Audio Engineering Society

Convention Paper

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

MPEG Unified Speech and Audio Coding – The ISO/MPEG Standard for High-Efficiency Audio Coding of all Content Types

Max Neuendorf[1], Markus Multrus[1], Nikolaus Rettelbach[1], Guillaume Fuchs[1], Julien Robilliard[1], Jérémie Lecomte[1], Stephan Wilde[1], Stefan Bayer[1], Sascha Disch[1], Christian Helmrich[1], Roch Lefebvre[2], Philippe Gournay[2], Bruno Bessette[2], Jimmy Lapierre[2], Kristofer Kjörling[3], Heiko Purnhagen[3], Lars Villemoes[3], Werner Oomen[4], Erik Schuijers[4], Kei Kikuiri[5], Toru Chinen[6], Takeshi Norimatsu[7], Chong Kok Seng[7], Eunmi Oh[8], Miyoung Kim[8], Schuyler Quackenbush[9], and Bernhard Grill[1]

[1]*Fraunhofer Institute for Integrated Circuits IIS, Erlangen, 91058, Germany* [2]*Université de Sherbrooke / VoiceAge Corp., Sherbrooke, Qc, J1K 2R1, Canada* [3]*Dolby Sweden, 113 30, Stockholm, Sweden* [4]*Philips Research Laboratories, Eindhoven, 5656AE, The Netherlands* [5]*NTT DOCOMO, INC., Yokosuka, Kanagawa, 239-8536, Japan* [6]*Sony Corporation, Shinagawa, Tokyo, 141-8610, Japan* [7]*Panasonic Corporation* [8]*Samsung Electronics, Suwon, Korea* [9]*Audio Research Labs, Scotch Plains, NJ, 07076, USA*

Correspondence should be addressed to Max Neuendorf (`max.neuendorf@iis.fraunhofer.de`)

ABSTRACT

In early 2012 the ISO/IEC JTC1/SC29/WG11 (MPEG) finalized the new MPEG-D Unified Speech and Audio Coding standard. The new codec brings together the previously separated worlds of general audio coding and speech coding. It does so by integrating elements from audio coding and speech coding into a unified system. The present publication outlines all aspects of this standardization effort, starting with the history and motivation of the MPEG work item, describing all technical features of the final system, and further discussing listening test results and performance numbers which show the advantages of the new system over current state-of-the-art codecs.

1. INTRODUCTION

With the advent of devices which unite a multitude of functionalities, the industry has an increased demand for an audio codec which can deal equally well with all types of audio content including both speech and music at low bitrates. In many use cases, e.g. broadcasting, movies, or audio books, the audio content is not limited to only speech or only music. Instead, a wide variety of content must be processed including mixtures of speech and music. Hence, a unified audio codec that performs equally well on all types of audio content is highly desired. Even though the largest potential for improvements is expected at the lower end of the bitrate scale, a unified codec requires, of course, to retain or even exceed the quality of presently available codecs at higher bitrates.

Audio coding schemes, such as MPEG-4 High Efficiency Advanced Audio Coding (HE-AAC) [1, 2], are advantageous in that they show a high subjective quality at low bitrates for music signals. However, the spectral domain models used in such audio coding schemes do not perform equally well on speech signals at low bitrates.

Speech coding schemes, such as Algebraic Code Excited Linear Prediction (ACELP) [3], are well suited for representing speech at low bitrates. The time domain source-filter model of these coders closely follows the human speech production process. State-of-the-art speech coders, such as the 3GPP Adaptive Multi-Rate Wideband (AMR-WB) [4, 5], perform very well for speech even at low bitrates, but show a poor quality for music. Therefore, the source-filter model of AMR-WB was extended by transform coding elements in the 3GPP AMR-WB+ [6, 7]. Still, for music signals AMR-WB+ is not able to provide an audio quality similar to that of HE-AAC(v2).

The following sections will first introduce the reader to the above-mentioned state of the art representatives of modern audio and speech coding, HE-AACv2 and AMR-WB+. Further, the ISO/IEC MPEG work item is described, followed by a technical description of the standardized coding system at a much higher level of detail than in previous publications [8]. Concluding, performance figures from the MPEG Verification Tests are presented and potential applications of the new technology are discussed.

2. STATE OF THE ART

2.1. General Codec Structure

Modern audio codecs and speech codecs typically exhibit a structure as shown in Figure 1. This scheme consists of three main components: (1) a core-coder (i.e. transform or speech coder) which provides a high quality and largely wave-form preserving representation of low frequency signal components; (2) a parametric bandwidth extension, such as Spectral Band Replication (SBR) [2], which reconstructs the high frequency band from replicated low frequency portions through the control of additional parameters; and (3) a parametric stereo coder, such as "Parametric Stereo" [1, 9], which represents stereo signals with the help of a mono downmix and a corresponding set of inter-channel level, phase, and correlation parameters. For low bitrates the parametric tools are able to reach much higher coding efficiency with a good quality / bitrate trade-off. At higher bitrates, where the core coder is able to handle a wider bandwidth and also discrete coding of multiple channels, the parametric tools can be selectively disabled.

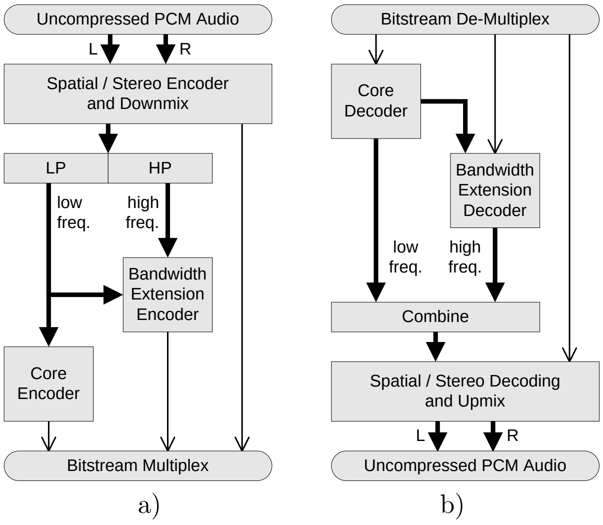

Fig. 1: General structure of a modern audio codec with a core codec accompanied by parametric tools for coding of bandwidth extension and stereo signals. USAC closely follows this coding paradigm. Figure 1 a) shows the encoder. Figure 1 b) shows the corresponding decoder structure. Bold arrows indicate audio signal flow. Thin arrows indicate side information and control data.

2.2. HE-AACv2

General transform coding schemes, such as AAC [1, 2], rely on a sink model motivated by the human auditory system. By means of this psychoacoustic model, temporal and simultaneous masking is exploited for irrelevance removal. The resulting audio coding scheme is based on three main steps: (1) a time/frequency conversion; (2) a subsequent quantization stage, in which the quantization error is controlled using information from a psychoacoustic model; and (3) an encoding stage, in which the quantized spectral coefficients and corresponding side information are entropy-encoded. The result is a highly flexible coding scheme, which adapts well to all types of input signals at various operating points.

To further increase the coding efficiency at low bitrates, HE-AACv2 combines an AAC core in the low frequency band with a parametric bandwidth and stereo extension. Spectral Band Replication (SBR) [2] reconstructs the high frequency content by replicating the low frequency signal portions, controlled by parameter sets containing level, noise, and tonality parameters. "Parametric Stereo" [1, 9] is capable of representing stereo signals by a mono downmix and corresponding sets of inter-channel level, phase, and correlation parameters.

2.3. AMR-WB+

Speech coding schemes, such as AMR-WB [4, 5], rely on a source model motivated by the mechanism of human speech production. These schemes typically have three major components: (1) a short-term linear predictive coding scheme (LPC), which models the vocal tract; (2) a long-term predictor (LTP) or "adaptive codebook", which models the periodicity in the excitation signal from the vocal chords; and (3) an "innovation codebook", which encodes the non-predictive part of the speech signal. AMR-WB follows the ACELP approach which uses an algebraic representation for the innovative codebook: a short block of excitation signal is encoded as a sparse set of pulses and associated gain for the block. The pulse codebook is represented in algebraic form. The encoded parameters in a speech coder are thus: the LPC coefficients, the LTP lag and gain, and the innovative excitation. This coding scheme can provide high quality for speech signals even at low bitrates.

To properly encode music signals, in AMR-WB+ the time domain speech coding modes were extended by a transform coding mode for the excitation signal (TCX). The AMR-WB+ standard also has a low rate parametric high frequency extension as well as parametric stereo capabilities.

3. THE MPEG UNIFIED SPEECH AND AUDIO CODING WORKITEM

Addressing the obvious need for an audio codec that can code speech and music equally well, ISO/IEC MPEG issued a Call for Proposal (CfP) on Unified Speech and Audio Coding (USAC) within MPEG-D [10] at the 82nd MPEG Meeting in October 2007. The responses to the Call were evaluated in an extensive listening test, with result that the joint contribution from Fraunhofer IIS and VoiceAge Corp. was selected as reference model zero (RM0) at the 85th MPEG meeting in summer 2008 [11]. Even at that point the system fulfilled all requirements for the new technology, as listed in the CfP [12].

In the subsequent collaborative phase the RM0 base system was further refined and improved within the MPEG Audio Subgroup until early 2011 when the technical development was essentially finished. The mentioned improvements were introduced by following a well defined core experiment process. In this manner further enhancements from Dolby Labs., Philips, Samsung, Panasonic, Sony and NTT Docomo were integrated into the system.

After technical completion of the standard, the MPEG Audio Subgroup conducted another comprehensive subjective Verification Test in summer 2011. The results of these tests are summarized in Section 5.

The standard reached International Standard (IS) stage in early 2012 by achieving a positive balloting vote from ISO's National Bodies voting for the standard [13].

4. TECHNICAL DESCRIPTION

4.1. System Overview

USAC preserves the same overall structure of HE-AACv2 as depicted in Figure 1. An enhanced SBR tool serves as a bandwidth extension module, while MPEG Surround 2-1-2 supplies parametric stereo coding functionality. The core coder consists of an AAC based transform coder enhanced by speech coding technology.

Figure 2 gives a more detailed insight into the workings of the USAC core decoder. Since in MPEG the encoder is not normatively specified, implementers are free to choose their own encoder architecture as long as it produces valid bitstreams. As a result,

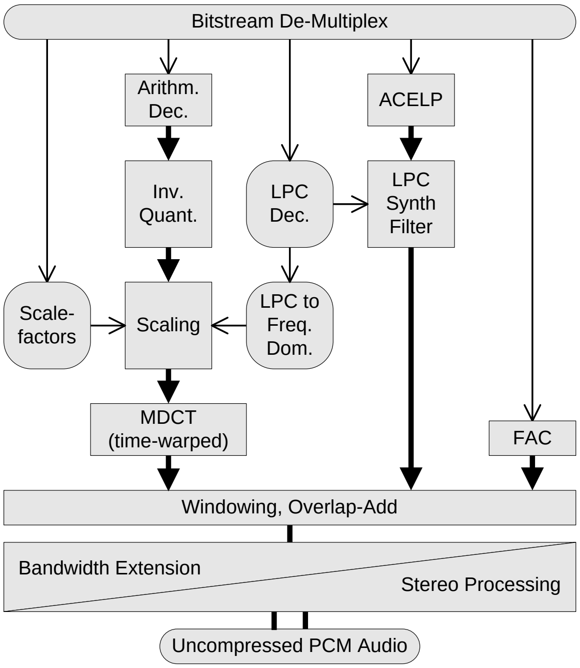

Fig. 2: Overview of USAC core decoder modules. The main decoding path features a Modified Discrete Cosine Transform (MDCT) domain coding part with scalefactor based or LPC based noise shaping. An ACELP path provides speech coder functionality. The Forward Aliasing Cancellation (FAC) enables smooth and flawless transitions between transform coder and ACELP. Following the core decoder, bandwidth extension and stereo processing is provided. Bold black lines indicate audio signal flow. Thin arrows indicate side information and control data.

USAC provides complete freedom of encoder implementation and - just like any MPEG codec - permits continuous performance improvement even years after finalization of the standardization process.

USAC retains all capabilities of AAC. In Figure 2 the left signal path resembles the AAC coding scheme. It comprises the function of entropy decoding (arithmetic decoder), inverse quantization, scaling of the spectral coefficients by the means of scalefactors and inverse MDCT transform. With respect to the MDCT, all flexibility inherited from AAC regarding the choice of the transform window, such as length, shape and dynamic switching is maintained. All AAC tools for discrete stereo or multi-channel operation are included in USAC. As a consequence, USAC can be operated in a mode equivalent to AAC.

In addition, USAC introduces new technologies which offer increased flexibility and enhanced efficiency. The AAC Huffman decoder was replaced by a more efficient context-adaptive arithmetic decoder. The scalefactor mechanism as known from AAC can control the quantization noise shaping with a fine spectral granularity. If appropriate, it can be substituted by a Frequency Domain LPC Noise Shaping (FDNS) mechanism which consumes fewer bits. The USAC MDCT features a larger set of window lengths. The 512 and 256 MDCT block sizes complement the AAC 1024 and 128 sizes, providing a more suitable time-frequency decomposition for many signals.

4.2. Core-Coder

In the following subsections each of the technologies employed in the core coder are described in more detail.

4.2.1. Arithmetic Coder

In the transform-coder path, a context adaptive arithmetic coder is used for entropy coding of the spectral coefficients. The arithmetic coder works on pairs of two adjacent spectral coefficients (2-tuples). These 2-tuples are split into three parts: (1) the sign bits; (2) the two most significant bit-planes; (3) the remaining least significant bit-planes. For the coding of the two most significant bit-planes, one out of 64 cumulative frequency tables is selected. This selection is derived from a context, which is modeled by previously coded 2-tuples (see Figure 3).

The remaining least significant bits are coded using one out of three cumulative frequency tables. This cumulative frequency table is chosen depending on the magnitude of the most-significant bits in the two uppermost bit-planes.

The signs are transmitted separately at the end of the spectral data. This algorithm allows a saving from 3 to more than 6% of the overall bitrate over AAC Huffman coding while showing comparable complexity requirements. [14]

4.2.2. Quantization Module

A scalar quantizer is used for the quantization of spectral coefficients. USAC supports two different quantization schemes, depending on the applied

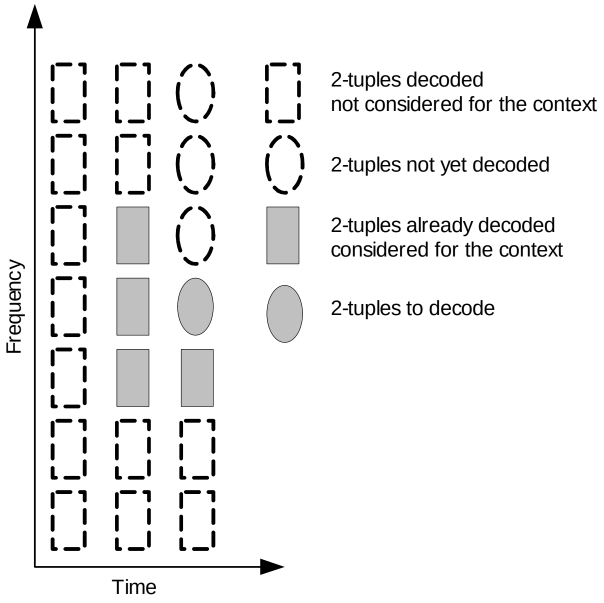

Fig. 3: Context Adaptive Arithmetic Coding.

noise shaping: (1) a non-uniform quantizer is used in combination with scalefactor based noise shaping. The scalefactor based noise shaping is performed on the granularity of pre-defined scalefactor bands. To allow for an additional noise shaping within a scalefactor band, a non-uniform quantization scheme is used [15]. In the non-uniform quantizer the quantization intervals get larger with higher amplitude. Thus, the increase in signal-to-noise ratio with rising signal energy is lower than in a uniform quantizer. (2) A uniform quantizer is used in combination with LPC based noise shaping. The LPC based noise shaping is able to model the spectral envelope continuously and without subdivision in fixed scalefactor bands. This alleviates the need for an extra intra-band noise shaping.

4.2.3. Noise Shaping using Scalefactors or LPC

USAC relies on two tools to shape the coding noise when encoding the MDCT coefficients. The first tool is based on a perceptual model and uses a set of scalefactors applied to frequency bands. The second tool is based on linear predictive modeling of the spectral envelope combined with a first-order filtering of the transform coefficients which achieves both frequency-domain noise shaping and sample-by-sample time-domain noise shaping. This second noise shaping tool, called FDNS for Frequency-Domain Noise Shaping, can be seen as a combination of perceptual weighting from speech coders and Temporal Noise Shaping (TNS). Both noise shaping tools are applied in the MDCT domain. The scalefactor approach is more adapted to stationary signals because the noise shaping stays constant over the whole MDCT frame whereas FDNS is more adapted to dynamic signals because the noise shaping evolves smoothly over time. Since the perceptual model using scalefactors is already well documented [15], only FDNS is described below.

When LPC based coding is employed, one LPC filter is decoded for every window within a frame. Depending on the decoded mode, there may be one up to four LPC filters per frame, plus another filter when initiating LPC based coding. Using these LPC coefficients, FDNS operates as follows (as seen at the decoder): for every window, the LPC parameters are converted into a set of M=64 gains $g_k[m]$ in the frequency domain, defining a coarse spectral noise shape at the overlap point between two consecutive MDCT windows. Then, in each of the M bands, a first-order inverse filtering is performed on the spectral coefficients $C_{mf}[k]$, as shown in Figure 4, to interpolate the noise level within the window boundaries noted as instants A and B in Figure 5. Therefore, instead of the conventional LPC coefficient interpolation and time-domain filtering as done in speech codecs, the process of noise shaping is applied only in the frequency domain. This provides two main advantages: first, the MDCT can be applied to the original signal (rather than the weighted signal as in speech coding), allowing proper TDAC on the transition between scalefactor and LPC based noise shaping; and second, because of the "TNS-like" feature of FDNS, the noise shape is finely controlled on a sample-by-sample basis rather than on a frame-by-frame basis.

4.2.4. (Time-Warped) MDCT

The Modified Discrete Cosine Transform (MDCT) is well suited for harmonic signals with a constant fundamental frequency F_0. In this case only a sparse spectrum with a limited number of relevant lines has to be coded. But when F_0 is rapidly varying, typically for voiced speech, the frequency modulation of the individual harmonic lines leads to a smeared spectrum and therefore a loss in coding

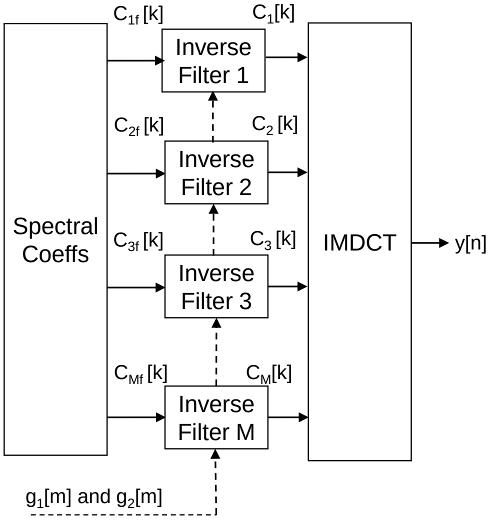

Fig. 4: FDNS processing.

gain. The Time Warped MDCT (TW-MDCT) [16] overcomes this problem by applying a variable resampling within one block prior to the transform. This resampling reduces or ideally completely removes the variation of F_0. The reduction of this variation causes a better energy compaction of the spectral representation and consequently an increased coding gain compared to the classic MDCT. Furthermore, a careful adaptation of the window functions and of the average sampling frequency retain the perfect reconstruction property and the constant framing of the classic MDCT. The necessary warp information needed for the inverse resampling at the decoder is efficiently coded and part of the side information in the bitstream.

4.2.5. Windowing

In terms of windows and transform block sizes, USAC combines the well-known advantages of the 50% overlap MDCT windows of length 2048 and 256 (transform core of 1024 and 128) from AAC with the higher flexibility of TCX with additional transform sizes of 512 and 256. The long transform windows allow optimal coding of distinctly tonal signals, while the shorter windows with shorter overlaps allow coding of signals with an intermediate and highly varying temporal structure. With this set of windows the codec can adapt its coding mode much more closely to the signal than possible before. Figure 6 shows the window and transform lengths and general shapes.

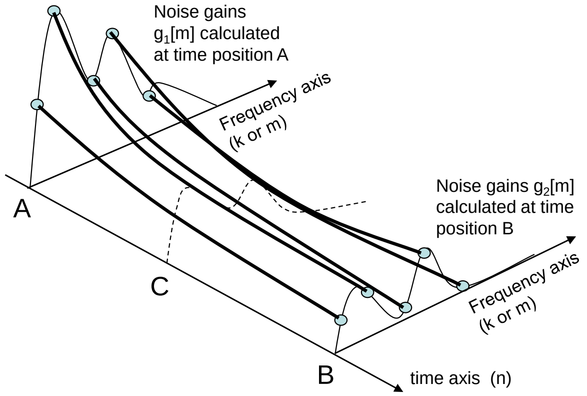

Fig. 5: Effect on noise shape of FDNS processing.

Similar to the start and stop windows of AAC, transitional windows accommodate the variation in transform length and coding mode [17]. In the special case of transitions to and from ACELP, the Forward Aliasing Cancellation takes effect (see Section 4.2.9).

Further flexibility is achieved by allowing a 768 sample based windowing scheme. In this mode all transform and windows sizes are reduced to $\frac{3}{4}$th of the above mentioned numbers. This allows even higher temporal resolution, which is particularly useful in situations where the codec runs on a reduced core sampling rate. This mode is combined with a 3 : 8 QMF filterbank upsampling in eSBR (see Section 4.3.2) such that a higher audio bandwidth can be achieved at the same time.

4.2.6. Quantization of LPC coefficients

USAC includes a new variable bitrate quantizer structure for the LPC filter coefficients. Rather than using trained codebooks which are memory-consuming, an extremely memory-efficient 2-stage approach based on algebraic vector quantization (AVQ, see Section 4.2.8) is used. An additional advantage of this approach is that the spectral distortion can be essentially maintained below a preset threshold by implicit bit allocation, thus making LPC quantization much less signal dependent. Another aspect of this variable bitrate quantizer is the application of LPC prediction within a frame.

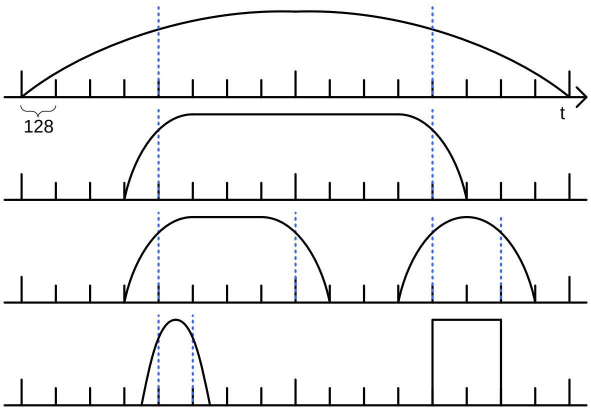

Fig. 6: Schematic overview over the allowed MDCT windows in USAC. Dotted lines indicate transform core boundaries. Bottom right shows ACELP window for reference. Transitional windows are not shown.

Specifically, if more than one LPC filter is transmitted within a frame, a subset of these filters are quantized differentially. This decreases significantly the bit consumption of the LPC quantizer in particular for speech signals. In this 2-stage quantizer, the first stage uses a small trained codebook as a first coarse approximation and the second stage uses a variable-rate AVQ quantizer in a split configuration (16-dimensional LPC coefficients quantized in 2 blocks of 8 dimensions).

4.2.7. ACELP

The time domain encoder in USAC is based on state-of-the-art ACELP speech compression technology. Several speech coding standards, in particular in cellular systems, integrate ACELP. The ACELP module in USAC uses essentially the same components as in AMR-WB+ [6] but with some improvements. LPC quantization was modified such that it is variable in bitrate (as described in Section 4.2.6). And the ACELP technology is more tightly integrated with other components of the codec. In ACELP mode, every quarter frame of 256 samples is split into 4 subframes of 64 samples (or for quarter frames of 192 samples it is split into 3 subframes of 64 samples). Using the LPC filter for that quarter frame (either a decoded filter or an interpolated filter depending on the position in a frame) each subframe is encoded as an excitation signal passed through the LPC filter. The excitation signal is encoded as the sum of two components: a pitch (or LTP) component (delayed, scaled version of the past excitation with properly chosen delay; also called adaptive codebook (ACB)) and an innovative component. The latter is encoded as a sparse vector formed by a series of properly placed non-zero impulses and corresponding signs and global gain. Depending on the available bitrate, the ACELP innovation codebook (ICB) size can be either of 20, 28, 36, 44, 52 or 64 bits. The more bits are spent for the codebook, the more impulses can be described and transmitted. Besides the LPC filter coefficients, the parameters transmitted in an ACELP quarter frame are:

Mean energy	2 bits
LTP pitch	9 or 6 bits
LTP filter	1 bit
ICB	20, 28, 36, 44, 52 or 64 bits
Gains	7 bits

All parameters are transmitted every subframe (every 64 samples), except the Mean energy which is transmitted once every ACELP quarter frame.

4.2.8. Algebraic Vector Quantization

Algebraic Vector Quantization (AVQ) is a structured quantization technique requiring very little memory and is intended to quantize signals with uniform distribution. The AVQ tool used in USAC is another component taken from AMR-WB+. It is used to quantize LPC coefficients and FAC parameters (see Section 4.2.9).

The AVQ quantizer is based on the RE8 lattice [18], which has a nice densely packed structure in 8 dimensions. An 8-dimensional vector in RE8 can be represented by a so-called "leader" along with a specific permutation of the leader components. Using an algebraic process, a unique index for each possible permutation can be calculated. Leaders with statistical equivalence can be grouped together to form base codebooks which will define the layers of the indexing. Three base codebooks have been defined: Q2, Q3 and Q4 where indexing all permutations of the selected leaders consumes 8, 12 and 16 bits respectively. To extend the quantizer to even greater size, instead of continuing to add bigger base codebook (Q5 and over), a Voronoi extension has been added to extend algebraically the base codebook. With each additional 8 bits (1 bit per dimension),

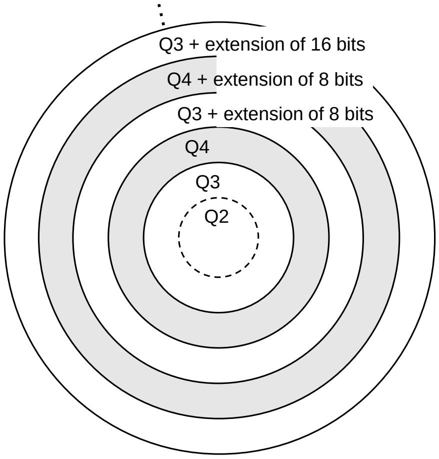

Fig. 7: Embedded structure of the AVQ quantizer.

the Voronoi extension doubles the size of the codebook. Therefore Q3 and Q4 extended by a factor of 2 will use 20 and 24 bits respectively, and for a factor of 4, they will use 28 and 32 bits respectively. Hence, although the first layer (Q2) requires 8 bits, each additional layer in the AVQ tool adds 4 bits to the indexing (1/2 bit resolution). It should be noted that Q2 is a subset of Q3. In the USAC bitstream, the layer number (Qn) is indexed separately using an entropy code since small codebooks are more probable than large codebooks.

4.2.9. Transition Handling

The USAC core combines two domains of quantization, the frequency domain which uses MDCT with overlapped windows and the time domain which uses ACELP with rectangular non-overlapping windows. To compute the synthesis signal, a decoded MDCT frame relies on time domain aliasing cancellation (TDAC) of adjacent windows whereas the decoded ACELP excitation uses the LPC filtering. To handle transitions in an effective way between the two modes, a new tool, called "Forward Aliasing Cancellation" (FAC) has been developed. This tool "Forwards" to the decoder the "Aliasing Cancellation" data required to retrieve the signal from the MDCT frame usually done by TDAC. Hence, at transitions between the two domains, additional parameters are transmitted, decoded and processed to obtain the FAC synthesis as shown in Figure 8. To recover the complete decoded signal, the FAC synthesis is merely combined with the windowed output of the MDCT. In the specific case of transitions from ACELP to MDCT, the ACELP synthesis and following zero-input response (ZIR) of the LPC filter is windowed, folded and used as a predictor, to reduce the FAC bit consumption.

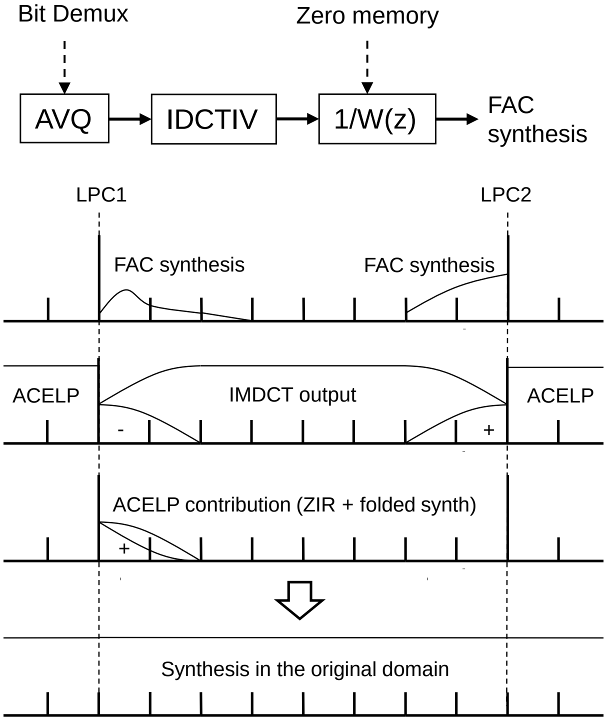

Fig. 8: Forward Aliasing Cancellation applied at transitions between ACELP and MDCT-encoded modes.

4.3. enhanced SBR Bandwidth Extension

4.3.1. Basic Concept of SBR

The Spectral Band Replication (SBR) technology was standardized in MPEG-4 in 2003, as an integral part of High Efficiency AAC (HE-AAC). The tool is a high frequency reconstruction tool that operates on a core coder signal, and extends the bandwidth of the output based on the available lowband signal and control data from the encoder. The principle of SBR and HE-AAC is elaborated on in [2, 19, 20].

The SBR decoder operating on the AAC as standardized in MPEG-4, is depicted in Figure 9. The

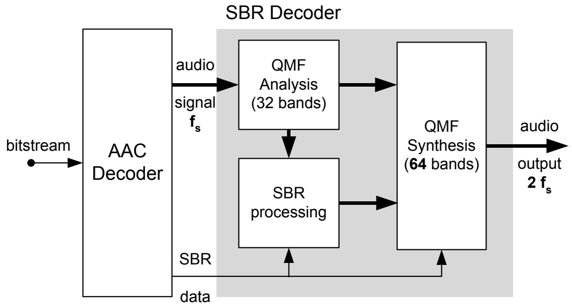

Fig. 9: Basic outline of SBR as used in MPEG-4 in combination with AAC.

system shown is a dual rate system where the SBR algorithm operates in a QMF domain and produces an output of wider bandwidth and twice the sampling rate of that of the core coded signal going into the SBR module.

The SBR decoder generates a high frequency signal by copy-up methods in the QMF domain as indicated in Figure 10. An inverse filtering is carried out within each QMF subband in order to adjust the tonality of the subband in accordance with parameters sent from the encoder.

The high frequency regenerated signal is subsequently envelope adjusted based on time/frequency tiles of envelope data transmitted from the encoder. During the envelope adjustment, additional noise and sinusoids are optionally added according to parametric data sent from the encoder.

4.3.2. Alternative Sampling Rate Ratios

MPEG-4 SBR was initially designed as a 2:1 system. Here, typically 1024 core coder samples are fed into a 32 band analysis QMF filterbank. The SBR tool performs a 2:1 upsampling in the QMF domain. After reconstructing the high frequency content, the signal is transformed back to time domain by means of 64 band synthesis QMF filterbank. This results in 2048 time domain samples at twice the core coder sampling rate.

For USAC, the traditional 2:1 system was extended by two additional operating modes. First, to cope with low core coder sampling rates, which are usually used at very low bitrates, a variation of the SBR module similar as standardized in DRM (Digital Radio Mondiale) has been adopted into the USAC standard. In this mode, the 32 band analysis QMF filterbank is replaced by a 16 band QMF analysis filterbank. Hence, the SBR module is also capable of operating as a 4:1 system, where SBR runs at four times the core coder sampling rate. In this case, the maximum output audio bandwidth the system can produce at low sampling rates is increased by a factor of two compared to that of the traditional 2:1 system. This increase in audio bandwidth results in a substantial improvement in subjective quality at very low bitrates.

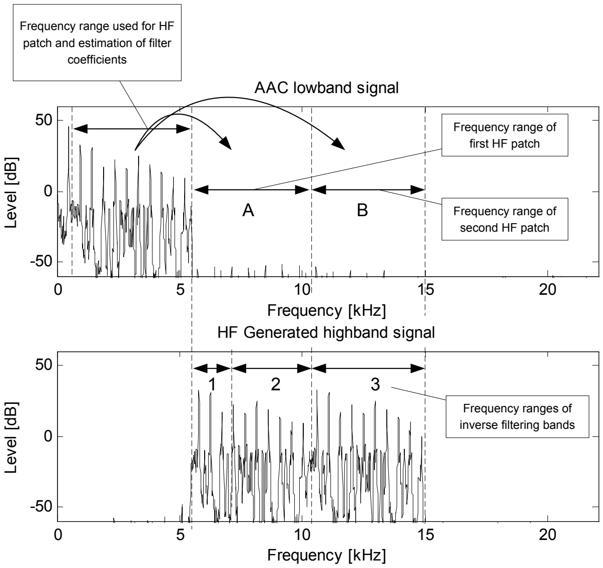

Fig. 10: Basic principle of copy-up based SBR as used in MPEG-4 in combination with AAC.

Second, USAC is also capable of operating in an 8:3 operating mode. In this case, a 24 band analysis QMF filterbank is used. In combination with a 768 core coder frame size, this mode allows for the best trade-off between optimal core-coder sampling rate and high temporal SBR resolution at medium bitrates, e.g. 24 kbit/s.

4.3.3. Harmonic Transposition

In USAC a harmonic transposer of integer order T maps a sinusoid with frequency ω into a sinusoid with frequency $T\omega$, while preserving signal duration. This concept was originally proposed for SBR in [21], and the quality advantage over the frequency shift method, especially for complex stationary music signals, was verified in [22].

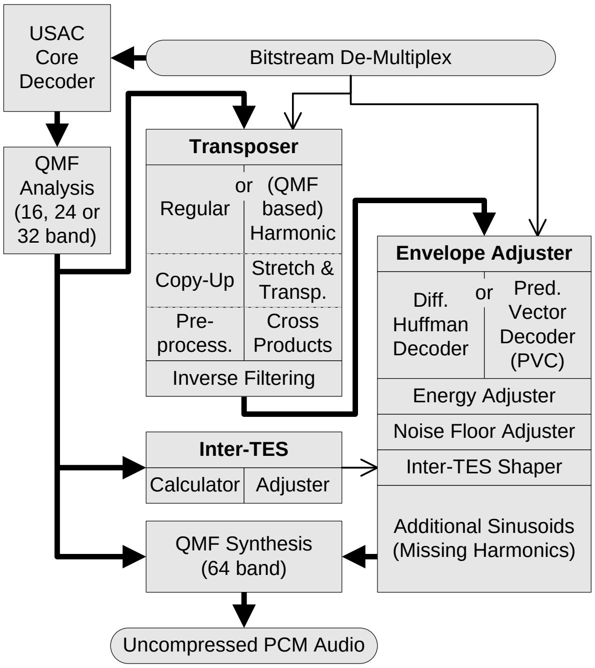

Fig. 11: Complete overview over the enhanced SBR of the USAC system. The figure shows the optional lower complexity QMF domain harmonic transposer, the PVC decoder and the Inter-TES decoder modules.

Three orders, $T = 2, 3, 4$, are used in sequence to produce each part of the desired output frequency range using the smallest possible transposition order. If output above the fourth order transposition range is required, it is generated by frequency shifts. When possible, near critically sampled baseband time domains are created for the processing to minimize computational complexity.

The benchmark quality transposer is based on a fine resolution sine windowed FFT. The algorithmic steps for $T = 2$ consist of complex filterbank analysis, subband phase multiplication by two, and filterbank synthesis with time stride twice of that of the analysis. The resulting time stretch is converted into transposition by a sampling rate change. The higher orders $T = 3, 4$ are generated in the same filterbank framework. For a given target subband, inputs from two adjacent source subbands are combined by interpolating phases linearly and magnitudes geometrically. Controlled by one bit per core coder frame, the FFT transposer adaptively invokes a frequency domain oversampling by 50% based on the transient improvement method of [23].

To allow the use of USAC in low-power applications such as portable devices, an alternate, low complexity, transposer that closely follows the bandwidth extension principle of the FFT transposer can be used as shown in Figure 11. This low complexity transposer operates in a QMF domain which allows for direct interfacing with the subsequent SBR processing. The coarse resolution QMF transposer suppresses intermodulation distortion by using overlapping block processing [24]. The finer time resolution of the QMF bank itself allows for a better transient response than that of the FFT without oversampling. Moreover, a geometrical magnitude weighting inside the subband blocks reduces potential time smearing.

The inherent spectral stretching of harmonic transposition can lead to a perceptual detachment of single overtones from periodic waveforms having rich overtone spectra. This effect can be attributed to the sparse overtone structure in the stretched spectral portions, since e.g. a stretching by the factor of two only preserves every other overtone. This is mitigated by the addition of cross products. These consist of contributions from pairs of source subbands separated by a distance corresponding to the fundamental frequency [23]. The control data for cross products is transmitted once per core coder frame and consists of an on/off flag and seven bits indicating the fundamental frequency in the case that the flag is set.

4.3.4. Predictive Vector Coding

In order to improve the subjective quality of the eSBR tool, in particular for speech content at low bitrates, Predictive Vector Coding (PVC) is added to the eSBR tool. Generally, for speech signals, there is a relatively high correlation between the spectral envelopes of low frequency bands and high frequency bands. In the PVC scheme, this is exploited by the prediction of the spectral envelope in high frequency bands from the spectral envelope in low frequency bands, where the coefficient matrices for the prediction are coded by means of vector quantization. The block diagram of the eSBR decoder including the PVC decoder is shown in Figure 11.

The analysis and synthesis QMF banks and HF generator remain unchanged, but the HF envelope adjuster is modified to process the high frequency envelopes generated by the PVC decoder. In the PVC decoder, the high frequency envelopes are generated by multiplying a prediction coefficient matrix with the low frequency envelopes. A prediction codebook in the PVC decoder holds 128 coefficient matrices. An index of the prediction coefficient matrix that provides the lowest difference between predicted and actual envelopes is transmitted as a 7 bit value in the bitstream.

4.3.5. Inter-subband-sample Temporal Envelope Shaping (Inter-TES)

For transient input signals audible distortion (pre/post-echoes) can occur in the high frequency components generated by eSBR due to its limited temporal resolution. Although splitting a frame into several shorter time segments can avoid the distortion, this requires more bits for eSBR information. In contrast, Inter-TES can reduce the distortion with smaller number of bits by taking advantage of the correlation between temporal envelopes in the low and high frequency bands. Inter-TES requires 1 bit for its activation and 2 bits for an additional parameter described below.

Figure 11 shows the Inter-TES module as part of the eSBR block diagram. When Inter-TES is activated, the temporal envelope of the low frequency signal is first calculated, and the gain values are then computed by adjusting the temporal envelope of the low frequency signal according to a transmitted parameter. Finally, the gain values are applied to the transposed high frequency signal including noise components. As shown in Figure 11, the shaped high frequency signal and the independent sinusoids are added if necessary, and then fed to the synthesis QMF bank in conjunction with the low frequency signal.

4.4. Stereo Coding

4.4.1. Discrete vs. Parametric Stereo Coding

There are two established approaches for coding of stereophonic audio signals. *Discrete stereo coding* schemes strive to represent the individual waveforms of each of the two channels of a stereo signal. They utilize joint stereo coding techniques such as mid/side (M/S) coding [25] to take inter-channel redundancy and binaural masking effects into account. *Parametric stereo coding* schemes [26, 27, 28], on the other hand, are designed to represent the perceived spatial sound image of the stereo signal. They utilize a compact parametric representation of the spatial sound image which is conveyed as side information in addition to a mono downmix signal and used in the decoder to recreate a stereo output signal. Parametric stereo coding is typically used at low target bitrates, where it achieves a higher coding efficiency than discrete stereo coding. USAC extends, combines, and integrates these two stereo coding schemes, thus bridging the gap between them.

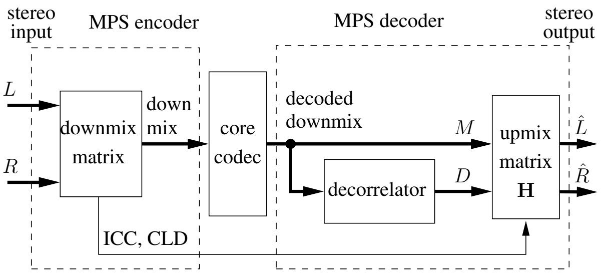

Fig. 12: Basic structure of the MPS 2-1-2 parametric stereo encoder and decoder used in USAC. Bold lines denote audio signal paths, whereas the thin arrow denotes the flow of parametric side information.

4.4.2. Parametric Stereo Coding with MPEG Surround 2-1-2

Parametric stereo coding in USAC is provided by an MPEG Surround 2-1-2 (MPS 2-1-2) downmix/upmix module which was derived from MPEG Surround (MPS) [9, 29]. The signal flow of the MPS 2-1-2 processing is depicted in Figure 12.

At the encoder, MPS calculates a downmix signal and parameters that capture the essential spatial properties of the input channels. These spatial parameters, namely the inter-channel level differences (CLDs) and inter-channel cross-correlations (ICCs), are only updated at a relatively low time-frequency resolution based on the limits of the human auditory system to perceive spatial phenomena, thus requiring a bitrate of only a few kbit/s.

In the decoder, a decorrelated signal D, generated from the downmixed input signal M, is fed along with M into the upmixing matrix $\mathbf{H}$, as depicted in the right dashed box in Figure 12. The coefficients of **H** are determined by the parametric spatial side information generated in the encoder.

Stereo sound quality is enhanced by utilizing phase parameters in addition to CLDs and ICCs. It is well-known that inter-channel phase differences (IPDs) can play an important role in stereo image quality, especially at low frequencies [26]. In contrast to parametric stereo coding in MPEG-4 HE-AAC v2 [1], phase coding in USAC only requires the transmission of IPD parameters, since it has been shown that the overall phase differences parameters (OPDs) can be analytically derived from the other spatial parameters on the decoder side [30, 31]. The USAC parametric stereo phase coding can handle anti-phase signals by applying an unbalanced weighting of the left and right channels during downmixing and upmixing processes. This improves stability for stereo signals where out-of-phase signal components would otherwise cancel each other in a simple mono downmix.

4.4.3. Unified Stereo Coding

In a parametric stereo decoder, the stereo signal is reconstructed by an upmix matrix from the mono downmix signal and a decorrelated version of the downmix, as shown in the right part of Figure 12. MPS enhances this concept by optionally replacing parts of the decorrelated signal with a residual waveform signal. This ensures scalability up to the same transparent audio quality achievable by discrete stereo coding, whereas the quality of a parametric stereo coder without residual coding might be limited by the parametric nature of the spatial sound image description. Unlike MPS, where the residual signals are coded independently from the downmix signals, USAC tightly integrates the coding of the downmix and residual signals.

As described above, in addition to CLD and ICC parameters, USAC also employs IPD parameters for coding the stereo image. The combination of parametric stereo coding involving IPD parameters and integrated residual coding is referred to as *unified stereo coding* in USAC. In order to minimize the residual signal, an encoder as shown in the left half of Figure 13 is used. In each frequency band, the left and right signals L and R are fed into a traditional mid/side (i. e. sum/difference) transform. The resulting signals are gain normalized by a factor c. A prediction of the scaled difference signal is made by multiplication of the mid (i. e. sum) signal M with a complex-valued parameter α. Both c and α are a function of the CLD, ICC, and IPD parameters. The resulting M and res signals are then fed into the 2-channel USAC core encoder which includes a correspondingly modified psychoacoustic model and can either encode the downmix and residual signal directly or can encode a mid/side transformed version known as *pseudo L/R* signal.

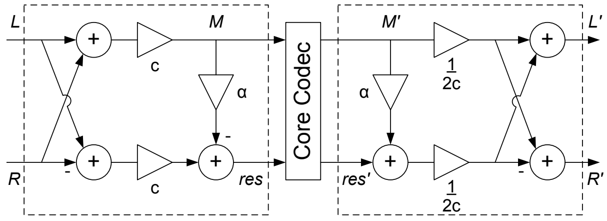

Fig. 13: Block diagram of unified stereo encoder (left) and decoder (right).

The decoder follows the inverse path, as depicted in the right half of Figure 13. The optimal order of MPS and SBR processing in the USAC decoder depends on the bandwidth of the residual signal. If no or a bandlimited residual signal is used, it is advantageous to apply mono SBR decoding followed by MPS 2-1-2 decoding. At higher bitrates, where the residual signal can be coded with the same bandwidth as the downmix signal, it is beneficial to apply MPS 2-1-2 decoding prior to stereo SBR decoding.

4.4.4. Transient Steering Decorrelator

Applause signals are known to be a challenge for parametric stereo coding. In a simple model, applause signals can be thought of as being composed of a quasi-stationary noise-like background sound originating from the dense, far-off claps, and a collection of single, prominently exposed claps. Both components have very different properties that need to be addressed in the parametric upmix [32].

Upmixed applause signals usually lack spatial envelopment due to the insufficiently restored transient distribution and are impaired by temporally smeared transients. To preserve a natural and convincing spatio-temporal structure, a decorrelating technique is needed that can handle both of the extreme signal characteristics as described by the applause model. The Transient Steering Decorrelator (TSD) is an im-

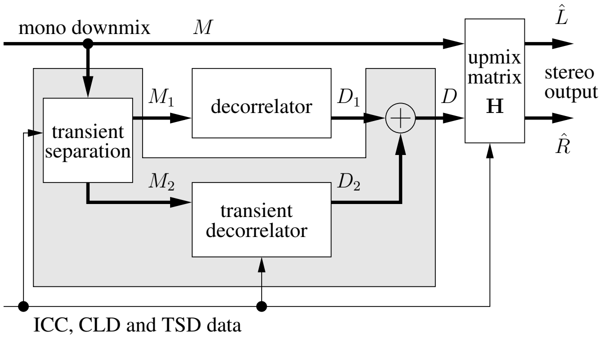

Fig. 14: TSD (highlighted by gray shading) within the MPS 2-1-2 module of the USAC decoder.

plementation of such a decorrelator [33]. TSD basically denotes a modification of the MPS 2-1-2 processing within USAC.

The block diagram of the TSD embedded in the upmix box of the MPS 2-1-2 decoder module is shown in Figure 14. The mono downmix is split by a transient separation unit with fine temporal granularity into a transient signal path and a non-transient signal path. Decorrelation is achieved separately within each signal path through specially adapted decorrelators. The outputs of these are added to obtain the final decorrelated signal. The non-transient signal path M_1 utilizes the MPS 2-1-2 late-reverb-type decorrelator. The transient signal path M_2 comprises a parameter-controlled transient decorrelator. Two frequency independent parameters that entirely guide the TSD process are transmitted in the TSD side information: a binary decision that controls the transient separation in the decoder, and phase values that spatially steer the transients in the transient decorrelator. Spatial reproduction of transient events does not require fine spectral granularity. Hence, if TSD is active, MPS 2-1-2 may use broadband spatial cues.

4.4.5. Complex Prediction Stereo Coding

MPS 2-1-2 and unified stereo coding employ complex QMF banks, which are shared with the SBR bandwidth extension tool. At high bitrates, however, the SBR tool is typically not operated, while unified stereo coding would still provide an improved coding efficiency compared to traditional joint stereo coding techniques such as mid/side coding. In order to achieve this improved coding efficiency without the computational complexity caused by the QMF banks, USAC provides a complex prediction stereo coding tool [34] that operates directly in the MDCT domain of the underlying transform coder.

Complex prediction stereo coding applies linear predictive coding principles to minimize inter-channel redundancy in mid signal M and side signal S. The prediction technique is able to compensate for inter-channel phase differences as it employs a complex-valued representation of either M or S in combination with a complex-valued prediction coefficient α. The redundant coherent portions between M and S signal are minimized - and the signal compaction maximized - by subtracting from the smaller of the two a weighted and phase-adjusted version of the larger one - the downmix spectrum D - leading to a residual spectrum E. Downmix and residual are then perceptually coded and transmitted along with prediction coefficients. Figure 15 shows the block diagram of a complex prediction stereo encoder and decoder.

The key here is to utilize a complex-valued downmix spectrum D obtained from a modulated complex lapped transform (MCLT) [35] representation for which the MDCT is the real part and whose imaginary part is the modified discrete sine transform (MDST). Given that in USAC, M and S are obtained via real-valued MDCTs, an additional real-to-imaginary (R2I) transform is required so that D can be constructed in both encoder and decoder [34]. In USAC, an efficient approximation of the R2I transform is utilized that operates directly in the frequency domain and does not increase the algorithmic delay of the coder.

4.5. System Aspects

4.5.1. Profiles

MPEG defines profiles as a combination of standardized tools. While all tools are always retained in the standard, the profile provides a subset or combination of tools that serve specific industry needs. A profile typically has several levels, where higher levels usually entail increasing complexity. Although there are many profiles defined in MPEG-4 Audio, the most successful and widely adopted ones are the "AAC family" of profiles, i.e. the "AAC profile", the "HE-AAC v1 profile" and the "HE-AAC v2 profile".

The AAC family of profiles, as outlined in Figure 16,

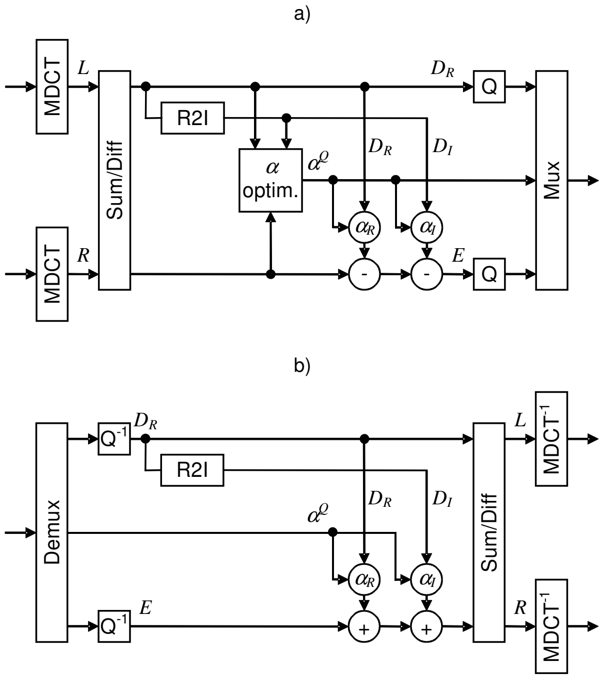

Fig. 15: Block diagram of complex prediction stereo encoder a) and decoder b).

are hierarchical. The structure of the profiles ensure that:

1. An AAC decoder plays: AAC LC (Low Complexity)
2. An HE-AAC decoder plays: AAC LC and SBR
3. An HE-AAC v2 decoder plays: AAC LC, SBR, and PS

In the MPEG USAC standard, two profiles are defined:

1. Extended HE-AAC Profile
2. USAC Baseline Profile

The Baseline USAC profile contains the complete USAC codec except for the following tools:

1. DFT transposer
2. Time-warped filterbank

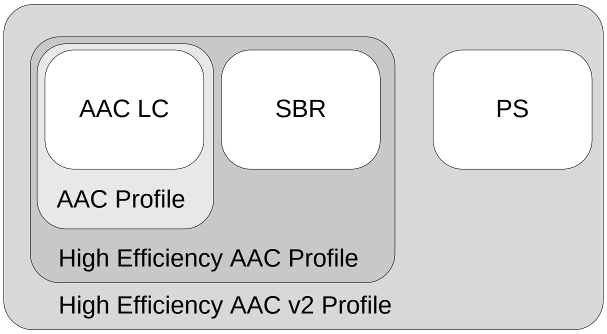

Fig. 16: The AAC Family of profiles.

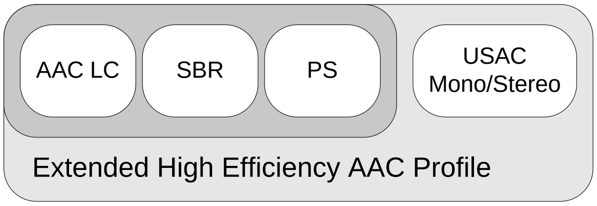

Fig. 17: The Extended HE AAC profile.

3. Fractional delay decorrelator

The Extended High Efficiency AAC profile contains all of the tools of the High Efficiency AAC v2 profile and is as such capable of decoding all AAC family profile streams. In addition, the profile incorporates mono/stereo capability of the Baseline USAC profile, as outlined in Figure 17.

The Extended High Efficiency AAC profile includes the mono/stereo capability of USAC since this part of USAC (when operated at low rates) provides the consistent performance across content types at low bitrates. This provides a natural evolution of one of the most successful families of profiles in MPEG Audio.

The Baseline USAC profile provides a clear standalone profile for applications and usage scenarios where the capability of supporting the AAC family is not relevant.

The worst case decoding complexity of both profiles is listed in Table 1 and 2. The complexity numbers are indicated in terms of Processor Complexity Units (PCU) and RAM Complexity Units (RCU). PCUs are specified in MOPS, and RCUs are expressed in kWords (1000 words).

Level	Max. channels	Max. sampling rate [kHz]	Max. PCU	Max. RCU
1	1	48	7	6
2	2	48	12	11
3	5.1	48	31	28
4	5.1	96	62	28

Table 1: USAC Baseline Profile processor and RAM complexity depending on decoder level.

Level	Max. channels	Max. sampling rate [kHz]	Max. PCU	Max. RCU
1	n/a	n/a	n/a	n/a
2	2	48 (Note 1)	12	11
3	2	48 (Note 1)	15	11
4	5.1 (Note 2)	48	25	28
5	5.1 (Note 2)	96	49	28

Table 2: Extended High Efficiency AAC Profile processor and RAM complexity depending on decoder level. Note 1: Level 2 and Level 3 differ for the decoding of HE-AACv2 bitstreams with respect to the max. AAC sampling rate in case Parametric Stereo data is present [1]. Note 2: USAC is limited to mono or stereo.

Each profile consists of several levels. The levels are defined hierarchically and denote the worst case complexity for a given decoder configuration. A higher level indicates an increased decoder complexity, which goes along with support for a higher number of channels or a higher output sampling rate.

First implementions of an Extended High Efficiency Profile decoder indicate similar complexity and memory requirements as for High Efficiency AAC v2 profile, when operated at the same level.

4.5.2. Transport

The way of signaling and transport of the USAC payload is very similar to MPEG-4 HE-AACv2. As for HE-AACv2, the concept of a signaling within MPEG-4 Audio is supported. For this purpose, a new Audio Object Type (AOT) for USAC is defined within the MPEG-4 *AudioSpecificConfig*. The *AudioSpecificConfig* can also carry the *UsacConfig* data, which is needed to properly set up the decoder.

The mandatory explicit signaling of all USAC decoder tools, such as SBR and PS, avoids several problems of HE-AACv2. For the reason of backward compatibility to decoders not supporting SBR or Parametric Stereo, an implicit signaling was introduced in HE-AACv2. As a consequence, a decoder at start-up was not able to clearly determine output sampling rate, channel configuration or number of samples per frame. In contrast to HE-AACv2, a USAC decoder unambiguously determines its configuration by reading the *UsacConfig* data at start-up. A set of *audioProfileLevelIndication* values allows for the signaling of the required decoder profile and level.

Like for HE-AACv2, the frame-wise payload (*UsacFrame*) directly corresponds to MPEG-4 access units. In combination with the MPEG-4 signaling, a multitude of transport formats natively supports the carriage of USAC. For streaming applications, the use of e.g. LATM/LOAS [1], IETF RFC 3016 [36] and RFC 3640 [37] is possible. For broadcasting applications, MPEG-2 transport stream [38] may be used. Finally, the MP4 and 3GPP file formats [39, 40] provide support for file-based applications.

MP4 and 3GPP file format based applications can now benefit from mandatory edit list support in USAC decoders to provide an exact time alignment: a decoder can reconstruct the signal with the exact starting and ending times, as compared to the original signal. Thus, additional samples at the beginning or end, introduced by frame-based processing and other buffering within the codec, are removed on the decoder side.

4.5.3. Random Access

Various tools in USAC may exploit inter-frame correlation to reduce the bit demand. In SBR and MPS 2-1-2, time differential coding relative to the previous frame may be used. The arithmetic coder may refer to a context based on the previous frame. Though these techniques improve coding efficiency for the individual frames, it comes at the cost of introducing a source of inter-frame dependencies. This means that a given frame may not be decoded without the knowledge of the previous frame.

In case of transmission over an error prone channel or in case of broadcasting where a continuously transmitted stream is received and shall be decoded starting with a randomly received first frame, these inter-frame dependencies can make the tune-in phase challenging.

For the reasons listed above, USAC audio streams contain random access frames that can be decoded entirely independent from any previous frame ("independent frames"). The information whether a frame acts as an "independent frame" is conveyed in the first bit of the USAC frame, and can be easily retrieved.

The frame independence is achieved by resetting the arithmetic coder context and forcing SBR and MPS 2-1-2 to frequency-differential coding only. The independent frame serves as safe starting points for random access decoding, also after a frame loss.

In addition to the indication of independent frames, great importance was paid to an explicit signaling of potential core-coder frame dependent information. Wherever window size, window shape or the need for FAC data is usually derived from the previous frame, this information can be unambiguously determined from the payload of any given independent frame.

5. PERFORMANCE

5.1. Listening Test Description

Three listening tests were performed to verify the quality of USAC. The objective of these verification tests was to confirm that the goals set out in the original Call for Proposals are met by the final standard [10, 41]. National Bodies could then take the test results as documented in the Verification Test report [42] into account when casting their final vote for USAC. Since the goal of USAC was the development of an audio codec that performs at least as good as the better of the best speech codec (AMR-WB+) and the best audio codec (HE-AACv2) around, the verification tests compared USAC to these codecs for mono and stereo at several bitrates. The results also provide the possibility to create a quality versus bitrate curve (a.k.a. rate-distortion curve) showing how the perceived quality of USAC progresses at different bitrates. The conditions included in each test are given in Tables 3 to 5. As stated above, along with USAC, two other audio coding standards were included as references: HE-AACv2 and AMR-WB+. These two reference codecs form the Virtual Codec (VC), which actually does not exist as a single technology, but is defined as the best of the HE-AACv2 or AMR-WB+ codecs on a given test item at a given operating point (bitrate and number of audio channels).

Condition	Label
Hidden reference	HR
Low pass anchor at 3,5 kHz[1]	LP3500
Low pass anchor at 7 kHz[1]	LP7000
USAC at 8 kbit/s	USAC-8
USAC at 12 kbit/s	USAC-12
USAC at 16 kbit/s	USAC-16
USAC at 24 kbit/s	USAC-24
HE-AAC v2 at 12 kbit/s	HE-AAC-12
HE-AAC v2 at 24 kbit/s	HE-AAC-24
AMR-WB+ at 8 kbit/s	AMR-8
AMR-WB+ at 12 kbit/s	AMR-12
AMR-WB+ at 24 kbit/s	AMR-24

Table 3: Conditions for Test 1 (mono at low bitrates).

Condition	Label
Hidden reference	HR
Low pass anchor at 3.5 kHz[1]	LP3500
Low pass anchor at 7 kHz[1]	LP7000
USAC at 16 kbit/s	USAC-16
USAC at 20 kbit/s	USAC-20
USAC at 24 kbit/s	USAC-24
HE-AAC v2 at 16 kbit/s	HE-AAC-16
HE-AAC v2 at 24 kbit/s	HE-AAC-24
AMR-WB+ at 16 kbit/s	AMR-16
AMR-WB+ at 24 kbit/s	AMR-24

Table 4: Conditions for Test 2 (stereo at low bitrates).

5.2. Test Items

Twenty-four test items were used in the test, consisting of 8 items from each of three content categories: Speech, Speech mixed with Music, and Music. Test items were stereo signals sampled at 48 kHz and were approximately 8 seconds in duration. A large number of relatively short test items were used so that the items could encompass a greater diversity of content. Mono items were derived from the stereo items

[1] Bandlimited but keeping the same stereo width as the original (hidden reference)

Condition	Label
Hidden reference	HR
Low pass anchor at 3.5 kHz[1]	LP3500
Low pass anchor at 7 kHz[1]	LP7000
USAC at 32 kbit/s	USAC-32
USAC at 48 kbit/s	USAC-48
USAC at 64 kbit/s	USAC-64
USAC at 96 kbit/s	USAC-96
HE-AAC v2 at 32 kbit/s	HE-AAC-32
HE-AAC v2 at 64 kbit/s	HE-AAC-64
HE-AAC v2 at 96 kbit/s	HE-AAC-96
AMR-WB+ at 32 kbit/s	AMR-32

Table 5: Conditions for Test 3 (stereo at high bitrates).

by either averaging left and right channel signals or taking only the left channel if an average resulting in significant comb filtering or phase cancellation. All items were selected to be challenging for all codecs under test.

5.3. Test Methodology

All tests used the MUSHRA methodology [43]. All tests were conducted in an acoustically controlled environment (such as a commercial sound booth) using reference quality headphones.

All items were concatenated to form a single file for processing by the systems under test. USAC processing was done using the Baseline USAC Profile encoder and decoder.

Fifteen test sites participated in the three tests. Of these, 13 test sites participated in test 1, 8 test sites in test 2 and 6 test sites in test 3. Listeners were post-screened and only those that showed consistent assessments were used in the statistical analysis. This post-screening consisted of checking whether, for a given listener in a given test, the Hidden Reference (HR) was always given a score larger or equal to 90 and whether the anchors are scored monotonic (LP3500<=LP7000<=HR). Only the scores of listeners having met these two post-screening conditions were retained for statistical analysis. After post-screening, tests 1, 2 and 3 had 60, 40 and 25 listeners, respectively.

5.4. Test Results

Figures 18 to 20 show the average absolute scores for each codec, including the Virtual Codec (VC), at the different operating points tested. Note that the scores between the tested points for a given codec are linearly interpolated in these Figures to show the trend of the quality/bitrate curve. The scores from all test sites, after listener post-screening, are pooled for this analysis. Vertical bars around each average score indicate the 95% confidence intervals using a t-distribution. The vertical axis in Figures 18 to 20 uses the MUSHRA subjective scale, shown in Table 6.

Descriptor	Range
EXCELLENT	80 to 100
GOOD	60 to 80
FAIR	40 to 60
POOR	20 to 40
BAD	0 to 20

Table 6: MUSHRA Subjective Scale.

Figures 18 to 20 show that, when averaging over all content types, the average score of USAC is significantly above that of the VC, with 95% confidence intervals not overlapping by a wide margin. Two exceptions are at 24 kbit/s mono and 96 kbit/s stereo where USAC and the VC have overlapping confidence intervals, but with the average score of USAC above that of the VC. Furthermore, Figures 18 to 20 show that when considering each signal content type individually (speech, music or speech mixed with music), the absolute score for USAC is always greater than the absolute score of the VC, and often by a large margin. This is most apparent in Test 2 (stereo operation between 16 and 24 kbit/s), with a 6 to 18 point advantage for USAC on the 100-point scale. A third observation from Figures 18 to 20 is that the quality for USAC is much more consistent across signal content types than the two state-of-the-art codecs considered (HE-AACv2 and AMR-WB+). This is especially apparent at medium and low rate operation (Figures 18 and 19).

The USAC verification test results show that not only does USAC match the quality of the best of HE-AACv2 and AMR-WB+ on all signal content types and at all bitrates tested (from 8 mono to 96 kbit/s stereo), but USAC actually exceeds that sound quality, and often by a large margin, in the bitrate range from 8 kbit/s mono to 64 kbit/s stereo. At higher

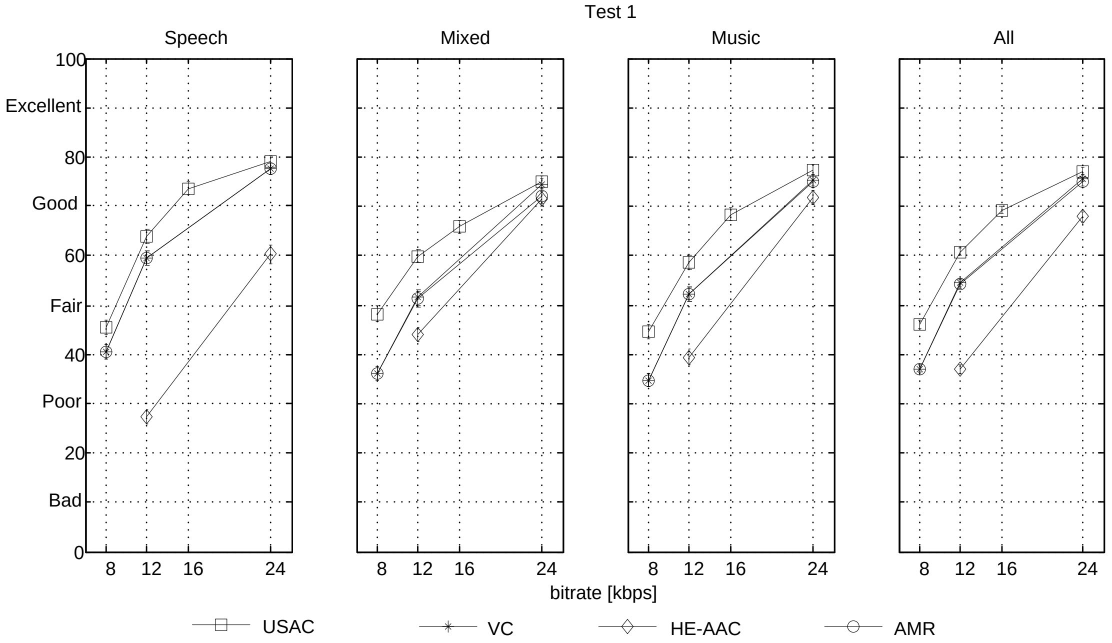

Fig. 18: Average absolute scores in Test 1 for USAC, HE-AACv2 (HE-AAC in the legend), AMR-WB+ (AMR in the legend) and the Virtual Codec (VC).

bitrates, the quality of USAC converges to that of HE-AACv2.

6. APPLICATIONS

USAC extends the HE-AACv2 range of use towards lower bitrates. As it additionally delivers at least the same quality as HE-AACv2 at higher rates, it also allows for applications requiring scalability over a large bitrate range. This makes USAC especially interesting for applications where bandwidth is limited or varying. Although mobile bandwidth is increasing with the upcoming 4G mobile standards, at the same time mobile data bandwidth usage increases dramatically. Moreover, multimedia streaming is accounting for a major part of today's growth in mobile bandwidth traffic.

In applications such as streaming multimedia to mobile devices, bandwidth scalability is a key requirement to ensure a pleasant user experience also under non-optimal conditions. Users want to receive the content without dropouts not only when being the only user in a cell and not moving. They want to listen to their favorite Internet radio station also when sitting in a fast traveling car or train, or while waiting for the very same train in a crowded station.

In digital radio, saving on transmission bandwidth reduces distribution costs and allows for a greater diversity of programs. Coding efficiency is most relevant for mobile reception, where robust channel coding schemes add to the needed transmission bandwidth.

Even in mobile TV, where video occupies the largest share of the transmission bandwidth, adding additional audio tracks like simulcasting stereo and multichannel audio or adding additional services like audio descriptive channels will significantly increase bandwidth demand. This raises the need for a highly efficient compression scheme, delivering good audio quality for both music and spoken material at low bitrates.

The situation is similar for audio books. Even though these contain mostly speech content, which may justify using dedicated speech codecs, background music and effects should be reproduced in high quality as well.

For all of the above-mentioned applications, the new USAC standard seems perfectly suited because of

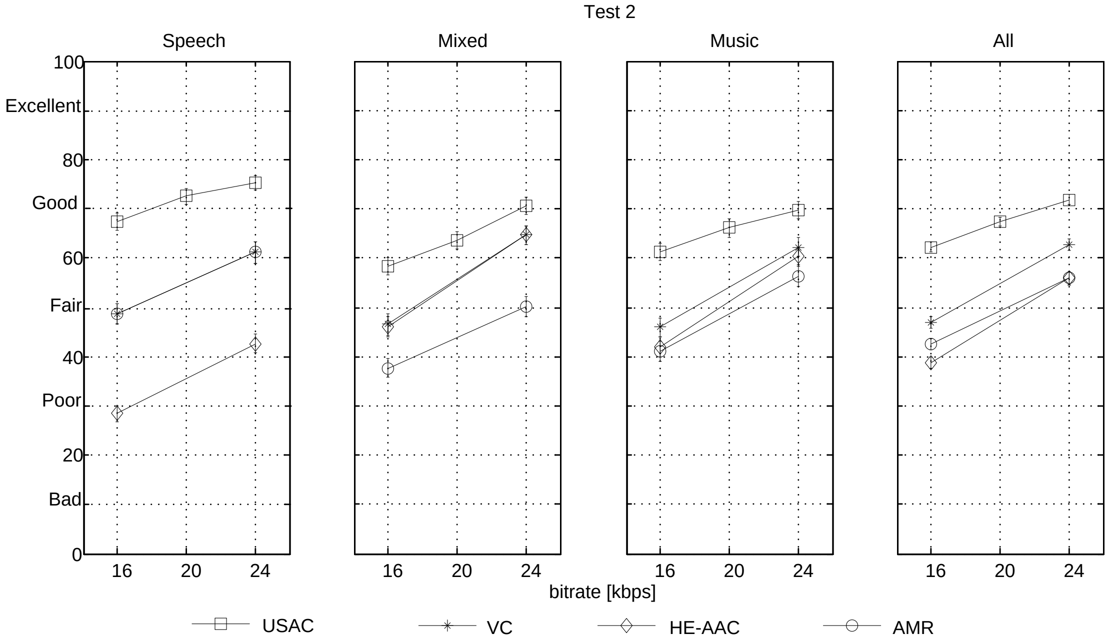

Fig. 19: Average absolute scores in Test 2 for USAC, HE-AACv2 (HE-AAC in the legend), AMR-WB+ (AMR in the legend) and the Virtual Codec (VC).

its extended bitrate range, quality consistency, and unprecedented efficiency.

7. CONCLUSION

The ISO/IEC 23003-3:2012 MPEG-D Unified speech and audio coding standard is the first codec that reliably merges the world of general audio coding and the world of speech coding into one solid design.

At the same time, USAC can be seen as the true successor of a long line of successful MPEG general audio codecs which started with MPEG-1 Audio and its most famous member, mp3. This was followed by AAC and HE-AAC(v2) which commercially share the success of mp3, as both codecs are present in virtually every mobile phone and many TV sets and mp3 players nowadays.

USAC now further builds on the technologies in mp3 and AAC and takes these one step further: It includes all the essential components of its predecessors in a further evolved form. It can, therefore, do everything mp3, AAC and HE-AAC can do, but is more efficient than its predecessors. Through the integration of the ACELP and TCX elements of AMR-WB+, USAC also represents a new state-of-the-art in low rate speech and mixed content coding.

This makes USAC today the most efficient codec for all signal categories, including speech signals. Starting at bitrates of around 8 kbit/s and up, it will deliver the best speech, music and mixed signal quality possible today for a given bitrate. Similar to AAC, USAC will scale towards perceptual transparency for higher bitrates.

During standardization, care was taken to keep the codec as lean as possible. As a result, the increase in implementation complexity over its predecessor is moderate and implementations for typical AAC and HE-AAC processing platforms are already available. All in all USAC can be considered the true 4th generation MPEG Audio codec, again setting a new state-of-the-art like its predecessors.

8. ACKNOWLEDGMENTS

A standardization project of this dimension can only ever be realized by a joint effort of a considerable number of highly skilled experts. Therefore, the authors would like to extend their appreciation to the

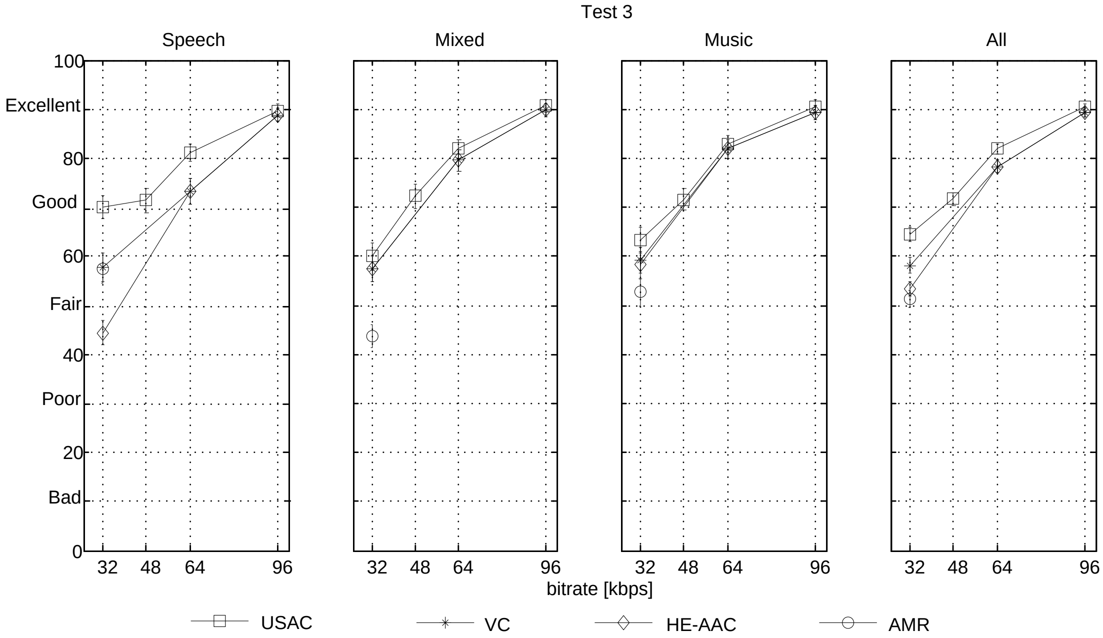

Fig. 20: Average absolute scores in Test 3 for USAC, HE-AACv2 (HE-AAC in the legend), AMR-WB+ (AMR in the legend) and the Virtual Codec (VC).

following list of contributors for their important and substantial work over the long duration of creation of the USAC standard:

T. Bäckström, P. Carlsson, Z. Dan, F. de Bont, B. den Brinker, S. Döhla, B. Edler, P. Ekstrand, D. Fischer, R. Geiger, P. Hedelin, J. Herre, M. Hildenbrand, J. Hilpert, J. Hirschfeld, J. Kim, J. Koppens, U. Krämer, A. Kuntz, F. Nagel, M. Neusinger, A. Niedermeier, M. Nishiguchi, M. Ostrovskyy, B. Resch, R. Salami, F. Lavoie, J. Samuelsson, G. Schuller, L. Sehlström, V. Subbaraman, M. Luis Valero, S. Wabnik, P. Warmbold, Y. Yokotani, H. Zhong, H. Zhou

Furthermore, the authors would like to express their gratitude to all members of the MPEG Audio Subgroup for their support and collaborative work during the standardization process.

9. REFERENCES

[1] ISO/IEC 14496-3:2009, "Coding of Audio-Visual Objects, Part 3: Audio," Aug. 2009.

[2] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A closer look into MPEG-4 High Efficiency AAC," in *115th Convention*, New York, NY, USA, Oct. 2003, Audio Eng. Soc., preprint 5871.

[3] C. Laflamme, J.-P. Adoul, R. Salami, S. Morissette, and P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic celp," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Toronto, Ont., Can., Apr. 1991, Inst. of Electrical and Electronics Eng., vol. 1, pp. 13–16.

[4] 3GPP, "Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description," 2002, 3GPP TS 26.171.

[5] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[6] 3GPP, "Audio codec processing functions; Extended Adaptive Multi-Rate - Wideband

(AMR-WB+) codec; Transcoding functions," 2004, 3GPP TS 26.290.

[7] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: a new audio coding standard for 3rd generation mobile audio services," in *Proc. IEEE ICASSP'05*, March 2005, vol. 2, pp. 1109–1112.

[8] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Taipei, Taiwan, Apr. 2009, Inst. of Electrical and Electronics Eng., number AE-L1.1.

[9] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*, John Wiley & Sons Ltd, West Sussex, England, 2007.

[10] ISO/IEC JTC1/SC29/WG11, "Call for Proposals on Unified Speech and Audio Coding," Shenzhen, China, Oct. 2007, MPEG2007/N9519.

[11] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, G. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, F. Nagel, J. Robilliard, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "A novel scheme for low bitrate unified speech and audio coding – MPEG RM0," in *126th Convention*, Munich, May 2009, Audio Eng. Soc., number 7713.

[12] ISO/IEC JTC1/SC29/WG11, "MPEG press release," Hannover, Germany, July 2008, MPEG2008/N9965.

[13] ISO/IEC 23003-3:2012, "MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding," 2012.

[14] G. Fuchs, V. Subbaraman, and M. Multrus, "Efficient Context Adaptive Entropy Coding for Real-Time Applications," in *Proc. IEEE ICASSP 2011*, May 2011, pp. 493–496.

[15] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, Davidson G., and Y. Oikawa, "Ise/iec mpeg-2 advanced audio coding," *J. Audio Eng. Soc*, vol. 45, pp. 789–814, 1997.

[16] B. Edler, S. Disch, S. Bayer, G. Fuchs, and R. Geiger, "A time-warped MDCT approach to speech transform coding," in *126th Convention*, Munich, May 2009, Audio Eng. Soc., number 7710.

[17] J. Lecomte, P. Gournay, R. Geiger, B. Bessette, and M. Neuendorf, "Efficient Cross-Fade Windows for Transitions Between LPC-Based and Non-LPC Based Audio Coding," in *126th Convention*, Munich, May 2009, Audio Eng. Soc., number 7712.

[18] S. Ragot, M. Xie, and R. Lefebvre, "Near-ellipsoidal Voronoi coding," *Information Theory, IEEE Transactions on*, vol. 49, no. 7, pp. 1815 – 1820, July 2003.

[19] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding," in *112th AES Convention*, Munich, 2002.

[20] A. C. den Brinker, J. Breebaart, P. Ekstrand, J. Engdegård, F. Henn, K. Kjörling, W. Oomen, and H. Purnhagen, "An overview of the coding standard mpeg-4 audio amendments 1 and 2: He-aac, ssc, and he-aac v2," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, Article ID 468971, 2009.

[21] L. Liljeryd, P. Ekstrand, F. Henn, and K. Kjörling, "Source coding enhancement using spectral band replication," 2004, EP0940015B1/ WO98/57436.

[22] F. Nagel and S. Disch, "A harmonic bandwidth extension method for audio codecs," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, April 2009, pp. 145–148.

[23] L. Villemoes, P. Ekstrand, and P. Hedelin, "Methods for enhanced harmonic transposition," in *IEEE Workshop on Appl. of Signal Proc. to Audio and Acoustics*, New Paltz, Oct. 2011.

[24] H. Zhong, L. Villemoes, P. Ekstrand, S. Disch, F. Nagel, S. Wilde, C. K. Seng, and T. Norimatsu, "Qmf based harmonic spectral band replication," in *131st Convention*, New York, Oct. 2011, AES, number 8517.

[25] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE ICASSP 1992*, March 1992, vol. 2, pp. 569–572.

[26] F. Baumgarte and C. Faller, "Binaural Cue Coding - Part I: Psychoacoustic Fundamentals and Design Principles," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 509 – 519, Nov. 2003.

[27] C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and Applications," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 520 – 531, Nov. 2003.

[28] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low complexity parametric stereo coding," in *116th Convention*, Berlin, May 2004, Audio Eng. Soc., number 6073.

[29] ISO/IEC 23003-1:2007, "MPEG-D (MPEG audio technologies), Part 1: MPEG Surround," 2007.

[30] J. Lapierre and R. Lefebvre, "On improving parametric stereo audio coding," in *120th Convention*, Paris, May 2006, Audio Eng. Soc.

[31] J. Kim, E. Oh, and J. Robilliard, "Enhanced stereo coding with phase parameters for MPEG unified speech and audio coding," in *127th Convention*, New York, Oct. 2009, Audio Eng. Soc.

[32] S. Disch and A. Kuntz, "A Dedicated Decorrelator for Parametric Spatial Coding of Applause-like Audio Signals," in *Microelectronic Systems: Circuits, Systems and Applications*, Albert Heuberger, Günter Elst, and Randolf Hanke, Eds., pp. 363–371. Springer Berlin Heidelberg, 2011.

[33] A. Kuntz, S. Disch, T. Bäckström, J. Robilliard, and C. Uhle, "The transient steering decorrelator tool in the upcoming MPEG unified speech and audio coding standard," in *131st Convention*. Audio Eng. Soc., Oct. 2011.

[34] C.R. Helmrich, P. Carlsson, S. Disch, B. Edler, J. Hilpert, M. Neusinger, H. Purnhagen, N. Rettelbach, J. Robilliard, and L. Villemoes, "Efficient Transform Coding of Two-Channel Audio Signals by Means of Complex-Valued Stereo Prediction," in *Proc. IEEE ICASSP 2011*, May 2011, pp. 497–500.

[35] F. Küch and B. Edler, "Aliasing reduction for modified discrete cosine transform domain filtering and its application to speech enhancement," in *IEEE Workshop on Appl. of Signal Proc. to Audio and Acoustics*, New Paltz, Oct. 2007, pp. 131–134.

[36] Y. Kikuchi, T. Nomura, S. Fukunaga, Y. Matsui, and H. Kimata, "RTP Payload Format for MPEG-4 Audio/Visual Streams," Nov. 2000, IETF RFC 3016.

[37] J. van der Meer, D. Mackie, V. Swaminathan, D. Singer, and P. Gentric, "RTP Payload Format for Transport of MPEG-4 Elementary Streams," Nov. 2003, IETF RFC 3640.

[38] ISO/IEC 13818-1:2007, "Generic Coding of Moving Pictures and Associated Audio Information, Part 1: Systems," 2007.

[39] ISO/IEC 14496-14:2003, "Coding of Audio-Visual Objects, Part 14: MP4 File Format," 2003.

[40] 3GPP, "Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)," 2011, 3GPP TS 26.244.

[41] ISO/IEC JTC1/SC29/WG11, "Evaluation Guidelines for Unified Speech and Audio Proposals," Antalya, Turkey, Jan. 2008, MPEG2008/N9638.

[42] ISO/IEC JTC1/SC29/WG11, "Unified Speech and Audio Coding Verification Test Report," Torino, Italy, Jul. 2011, MPEG2011/N12232.

[43] International Telecommunication Union, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," 2001, ITU-R, Recommendation BS. 1543-1, Geneva, Switzerland.