $2L$ samples long. At the same time the permutation matrix S is generated and stored in the form of a vector which is of the same length as the test sequence. The measurement process may then be carried out:

(i) The test sequence is applied to the system and the observation vector y is permuted by the permutation S, reducing it in size to L samples.

(ii) The elements in g for which sets tagging the rows contain only one value are solved using the same rows in y by the relationship $g_{(new)1n} = A^{-1}_{(new)1} y_{1n}$, stage 1 in the solution.

(iii) The elements in g for which sets tagging the rows contain two different values are solved using the same rows in y and the elements solved in stage 1.

(iv) The previous process is carried out for rows with three to N different values stage by stage using elements solved in the previous stages until all of g has been solved.

The Volterra kernels are obtained from the Wiener model coefficients by a change of basis eqn. 19. This requires the vector $Q_r g_r$ which has been measured to be expanded to the symmetrical form g_r which will be much larger. This gives us the largest vector that has to be operated on as g_N of dimension M^N. The implemented method uses a minimal set of excitations and makes use of a sparse matrix method to achieve an efficient solution. However, applying any Volterra filter identification scheme to actual system modelling results in a very large number of coefficients that have to be determined with a corresponding large number of test sequence samples and number of operations to achieve the result.

A solution to the coefficient size problem has been developed by Reed and Hawksford [20]. The reduction technique makes use of properties of the device under test, determined before the modelling procedure, to limit the search space for a Volterra test algorithm. The

technique described [20] is independent of the Volterra modelling algorithm used so that it may be implemented using the method described in this paper to give practical system measurements.

6 Example results

The nonlinear system shown in Fig. 4 was simulated by a computer using multipliers and FIR filters. The total memory of the system is 33 samples and the maximum order of nonlinearity 3. The model to be used has $M = 63$ and $N = 3$ giving 45759 coefficients to be solved and a test sequence of 90875 samples. It takes of the order 10^6 operations to obtain the coefficients from the measurement and a further order 10^7 to transform the coefficients into the Volterra kernel values.

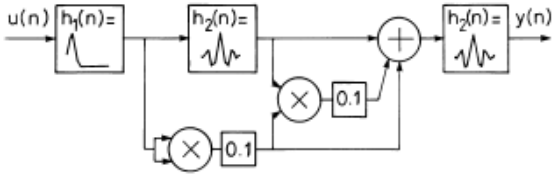


Fig. 4 *Simulated device under test*

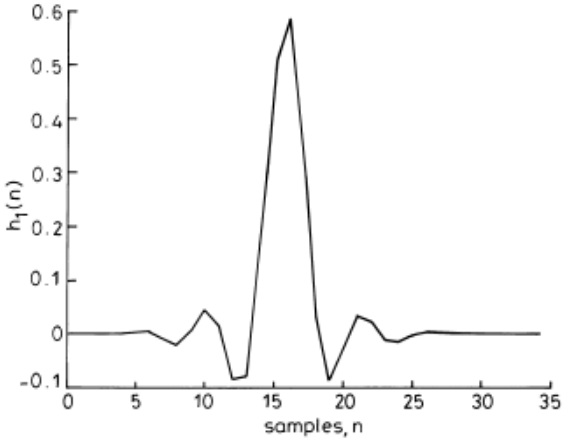


Fig. 5 *First-order Volterra kernel of example system*

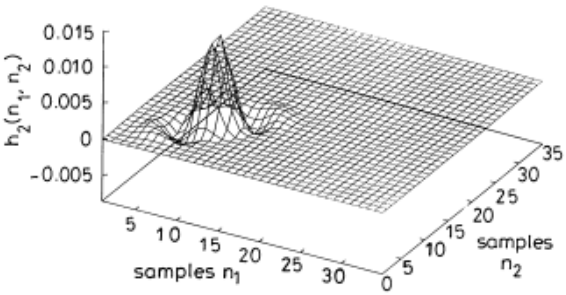


Fig. 6 *Second-order Volterra kernel of example system*

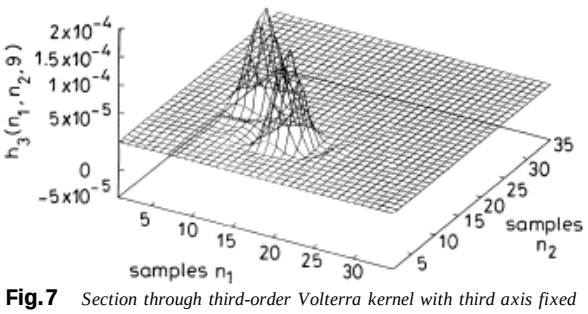


Fig. 7 *Section through third-order Volterra kernel with third axis fixed*

The first two kernels of the system are shown in Figs. 5 and 6. The third-order kernel cannot be easily represented so a section through it is shown in Fig. 7 with the third time axis kept fixed to display the highest points in the kernel.

The results given are for an ideal system without noise. To simulate more realistic process conditions the modelling procedure was repeated with noise added to the output of the system. The noise was simulated with a bandlimited, sampled, gaussian random variable. The modelling parameters were the same as the first example with a peak test signal amplitude of 1.0. The units of all the signal powers and results are arbitrary but all are relative to the test signal amplitude of 1.0. Table 3 gives the levels of all of the components of the output of the system and the ratio of noise to signal components.

Table 3: Levels of components of output from simulated system with added noise (arbitrary units)

Component	RMS level	Peak level	RMS noise/ RMS comp.	Peak noise/ Peak comp.
Linear	0.49	1.7	0.0010	0.0015
2nd order	0.017	0.144	0.0029	0.018
3rd order	0.00062	0.013	0.81	0.20
Noise	0.00050	0.0026		

Table 4: Error in kernel measurements when bandlimited gaussian distributed noise is added with levels given in Table 3 (arbitrary units)

Kernel	Peak in Kernel	RMS error	Peak error	Peak error/ Kernel peak
1	0.59	1.7×10^{-4}	3.9×10^{-4}	0.00066
2	0.012	1.9×10^{-5}	3.7×10^{-5}	0.0031
3	0.00016	7.4×10^{-6}	4.2×10^{-5}	0.26

The results for the system with added noise were subtracted from the first set of results to give the errors which are recorded in Table 4. As the kernel spaces have information which is quite localised it is most useful to use the ratio of the peak error and the peak in the kernel, given in the Table, as a comparative measure. It can be seen from Tables 3 and 4 that the order of the ratios of peak errors corresponds to the ratio of the noise to signal components. This result shows that the solution process does not decrease the sensitivity of the measurement to noise for the example given. The conditioning of the problem would suggest that the error in the results due to noise should be increased. For example, Table 2 shows there could be an increase in error from noise for the first kernel result of up to a factor of 7.56, while the actual result shows a decrease in sensitivity of error owing to noise. This is not unusual as the condition number is known to give the upper limit for the sensitivity of a problem to noise and can often be pessimistic [21].

7 Conclusions

We have shown that the discrete Volterra filter is equivalent to a Wiener model that uses a parallel set of discrete linear filters to represent the memory of the system. The impulse responses of these filters are a set of functions that form an orthonormal basis so

allowing the Volterra kernels to be transformed directly to the coefficients of the Wiener model.

A maximum length sequence was used to construct the set of orthonomal functions by suitably scaling and adding a constant to the sequence. When the same functions are applied in a time reversed form to the Wiener model they form a set of delta functions at the output of the linear sections. By using a test sequence made up of combinations of these functions summed together it is possible to stimulate the coefficients of the Wiener model in such a way as to solve them efficiently. It has been shown that this can be implemented as a sparse matrix operation. The sparse matrix is broken down into many identical sets of small matrices that can be inverted by a standard matrix inversion package. The coefficients of the Wiener model are then obtained in stages by back substituting solutions from earlier stages until every coefficient is determined.

The coefficients of the Wiener model can be used directly to model the system however, unlike the Volterra kernels, they have no readily interpretable physical significance. The Volterra kernels are obtained from the Wiener model coefficients by a change of basis that uses existing algorithms to improve the efficiency of the operation.

A simulated nonlinear system has been given to demonstrate the method. The example shows that the process to change basis between the Wiener model coefficients and the Volterra kernels dominates over the determination of the coefficients both in terms of operation count and storage requirements.

The measurement method has only been discussed for strictly discrete systems. It is expected that the conditioning of the LMS solution will limit the accuracy of practical measurements. This problem has been reduced by applying a minimisation scheme that adjusts the relative amplitudes of the test sequence to give the optimal conditioning for the solution.

8 References

1 VOLTERRA, V.: 'Theory of Functionals' (Dover, New York, 1959)

2 KAIZER, A.J.M.: 'Modeling of the nonlinear response of an electrodynamic loudspeaker by a Volterra series expansion', *J. Audio Eng. Soc.*, 1987, 35, pp. 421–432

3 SHI, Y., and HECOX, K.E.: 'Nonlinear system identification by m-pulse sequences: Application to brainstem auditory evoked responses', *IEEE Trans.*, September 1991, BE–38, pp. 834–845

4 SANDBERG, I.W.: 'Uniform approximation with doubly finite Volterra series', *IEEE Trans.*, June 1992, SP–40, pp. 1438–1441

5 LEE, Y.W., and SCHETZEN, M.: 'Measurement of the Wiener kernels of a nonlinear system by crosscorrelation', *Int. J. Control*, 1965, 23, pp. 237–254

6 NOWAK, R.D., and VAN VEEN, B.D.: 'Random and pseudorandom inputs for Volterra filter identification', *IEEE Trans.*, August, 1994, SP–42, pp. 2124–2134

7 SCHETZEN, M.: 'The Volterra and Wiener theories of nonlinear systems' (Wiley, New York, 1980)

8 BREWER, J.W.: 'Kronecker products and matrix calculus in system theory', *IEEE Trans.*, September 1978, CAS–25, pp. 772–781

9 WILL-HANS STEEB.: 'Kronecker product of matrices and applications' (BI Wissenschaftsverlag, 1991)

10 EYKHOFF, P.: 'System identification' (Wiley, 1974)

11 KREYSZIG, E.: 'Advanced engineering mathematics' (Wiley, 1983, 5th edn.)

12 MACWILLIAMS, F.J., and SLOANE, N.J.A.: 'Pseudorandom sequences and arrays', *Proc. IEEE*, December 1976, 64, pp. 1715–1729

13 RIFE, D.D., and VANDERKOOY, J.: 'Transfer-function measurements with maximum length sequences', *J. Audio Eng. Soc.*, June, 1989, 37, pp. 419–444

14 BARKER, H.A., and PRADISTHAYON, T.: 'Higher-order autocorrelation functions of pseudorandom signals based on m sequences', *Proc. IEEE*, September 1970, 117, pp. 1857–1863

15 VANDERKOOY, J.: 'Aspects of MLS measuring systems', *J. Audio Eng. Soc.*, April 1994, 42, pp. 219–231

16 SEBER, G.A.F.: 'Linear regression analysis' (Wiley, 1977)

17 PRESS, W.H., TEUKOLSKY, S.A., *et al.*: 'Numerical recipes in C: The art of scientific computing' (CUP, Cambridge, 1992)

18 BORISH, J., and ANGELL, J.B.: 'An efficient algorithm for measuring the impulse response using pseudorandom noise', *J. Audio Eng. Soc.*, July/August 1983, 31, pp. 478–487

19 SUTTER, E.E.: 'The fast m-transform: A fast computation of cross-correlations with binary m-sequences', *SIAM J. Computation*, 1991, 20, pp. 686–694

20 REED, M.J., and HAWKSFORD, M.O.: 'Practical modelling of nonlinear audio systems using the Volterra series'. Presented at the 100th convention of the Audio Engineering Society, Copenhagen, May 1996, (preprint 4264)

21 JENNINGS, A., and MCKEOWN, J.: 'Matrix computation' (Wiley, 1992, 2nd edn.)

Efficient implementation of the Volterra filter

M.J.Reed and M.O.J.Hawksford

Abstract: An efficient implementation of the Volterra filter is presented which uses a frequency domain representation to reduce the number of computations. The multidimensional convolution of the Volterra filter is transformed to the frequency domain giving a transformed input matrix which is sparse and obtained directly from a one-dimensional Fourier transform. In addition to the sparse nature of the transformed input matrix, symmetries in both the Volterra filter and the frequency domain representation are exploited to increase the efficiency of the algorithm. The computational saving is demonstrated by comparing it with the direct implementation of the time domain representation and another technique which uses a frequency domain representation but does not utilise symmetry.

1 Introduction

The Volterra series represents a nonlinear system as a set of multidimensional convolutions [1] and is used to model a range of nonlinear systems such as the auditory system [2], mechanical systems [3, 4] and electromechanical loudspeakers [5]. Possible applications of such a Volterra series model include system prediction due to arbitrary excitation and nonlinear correction [1]. For either of the mentioned applications it is required to implement the Volterra model of the nonlinear system and this is performed using a signal processing algorithm termed a Volterra filter. However, the Volterra filter requires the implementation of computationally intensive multidimensional convolutions that make it unsuitable for applications requiring real-time performance. This paper describes a new technique for the implementation of the Volterra filter that improves implementation efficiency compared with existing techniques. One target application for this work is towards real-time implementation of nonlinear correction for loudspeaker transducers [6]. However, the technique can be generally applied to improve the efficiency of any Volterra filter implementation.

Sandberg has shown that a good approximation to a wide range of nonlinear systems is given by a Volterra filter which is a doubly finite and discrete implementation of the Volterra series that is expressed as [7]

$$y(n) = \sum_{r=1}^{N} \sum_{i_1=0}^{M-1} \cdots \sum_{i_r=0}^{M-1} h_r(i_1, i_2, \ldots, i_r) \\ u(n-i_1)u(n-i_2)\ldots u(n-i_r) \quad (1)$$

where u is an arbitrary input signal to the model, y the output, h_r is the rth Volterra kernel, N the maximum order of nonlinearity and M is the memory of the filter. The only general condition placed upon the input signal u is that the model is usually only valid over a finite input amplitude [7], for example in many systems the model is only valid at amplitudes below the occurrence of clipping.

The implementation of eqn. 1 in the direct form for models of real systems is computationally burdensome as the number of points in the kernel spaces is usually large [8]. It is possible to improve the efficiency of the Volterra filter by approximating part of the kernel spaces using a mirror filter [9] constructed from linear filters combined with multipliers. However, then the model is only valid for a limited set of signals that excite the kernel spaces for which the approximation is valid. Consequently, for accurate system prediction or correction with unconstrained input signals the whole kernel space rather than partial approximations to it are required [6]. This paper presents an efficient frequency domain implementation of the Volterra filter incorporating the whole kernel space. Frequency domain techniques already exist [10, 11] but, unlike the method presented, they do not make use of symmetries in the filter to improve efficiency and do not extend the result to a general order of nonlinearity.

2 Expressing the Volterra filter in vector notation

Define a column vector $\boldsymbol{h}_r$ which represents all permutations of points in the kernel h_r from eqn. 1

$$\boldsymbol{h}_r = \begin{bmatrix} h_r(0,0,\ldots,0) \\ h_r(0,0,\ldots 1) \\ \vdots \\ h_r(M-1,M-1,\ldots,M-2) \\ h_r(M-1,M-1,\ldots,M-1) \end{bmatrix} \quad (2)$$

The format of eqn. 2 represents the multidimensional points of a Volterra kernel (a multidimensional impulse response function) as a single vector by representing the

IEE Proceedings online no. 20000183

DOI: 10.1049/ip-vis:20000183

Paper first received 9th July and in revised form 25th November 1999

The authors are with the Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

points of the kernel in natural order. For example, a second-order kernel

$$\begin{bmatrix} h_2(0,0) & h_2(0,1) & h_2(0,2) \\ h_2(1,0) & h_2(1,1) & h_2(1,2) \\ h_2(2,0) & h_2(2,1) & h_2(2,2) \end{bmatrix}$$

is represented as

$$\boldsymbol{h}_2 = [h_2(0,0), h_2(0,1), h_2(0,2), h_2(1,0), h_2(1,1), h_2(1,2), h_2(2,0), h_2(2,1), h_2(2,2)]^T$$

Define a row vector $\boldsymbol{u}_n$ representing the input u

$$\boldsymbol{u}_n = [u(n), u(n-1), u(n-2), \ldots, u(n-M+1)]$$

All the rth-order products of the input from $u(n)$ to $u(n-M+1)$ may be expressed as the rth-order Kronecker product (denoted as the superscript $^{[r]}$) of $\boldsymbol{u}_n$ with itself

$$\underbrace{\boldsymbol{u}_n \otimes \boldsymbol{u}_n \otimes \ldots \otimes \boldsymbol{u}_n}_{r \text{ times}} = \boldsymbol{u}_n^{[r]}$$

where $\otimes$ is the Kronecker product between two vectors or matrices which is defined as [12, 13]

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{0,0}\boldsymbol{B} & a_{0,1}\boldsymbol{B} & \cdots & a_{0,n-1}\boldsymbol{B} \\ a_{1,0}\boldsymbol{B} & \ddots & & \\ \vdots & & & \vdots \\ a_{m-1,0}\boldsymbol{B} & a_{m-1,1}\boldsymbol{B} & \cdots & a_{n-1,m-1}\boldsymbol{B} \end{bmatrix}$$

where $\boldsymbol{A}$ is of dimension $m \times n$ and $a_{i,j}$ is an element from $\boldsymbol{A}$.

Thus, eqn. 1 can be represented in vector notation as

$$y(n) = \sum_{r=1}^{N} \boldsymbol{u}_n^{[r]} \boldsymbol{h}_r \tag{3}$$

Define a vector $\boldsymbol{y} = [y(n), y(n-1), y(n-2), \ldots, y(n-(M-1))]^T$ and a corresponding partitioned input matrix $\boldsymbol{U}^{\{r\}}$

$$\boldsymbol{U}^{\{r\}} = \begin{bmatrix} \boldsymbol{u}_n^{[r]} \\ -- \\ \boldsymbol{u}_{n-1}^{[r]} \\ -- \\ \vdots \\ -- \\ \boldsymbol{u}_{n-(M-1)}^{[r]} \end{bmatrix}$$

where the symbol $^{\{r\}}$ signifies performing the rth-order Kronecker product of each row of $\boldsymbol{U}$ with itself. The resulting matrix $\boldsymbol{U}^{\{r\}}$ is of dimension $M \times M^r$. The output of a Volterra filter over M samples can now be given by

$$\boldsymbol{y} = \sum_{r=1}^{N} \boldsymbol{U}^{\{r\}} \boldsymbol{h}_r \tag{4}$$

Thus, for a block of M samples, each order r of the direct form of the Volterra filter requires rM^{r+1} multiplications.

3 Implementation of the Volterra filter using Fourier transform relationships

We will now show how Fourier transform relationships often applied to linear convolution can be extended to the multidimensional convolution of the Volterra filter giving rise to a sparse matrix operation.

Define a circulant input signal matrix $\boldsymbol{C}$ of the form

$$\boldsymbol{C} = \begin{bmatrix} u(n) & u(n-(M-1)) & & \\ u(n-1) & u(n) & & \\ \vdots & & & \\ u(n-(M-1)) & u(n-(M-2)) & & \\ & u(n-(M-2)) & \ldots & u(n-1) \\ & u(n-(M-1)) & \ldots & u(n-2) \\ & & \vdots & \\ & u(n-(M-3)) & \ldots & u(n) \end{bmatrix}$$

Define the Fourier transform as the matrix $\boldsymbol{F}$ of dimension $M \times M$ with the element f_{ik} of the ith row and kth column given by

$$f_{ik} = \frac{1}{\sqrt{M}} \omega_M^{ik} \qquad i, k = 0, 1, \ldots, M-1$$

where $\omega_M = \mathrm{e}^{-2\pi j/M}$. Define the inverse Fourier transform as $\boldsymbol{F}^H$, the conjugate transpose of $\boldsymbol{F}$. To make use of the orthonormality of the $M \times M$ transform $\boldsymbol{F}^H\boldsymbol{F} = \boldsymbol{I}$ with the larger dimensional Volterra filter we make use of the corollary from the following theorem. Note that while the authors do not claim that the following theorem and corollary are novel, they are not aware of related published work.

Theorem 1: $\boldsymbol{A}^{[r]}\boldsymbol{B}^{[r]} = (\boldsymbol{AB})^{[r]}$

Proof:

$$\boldsymbol{A}^{[r]}\boldsymbol{B}^{[r]} = \left(\boldsymbol{A}^{[r-1]} \otimes \boldsymbol{A}\right)\left(\boldsymbol{B}^{[r-1]} \otimes \boldsymbol{B}\right) \tag{5}$$

By application of the mixed product rule [12]

$$\left(\boldsymbol{A}^{[s]} \otimes \boldsymbol{A}\right)\left(\boldsymbol{B}^{[s]} \otimes \boldsymbol{B}\right) = \left(\boldsymbol{A}^{[s]}\boldsymbol{B}^{[s]}\right) \otimes (\boldsymbol{AB}) \tag{6}$$

By repeated application of eqn. 6 on the right side of eqn. 5 and eqn. 5 on the first term on the right side of eqn. 6 for $s = r-1, \ldots, 2$ the following is obtained

$$\begin{aligned} \boldsymbol{A}^{[r]}\boldsymbol{B}^{[r]} &= (\boldsymbol{AB}) \otimes \ldots \otimes (\boldsymbol{AB}) \\ &= (\boldsymbol{AB})^{[r]} \end{aligned}$$

From theorem 1 the following corollary is obtained:

Corollary 1: If $\boldsymbol{A}$ is orthonormal then $\boldsymbol{A}^{[r]}$ is also orthonormal.

Corollary 1 follows from theorem 1 as if $\boldsymbol{A}$ is orthonormal $\boldsymbol{A}\boldsymbol{A}^H = \boldsymbol{I}_M$ where $\boldsymbol{I}_M$ is the $M \times M$ identity matrix and by applying theorem 1

$$\begin{aligned} \boldsymbol{A}^{[r]}(\boldsymbol{A}^H)^{[r]} &= (\boldsymbol{A}\boldsymbol{A}^H)^{[r]} \\ &= \boldsymbol{I}_M^{[r]} \\ &= \boldsymbol{I}_{M^r} \end{aligned}$$

As $\boldsymbol{I}_M^{[r]}$ is the $M^r \times M^r$ identity matrix then $\boldsymbol{A}^{[r]}$ is orthonormal.

As the defined Fourier transform relationship is orthonormal then, from corollary 1, eqn. 4 can be expressed as

$$\boldsymbol{y} = \sum_{r=1}^{N} \boldsymbol{F}^H \boldsymbol{F} \boldsymbol{C}^{\{r\}} (\boldsymbol{F}^H)^{[r]} \boldsymbol{F}^{[r]} \boldsymbol{h}_r \tag{7}$$

where the input matrix $\boldsymbol{U}$ is replaced by the circulant matrix $\boldsymbol{C}$.

To refer to the elements in the matrices or vectors which arise from the Kronecker product we define the following notation. The ith row in a matrix $\boldsymbol{B}$ is defined as the row vector $\boldsymbol{b}_{i\bullet}$, the kth column as a column vector $\boldsymbol{b}_{\bullet k}$ and the ith element in the kth column as b_{ik}.

If the row vector $\boldsymbol{a}=\boldsymbol{b}\otimes\boldsymbol{b}$, where $\boldsymbol{b}$ is of dimension M, then define the vector $\boldsymbol{a}$ to be partitioned as

$$\boldsymbol{a}=[b_0b_0, b_0b_1, \ldots, b_0b_{M-1}|_1^1 b_1b_0, b_1b_1, \ldots, b_1b_{M-1}|_1^1 \ldots]$$

so that the k_2th element in the k_1th partition is given by $b_{k_1}b_{k_2}$. If $\boldsymbol{a}=\boldsymbol{b}\otimes\boldsymbol{b}\otimes\ldots\otimes\boldsymbol{b}=\boldsymbol{b}^{[r]}$ then the vector $\boldsymbol{a}$ is recursively partitioned r times with each partition being partitioned M times. Now the kth element in $\boldsymbol{a}$ is labelled by defining the ordered set $K=\langle k_1\ldots k_r\rangle=\langle k:0\le k<M\rangle$ such that an element in $\boldsymbol{a}$ is given as

$$a_K=a_{\langle k_1\ldots k_r\rangle}=b_{k_1}b_{k_2}\ldots b_{k_r} \tag{8}$$

This notation is very useful for the Volterra filter as for example the point in the rth Volterra kernel $h_r(k_1\ldots k_r)$ is given by the element h_K in the vector $\boldsymbol{h}_r$ defined in eqn. 2.

Define the Fourier transform of the input signal as the column vector Θ

$$\Theta=\boldsymbol{F}\boldsymbol{c}_{\bullet 0}=\begin{bmatrix}\theta_0\\ \vdots\\ \theta_{M-1}\end{bmatrix} \tag{9}$$

Consider the operation $\boldsymbol{F}\boldsymbol{C}^{\{r\}}\ (\boldsymbol{F}^H)^{[r]}$ given in eqn. 7. Let

$$\boldsymbol{A}_r=\boldsymbol{F}\boldsymbol{C}^{\{r\}}(\boldsymbol{F}^H)^{[r]} \tag{10}$$

and

$$\boldsymbol{B}=\boldsymbol{C}^{\{r\}}(\boldsymbol{F}^H)^{[r]} \tag{11}$$

The first row of $\boldsymbol{B}$ is given by

$$\boldsymbol{b}_{0\bullet}=\boldsymbol{c}_{0\bullet}^{[r]}(\boldsymbol{F}^H)^{[r]} \tag{12}$$

and by applying theorem 1

$$\boldsymbol{b}_{0\bullet}=(\boldsymbol{c}_{0\bullet}\boldsymbol{F}^H)^{[r]} \tag{13}$$

As $\boldsymbol{c}_{0\bullet}^T$ is the time reversed form of the input signal $\boldsymbol{c}_{\bullet 0}$, which we define as always real, then by the Fourier transform theorem for reversed time signals [14]

$$\Theta=\boldsymbol{F}\boldsymbol{c}_{\bullet 0}=(\boldsymbol{c}_{0\bullet}\boldsymbol{F}^H)^T \tag{14}$$

and together with eqn. 13 the following is obtained

$$\boldsymbol{b}_{0\bullet}=(\Theta^T)^{[r]} \tag{15}$$

The Fourier transform theorem for time shifted signals states that for a signal $x(n)$ with a Fourier transform $X(\mathrm{e}^{\mathrm{j}w})$ [14]

$$x(n-i)\overset{F}{\leftrightarrow}\mathrm{e}^{\mathrm{j}wi}X(\mathrm{e}^{\mathrm{j}w}) \tag{16}$$

Thus, as the ith row of $\boldsymbol{C}$ is a time shifted version of $\boldsymbol{c}_{0\bullet}$ by i places forward, eqn. 16 applied to eqn. 14 gives

$$\boldsymbol{c}_{i\bullet}\boldsymbol{F}^H=[\theta_0\overline{w}_M^{0i}, \theta_1\overline{w}_M^{1i}, \theta_2\overline{w}_M^{2i}, \ldots, \theta_{M-1}\overline{w}_M^{(M-1)i}] \tag{17}$$

where $\bar{w}$ is the conjugate of w. Substituting eqn. 17 into eqn. 13 gives

$$\boldsymbol{b}_{i\bullet}=[\theta_0\overline{w}_M^{0i}, \theta_1\overline{w}_M^{1i}, \theta_2\overline{w}_M^{2i}, \ldots, \theta_{M-1}\overline{w}_M^{(M-1)i}]^{[r]} \tag{18}$$

Hence, applying the notation of eqn. 8 to eqn. 18 an element b_{iK} of $\boldsymbol{B}$ is given by

$$b_{iK}=\theta_{k_1}\overline{w}_M^{k_1 i}\times\theta_{k_2}\overline{w}_M^{k_2 i}\times\ldots\times\theta_{k_r}\overline{w}_M^{k_r i} \tag{19}$$

$$=\theta_{k_1}\times\theta_{k_2}\times\ldots\times\theta_{k_r}\overline{w}_M^{k_{sum} i} \tag{20}$$

where $k_{sum}=(k_1+k_2+\ldots+k_r)_M$ using modulus M addition. Thus, the Kth column of $\boldsymbol{B}$ is given by

$$\boldsymbol{b}_{\bullet K}=\theta_{k_1}\times\ldots\times\theta_{k_r}\begin{bmatrix}\overline{w}_M^{k_{sum}0}\\ \overline{w}_M^{k_{sum}1}\\ \vdots\\ \overline{w}_M^{k_{sum}(M-1)}\end{bmatrix}$$

$$=\sqrt{M}\theta_{k_1}\times\ldots\times\theta_{k_r}\boldsymbol{f}_{\bullet k_{sum}}^H \tag{21}$$

where $\boldsymbol{f}_{\bullet k_{sum}}^H$ is a column from $\boldsymbol{F}^H$. From the definitions given in eqns. 10 and 11

$$\boldsymbol{A}_r=\boldsymbol{F}\boldsymbol{B} \tag{22}$$

so that using the set notation for the column number the element a_{iK} in $\boldsymbol{A}_r$ is given by

$$a_{iK}=f_{i\bullet}b_{\bullet K} \tag{23}$$

Substituting eqn. 21 into the above gives

$$a_{iK}=\sqrt{M}\theta_{k_1}\times\ldots\times\theta_{k_r}f_{i\bullet}f_{\bullet k_{sum}}^H \tag{24}$$

As $\boldsymbol{F}$ is orthonormal

$$f_{i\bullet}f_{\bullet j}^H=\begin{cases}1 & \text{if}\quad i=j\\ 0 & \text{if}\quad i\ne j\end{cases}$$

where $\boldsymbol{f}_{\bullet j}^H$ is the jth column of $\boldsymbol{F}^H$. Therefore the elements of the matrix $\boldsymbol{A}_r$ are given by

$$a_{iK}=\begin{cases}\sqrt{M}\theta_{k_1}\times\theta_{k_2}\times\ldots\times\theta_{k_r} & \text{if}\quad k_{sum}=i\\ 0 & \text{if}\quad k_{sum}\ne i\end{cases} \tag{25}$$

The equation for the Volterra filter is now expressed as

$$\boldsymbol{y}=\sum_{r=0}^{N}\boldsymbol{F}^H\boldsymbol{A}_r\boldsymbol{F}^{[r]}\boldsymbol{h}_r \tag{26}$$

where $\boldsymbol{F}^{[r]}\boldsymbol{h}_r$ is the r-dimensional Fourier transform of $\boldsymbol{h}_r$. The matrix $\boldsymbol{A}_r$ is of dimension $M\times M^r$ and is partitioned, by columns, into M^{r-1} parts. From eqn. 25 the matrix $\boldsymbol{A}_r$ is sparse as each column is non-zero in only one element. Additionally, note that each element in $\boldsymbol{A}_r$ is formed from $r-1$ multiplications from points in the one-dimensional Fourier transform of the input signal. Thus, for each order r in the implementation of eqn. 26 there are two one-dimensional Fourier transforms of order M required and rM^r complex multiplications. This can be compared to rM^{r+1} multiplications required for the direct form of the Volterra series given in eqn. 4 so that, assuming the Fourier transforms require comparably little computation, the implementation of eqn. 26 will require less computation than the direct form of eqn. 4.

4 Reduction due to symmetries

There are two symmetries in eqn. 26 which can be used to reduce the complexity of the implementation as illustrated in Fig. 1 for an example of the matrix $\boldsymbol{A}_2$ for $M=4$. To simplify the diagram the product $\sqrt{4}\,\theta_i\theta_k$ is replaced by the corresponding subscript numbers. The first symmetry used

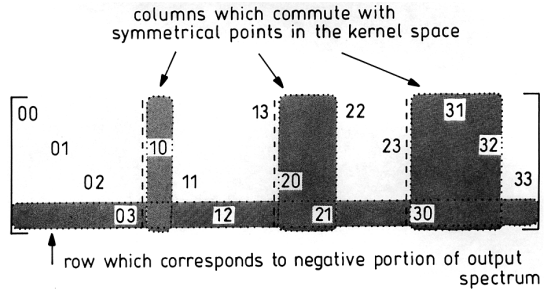


Fig. 1 *Example of matrix* A_2 *for* $M=2$ *showing only the non-zero points*

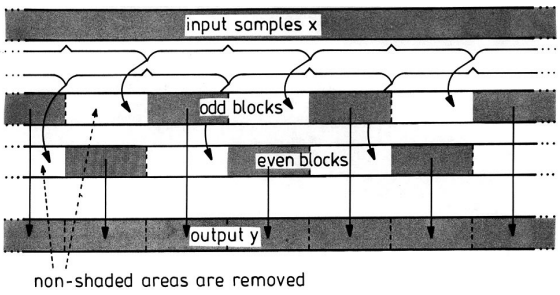


Fig. 2 *Block structure of the overlap and save technique*

is that of each Volterra kernel which can always be represented as a symmetrical space [15]. This symmetry can be expressed using the set notation for a point in the Volterra kernel vector $\boldsymbol{h}_r$ or a corresponding row in $\boldsymbol{A}$ represented by the set $K=\langle k_1,\ldots,k_r\rangle$. Any points in $\boldsymbol{h}_r$ or columns in $\boldsymbol{A}$ which are represented by a set which contains the same elements, in any order, are points in symmetry. An example of the symmetry in a Volterra kernel is that it is always possible to make $h_2(i,j)=h_2(j,i)$ without loss of generality. The symmetrical points in a kernel can be summed and the repetitions removed [16], the corresponding columns in A_r can then also be removed. This changes the dimension of h_r from M^r to

$$\binom{M+r-1}{r}$$

which gives a significant reduction for $r>1$.

A second reduction in the number of points can be achieved due to the fact that the Fourier transform of a real sequence is conjugate-symmetric [14]. This implies that (for even M) only $M/2+1$ of the rows of $\boldsymbol{A}_r$ are required to construct the output $\boldsymbol{y}$ giving approximately a saving of one half for large M.

The implementation of the rth-order stage of the efficient Volterra filter in eqn. 26 is carried out using two 'real-valued' one-dimensional Fourier transforms and an additional

$$\frac{r}{2}\binom{M+r-1}{r}$$

complex multiplications. If the degree of the filter M is a power of two then the aforementioned 'real-valued' one-dimensional Fourier transforms can each be performed using a fast Fourier transform (FFT) in $2M\log_2 M$ real multiplications [17]. Thus, the number of operations (NOP) for a block of M samples using eqn. 26, with M a power of two and taking into account symmetries, is

$$\mathrm{NOP}=4M\log_2 M+\sum_{r=1}^{N}2r\binom{M+r-1}{r} \qquad (27)$$

The above assumes that a complex multiplication is performed using four real multiplications and a multiply accumulate is carried out in one operation.

5 Implementation using overlap and save techniques

The efficient implementation of the Volterra filter can be extended from a circulant input matrix to a generalised input signal by the overlap-and-save technique [14, 18]. First the memory of the filter, originally M, is doubled maintaining all the original points and setting any additional points for time samples greater than $M-1$ to zero. The input matrix size must also be doubled by including the previous M samples and these $2M$ samples are the elements of the column $\boldsymbol{c}_{\bullet 0}$ in eqn. 7. By doubling the memory of the operation of eqn. 7 the effect of wrapping round the input signal for the last M samples is removed. Thus, although the first M samples of the block are not a valid output for the continuous time signal the last M samples are. By overlapping blocks by M samples and 'saving' only the last M samples the output signal is generated. The process is shown diagrammatically in Fig. 2. For linear systems the overlap and save technique has an alternative method termed 'overlap and add' which uses the superposition property of linear systems. As the superposition property does not apply to nonlinear systems the overlap-and-add technique is not applicable to the technique in this paper.

The implementation of the overlap-and-save technique increases the NOP for a block of size M (M a power of two) to

$$\mathrm{NOP}=8M\log_2 2M+\sum_{r=1}^{N}2r\binom{2M+r-1}{r} \qquad (28)$$

6 Comparison of direct and frequency domain methods

The frequency domain technique developed in this paper will now be compared to the direct form of the Volterra filter given in eqn. 1 and another frequency domain technique developed by Im and Powers [11]. As all three methods are mathematically equivalent the comparison is only made on the basis of computational complexity.

To compare the direct form and the frequency domain methods a second- and third-order filter section will be considered separately; the linear case, which is a first-order Volterra system, has already been widely reported and investigated [14]. The direct form of an rth-order Volterra filter which makes use of the kernel symmetry requires $r\,M$

$$\binom{M+r-1}{r}$$

operations for an input sample block of M samples. For the frequency domain techniques it is assumed that the multi-dimensional Fourier transform of the kernel spaces required for eqn. 26 has been performed as this can be computed in advance of the filtering operation. Additionally, for the frequency domain technique, M is set to the next largest power of two thus enabling the use of FFT techniques using block processing. The number of operations for the direct form and the frequency domain technique per input sample are given in Table 1 for a second- and third-order Volterra filter. The Table also gives the

Table 1: Number of operations per output sample for *M*-length second- and third-order sections

Filter memory	Number of operations per sample		
M	Direct	Frequency domain no use of symmetry [11]	Frequency domain using symmetry
Second-order section			
2	6	68	32
4	20	131	47
8	72	258	78
16	272	513	141
32	1056	1024	268
64	4160	2048	524
128	16512	4096	1036
256	65792	8192	2060
Third-order section			
2	12	388	106
4	60	1539	249
8	360	6146	728
16	2448	24577	2455
32	17952	98304	8982
64	137280	393216	34326
128	1.07×10^6	1.57×10^6	134166
256	8.48×10^6	6.29×10^6	530454

number of operations for a method by Im and Powers [11] which uses the frequency domain but has not utilised symmetries. Note that the use of symmetries gives a significant reduction in the number of operations: for the third-order case with $M \geq 16$ it gives a factor of ten or more improvement over the other frequency domain method.

To test the efficiency of the technique the direct form and the frequency domain form using symmetry reductions were implemented on a workstation in C++ running under a UNIX environment. The processor time for each method was estimated using the time() function and the result was scaled to give the equivalent number of operations per sample. The results for a second- and third-order filter section are shown in Figs 3*a* and 3*b*, respectively thus demonstrating that the theoretical computational requirement accurately predicts the operation count of the practical implementation. Figs. 3*a* and 3*b* show the effect of imposing a block structure for the frequency domain techniques which utilise FFTs. The consequence of the block structure is that the NOP for the frequency domain techniques only changes as *M* equals a power of two whereas the direct form increases with each increase in *M*.

7 Conclusions

An efficient algorithm for performing the Volterra filter using a frequency domain representation is presented. The method is developed from utilising the Kronecker product of matrices to manipulate the multidimensional Volterra filter equation. By transforming the input matrix of the Volterra filter to the frequency domain a sparse matrix is obtained. The transformed input matrix can be obtained solely from the one-dimensional Fourier transform of the input signal so that multidimensional transforms of the input are not required.

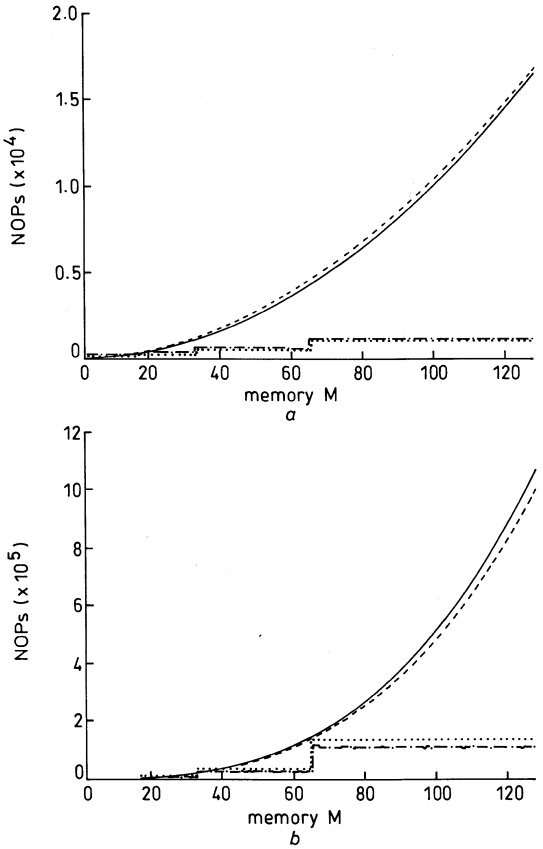


Fig. 3 *Number of operations per sample for the direct and frequency domain forms of second- and third-order sections*

—— direct calculated
····· frequency calculated
– – – direct measured
– – · frequency measured
a Second-order sections
b Third-order sections

The kernels of the Volterra filter and the Fourier transform of a real-time signal exhibit symmetries which can be exploited to further reduce the computational complexity of the technique. To demonstrate the efficiency of the technique it is compared with the direct time domain form and another frequency domain technique which does not exploit the symmetry properties. The comparison shows that using the symmetry properties gives approximately fourfold improvement for the second-order filter and tenfold for the third-order filter over existing frequency domain techniques for filter memories of 16 samples or above. Applications for this efficient filter include simulation of a nonlinear system and real-time nonlinear correction.

8 References

1 SCHETZEN, M.: 'The Volterra and Wiener theories of nonlinear systems' (Wiley, New York, 1980)
2 SHI, Y., and HECOX, K.E.: 'Nonlinear system identification by m-pulse sequences: Application to brainstem auditory evoked responses', *IEEE Trans. Biomed. Eng.*, 1991, **38**, (9), pp. 834–845
3 PARKER, G.A., and MOORE, E.L.: 'Practical nonlinear system identification using a modified Volterra series approach', *Automatica*, 1982, **18**, (1), pp. 85–91
4 WINTERSTEIN, S.R., UDE, T.C., and MARTHINSEN, T.: 'Volterra model of ocean structures –extreme and fatigue reliability', *J. Eng. Mech., ASCE*, 1994, **120**, (6), pp. 1369–1385
5 KAIZER, A.J.M.: 'Modeling of the nonlinear response of an electrodynamic loudspeaker by a Volterra series expansion', *J. Audio Eng. Soc.*, 1987, **35**, (6), pp. 421–432

6 REED, M.J., and HAWKSFORD, M.O.J.: 'Nonlinear error correction of horn transducers using a Volterra filter', *J. Audio Eng. Soc.*, 1997, **45**, (5), pp. 414, (abstract only), presented at the 102nd Convention of the Audio Eng. Soc., Munich March 1997, preprint 4468
7 SANDBERG, I.W.: 'Uniform approximation with doubly finite Volterra series', *IEEE Trans. Signal Process*, 1992, **40**, (6), pp. 1438–1441
8 REED, M.J., and HAWKSFORD, M.O.: 'Practical modelling of nonlinear audio systems using the Volterra series', *J. Audio Eng. Soc.*, 1996, **44**, (7/8), pp. 649–650 (abstract only), presented at the 100th Convention of the Audio Eng. Soc., Copenhagen May 1996, preprint 4264
9 KLIPPEL, W.: 'The mirror filter –a new basis for reducing nonlinear distortion and equalizing response in woofer systems', *J. Audio Eng. Soc.*, 1992, **40**, (9), pp. 675–691
10 MORHAC, M.: 'A fast algorithm of nonlinear Volterra filtering', *IEEE Trans. Signal Process.*, 1991, **39**, (10), pp. 2353–2356
11 IM, S., and POWERS, E.J.: 'A fast method of discrete third-order Volterra filtering', *IEEE Trans. Signal Process.*, 1996, **44**, (9), pp. 2195–2208
12 BREWER, J.W.: 'Kronecker products and matrix calculus in system theory', *IEEE Trans. Circuits Syst.*, 1978, **CAS-25**, (9), pp. 772–781
13 STEEB, W.-H.: 'Kronecker product of matrices and applications' (BI, Wissenschaftsverlag, 1991)
14 OPPENHEIM, A.V., and SCHAFER, R.W.: 'Discrete-time signal processing' (Prentice Hall, 1989)
15 EYKHOFF, P.: 'System identification' (Wiley, 1974)
16 REED, M.J., and HAWKSFORD, M.O.J.: 'Identification of discrete Volterra series using maximum length sequences', *IEE Proc., Circuits Devices Syst.*, 1996, **143**, (5), pp. 241–248
17 SORENSEN, H.V., JONES, D.L., HEIDEMAN, M.T., and BURRUS, C.S.: 'Real-valued fast Fourier transform algorithms', *IEEE Trans. Acoust. Speech Signal Process*, 1987, **ASSP-35**, (6), pp. 849–863
18 IFEACHOR, E.C., and JERVIS, B.W.: 'Digital signal processing, a practical approach' (Addison-Wesley, 1993)

PAPERS

System Measurement and Identification Using Pseudorandom Filtered Noise and Music Sequences*

M. O. J. HAWKSFORD, *AES Fellow*

Centre for Audio Research and Engineering, University of Essex, Colchester, C043SQ, UK

System measurement using pseudorandom filtered noise and music sequences is investigated. A single-pass technique is used to evaluate simultaneously the transfer function and the spectral-domain signal-to-distortion ratio that is applicable to amplifiers, signal processors, digital-to-analog converters, loudspeakers, and perceptual coders. The technique is extended to include a simplified Volterra model expressed as a power series and linear filter bank where for compliant systems, nonlinear distortion can be estimated for an arbitrary excitation without a need for remeasurement.

0 INTRODUCTION

This paper explores methods of determining a system's transfer function using pseudorandom noise applied in a single-pass process and builds on earlier work by Borish and Angell [1] and, later, Vanderkooy and Rife [2]. Both linear and nonlinear distortion is considered, and a simplified method of system identification is introduced that models a class of system based on a Taylor series, but where each power term in the series is filtered by a unique transfer function. The nonlinear kernels of this model form a subset of a full Volterra model and are extracted here using concatenated finite noise sequences, a method that may be considered the dual of the time-domain spectrometry (TDS) approach of Farina [3]. Although this model is not universal, it can be applied to a range of audio systems and forms a bridge between input-specific measurements and formal methods of identification [4]–[6].

As a further extension, techniques are presented using comb-filtered pseudorandom noise, which allows the simultaneous estimation of both linear and nonlinear distortion by determining the distortion residue falling within the spectral nulls of the excitation. In addition to noise, alternative audio signals can be substituted to enable the relationship between signal and distortion to be explored. Thus by extracting the distortion waveform and using classic block transform analysis the measurement technique can be extended to include systems exhibiting dynamic and perceptually motivated nonlinearity. Finally a graph depicting both the magnitude transfer function and an estimate of the linear error function is presented that offers a more holistic picture of linear system performance. This graph preserves the small differences in frequency response that are often lost because of limited display resolution and system noise. It also expresses the linear dynamic range (LDR) based upon the target signal and residue resulting from linear system error. Although there are caveats to this approach, additional insight into system behavior may be gained where, as an example, a high-performance CD player is assessed.

The exploitation of noise signals has had a long history in the field of system measurement [7], [8]. For example, continuous Gaussian white noise has been used to evaluate a system's transfer function and has proved effective because of its persistent nature and improved measurement signal-to-noise ratio (SNR). More recently maximum-length binary sequences (MLSs) [1], [2] have been used especially for loudspeaker and acoustic measurements. By selecting a periodic MLS of appropriate length to minimize time aliasing distortion [2], a system's periodic impulse response can be calculated directly by circular cross correlation of the measured sequence and the excitation sequence [1], [2]. An interesting observation when performing MLS-based measurements on a nonlinear system is that the resultant impulse response includes a broad distribution of minor impulses [9]. Hence part of the motivation for the present work is to exploit this phenomenon and to construct a simplified nonlinear Volterra model of the system being measured, derived from data extracted using noise sequences.

*Presented at the 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2003 March 22–25; revised 2005 January 14 and February 9.

The techniques described can achieve state-of-the art measurements that compare favorably with MLS methods and multitone methods. Indeed by suitable spectral weighting of the excitation, some of the methods presented here may be considered a generalization of multitone testing. The principal limitations, as with most systems, lie with the quality of the converters used; however, the core processing easily exceeds the resolution of practical audio equipment. The system was developed from a need to explore the performance limits of high-performance CD players, where the aim was to determine the limits of both linear and nonlinear performance, preferably using a single-pass measurement. In this application it was also desirable to have a noise excitation with a uniform probability density function in order to fully exercise the converters; thus binary MLS was rejected (although it is recognized that there are filtering techniques that can address this limitation). Also, it was recognized that Fourier transform techniques were just as effective as the Hadamard transform often used with MLS. In addition, the methods have also been used to assess algorithms in desktop audio editors, such as sample rate converters, of which an example is given in Section 4.3 of course in these applications there is no limitation imposed by converters. The nature of the techniques applied here is that they are easily adapted where, for example, the decision to include fine frequency filtering together with broad-band excitation was seen as pivotal. Such an approach can be configured to push a system under test to its performance boundary so that extremely complex patterns of distortion can be exposed and where, to introduce greater reality, even music-derived signals may be employed. The work also allowed several research threads to be merged, including a long-term personal interest in errors related to linear distortion and to nonlinear modeling. For example, the measurement system can be adapted both to excitation-specific system measurements and to a simplified method of Volterra identification, where it acts as a bridge between the two approaches. It also became evident that once spectral-domain filtering was incorporated to facilitate the segregation of excitation and the resulting distortion, then the problem of measuring both time-varying and non-time-varying systems should be addressed. As a result, the paper offers a contribution to both approaches, where by translating the comb-filtering methodology into the z domain it becomes possible to construct a system using short-term spectral analysis to track dynamic distortion as encountered, for example, with perceptual codecs.

In the measurement procedures described in this study, a rectangular windowed noise (or in some cases music) segment constitutes a frame,[1] with several frames then concatenated to approximate a continuous sequence. An individual frame may also be equalized to have constant-magnitude spectrum but random phase over the length $N = 2^K$ ($2^K - 1$ for MLS). Here K is a positive integer for compatibility with fast Fourier transform (FFT) procedures. Thus most measurement advantages of MLS are retained. Parallels can also be drawn with multi-sine-wave testing [10], [11] since a repetitive noise sequence constitutes a multitone signal, where the fundamental frequency is the frame repetition rate with harmonics forming the tones. Consequently if a noise frame is equalized for a flat-magnitude spectrum, then the multiple tones are of equal amplitude but with random phase relationships.

All processing described in this paper was written in Matlab[2] running on a PC interfaced to high-quality converters. The paper therefore adopts Matlab notation to describe vector operations. Also, to evaluate DVD/CD players (see example measurement, Section 5), the test signals can be burnt to CD/DVD, thus eliminating the need for dedicated test equipment, with further benefits accrued in terms of convenience as all tests employ a single-pass procedure.

The study commences by describing the noise sequence, its equalization and transfer function derivation. Consideration is also given to the formation of a composite test sequence and factors pertaining to the selection of frame length. In all measurement variants discussed it is assumed that analog-to-digital converter (ADC) and digital-to-analog converter (DAC) sampling rates are synchronized as this is critical to proper transform analysis in relation to frame size. Nonsynchronous operation is not considered in this study since both incurred processing errors and remedial windowing artifacts reduce measurement precision. Where a CD/DVD player is used either as an excitation source or for its evaluation, the ADC is slaved to the player sampling rate via the standard Sony/Philips digital interface (S/PDIF).

1 LINEAR SYSTEM IDENTIFICATION USING PSEUDORANDOM NOISE

The core technique exploited this study to measure a system's transfer function is based on a repetitive equalized noise sequence (that is, pseudorandom noise), where a noise sequence must be generated with a duration greater than the time over which a system's impulse response $h(t)$ remains significant. The noise sequence is defined in discrete time, where Nyquist sampling theory determines the sampling rate as a function of bandwidth. Consequently measurement accuracy is bounded by both time [2] and frequency [12] aliasing distortion. The noise sequence is concatenated to form repetitive frames with no interframe guard bands. It then follows that to extract spectral information, only sampling-rate synchronization is required; exact frame synchronization, although beneficial, is not mandatory, provided the frame size is known, as the transforms used are circular. Consequently a sample-rate synchronized ADC captures the output response of the system being tested and the frame detection achieved both by counting the frames of 2^K samples and using a synchronization preamble embedded in the test sequence.

Consider a noise vector noise(n) with rectangular probability density function, generated over $N = 2^K$ samples,

[1] A frame is defined here as a finite set of uniform samples represented as a vector.

[2] Matlab is a trade name of Mathworks, Inc.

where K is a positive integer and n is the vector [1:N]. Expressed in Matlab notation,

$$\text{noise}(n) = \text{rand}(1, N). \quad (1)$$

The frequency-domain noise sequence noisef(n) is calculated using a one-dimensional FFT. Thus if fft(noise(n)) is the length-N discrete Fourier transform of the sequence noise(n), where N is the length of the sequence noise(n), then

$$\text{noisef}(n) = \text{fft}(\text{noise}(n)). \quad (2)$$

A time-domain excitation sequence test(n) with constant-magnitude spectrum but random phase is then determined using spectral normalization and the inverse Fourier transform, where[3]

$$\text{test}(n) = \text{real}(\text{ifft}(\text{noisef}(n)./(\gamma + \text{abs}(\text{noisef}(n))))). \quad (3)$$

Although processing using the inverse Fourier transform should in this application return only real numbers, computational errors result in small but finite imaginary terms, which is not the norm for time-domain sampled data and is unacceptable when writing a wav file. Hence only the real part of the transform is selected both here and in later inverse transform operations. In addition, the real function also halves the vector storage requirement as the small nonzero imaginary elements are deleted. Also, a constant γ (say 10^{-12}) is introduced to eliminate small-number division anomalies in the spectral normalization process. Alternatively this potential problem can be avoided completely by using a complex exponential function testf(n) with random phase to guarantee an exactly constant magnitude spectrum,

$$\text{testf}(n) = \exp(\text{i}*\text{angle}(\text{fft}(\text{noise}(n)))) \quad (4)$$

that is, $|\text{testf}(n)| = 1$. The corresponding time-domain vector test(n) then follows as

$$\text{test}(n) = \text{real}(\text{ifft}(\text{testf}(n))). \quad (5)$$

In practice, because the test sequences have to be generated only once, it is prudent to sift a number of computed examples in order to seek a sequence with low crest factor such that the measurement SNR can be enhanced. A composite repetitive excitation pattern of test(n) is then constructed, as shown in Fig. 1, which includes both a zero pulse preamble and an embedded synchronization sequence defined as [0 0 0 . . . 0 0 0 0 1 1 −1 −1 0 0 0 0]. Hence by using cross-correlation-based detection the commencement of the recovered test sequence can be detected to sample accuracy. Finally the composite sequence is peak amplitude normalized, quantized to the required bit depth, and converted to a two-channel linear pulse code modulation (LPCM) wav file for subsequent outputting to the system under test, typically via a 96-kHz, 24-bit DAC.

The measured data are captured using a sample synchronized ADC, where following frame synchronization a data frame output(n) is extracted. Taking testf(n) from Eq. (4), the complex transfer function TF(n) of the system is then calculated using element-by-element division,

$$\text{TF}(n) = \text{fft}(\text{output}(n))./\text{testf}(n). \quad (6)$$

The magnitude response $M(n)$ and the phase response $P(n)$ then follow as

$$M(n) = \text{abs}(\text{TF}(n)) \quad \text{and} \quad P(n) = \text{angle}(\text{TF}(n)) \quad (7)$$

and the system impulse response $h(n)$ as

$$h(n) = \text{real}(\text{ifft}(\text{TF}(n))). \quad (8)$$

Alternatively, if excess phase information is not required, then the magnitude response of the spectrum may be calculated directly from output(n) since the excitation was normalized to a constant-magnitude spectrum. The minimum-phase impulse response $h_{\text{min}}(n)$ then follows from the Hilbert transform [13],

$$h_{\text{min}}(n) = \text{real}(\text{ifft}(\exp(\text{conj}(\text{hilbert}(\log(\text{abs}(\text{fft}(\text{output}(n)))))))))). \quad (9)$$

However, if frame synchronization is not achieved, then because of circularity and repetitive noise frames, the true impulse response can still be derived, but within an arbitrary time shift. A key factor in this process is for the noise frame to exceed the duration of $h(n)$. Otherwise the circular nature of the test procedure allows time aliasing distortion, which is a fundamental and irreversible measurement distortion. With this proviso then, for a linear system, this procedure creates an exact model within the constraints of measurement bandwidth and sampling rate. In the next section the process is extended to include approximate nonlinear identification employing a simplified Volterra model.

2 NONLINEAR SYSTEM IDENTIFICATION USING PSEUDORANDOM NOISE

Farina [3] has reported a TDS-based scheme to identify mildly nonlinear systems in terms of a simplified Volterra model. An alternative measurement procedure is described here using pseudorandom noise similar to that presented in Section 1. The model is appropriate for stationary nonlinear systems with memory where Volterra kernels expressed as impulse responses encapsulate higher order frequency dependence. However, only powers of the input

[3]See Matlab glossary in Appendix A for a definition of operator "."

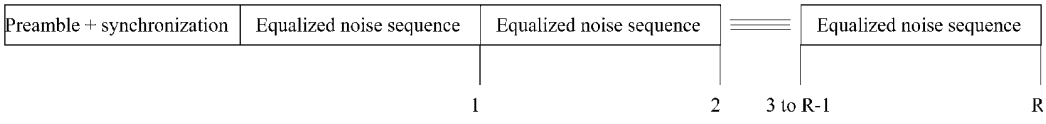


Fig. 1. Test signal structure with preamble and synchronization sequence.

sequence are included whereas cross-product terms inherent in the full Volterra model are ignored. As such the method is positioned between input-specific measurements and a fully populated Volterra model that can predict the output response to a generalized excitation [14]–[16]. Also as a corollary, linearity can be tested as linear and nonlinear responses are segregated.

2.1 Volterra Modeling

The general Volterra model linking input vector $x(n)$ to output vector $y(n)$ is described by M-dimensional convolution,

$$y(l)\Big|_{l=1}^{N} = \sum_{r=1}^{M}\sum_{r=1}^{N-1}\cdots\sum_{r=1}^{N-1} h_r(i_1, i_2, \ldots, i_r) \times x(l-i_1)x(l-i_2)\cdots x(l-i_r) \quad (10)$$

where N is the memory length of the filters. Consequently the general Volterra model requires a large number of coefficients to populate a multidimensional space as unique impulse responses are associated with all the convolutional combinations of power and cross-product terms. However, in the simplified representation the only convolutions included are those associated with powers of the input sequence.

Consider a nonlinearity where output vector $y(n)$ is related to input vector $x(n)$ by a power series of order M,

$$\begin{aligned} y(n) &= a_0 + a_1x(n) + a_2x(n).\hat{}2 + \cdots + a_rx(n).\hat{}r + \cdots \\ &\quad + a_Mx(n).\hat{}M \\ &= [a_0\ a_1\ a_2 \ldots a_M].*[1\ x(1)\ x(2)^2 \ldots x(n)^r \ldots x(n)^M] \\ &= [a].*[1\ x(1)\ x(2)^2 \ldots x(n)^r \ldots x(n)^M]. \end{aligned} \quad (11)$$

To identify this memoryless system fully, only the M coefficients $[a]$ need to be determined. However, for the simplified Volterra model with memory the $[a]$ coefficients associated with each term in the power series translate to a set of M impulse responses $[h(n)]$, reducing the M-dimensional convolution[4] in Eq. (10) to just

$$\begin{aligned} y(n) &= h_0 + h_1(n) \otimes x(n) + h_2(n) \otimes (x(n).\hat{}2) + \cdots \\ &\quad + h_r(n) \otimes (x(n).\hat{}r) + \cdots \\ &\quad + h_M(n) \otimes (x(n).\hat{}M). \end{aligned} \quad (12)$$

In Eq. (12) h_0 is the dc term and $h_1(n)$ is the linear system impulse response, while for $r = 2, \ldots, M$, $h_r(n)$ describes the respective impulse responses relating to the power terms $x(n).\hat{}r$. Because the Volterra model described by Eq. (12) contains M impulse responses, M independent noise sequences are required in the identification procedure, although vectors are transformed into the frequency domain to allow simpler element-by-element multiplication rather than time-domain convolution.

2.1.1 Vector and Transform Notation

For an M-dimensional system $x_r^s(n)$ represents an input vector r where each element is raised to the power s, $y_r(n)$ is output vector r, and $h_r(n)$ is an impulse response r. The corresponding Fourier transforms Y_r, $X_{r,s}$, H_r are then defined,

$$Y_r = \mathrm{fft}(y_r(n)), \qquad X_{r,s} = \mathrm{fft}(x_r^s(n)), \qquad H_r = \mathrm{fft}(h_r(n)).$$

In the following analysis vectors are transformed between time and frequency domains to transmute convolution to element-by-element multiplication. Consider M Fourier-transformed vectors Y_r derived from M uncorrelated excitation noise vectors $X_{r,1}$ and applied successively to Eq. (12) to form M equations, that is,

$$\begin{aligned} Y_1 &= H_0 + H_1.*X_{1,1} + H_2.*X_{1,2} + \cdots + H_r.*X_{1,r} + \cdots \\ &\quad + H_M.*X_{1,M} \\ Y_2 &= H_0 + H_1.*X_{2,1} + H_2.*X_{2,2} + \cdots + H_r.*X_{2,r} + \cdots \\ &\quad + H_M.*X_{2,M} \\ &\vdots \\ Y_M &= H_0 + H_1.*X_{M,1} + H_2.*X_{M,2} + \cdots + H_r.*X_{M,r} + \cdots \\ &\quad + H_M.*X_{M,M}. \end{aligned}$$

Rewriting in matrix form,

$$\begin{bmatrix} Y_1 - H_0 \\ Y_2 - H_0 \\ \vdots \\ Y_M - H_0 \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,M} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,M} \\ \vdots & \vdots & & \vdots \\ X_{M,1} & X_{M,2} & \cdots & X_{M,M} \end{bmatrix} .* \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_M \end{bmatrix} \quad (13)$$

that is,

$$[Y] = [X].*[H]. \quad (14)$$

In Eq. (13) H_0 is the output dc offset. It is measured when the input signal is zero and the system quiescent. To determine impulse responses $[H]$, define $[Z] = [X]^{-1}$ (see Appendix B.1 for inversion), whereby the decoding matrix equation expressed in terms of M measured output vectors becomes

$$[H] = [X]^{-1}.*[Y] = [Z].*[Y]. \quad (15)$$

Eq. (15) describes an input-specific decoding key, where $[Z]$ is related uniquely to the set of M noise vectors and has to be calculated only once. This simplifies computation, since a typical $[M, M, N]$ matrix for $M = 8$ and $N = 2^{14}$ contains 2^{20} complex elements. To complete the analysis the set of Volterra impulse responses $[h]$ follow from the inverse fast Fourier transform of matrix $[H]$, where

$$[h] = \mathrm{real}(\mathrm{ifft}([H])). \quad (16)$$

2.2 Test Sequence Generation

In performing a system measurement it is critical for each of the M noise vectors to have the same relative level. To facilitate this requirement, a composite signal is constructed where each noise vector is repeated four times to

[4]Although not a Matlab operator, the symbol $\otimes$ represents circular convolution.

form a subframe, then all M subframes are concatenated into a single sequence. Subframes containing repeated sequences allow convergence to a pseudoperiodic output signal and introduce a margin against subframe misalignment within the decoder. Repeated frames also enable noise averaging to be applied to improve SNR. A preamble and synchronization sequence is then added similar to that described in Section 1 to facilitate demultiplexing of the M sequential system responses. Consequently all M sequences are processed almost simultaneously in a single measurement so differential gain errors are eliminated.

2.3 Volterra Modeling Validation

To validate the Volterra modeling scheme the computational process is divided into three routines (discussed in Appendix B.2 and B.3) and then applied to two nonlinear examples.

- An M-vector composite test sequence is generated and the three-dimensional decoding matrix $[X]$ inverted to $[Z]$ as a one-off calculation.
- Two example simulations are performed on a stationary nonlinearity, with and without memory.
- Output data are analyzed and Volterra responses computed using Eq. (15).

1) *Nonlinearity without Memory.* The first example system employs just a power series as described by Eq. (11), where the excitation is processed sequentially, sample by sample. The coefficient matrix $[a]$ of the series is selected arbitrarily and the performance of the decoding algorithm evaluated by comparing the amplitude of the Volterra frequency-domain responses to $[a]$. For this memoryless case each Volterra response has constant amplitude, assuming no measurement channel filtering. The selected coefficient values for $M = 8$ are

$$[a] \equiv [0 \quad 1 \quad 0.001 \quad 0.05 \quad 0.0002 \quad 0.02 \quad 0.0005 \quad 0.05 \quad 0.001].$$

2) *Nonlinearity with Memory.* The second example, depicted in Fig. 2 for $M = 8$, adds a set of linear low-pass output filters applied to each power term of the nonlinear series described by Eq. (12). All eight filters have brick-wall responses with respective cutoff frequencies of 16, 14, 12, 10, 8, 6, 4, and 2 kHz.

Volterra analysis was applied to each nonlinear system, where $N = 2^{15}$. For the memoryless case the eight derived Volterra frequency-domain responses are shown in Fig. 3, those for the second example with memory are shown in Fig. 4. Results correspond to theory within the bounds of measurement noise where all responses match those preselected, including correct identification of the eight brick-wall filter responses.

3 NONLINEAR DISTORTION ESTIMATION USING COMB-FILTERED NOISE SEQUENCES

It is known that nonlinear systems when excited by broad-band signals produce complicated spectral patterns of intermodulation distortion [9]. Section 2 exploited this phenomenon in Volterra identification. In this section it is shown that a combination of noise excitation and comb filtering enable distortion and signal to be partially separated, thus allowing simultaneous estimates of both linear and nonlinear distortion. The proposed spectral interleave technique takes inspiration from both Belcher [17] and techniques of multitone testing [10], [11].

3.1 Evaluation of Spectral Interleave Measurement System

The measurement procedure developed in Section 1 is extended to include frequency-domain comb filtering by introducing an interleave-mask function intf0(n). The mask forces regions of zero spectral energy in the excitation sequence and is also used in measurement analysis to segregate signal and distortion. Applying the frequency-domain vector testf(n) with constant magnitude and random phase from Eq. (4), the frequency-normalized and comb-filtered test sequence $\text{test}_{\text{int}}(n)$ becomes

$$\text{test}_{\text{int}}(n) = \text{real}(\text{ifft}(\text{intf0}(n).*\text{testf}(n))). \tag{17}$$

Two interleave-mask options were incorporated for frequency domain filtering:

1) Alternating binary sequence . . . 010101 . . . to create a regular pattern of active and zero frequency bins.

2) MLS to create a random pattern of active and zero spectral bins.

To construct a frequency-domain mask that takes proper account of the sampled-data format of the excitation, appropriate spectral symmetry is required about the half-sampling frequency, where if N is the number of samples in the noise sequence,

$$\text{intf0}(1{:}N) = [0 \quad \text{intf0}(1{:}N/2-1) \quad 0 \quad \text{intf0}(N/2-1{:}{-}1{:}1)]. \tag{18}$$

Hence if mdata(n) is measured data windowed precisely to match the excitation sequence, then applying testf(n) from Eq. (4), the system frequency response GF(n) is extracted as

$$\text{GF}(n) = \text{intf0}(n).*[\text{fft}(\text{mdata}(n))./\text{testf}(n)] \tag{19}$$

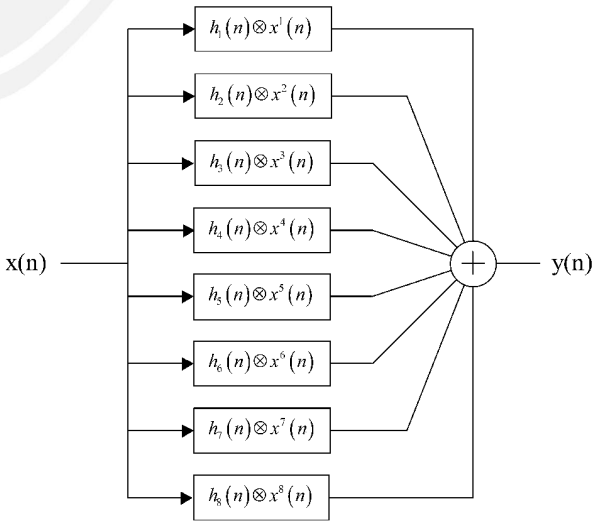


Fig. 2. Simplified Volterra model based on Eq. (12).

while the distortion spectrum DF(n) is obtained using the complementary mask, where

$$\mathrm{DF}(n) = [\mathrm{ones}(1,N) - \mathrm{intf0}(n)].*[\mathrm{fft}(\mathrm{mdata}(n))./\mathrm{testf}(n)]. \quad (20)$$

In practice vector lengths up to $N = 2^{20}$ have been used successfully without experiencing computational problems in the Matlab FFT. This represents an excitation repetition period of approximately 23.77 second at a sampling rate of 44.1 kHz, corresponding to a frequency-bin

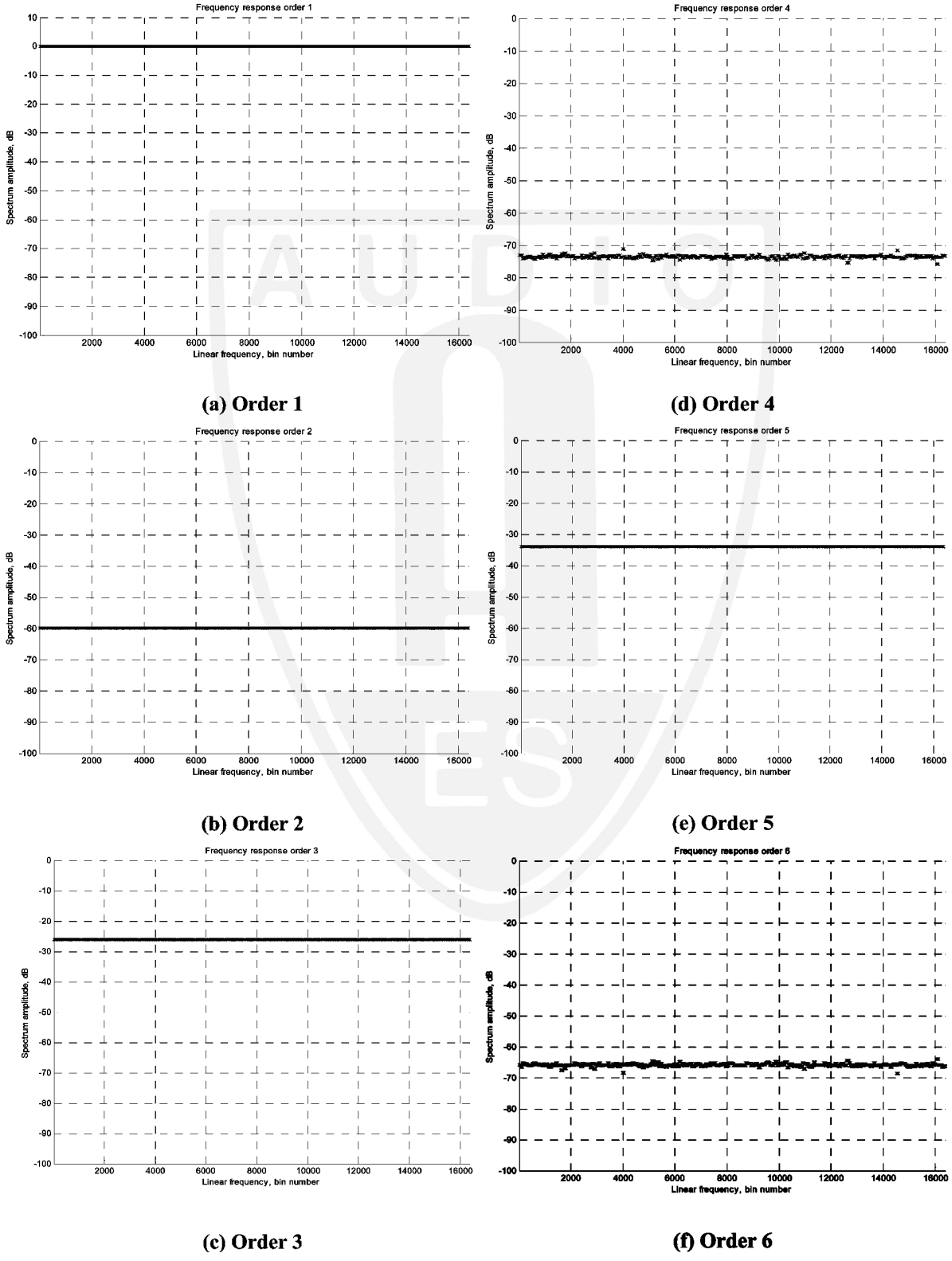


Fig. 3. Volterra frequency-domain responses for nonlinear system without memory.

spacing of 0.042 Hz. Because of the fine frequency resolution achievable, precise sampling rate synchronization is mandatory; otherwise spectral leakage into adjacent bins introduces false estimates of distortion.

3.2 Noise Averaging and Full Spectral Resolution Using Multiple Frames

Because excitation sequences are quasi repetitive, it is straightforward to capture a number of consecutive frames and perform noise averaging to improve the measurement SNR, where the theoretical noise improvement $\mathrm{SNR}(N)|_{\mathrm{imp}}$ for λ averaged frames is

$$\mathrm{SNR}(N)|_{\mathrm{imp}} = 10\log_{10}(\lambda). \tag{21}$$

However, once multiple frames are used, two sets of frames can be constructed using complementary interleave masks, where, for example, four frames are assigned a normal interleave mask while a subsequent four frames are assigned a complementary interleave mask. Hence a single compound measurement sequence enables measured data to be gathered where the respective frequency bins assigned to frequency response and distortion estimation are interchanged. The two sets of data from successive groups of frames can then be segregated temporally and subsequently merged to produce noninterleaved full-resolution spectra for both the transfer function $\mathrm{GF}_{\mathrm{full}}(n)$ and distortion $\mathrm{DF}_{\mathrm{full}}(n)$ as follows.

Let the transfer function and the distortion spectra derived from the first set of measured sequences mdata_1 be

$$\mathrm{GF}_1(n) = \mathrm{intf0}(n).*[\mathrm{fft}(\mathrm{mdata}_1(n))./\mathrm{testf}(n)] \tag{22}$$

$$\mathrm{DF}_1(n) = [\mathrm{ones}(1,N) - \mathrm{intf0}(n)].*[\mathrm{fft}(\mathrm{mdata}_1(n))./\mathrm{testf}(n)] \tag{23}$$

and from the second set of measured sequences mdata_2,

$$\mathrm{GF}_2(n) = [\mathrm{ones}(1,N) - \mathrm{intf0}(n)].*[\mathrm{fft}(\mathrm{mdata}_2(n))./\mathrm{testf}(n)] \tag{24}$$

$$\mathrm{DF}_2(n) = \mathrm{intf0}(n).*[\mathrm{fft}(\mathrm{mdata}_2(n))./\mathrm{testf}(n)]. \tag{25}$$

The full-resolution spectra $\mathrm{GF}_{\mathrm{full}}(n)$ and $\mathrm{DF}_{\mathrm{full}}(n)$ are then calculated,

$$\mathrm{GF}_{\mathrm{full}}(n) = \mathrm{GF}_1(n) + \mathrm{GF}_2(n) \tag{26}$$

$$\mathrm{DF}_{\mathrm{full}}(n) = \mathrm{DF}_1(n) + \mathrm{DF}_2(n). \tag{27}$$

The corresponding non-comb-filtered time-domain periodic sequences $\mathrm{out}_f(n)$ and $\mathrm{dist}_f(n)$ follow from the inverse fast Fourier transform as

$$\mathrm{out}_f(n) = \mathrm{real}(\mathrm{ifft}(\mathrm{GF}_{\mathrm{full}}(n))) \tag{28}$$

$$\mathrm{dist}_f(n) = \mathrm{real}(\mathrm{ifft}(\mathrm{DF}_{\mathrm{full}}(n))). \tag{29}$$

To gain additional insight into this process, including inherent time smearing of both excitation and retrieved distortion sequences which results from comb filtering, Appendix B.4 presents a time-domain description of the measurement process.

3.3 Examples Using Comb-Filter Measurement Procedure

To evaluate the measurement system incorporating comb filtering, three example systems were simulated. In each of these tests the excitation sequence used 44.1-kHz sampling with 16-bit resolution.

3.3.1 Linear Filter

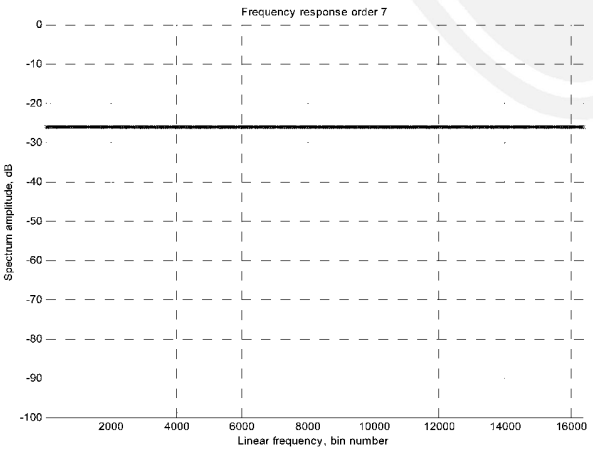
A linear filter was simulated, where the frequency response had two linear segments located above and below 1 kHz, with the attenuation peaking at 4.5 dB. Fig. 5 shows the simulation results, where the correct amplitude response has been obtained and where all "zero bins" contain only quantization noise, thus confirming system linearity.

3.3.2 Nonlinearity, No Filtering

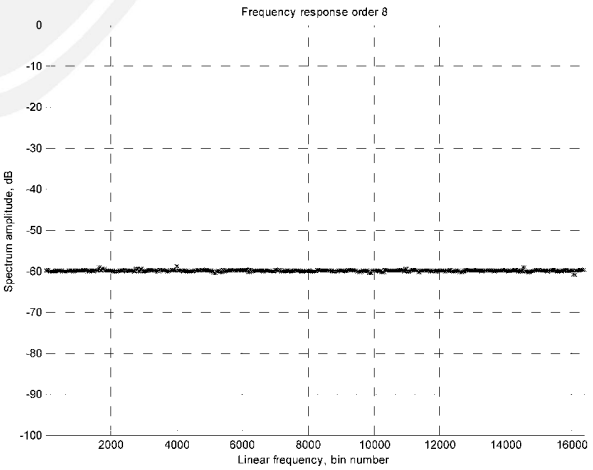
The second example used a memoryless nonlinearity defined by

$$y(n) = x(n) + 0.01x^2(n). \tag{30}$$

Fig. 6(a) shows both the magnitude frequency response and the intermodulation distortion, where the peak distortion is



(g) Order 7



(h) Order 8

Fig. 3. *Continued*

about 40 dB below the input sequence. In this diagram the frequency response appears almost as a straight line, although with closer scrutiny Fig. 6(b) reveals the magnitude spectrum to have noiselike deviation about unity created by the noise excitation interacting with the nonlinearity.

3.3.3 Nonlinearity with Filtering

The third example employs the same nonlinearity as used in Section 3.3.2, but with the inclusion of a sixth-order Chebyshev low-pass filter with a 5-kHz bandwidth.

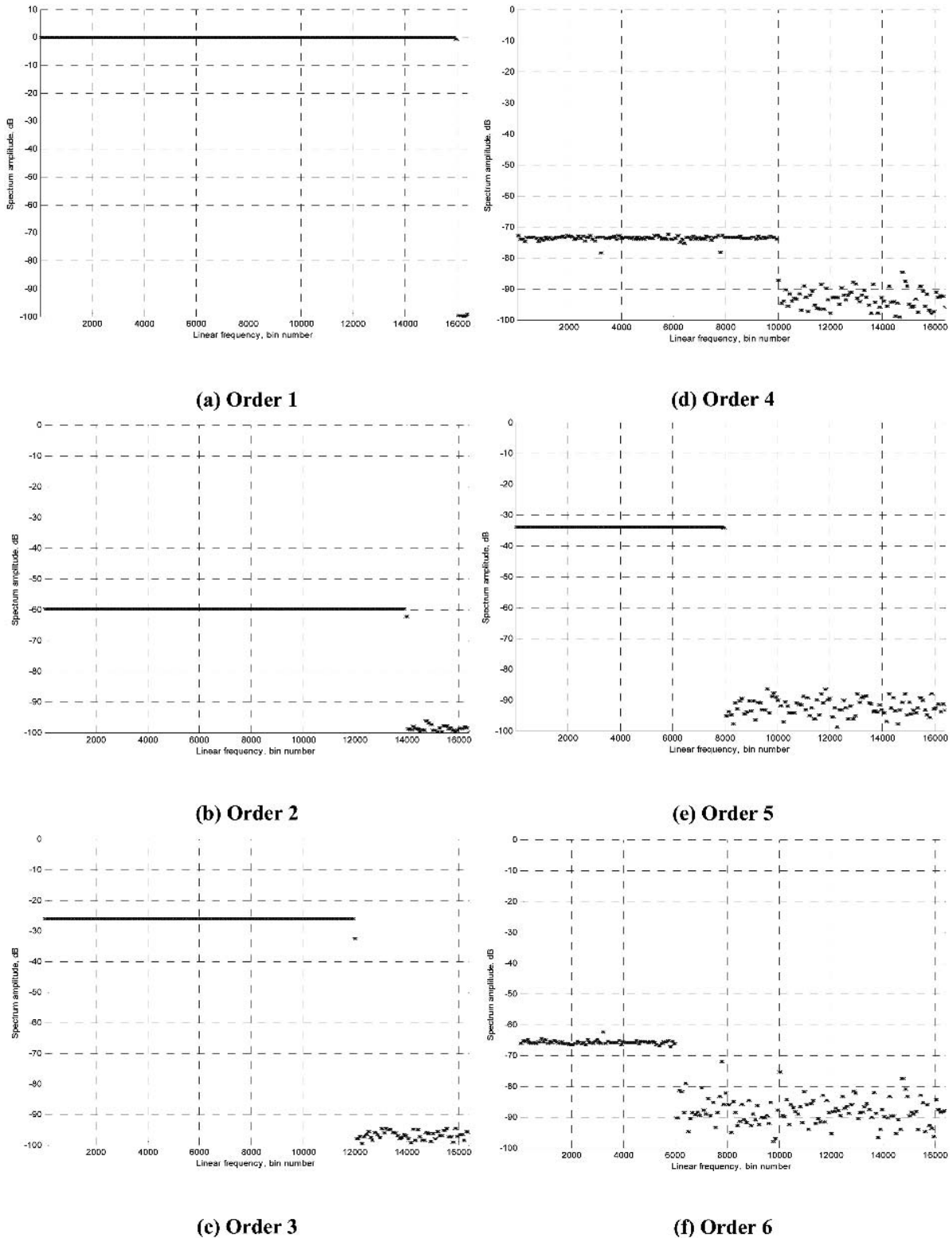


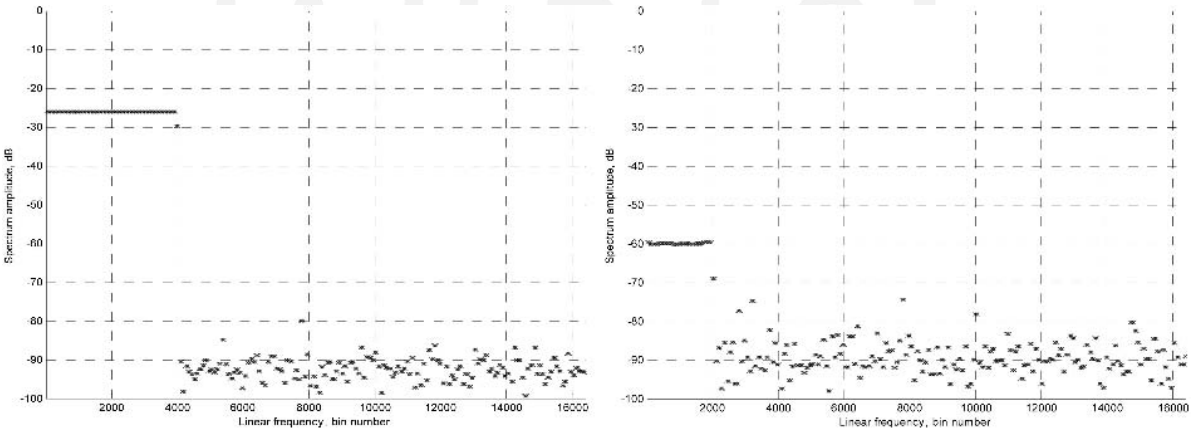
Fig. 4. Volterra frequency-domain responses for nonlinear system with memory.

Three cases were simulated: 1) filter only without nonlinearity; 2) filter located before the nonlinearity; 3) filter located after the nonlinearity.

Derived magnitude responses and distortion spectra for these three cases are shown in Figs. 7–9. The results for just the filter confirm accurate identification of the transfer function, with low distortion levels and only mild levels of noise shaping and progressive spectral corruption as the filter gain approaches the measurement noise floor. When the nonlinearity is positioned after the filter, Fig. 8 now reveals frequency-shaped distortion that follows the attenuation characteristic of the low-pass filter. Finally, Fig. 9 shows again the filter response, but here the filter predictably band-limits the broad-band distortion created by the nonlinearity.

4 SYSTEM TESTING USING MUSIC SIGNALS

This section investigates three nonlinear system examples using a periodic music excitation combined with the procedures described in Sections 1 and 3. The first is a memoryless nonlinearity, the second an MP3 codec, and the third a desktop editor sampling-rate converter. The rationale for choosing music is that certain nonlinear audio systems such as perceptual codecs produce excitation-specific distortion critical to their operational philosophy. Also, because system modeling is nonfeasible for many classes of nonlinearity, the relationship between excitation and distortion to auditory masking [18] establishes the foundation for perceptually motivated objective analysis (see Holler et al. [19], [20]). Two variations of the mea-



(g) Order 7 (h) Order 8

Fig. 4. *Continued*

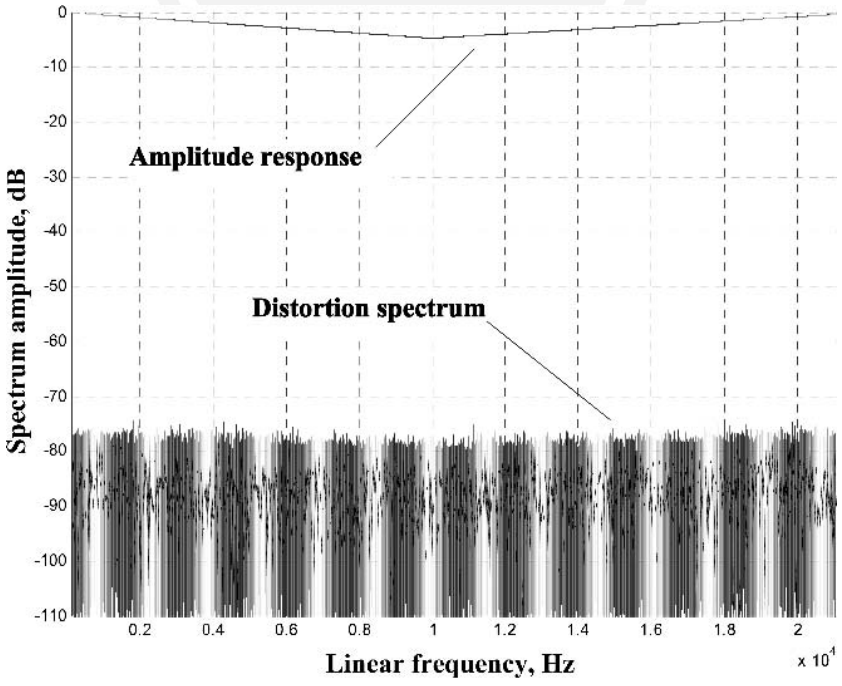


Fig. 5. Measured amplitude response and distortion spectrum of linear filter.

surement system are presented. The first, described in Section 4.1, uses a full-frame music sequence combined with comb filtering, with the spectral analysis applied full frame, a technique better suited to non-time-varying systems. In the second variation, presented in Section 4.2, the test sequence is modified so that signal excitation and distortion generation become uniquely localized in time. Also spectral analysis of both excitation and distortion is applied to overlapping data blocks of a duration of typically 25 ms. These two expedients enable the measurement system to be applied to time-varying systems and thus include, for example, perceptual codecs. The objective is to display how the short-term excitation spectrum tracks the short-term signal distortion, thus facilitating the inclusion of more sophisticated masking models to enable formal perceptually motivated analysis. In the example presented in Section 4.2, a short zero signal segment is embedded in the input in order that a corresponding null in the distortion spectrum can be observed to validate correct temporal linkage between excitation and distortion. Also, in order to gain greater insight into how the comb filters are visualized in the time domain, Appendix B.4 presents a z-domain analysis of the overall process, including signal generation, comb filter-

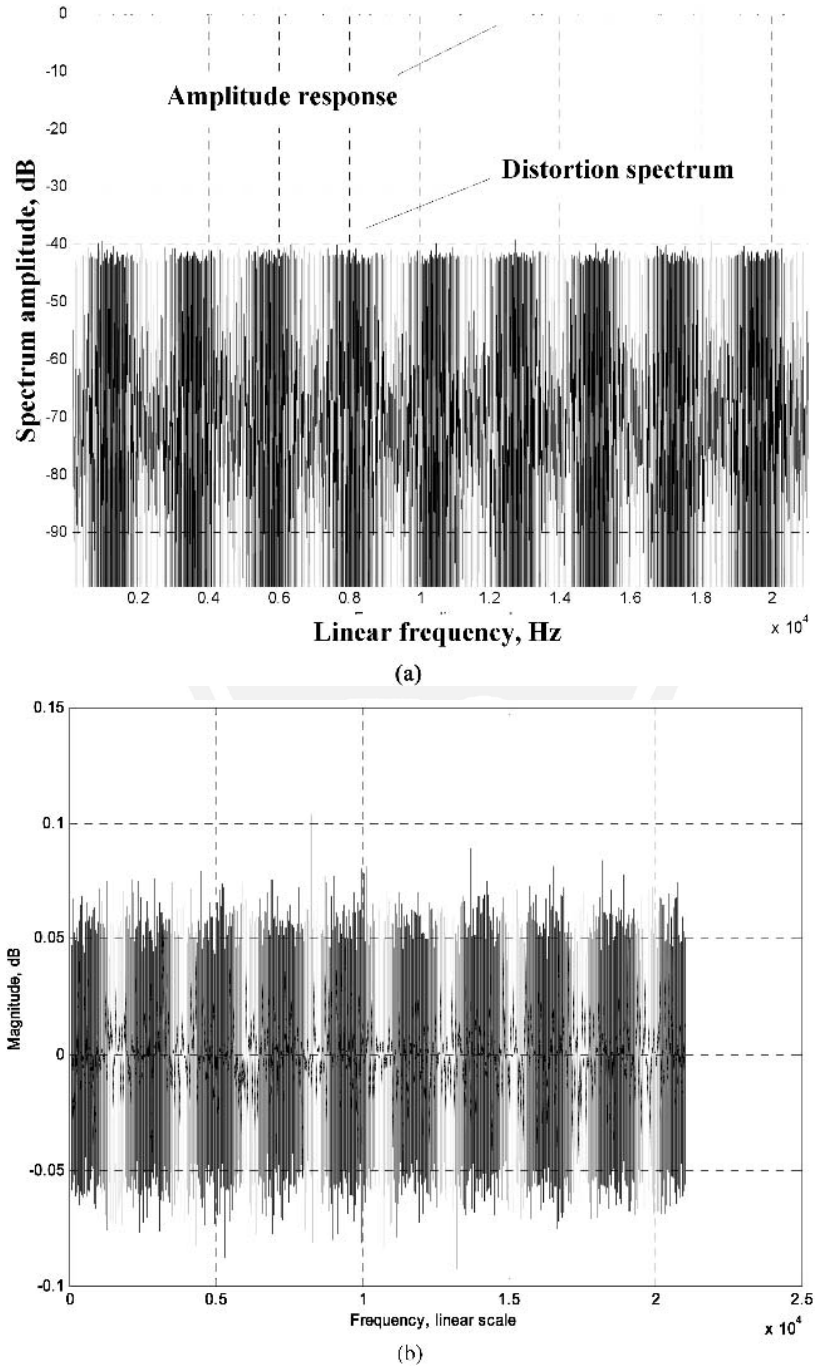


Fig. 6. (a) Distortion derived using memoryless nonlinearity. (b) Noiselike magnitude frequency response for memoryless nonlinearity.

ing, and data analysis, whereas Appendix B.5 describes the method of block-based Fourier analysis used to form the three-dimensional spectral temporal-frequency output displays.

4.1 Memoryless Nonlinearity Evaluated Using Music Excitation

For comparison, the same nonlinearity as used in Section 3 was tested with music [see Eq. (30)] and comb-filter processing. However, the excitation consisted now of an $N = 2^{20}$ sample sequence of music, requantized with dither[5] from 16 to 24 bit to extend measurement resolution. Fig. 10 shows both full-frame signal and distortion spectra. Interestingly here the distortion spectral envelope is revealed to be similar to that of the music signal.

[5]Dither is set at the 16-bit level, generated to 32-bit resolution.

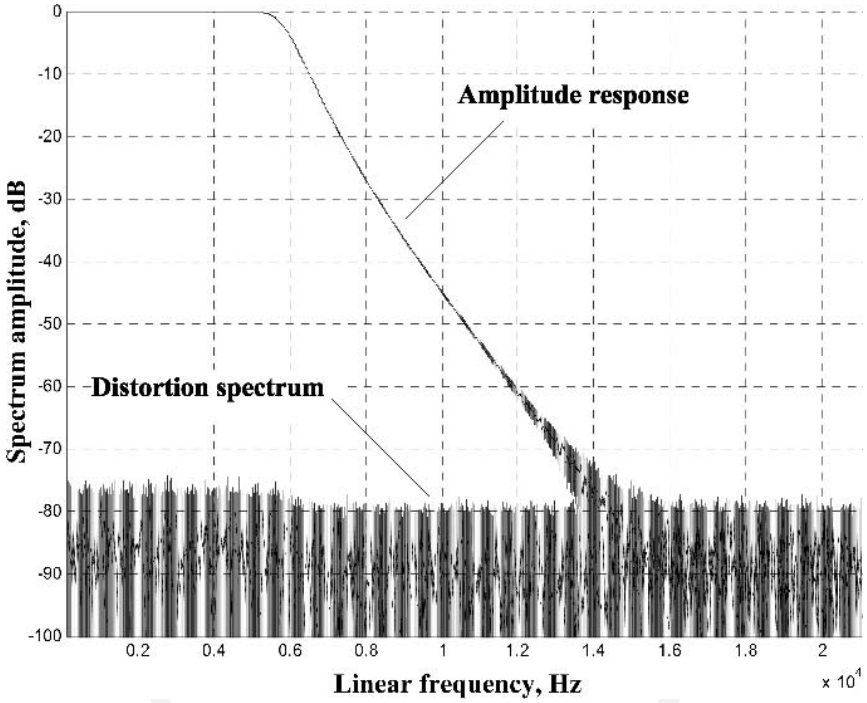


Fig. 7. Spectral results derived using full-resolution interleave procedure for low-pass filter.

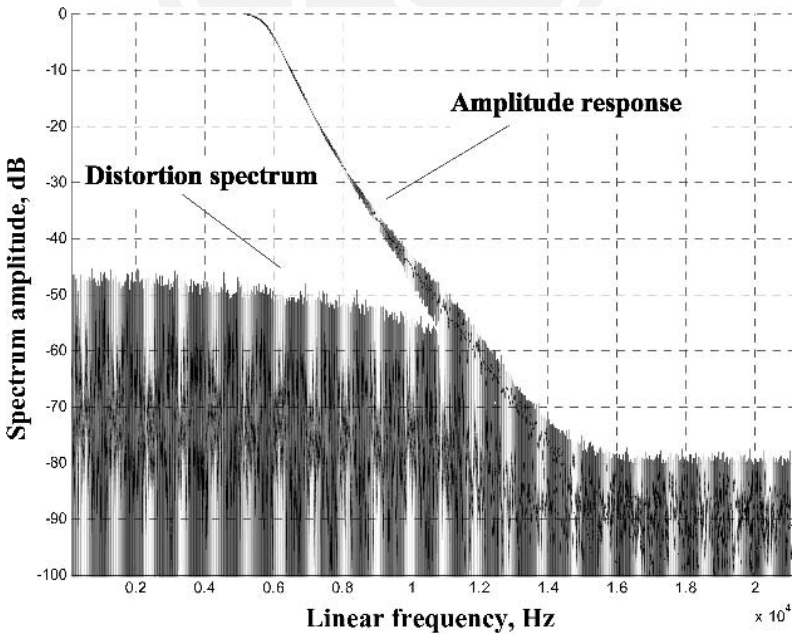


Fig. 8. Spectral results derived using full-resolution interleave procedure for low-pass filter followed by memoryless nonlinearity.

4.2 Perceptual Codec Evaluated Using Music Excitation

Two illustrative measurements were performed on an MP3 codec operating at 192 kbit/s. The first was based on spectral interleave analysis, whereas the second adapted the nonfiltered procedure of Section 1 to extract true distortion using a finely calibrated difference technique. When applying spectral analysis across a whole music frame, as in Section 4.1, although the magnitude response is a faithful average assessment, the distortion spectrum is unrealistic because with perceptually motivated coding the error spectrum undergoes dynamic modulation in an attempt to match the masking behavior of the human auditory system. However, a more representative performance evaluation can be solicited by applying short-term spectral

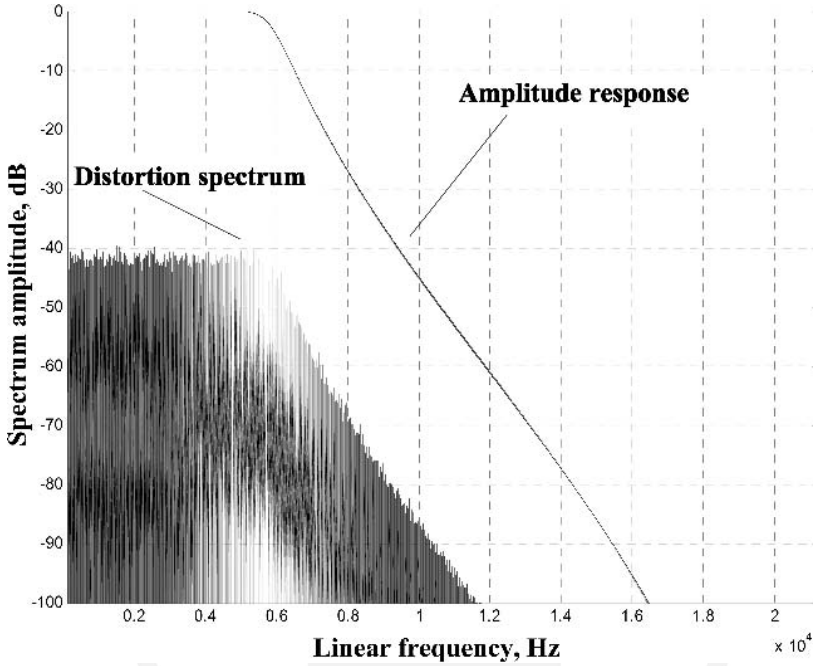


Fig. 9. Spectral results derived using full-resolution interleave procedure for low-pass filter preceded by memoryless nonlinearity.

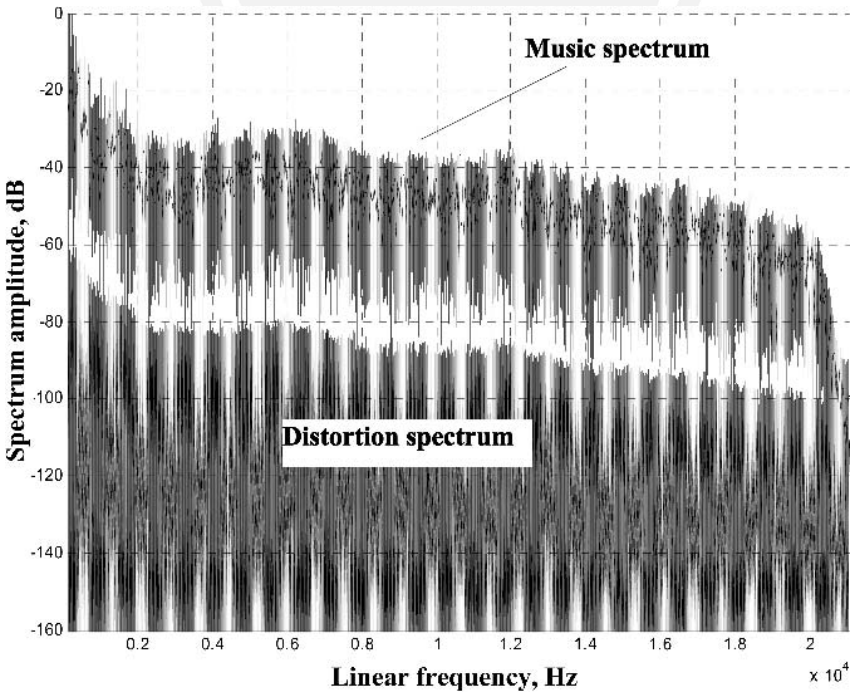


Fig. 10. Music signal and distortion spectra for memoryless nonlinearity.

analysis (see Appendix B.5) to the measurement time-domain output sequences $out_f(n)$ and $dist_f(n)$ derived from Eqs. (28) and (29).

4.2.1 MP3 Codec Example Evaluation Using Comb-Filter Analysis

For a measurement to have relevance in the context of a perceptual codec, distortion and signal must be coherent in time. In practice this condition is not met when spectral interleaving and full-length test sequences are used as comb filtering introduces circular time dispersion to both the excitation and the subsequently recovered distortion (see Appendix B.4). However, by applying a rectangular window to force frame samples in the range $N/2$ to N to zero in the music source, the following changes arise:

1) Eq. (47) forces $test_1(1{:}m/2) = test_1(1 + m/2{:}m)$ without time-aliasing distortion.

2) The codec now experiences the test sequence twice in each analysis frame, but because of the stochastic processes within a perceptual codec the distortion generated in each repeated sequence should be similar in terms of its spectral envelope but lacking coherence due to phase noise; thus $dist_1(1{:}m/2) \neq dist_1(1 + m/2{:}m)$.

3) In decoding distortion, although Eq. (50) implies a zero result if the distortion waveform were repeated precisely when the excitation is repeated, noncoherence implies that a significant fraction of the distortion is retrieved.

4) As a corollary, if Eq. (50) yields a low output incompatible with the expected level of distortion, this implies coherence and actually reveals poor randomization of coding artifacts.

By way of example, Fig. 11 shows spectral-domain results for output and distortion for an MP3 codec at 192 kbit/s. A short zero-level gap was included within the music sequence to confirm the temporal coincidence of signal and distortion. This zero signal segment is clearly resolved in both spectral displays.

4.2.2 MP3 Codec Example Evaluation Using Difference Test

To eliminate problems of time dispersion using comb filters and to allow the full vector sequence to be used in analysis, the system was adapted to enable the input–output error to be determined. The measured signal vector of length N was recovered as described in Section 1, and both excitation and measured vectors were normalized to have identical standard deviation. Circular correlation together with circular data shifting was then used to achieve precise time alignment and to correct for time delay in the codec, allowing true distortion to be calculated by subtraction.

The measured data and derived distortion are finally processed to produce a dynamic spectrum using the same block analysis as in Section 4.2.1 (see Appendix B.5). The spectral-domain result for the distortion (using the same music segment with a zero gap to facilitate comparison) is shown in Fig. 12(a) whereas Fig. 12(b) presents the corresponding plot of signal spectrum minus distortion spectrum [see Eq. (60)]. Although the details of the distortion spectra derived using the two techniques differ, a similar form is evident.

4.3 Desktop Audio Editor Sample-Rate Conversion

As an example of the use of the comb-filter-based measurement system to evaluate algorithms within a desktop audio editor, the procedure was applied to both integer and noninteger sample-rate conversion. The exploration had two stages of sample-rate conversion and converted audio data initially at 44.1 kHz in two directions, such that the output file sampling rate returned the same rate as the input file. The input frame length had 2^{20} samples and thus gave a frequency resolution of about 0.042 Hz. The results are shown in Fig. 13. It can be seen that for integer sample-rate conversion there is virtually no distortion evident whereas for the noninteger conversion, apparent high-level distortion has been generated. In fact the distortion remained low in both cases, but the algorithm introduced small block-based frequency shifting errors during conversion due to the noninteger conversion ratio. Thus signal frequency-dependent spillage into the adjacent null bands was resolved, as shown in Fig. 13(b).

5 GRAPHICAL DISPLAY OF SMALL RESPONSE ERRORS

To conclude the discussion on system measurement, this section describes a means of representing small frequency-response deviations more accurately. As an illustration, consider a discrete echo of time delay τ_{echo} and relative amplitude γ_{echo}, where the discrete frequency-domain transfer function $G(n)$ of the system is

$$G(n) = 1 + \gamma_{echo} * \exp(-i2\pi n f_0 \tau_{echo}) \quad (31)$$

with f_0 being the block repetition frequency of the excitation sequence and $n = 1{:}N$. The magnitude response shows periodic frequency variation, where the peak-to-peak response variation Dev_{dB} about the target response is

$$Dev_{dB} = 20 * \log_{10}((1 + \gamma_{echo})/(1 - \gamma_{echo})). \quad (32)$$

Table 1 expresses Dev_{dB} as a function of γ_{echo}. If presented graphically, as $\gamma_{echo} << 0$ dB, it becomes progressively more difficult to discern Dev_{dB} and therefore to extract the true character of the error, implying that fine detail may be lost and measurement data misrepresented. To improve the representation of small response deviations, an error function $E(n)$ is defined in terms of the measured transfer function $G(n)$ and the target function $T(n)$,

$$G(n) = T(n).*(1 + E(n)). \quad (33)$$

Fig. 14 shows a representation of Eq. (33), where the error function can be referred to either the input or the output of the system. In practice, to impart more performance information, both amplitude response and error function can be plotted on the same graph, where the distance between traces forms a measure of LDR. As an example, Fig. 15 shows a three-dimensional plot of the frequency response and the error response resulting from the single echo. It

can be seen that although frequency response deviations become harder to discern as the echo level is reduced, the error information is retained in the error function plot.

To use the LDR display when the target system is unknown requires frequency response estimation. It is assumed that the target response, although not necessarily flat, is characterized by a smooth curve allowing, for example, spline interpolation to form a smooth curve fit to a frequency subsampled (typically a factor in the range 32 to 256) version of the measured amplitude response. To circumvent phase problems, a magnitude-based error spectrum referred either to the input or to the output may be defined in terms of the actual measured spectrum and smoothed spectrum as follows:

$$\text{magnitude error spectrum}|_{\text{dB,input}} = 20 * \log_{10}(10^{-10} + \text{abs}(\text{abs}(G(n)) - \text{abs}(T(n)))) \quad (34)$$

$$\text{magnitude error spectrum}|_{\text{dB,output}} = 20 * \log_{10}(10^{-10} + \text{abs}(\text{abs}(G(n))./\ \text{abs}(T(n)) - 1)). \quad (35)$$

As a final evaluation example, a CD player with integral upsampling was measured [21] using the test procedures described in Section 1. Fig. 16 shows both the magnitude

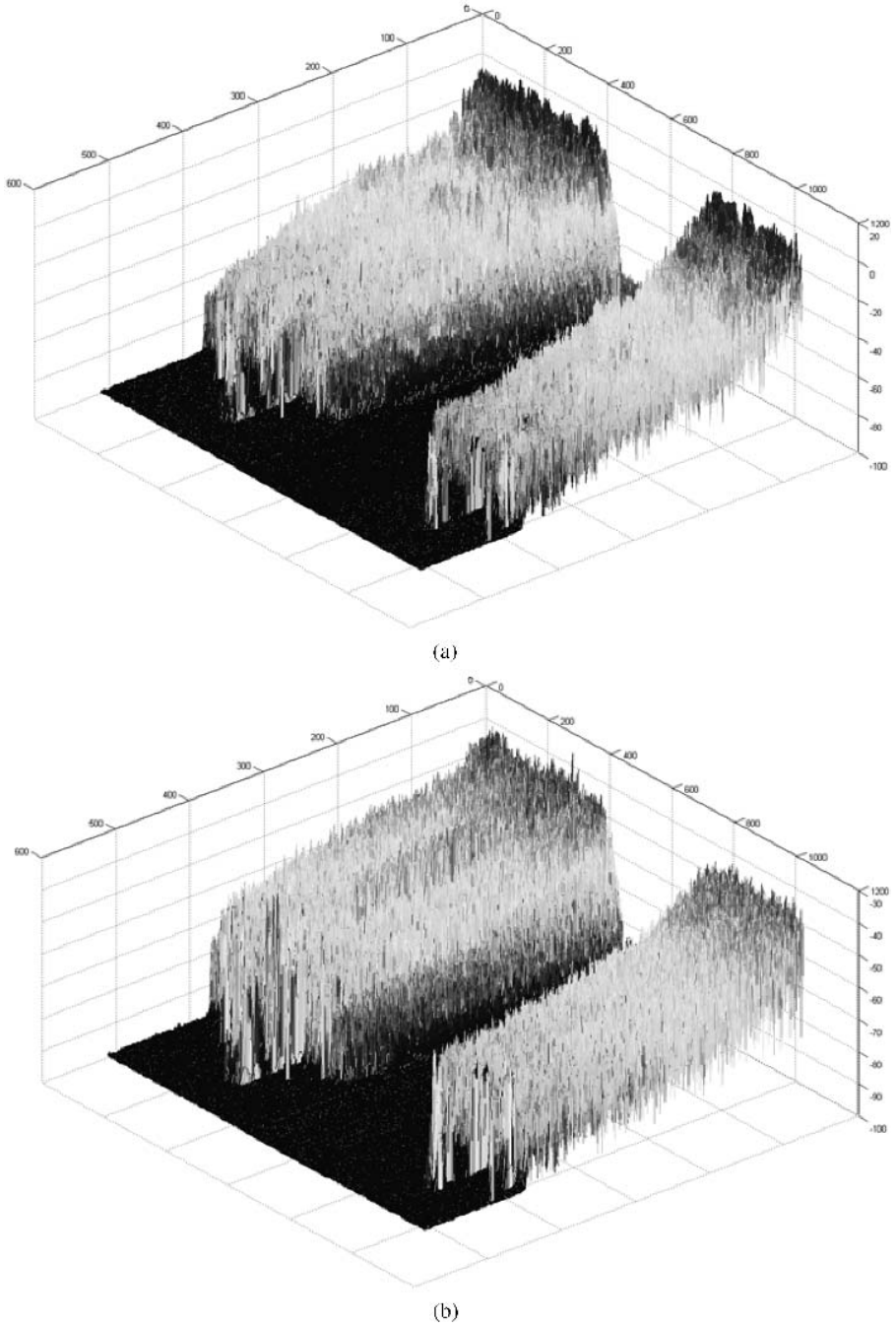


Fig. 11. MP3 codec performance at 192 kbit/s derived using interleave filter. (a) Codec output spectrum. (b) Codec distortion spectrum.

frequency response (top trace) and the error function (lower trace) presented on a common graph, where the target response was estimated using spline interpolation. The space between the two traces defines the LDR of the CD player. Response errors result from both quantization artifacts and frequency ripple within the interpolation filters used within the DAC. Fig. 17 shows the actual input minus output spectral error, confirming that ripple close to the noise floor is resolved and also demonstrating the accuracy achieved by the measurement system. (Note that some spectral lines below 2 kHz are believed to be low-level interference and not related to the system under test.)

6 CONCLUSIONS

A PC-based measurement system has been described that exploits either pseudorandom noise or music and where measurement accuracy is bounded mainly by external converter performance. In addition to transfer function and distortion measurements, the scheme included a simplified Volterra model as a method of nonlinear modeling where, although not universal, it forms a compromise between full system identification and measurement-specific assessment techniques. The procedures were examined using a number of linear and nonlinear examples, where

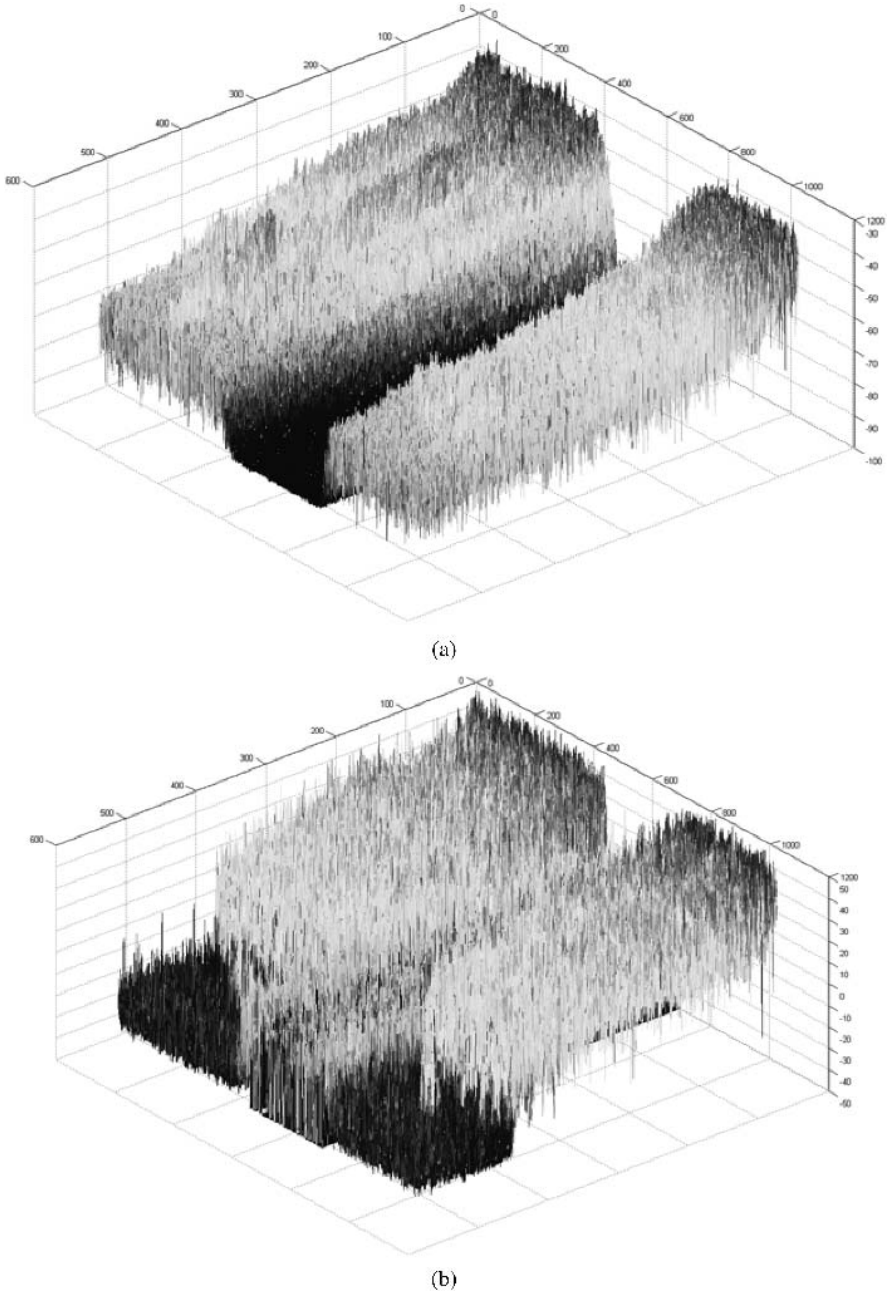


Fig. 12. MP3 codec performance at 192 kbit/s derived using true difference test. (a) Codec distortion spectrum. (b) Codec signal spectrum minus distortion spectrum.

resolution and accuracy were confirmed and insight was gained into the way the distortion was spectrally shaped according to excitation spectrum and nonlinearity.

The testing regime was extended to include comb filters in both signal generation and data analysis. This allows signal and distortion to be separated within the frequency domain, enabling estimates of frequency response and distortion to be made with a single-pass measurement. Also, by including complementary comb filters and excitation virtually full measurement resolution is achieved with minimal filtering artifacts present in the recovered output signal. The use of a single-pass test signal is important, not only to save time, but to minimize effects of gain drift that otherwise contribute to measurement error.

Methods were also reported using music sequences for system evaluation, and two example applications were presented. In particular, the opportunity to evaluate time-varying systems such as perceptual-based codecs was described, where standard block-based analysis was included to enable time-varying spectral distortion and signal-to-distortion information to be displayed.

In applying these techniques there are three principal caveats to be observed.

Caveat 1 As with MLS measurement systems, the analysis is based on circular transforms, and it is critical for the excitation sequences to have a duration that ex-

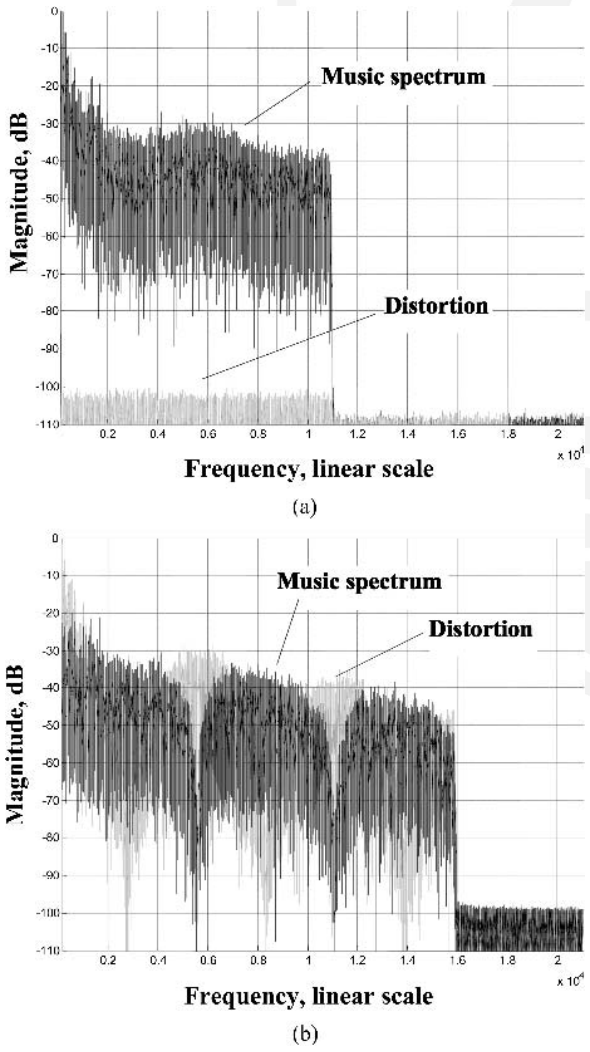


Fig. 13. Error spectra derived using music signal excitation and interleave processing for sampling-rate conversion. (a) Integer-ratio sampling-rate conversion. (b) Non-integer-ratio sampling-rate conversion.

Table 1. Peak-to-peak (dB) deviation as a function of echo amplitude.

Echo (Error) Level (dB)	Peak-to-peak Frequency Response Variation Dev_{dB}
−10	5.6884
−20	1.7430
−30	0.5495
−40	0.1737
−50	0.0549
−60	0.0174
−70	0.0055
−80	0.0017
−90	5.4934e-004
−100	1.7372e-004

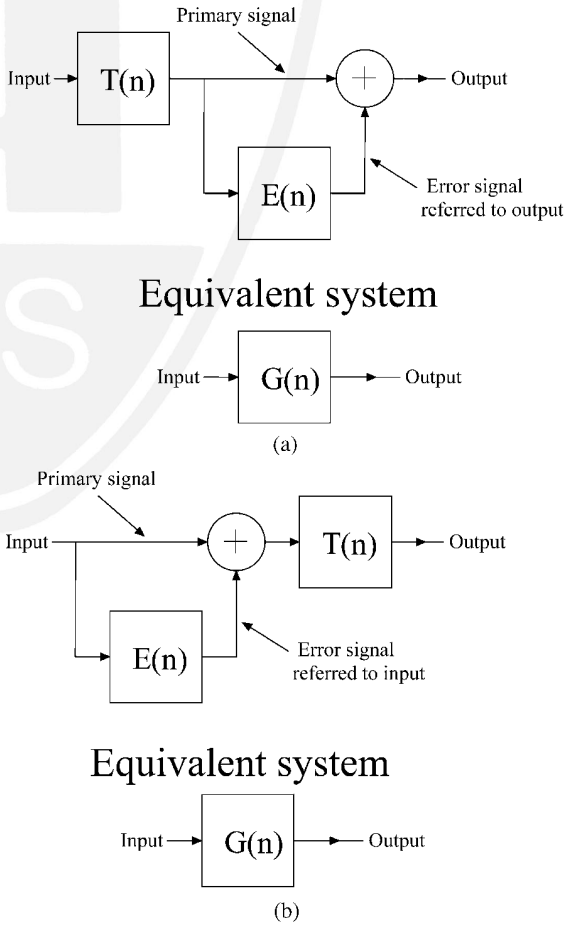


Fig. 14. Transfer function with error referred to both output and input.

ceeds the time over which the measured system's impulse response remains significant. Because of time-aliasing distortion it is not possible to reduce the excitation period and then apply window functions as the source is repetitive. However, the recorded measurements do capture the full impulse response and as such do not show windowing artifacts. Of course, if a loudspeaker system is measured, then the standard practice of windowing the derived total impulse response to eliminate reflections can be applied. However, the duration of the test sequence must exceed not only the loudspeaker impulse response but also the effect of room reflections and subsequent reverberation.

Caveat 2 A second factor is the requirement for exact sampling-rate synchronization of the test source and the ADC used to capture the measured response. This is especially critical when comb filters are incorporated as otherwise spectral spillage degrades the separation of distortion and transfer function data. All analyses and discussions presented have assumed sampling-rate synchronization. However, the need for precise frame alignment is less important as multiple sequences are output and transform circularity applies. Thus framing error just adds uncertain delay to measured impulse responses but has no effect on magnitude transfer functions.

Caveat 3 The process of comb filtering allows the separation of some of the intermodulation distortion. However, the filters viewed in the time domain introduce circular time dispersion where the excitation is effectively repeated twice and also overlaid with the existing sequence. This is not a problem when just making an input-

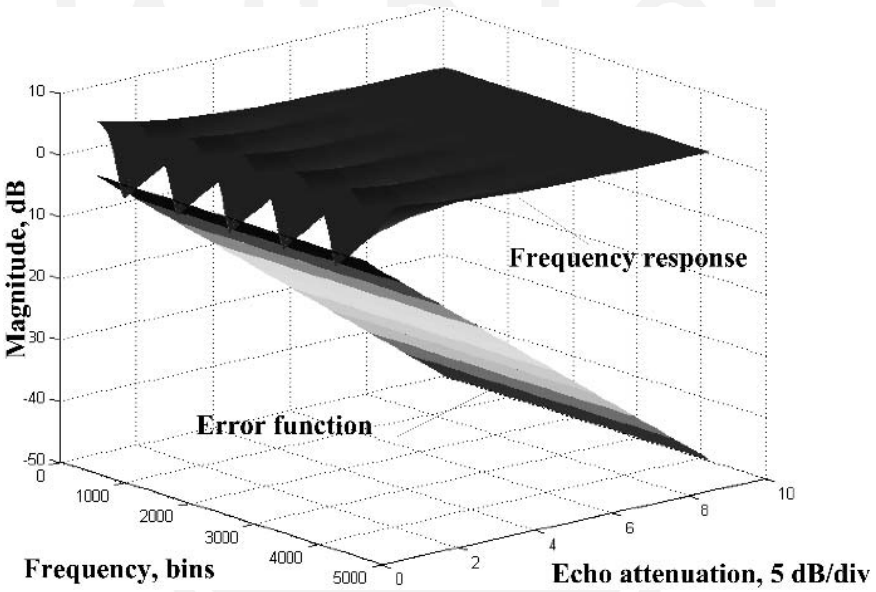


Fig. 15. LDR spectral plot for system with single echo path.

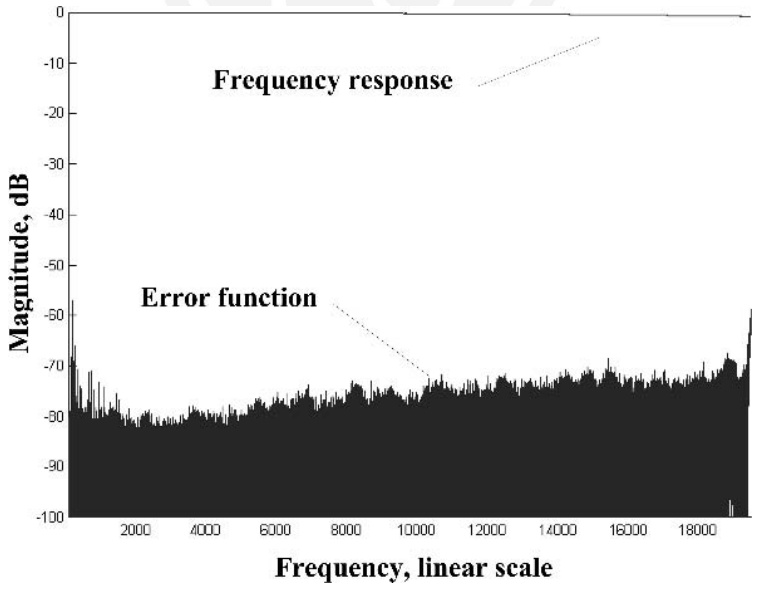


Fig. 16. Frequency response and magnitude error function for CD player.

specific distortion estimate. However, where time coherence is important, as with perceptual coding, this approach must be used with caution. This aspect was discussed in detail, and an absolute difference method was included for precise distortion analysis.

Finally, to complement the measurement schemes, two methods of enhanced data display were discussed. For linear distortion, a graph combining both frequency response and error function was proposed. In Section 5 this was shown especially suitable for cases where small response deviations occur, whereas for systems such as loudspeakers it is less suitable as the error function is relatively large. A classic three-dimensional spectral display was also included because of its relevance to perceptual codecs.

Using the system in a number of applications, the technique has proved to be an accurate and sensitive instrument to extract performance parameters. Also, it has enabled insight to be gained about the relationship between excitation and distortion spectra for a range of nonlinearities. For example, when a nonlinear system is tested using noise, then small frequency-response irregularities appear because the distortion is noiselike. When these small deviations are analyzed using the error function and LDR display described in Section 5, then the assessment of transfer function and distortion compares favorably with that derived using the comb-filter process in Section 3.

7 Acknowledgment

The author would like to offer his appreciation for the helpful feedback given by the reviewers, which has led to a number of significant enhancements of this paper.

8 References

[1] J. Borish and J. B. Angell, "An Efficient Algorithm for Measuring the Impulse Response Using Pseudorandom Noise," *J. Audio Eng. Soc.,* vol. 31, pp. 478–488 (1983 July/Aug.).

[2] D. D. Rife and J. Vanderkooy, "Transfer-Function Measurement with Maximum-Length Sequences," *J. Audio Eng. Soc.,* vol. 37, pp. 419–444 (1989 June).

[3] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," presented at the 108th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 48, p. 350 (2000 Apr.), preprint 5093.

[4] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems* (Wiley, New York, 1980).

[5] S. Boyd, Y. S. Tang, and L. O. Chua, "Measuring Volterra Kernels," *IEEE Trans. Circuits Sys.,* vol. CAS-30, pp. 571–577 (1983 Aug.).

[6] A. J. M. Kaizer, "The Modeling of the Nonlinear Response of an Electrodynamic Loudspeaker by a Volterra Series Expansion," presented at the 80th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 34, p. 388 (1986 May), preprint 2355.

[7] A. J. Berkhout, M. M. Boone, and C. Kesselman, "Acoustic Impulse Response Measurement: A New Technique," *J. Audio Eng. Soc.,* vol. 30, pp. 740–746 (1984 Oct.).

[8] R. C. Cabot, "Acoustic Applications of Cross Correlation," presented at the 64th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 27, p. 1020 (1979 Dec.), preprint 1544.

[9] J. Vanderkooy, "Aspects of MLS Measuring Systems," presented at the 93rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 40, p. 1048 (1992 Dec.), preprint 3398.

[10] E. Czerwinski, A. Voishvillo, S. Alexandrov, and A. Terekhov, "Multitone Testing of Sound System Components—Some Results and Conclusions, Part 1: History and Theory," *J. Audio Eng. Soc.,* vol. 49, pp. 1011–1048 (2001 Nov.).

[11] E. Czerwinski, A. Voishvillo, S. Alexandrov, and A. Terekhov, "Multitone Testing of Sound System Components—Some Results and Conclusions, Part 2: Model-

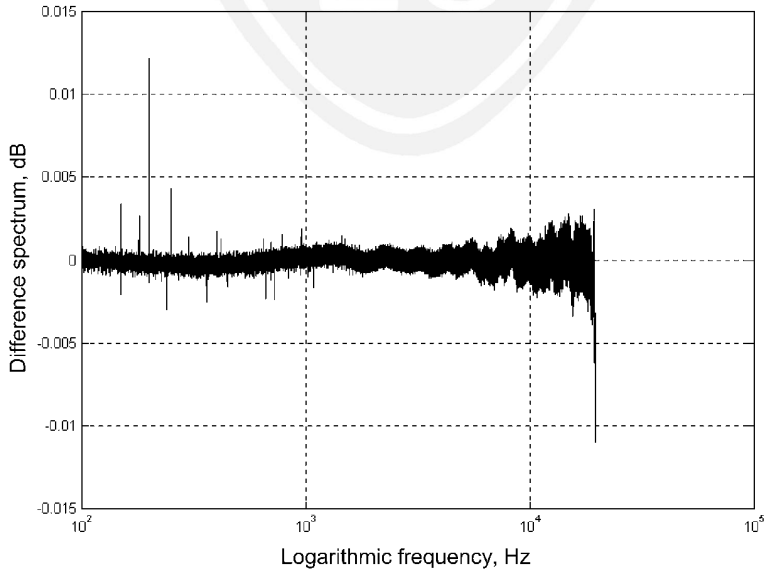


Fig. 17. Difference between estimated target response and actual response.

ing and Application," *J. Audio Eng. Soc.,* vol. 49, pp. 1181–1192 (2001 Dec.).

[12] A. I. Zayed, *Advances in Shannon's Sampling Theory* (CRC Press, Boca Raton, FL, 1993).

[13] M. O. J. Hawksford, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," *J. Audio Eng. Soc.,* vol. 45, pp. 37–62 (1997 Jan./Feb.).

[14] M. J. Reed and M. O. J. Hawksford, "Identification of Discrete Volterra Series Using Maximum Length Sequences," *IEE Proc. Circuits, Devices, Syst.,* vol. 143, pp. 241–248 (1996 Oct.).

[15] M. O. J. Reed and M. O. J. Hawksford, "Efficient Implementation of the Volterra Filter," *IEE Proc. Vis., Image, Signal Process.,* vol. 147, pp. 109–114 (2000 Apr.).

[16] M. J. Reed and M. O. J. Hawksford, "Comparison of Audio System Nonlinear Performance in Volterra Space," presented at the 103rd Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 45, p. 1026 (1997 Nov.), preprint 4606.

[17] R. A. Belcher, "Audio Nonlinearity: A Comb Filter Method for Measuring Distortion," Rep. 2, BBC Research Dept. (1976).

[18] J. R. Stuart, "Noise: Methods for Estimating Detectability and Threshold," *J. Audio Eng. Soc.,* vol. 42, pp. 124–140 (1994 Mar.).

[19] M. P. Hollier, M. O. J. Hawksford, and D. R. Guard, "Characterization of Communications Systems Using a Speechlike Test Stimulus," *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 41, pp. 1008–1021 (1993 Dec.).

[20] M. P. Hollier, M. O. J. Hawksford, and D. R. Guard, "Error Activity and Error Entropy as a Measure of Psychoacoustic Significance in the Perceptual Domain," *IEE Proc. Vis., Image, Signal Process.,* vol. 141, pp. 203–208 (1994 June).

[21] M. O. J. Hawksford, "Clockless and Bible Black," (Linn CD12 review), *Hi-Fi News* and *Record Review,* vol. 43, no. 2, pp. 66–71 (1999 Aug.).

APPENDIX A
COMMON MATLAB NOTATION AND OPERATORS

$a + \mathrm{i}b$, complex number, where $\mathrm{i} = \sqrt{-1}$

$x(n) = [x(1)\ x(2)\ \ldots\ x(r)\ \ldots\ x(N)]$, vector definition

$1{:}N = [1\ \ 2\ \ \ldots\ \ r\ \ \ldots\ \ N]$, vector with simple arithmetic progression

Element-by-Element Processing

The following three functions use the dot operator to describe an element-by-element process as distinct from conventional matrix operators. The definitions define the operations:

$x(n).*y(n) = [x(1)\,y(1)\ \ x(2)\,y(2)\ \ \ldots\ \ x(r)\,y(r)\ \ \ldots\ \ x(N)\,y(N)]$, element-by-element multiplication of vectors $x(n)$ and $y(n)$

$x(n)./y(n) = [x(1)/y(1)\ \ x(2)/y(2)\ \ \ldots\ \ x(r)/y(r)\ \ \ldots\ \ x(N)/y(N)]$, element-by-element division of vectors

$x(n).\hat{}M = [x(1)^M\ x(2)^M\ \ \ldots\ \ x(r)^M\ \ \ldots\ \ x(N)^M]$, each element in $x(n)$ is raised to the power of M

fft($x(n)$), fast Fourier transform of vector $x(n)$

ifft($x(n)$), inverse fast Fourier transform of vector $x(n)$

real($a + \mathrm{i}*b$) = a, real part of a complex number

abs($x(n)$) = $[|x(1)|\ |x(2)|\ \ \ldots\ \ |x(r)|\ \ \ldots\ \ x(N)|]$, magnitude value of each element

rand(1,N), vector, N random elements 0 to 1, rectangular probability distribution function

ones(1,N), vector of length N with unit elements

zeros(1,N), vector of length N with zero elements

APPENDIX B

B.1 Three-Dimensional Matrix [*X*] Inversion

Because [X] is a three-dimensional complex matrix (with, for example, $8*8*2^{15}$ coefficients), inversion is performed individually for each discrete frequency in the Fourier transforms. This sequential process is relatively time consuming. However, it is undertaken only once for a given set of M noise vectors. The inversion is required for decoding by Eq. (15). Recalling that the inverse $M \times M$ matrix $[X]^{-1}$ is [Z], then for frequency-domain bin x, [$Z(x)$] follows,

$$[Z(x)] = [X(x)]^{-1} = \mathrm{inv}\begin{bmatrix} X_{1,1}(x) & X_{1,2}(x) & \cdots & X_{1,s}(x) & \cdots & X_{1,M}(x) \\ X_{2,1}(x) & X_{2,2}(x) & \cdots & X_{2,s}(x) & \cdots & X_{2,M}(x) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ X_{r,1}(x) & X_{r,2}(x) & \cdots & X_{r,s}(x) & \cdots & X_{r,M}(x) \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ X_{M,1}(x) & X_{M,2}(x) & \cdots & X_{M,s}(x) & \cdots & X_{M,M}(x) \end{bmatrix}. \quad (36)$$

The inverse described in Eq. (36) is repeated for each frequency bin x over the range 1 to N (where N is the frequency vector length) and elements are concatenated to form vectors $Z_{r,s}$ that make up matrix [Z]. The input-specific inverted matrix [Z] is then available for decoding.

B.2 Volterra Test Sequence Formation

Initially the noise sequence generator program creates a preamble consisting of 2^{10} zero elements followed by a synchronization bit pattern that are used for frame locking following a measurement. They are defined by the vectors preamble_y and preamble_x, where

$$\text{preamble}_y = \text{zeros}(1,2\text{^}10) \tag{37}$$

$$\text{preamble}_x = [\text{preamble}_y\,[0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1 \\ -1\ -1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]]. \tag{38}$$

The M uncorrelated excitation noise vectors (each of length N) used for identification are calculated [see also Section 1, Eqs. (4) and (5)] using complex exponentials, each with individual random phase, to form a set of time-domain vectors with constant-magnitude frequency responses. The individual vectors $x_r(n)$ are derived as follows. Let

$$\text{NF}_r(n) = \exp(\text{i*angle}(\text{fft}(\text{rand}(1,N)))) \tag{39}$$

where transforming to the time domain,

$$x_r(n)|_{r=1}^{M} = \text{real}(\text{ifft}(\text{NF}_r(n))) \tag{40}$$

and forcing the mean of the vector to zero, gives the set of vectors as

$$x_r(n)|_{r=1}^{m} = x_r(n) - \text{mean}(x_r(n)). \tag{41}$$

The ensemble of M noise sequences is then amplitude normalized by the peak absolute value of the whole ensemble such that when the composite sequence is formed, signal excursion is bounded within the range −1 to 1. Also, normalization is combined with amplitude quantization and appropriate dither to set vector resolution (typically 24 bit). M subframes are then assembled using each noise vector repeated four times. Finally the test sequence gen is formed by concatenating the M subframes and the preamble,

$$\text{gen} = [\text{preamble}_x\,[x_1(n)\,x_1(n)\,x_1(n)\,x_1(n)][x_2(n)\,x_2(n)\,x_2(n)\,x_2(n)] \\ \cdots [x_r(n)\,x_r(n)\,x_r(n)\,x_r(n)] \\ \cdots [x_M(n)\,x_M(n)\,x_M(n)\,x_M(n)]\,\text{preamble}_y]'. \tag{42}$$

The stereo wavfile of gen with the sampling rate f_{sam} and resolution bit is realized in Matlab as

wavwrite([gen, gen], fsam,bit,′file name′).

B.3 Decoding and Analysis of Measurement Data

Following data acquisition using a sample-rate synchronized ADC, M measured data sequences designated $y_1(n)$, $y_2(n)$, . . . , $y_M(n)$ are extracted by sample counting and transformed to the frequency domain. The frequency-domain vector Y_r corresponding to excitation r is

$$Y_r|_{r=1}^{M} = \text{fft}(y_r(n))/N. \tag{43}$$

The Volterra frequency-domain responses $[H]$ are determined by applying Eq. (15). Noting that $Z_{r,s}$ is the vector (r, s) of matrix $[Z]$, then the rth row of $[H]$ is calculated as

$$H_r|_{r=1}^{M} = Z_{r,1}.*Y_1 + Z_{r,2}.*Y_2 + \quad \cdots \quad + Z_{r,3}.*Y_r + \quad \cdots \\ + Z_{r,4}.*Y_M \tag{44}$$

where both Y_r and $Z_{r,s}$ are frequency-domain vectors of length N. The set of M vectors forming H are subsequently transformed using Eq. (16) to determine the Volterra time-domain impulse-response matrix $[h]$. To correct for gain error between test sequence and redigitized measured data, $[h]$ is normalized to set the peak absolute value of the linear time-domain impulse response to unity.

B.4 *z*-Domain Description of Interleave Filtering

In this appendix time dispersion resulting from comb filtering is analyzed to give additional insight into the measurement procedure and also to inform how the system can be modified to cope with measurements of systems such as perceptual coders. The use of frequency interleaving as described in Section 3 offers a method to segregate signal and distortion and where by using two test segments with complementary interleave functions, full frequency resolution is obtainable. However, the employment of interleave filtering has consequences in the time domain in terms of both signal excitation and subsequent processing used to extract signal and distortion. In practice comb filtering introduces time dispersion, which modifies both excitation and recovery of distortion such that they are no longer time coherent, which can invalidate the results for nonstationary systems. For example, if a full-frame music signal is used, then a time-delayed version is overlaid, thus corrupting the excitation. Postprocess filtering also smears the resulting distortion such that signal and distortion are no longer linked correctly in time. For a perceptual codec evaluation this linkage is critical.

The even and odd frequency raised-cosine comb filters can be defined in the z-domain by finite impulse response filters (FIRs), where the respective filter functions $C_{\text{even}}(z)$ and $C_{\text{odd}}(z)$ are

$$C_{\text{even}}(z) = 0.25_Z{}^{0.5N} + 0.5 + 0.25z^{-0.5N} \rightarrow 0.5(1 + \cos(\pi fT))$$

$$C_{\text{odd}}(z) = 0.25_Z{}^{0.5N} + 0.5 - 0.25z^{-0.5N} \rightarrow 0.5(1 - \cos(\pi fT)).$$

However, circularity implies over period N that $z^{0.5N} \equiv z^{-0.5N}$, whereby

$$C_{\text{even}}(z) = 0.5(1 + z^{-0.5N}) \tag{45}$$

$$C_{\text{odd}}(z) = 0.5(1 - z^{-0.5N}). \tag{46}$$

The measurement process uses up to two sequential data sets (that is, $h = 1$, $h = 2$) using complementary comb filters, where $\text{source}(z)$, $\text{test}_h(z)$, $\text{dist}_h(z)$, and $\text{mdata}_h(z)$ are respective samples of source sequence (noise or music), test sequence, distortion, and captured data, and $\text{decode}_{h1}(z)$, $\text{decode}_{h2}(z)$ are data decoded by the respective comb filters. All samples form elements within frames of length N.

The following z-domain system functions describe the measurement process with frame sequential, complementary comb filtering, $h = 1, 2$,

- Test sequence:

$$\text{test}_1(z) = C_{\text{even}}(z)\text{source}(z)$$

$$\text{test}_2(z) = C_{\text{odd}}(z)\text{source}(z)$$

- Measured response:

$$\text{mdata}_1(z) = \text{test}_1(z) + \text{dist}_1(z)$$

$$\text{mdata}_2(z) = \text{test}_2(z) + \text{dist}_2(z)$$

- Decoded response 1:

$$\text{decode}_{11}(z) = \text{mdata}_1(z)C_{\text{even}}(z)$$

$$\text{decode}_{21}(z) = \text{mdata}_2(z)C_{\text{odd}}(z)$$

- Decoded response 2:

$$\text{decode}_{12}(z) = \text{mdata}_1(z)C_{\text{odd}}(z)$$

$$\text{decode}_{22}(z) = \text{mdata}_2(z)C_{\text{even}}(z).$$

Rearranging and substituting $C_{\text{even}}(z)$ and $C_{\text{odd}}(z)$ from Eqs. (45) and (46),

$$\begin{aligned}\text{decode}_{11}(z) &= 0.25(1 + z^{-0.5N})^2\text{source}(z) \\ &\quad + 0.5(1 + z^{-0.5N})\text{dist}_1(z) \\ \text{decode}_{12}(z) &= 0.25(1 + z^{-0.5N})(1 - z^{-0.5N})\text{source}(z) \\ &\quad + 0.5(1 - z^{-0.5N})\text{dist}_1(z) \\ \text{decode}_{21}(z) &= 0.25(1 - z^{-0.5N})^2\text{source}(z) \\ &\quad + 0.5(1 - z^{-0.5N})\text{dist}_2(z) \\ \text{decode}_{22}(z) &= 0.25(1 - z^{-0.5N})(1 + z^{-0.5N})\text{source}(z) \\ &\quad + 0.5(1 + z^{-0.5N})\text{dist}_2(z).\end{aligned}$$

Observing circularity over N, where $z^{-N} \equiv 1$,

$$(1 + z^{-0.5N})^2 = 1 + 2z^{-0.5N} + z^{-N} \equiv 2(1 + z^{-0.5N})$$

$$(1 - z^{-0.5N})^2 = (1 - 2z^{-0.5N} + z^{-N}) \equiv 2(1 - z^{-0.5N})$$

$$(1 + z^{-0.5N})\,(1 - z^{-0.5N}) = (1 - z^{-N})^2 \equiv 0$$

and the equations simplify to

$$\text{test}_1(z) = 0.5(1 + z^{-0.5N})\text{source}(z) \tag{47}$$

$$\text{test}_2(z) = 0.5(1 - z^{-0.5N})\text{source}(z) \tag{48}$$

$$\begin{aligned}\text{decode}_{11}(z) &= 0.5(1 + z^{-0.5N})\text{source}(z) \\ &\quad + 0.5(1 + z^{-0.5N})\text{dist}_1(z)\end{aligned} \tag{49}$$

$$\text{decode}_{12}(z) = 0.5(1 - z^{-0.5N})\text{dist}_1(z) \tag{50}$$

$$\begin{aligned}\text{decode}_{21}(z) &= 0.5(1 - z^{-0.5N})\text{source}(z) \\ &\quad + 0.5(1 - z^{-0.5N})\text{dist}_2(z)\end{aligned} \tag{51}$$

$$\text{decode}_{22}(z) = 0.5(1 + z^{-0.5N})\text{dist}_2(z). \tag{52}$$

Combining the two complementary filtered sets of measurement, we make the following observations.

$$T(z) = \sum_{h=1}^{2} \text{test}_h(z).$$

From Eqs. (47) and (48),

$$T(z) = \text{test}_1(z) + \text{test}_2(z) = \text{source}(z). \tag{53}$$

$$\text{DEC}_1(z) = \sum_{h=1}^{2} \text{decode}_{h1}(z).$$

From Eqs. (49) and (51),

$$\begin{aligned}\text{DEC}_1(z) &= \text{decode}_{11}(z) + \text{decode}_{21}(z) \\ &= \text{source}(z) + 0.5(1 + z^{-0.5N})\text{dist}_1(z) \\ &\quad + 0.5(1 - z^{-0.5N})\text{dist}_2(z).\end{aligned} \tag{54}$$

$$\text{DEC}_2(z) = \sum_{h=1}^{2} \text{decode}_{h2}(z).$$

From Eqs. (50) and (52),

$$\begin{aligned}\text{DEC}_2(z) &= \text{decode}_{12}(z) + \text{decode}_{22}(z) \\ &= 0.5(1 - z^{-0.5N})\text{dist}_1(z) \; + 0.5(1 + z^{-0.5N})\text{dist}_2(z).\end{aligned} \tag{55}$$

Eqs. (47)–(52) show how time dispersion affects both the test signal and the extracted distortion, where in all cases filtering superimposes a copy of the specific sequence, but with a half-frame circular delay. These results show that both the excitation and the recovered distortion exhibit time dispersion based on half-frame circular repetition. Finally, Eq. (53)–(55) describe the z-domain process applied over two sets of frames with complementary comb filtering. It is shown in Section 4 that by knowing the time-domain structure, the test sequences can be modified so that, for example, music signals remain intact (though of shorter duration), thus allowing proper analysis of perceptual codecs. This is verified by introducing a short gap in the music so that a corresponding gap in the distortion can he confirmed.

B.5 Block Analysis to Derive Spectral Envelope as a Function of Time

To display spectral distortion as a function of time, Fourier analysis is applied to a windowed segment of measured data. A raised-cosine window is used with 50% block overlap and a block length of approximately 25 ms to match the data structure common to perceptual codecs. Data are then displayed on a three-dimensional graph with respective axes of spectral amplitude, frequency, and block number (corresponding to half-block time increments of 12.5 ms). The postprocessing analysis is described for vector length N as follows.

For compatibility with the FFT, the block length is 2^W, where W is a positive integer. Thus if the measurement sampling rate is f_s Hz, we have $2^W/f_s \approx 25$ ms, where w is calculated as

$$w = 2\text{^round}(\log_2(0.025\, f_s)). \quad (56)$$

The raised-cosine window function winc for samples 1:w is defined as

$$\text{winc}(1{:}w) = 0.5(1 - \cos(2\pi((1{:}w) - 0.5)/w)). \quad (57)$$

Data blocks of length w and spaced at $w/2$ sample increments are extracted sequentially from both $\text{out}_f(n)$ and $\text{dist}_f(n)$ then weighted by winc(1:w). For each weighted block, Fourier transformation is performed and a matrix compiled to represent the spectral surface as a function of discrete frequency and block number. Hence for analysis block b, the respective transforms $\text{OUT}_{bk}(b,1{:}w)$ and $\text{DIST}_{bk}(b,1{:}w)$ for signal and distortion are

$$\text{OUT}_{bk}(b,1{:}w) = \text{fft}(\text{out}_f(b{*}w/2{:}b{*}w/2 + w - 1).{*}\text{winc}(1{:}w)) \quad (58)$$

$$\text{DIST}_{bk}(b,1{:}w) = \text{fft}(\text{dist}_f(b{*}w/2{:}b{*}w/2 + w - 1).{*}\text{winc}(1{:}w)). \quad (59)$$

The difference spectrum $\text{DIFF}_{bk}(b,1{:}w)$ can be calculated on a logarithmic basis between signal and distortion spectral surfaces, and forms a measure of the dynamic signal-to-distortion ratio,

$$\text{DIFF}_{bk}(b,1{:}w) = 20\log_{10}\left(\frac{\text{abs}(\text{OUT}_{bk}(b,1{:}w)) + \alpha}{\text{abs}(\text{DIST}_{bk}(b,1{:}w)) + \alpha}\right) \quad (60)$$

where α bounds the display range. Hence surfaces can be plotted that represent the dynamic spectral behavior as a function of time for the codec output signal, distortion, and the difference between signal and distortion.

THE AUTHOR

Malcolm Hawksford received a B.Sc. degree with First Class Honors in 1968 and a Ph.D. degree in 1972, both from the University of Aston in Birmingham, UK. His Ph.D. research program was sponsored by a BBC Research Scholarship and he studied delta modulation and sigma-delta modulation (SDM) for color television applications. During this period he also invented a digital time-compression/time-multiplex technique for combining luminance and chrominance signals, a forerunner of the MAC/DMAC video system.

Dr. Hawksford is director of the Centre for Audio Research and Engineering and a professor in the Department of Electronic Systems Engineering at Essex University, Colchester, UK, where his research and teaching interests include audio engineering, electronic circuit design, and signal processing. His research encompasses both analog and digital systems, with a strong emphasis on audio systems including signal processing and loudspeaker technology. Since 1982 his research into digital crossover networks and equalization for loudspeakers has resulted in an advanced digital and active loudspeaker system being designed at Essex University. The first one was developed in 1986 for a prototype system to be demonstrated at the Canon Research Centre and was sponsored by a research contract from Canon. Much of this work has appeared in *JAES,* together with a substantial number of contributions at AES conventions. He is a recipient of the AES Publications Award for his paper, "Digital Signal Processing Tools for Loudspeaker Evaluation and Discrete-Time Crossover Design," for the best contribution by an author of any age for *JAES,* volumes 45 and 46.

Dr. Hawksford's research has encompassed oversampling and noise-shaping techniques applied to analog-to-digital and digital-to-analog conversion with special emphasis on SDM and its application to SACD technology. In addition, his research has included the linearization of PWM encoders, diffuse loudspeaker technology, array loudspeaker systems, and three-dimensional spatial audio and telepresence including scalable multichannel sound reproduction.

Dr. Hawsford is a chartered engineer and a fellow of the AES, IEE, and IOA. He is currently chair of the AES Technical Committee on High-Resolution Audio and is a founder member of the Acoustic Renaissance for Audio (ARA). He is also a technical consultant for NXT, UK and LFD Audio UK and a technical adviser for *Hi-Fi News* and *Record Review.*

Section 7: Appendix 1: Conference paper listing

Appendix 1: Conference paper listing

(*excluding papers subsequently published in a journal*)

C2 A MULTIPLEX STEREO DECODER WITH AUTOMATIC PHASE-ERROR CORRECTION, Hawksford, M.J., 50th Convention of the Audio Engineering Society, March 1975

C10 POWER AMPLIFIER OUTPUT STAGE DESIGN INCORPORATING ERROR FEEDBACK CORRECTION WITH CURRENT DUMPING ENHANCEMENT, Hawksford, M.J., AES 74th Convention, New York, *preprint 1993 (B-4),* October 1983

C11 PONTOON AMPLIFIER CONSTRUCTIONS INCORPORATING ERROR-FEEDBACK LOCATION OF FLOATING POWER SUPPLIES, Hawksford, M.J., 78th Convention of the AES, *preprint No.2247 (A-14*), May 3-6, 1985

C12 N^{th}-ORDER RECURSIVE SIGMA-ADC MACHINERY AT THE ANALOGUE-DIGITAL GATEWAY, Hawksford, M.J., 78th Convention of the AES, *preprint No.2248 (A-15),* May 3-6, 1985

C14 A FAMILY OF CIRCUIT TOPOLOGIES FOR THE LINKWITZ-RILEY (LR-4) CROSSOVER ALIGNMENT, Hawksford, M.J., 82nd Convention of the AES, London, *preprint 246* (*J-2*), 1987

C16 OVERSAMPLING AND NOISE SHAPING FOR DIGITAL TO ANALOGUE CONVERSION, Hawksford, M.J., *Reproduced Sound* 3, Institute of Acoustics, Windermere, pp.151-163, November 1987

C17 Patent application: LOUDSPEAKER SYSTEMS, M.O.J. Hawksford, P.G.L. Mills, British Patent Application No. 86-05400

C19 MULTI-LEVEL TO 1 BIT TRANSFORMATIONS FOR APPLICATIONS IN DIGITAL-TO-ANALOGUE CONVERTERS USING OVERSAMPLING AND NOISE SHAPING, Hawksford, M.O.J., *Proc. Institute of Acoustics, Reproduced Sound 4*, vol.10, part 7, pp. 129-143, 1988

C21 A TUTORIAL GUIDE TO NOISE SHAPING AND OVERSAMPLING IN ADC AND DAC SYSTEMS, Hawksford M.O.J., *Proc. Institute of Acoustics, Reproduced Sound 5*, vol. 11, part 7, pp. 289-302, 1989

C27 INTRODUCTION TO DIGITAL AUDIO, (tutorial paper), Hawksford, M.O.J., *Images of Audio, Proceedings of the 10th International AES Conference*, London September, 1991

C28 A COMPARISON OF TWO-STAGE 4TH-ORDER AND SINGLE-STAGE 2ND-ORDER DELTA-SIGMA MODULATION IN DIGITAL-TO-ANALOGUE CONVERSION, Hawksford, M.O.J., *IEE Conference on Analogue to Digital and Digital to Analogue Conversion*, Conference publication 343, pp 148-152, Swansea, September 1991

C29 BANDPASS IMPLEMENTATION OF THE SIGMA-DELTA A-D CONVERSION TECHNIQUE, Thurston, A.M., Pearce, T.H. and Hawksford, M.O.J., *IEE Conference on Analogue to Digital and Digital to Analogue Conversion*, Conference publication 343, pp 81-86, Swansea, September 1991

C30 TOWARDS A DEFINITIVE ANALYSIS OF AUDIO SYSTEM ERRORS, Dunn, C., and Hawksford, M.O.J., *91st Convention of the Audio Engineering Society*, New York, October 1991, preprint 3137 (P-1)

C34 TANDEM QUADRUPLET VCA TOPOLOGY, Hawksford, M.O.J., *Proc. Institute of Acoustics, Reproduced Sound 7,* vol. 13, pt 7, pp 227-237, Nov 1991

C35 TOWARDS A GENERALISATION OF ERROR CORRECTION AMPLIFIERS, Hawksford, M.O.J., *Proc. Institute of Acoustics, Reproduced Sound 7*, vol. 13, pt 7, pp 167-190, Nov 1991

C37 BANDPASS SIGMA DELTA A-D CONVERSION, Thurston, A.M., Pearce, T.H., Higman, M. and Hawksford, M.O.J., *Proc. Workshop Advances in Analog Circuit Design*, Delft University of Technology, pp 266-297, April 1992

C38 BANDPASS DELTA-SIGMA CONVERSION, Thurston, A. and Hawksford, M.O.J., *IEE Colloquium on Implementation of Novel Hardware for Radio Systems*, Savoy Place, London, 27th May 1992

C41 IS THE AES/EBU/SPDIF DIGITAL AUDIO INTERFACE FLAWED?, Dunn, C. and Hawksford, M.O.J., *93rd AES Convention*, San Francisco, preprint 3360, October 1992

C43 CHARACTERIZATION OF COMMUNICATIONS SYSTEMS USING A SPEECH-LIKE TEST STIMULUS, Hollier, M.P., Guard, D.R. and Hawksford M.O.J., *93rd AES Convention*, San Francisco, preprint 3395, October 1992

C48 INTEGRATING FINITE IMPULSE RESPONSE FILTER, Heylen, R.L.M., and Hawksford, M.O.J., *94th AES Convention*, Berlin, preprint 3587, February 1993

C49 BANDPASS SIGMA-DELTA A-D CONVERSION FOR RADIO SYSTEM APPLICATIONS, Thurston A. and Hawksford M.O.J., International Defense and Technology, no 12, ISSN 1155-3480, March 1993, pp74-80

C52 DYNAMIC JITTER FILTERING IN HIGH-RESOLUTION DSM AND PWM DIGITAL-TO-ANALOGUE CONVERSION, Hawksford M.O.J., 96th Convention of the Audio Engineering Society, Amsterdam, preprint 3811, February 1994

C53 INTEGRATING FILTERS FOR PROFESSIONAL AND CONSUMER APPLICATIONS, Heylen R. and Hawksford M.O.J., 96th Convention of the Audio Engineering Society, Amsterdam, preprint 3834, February 1994

C55 ADVANCES IN COMPUTER MODELLING OF RIBBON LOUDSPEAKERS, Bank G. and Hawksford M.O.J., 96th Convention of the Audio Engineering Society, Amsterdam, preprint 3837, February 1994

C60 LINEARIZATION OF CLASS D OUTPUT STAGES FOR HIGH PERFORMANCE AUDIO POWER AMPLIFIERS, Hawksford M.O.J. and Logan S., 2nd International IEE Conference on Analogue-to-digital and digital-to-analogue conversion, Cambridge University, July 1994

C61 DYNAMIC OVERLOAD RECOVERY MECHANISM FOR SIGMA-DELTA MODULATION, Thurston A. and Hawksford M.O.J., 2nd International IEE Conference on Analogue-to-digital and digital-to-analogue conversion, Cambridge University, July 1994

C72 DIGITAL AND ACTIVE LOUDSPEAKER SYSTEMS FOR HIGH-QUALITY MONITORING, Hawksford M.O.J., Proceedings of Active 95, The 1995 International Symposium on Active Control of Sound and Vibration, Newport Beach, CA, USA, pp 1247-1258, July 6th - 8th, 1995

C80 ***INVITED PAPER:*** MULTI-CHANNEL HIGH-DEFINITION DIGITAL AUDIO SYSTEMS FOR HIGH-DENSITY COMPACT DISK, Hawksford, M.O.J., 101st AES Convention, Los Angeles, November 1996, preprint 4362 (J-2)

C82 ERROR CORRECTION AND NON-SWITCHING POWER AMPLIFIER OUTPUT STAGES, Hawksford, M.O.J., 102nd AES Convention, Munich, March 1997, preprint 4492 (M5)

C84 OBJECTIVE ASSESSMENT OF PHANTOM IMAGES IN A 3-DIMENSIONAL SOUND FIELD USING A VIRTUAL LISTENER, Theiβ, B. and Hawksford, M.O.J., 102nd AES Convention, Munich, March 1997, preprint 4462

C87 ***INVITED PAPER:*** HIGH-DEFINITION DIGITAL AUDIO IN 3-DIMENSIONAL SOUND REPRODUCTION, Hawksford, M.O.J., 103rd AES Convention, New York, September 1997, preprint 4560

C94 ***INVITED PAPER:*** DVD-AUDIO: HIGH-RESOLUTION, MULTI-CHANNEL DIGITAL FORMATS, Hawksford, M.O.J., PALA'98 - AES Singapore Section Seminar, World Trade Centre, Singapore, 17th July 1998

C95 ERROR AND CONVERGENCE PROPERTIES OF THE FAST AFFINE PROJECTION AND LEAST MEAN SQUARE ECHO CANCELLATION ALGORITHMS, Reed, M.J., Hawksford, M.O.J.and Hollier, M.P., 105th AES Convention, San Francisco, September 1998, preprint 4765 (B-2)

C101 ***INVITED PAPER**:* CROSSOVER ALIGNMENTS FOR ANALOGUE AND DIGITAL ACTIVE LOUDSPEAKERS, Hawksford, M.O.J., PALA'99 - AES Singapore Section Seminar, Singapore, 8th July 1999

C103 TIME DOMAIN AUDITORY MODEL FOR THE ASSESSMENT OF HIGH QUALITY CODED AUDIO, Robinson, D.J.M. and Hawksford, M.J., 107th AES Convention, New York, September 1999, preprint 5017 (G-2)

C106 AUDITORY MODEL FOR LATERALIZATION AND THE PREDICTIVE EFFECT, Theiβ, B. and Hawksford, M.J., 107th AES Convention, New York, September 1999, preprint 5048 (N-5)

C108 TIME-DOMAIN AUDITORY MODEL FOR THE ASSESSMENT OF HIGH-QUALITY CODED AUDIO, Robinson, D.J.M. and Hawksford, M.J., EPSRC Conference Prep 2000, 2000, [Awarded best conference paper award]

C111 CURRENT-STEERING TRANSIMPEDANCE AMPLIFIERS FOR HIGH-RESOLUTION DIGITAL-TO-ANALOGUE CONVERTERS, Hawksford, M.O.J., 109th AES Convention, Los Angeles, September 2000, preprint 5192

C113 PSYCHOACOUSTIC MODELS AND NON-LINEAR HUMAN HEARING, Robinson, D. and Hawksford, M.O.J., 109th AES Convention, Los Angeles, September 2000, preprint 5228

C116 ***KEYNOTE ADDRESS:*** ULTRA HIGH RESOLUTION SPATIAL AUDIO TECHNOLOGY FOR HDTV ON DVD, Hawksford, M.O.J., 10th AES Japanese Regional Convention, Tokyo, June 2001, Convention proceedings pp 1-16

C120 DIFFUSE SIGNAL PROCESSING AND ACOUSTIC SOURCE CHARACTERIZATION FOR APPLICATIONS IN SYNTHETIC LOUDSPEAKER ARRAYS, Hawksford, M.O.J., 112th AES Convention, Munich, May 2002, paper 5612

C126 PERCEPTUALLY MOTIVATED PROCESSING FOR SPATIAL AUDIO MICROPHONE ARRAYS, Reller, C.P.A. and Hawksford, M.O.J., 115th AES Convention, New York, October 2003, paper 5933

C128 PERFORMANCE PREDICTION OF DAB MODULATION AND TRANSMISSION USING MATLAB MODELLING, Gaetzi, L.M. and Hawksford, M.O.J., IEEE International Symposium on Consumer Electronics, Reading, UK, Hard-copy ISBN: 0-7803-8526-8, CD-ROM ISBN: 0-7803-8527-6, pp. 272-277, 2004

C129 SCALEABLE MULTICHANNEL DSD CODING, Hawksford, M.O.J., 117th AES Convention, San Francisco, October 2004, paper 6297

C132 NOISE SHAPING IN TIME-DOMAIN QUANTIZED LFM, Hawksford, M.O.J., 119th AES Convention, New York, October 2005, paper 6617

C134 JITTER SIMULATION IN HIGH RESOLUTION DIGITAL AUDIO, Hawksford, M.O.J., 121st AES Convention, San Francisco, October 2006, paper 6864