Structures for SNR Scalable Speech Coding

Hui Dong and Jerry D. Gibson, Fellow, IEEE

Abstract—SNR scalable speech coding is desirable for a number of network multimedia applications, but relatively few SNR-scalable speech coders exist for operation at rates below 16 kbps. We investigate several SNR scalable source coding structures and define the new concepts of dependent and independent SNR scalability, where independent SNR scalable coders depend on the core layer coder only through the core layer output. Independent SNR scalable structures offer the possibility of providing bit rate scalable functionality to existing nonscalable coders and standards. We show that the MPEG-4 scalable coders are examples of dependent SNR scalable coders, and we introduce a new independent SNR scalable coder called CELPTree, which has the additional advantage of being low delay. We compare the performance of the MPEG-4 coders and CELPTree for both clean and noisy speech, and we examine the effects of frequency-weighted distortion measures in the enhancement layers of SNR scalable speech coders.

Index Terms—CELP coding, scalable speech coding, speech coding.

I. INTRODUCTION

A host of speech coders have been standardized over the past few decades. However, the rapid evolution of wireless and wired communications networks motivates new, very flexible speech coding systems that not only have the ability to operate simultaneously under multiple constraints of bit rate, complexity, delay, and robustness, but that also offer enhanced functionalities such as SNR and bandwidth scalability [1]–[4].

SNR scalable coding consists of a minimum rate bit stream that provides acceptable coded speech quality, along with one or more enhancement bit streams, which when combined with a lower rate coded bit stream, provide improved speech quality. SNR scalability is implemented in a layered structure with a core bit-stream and enhancement bit-streams, so it may also be denoted as layered coding, embedded coding, or bit rate scalable coding [5]–[15]. From the viewpoint of rate distortion theory, SNR scalability is referred to as “successive refinement of information” [16]–[21].

Scalable coding has a number of advantages. In speech communications, for example, it improves interoperability of networks and services. Using a scalable approach, users can receive different quality versions of the same source according to their individually available resources. It also offers flexibility for error resilience since the high-priority information can be transmitted over a more reliable transmission path, or we can use unequal error protection of the core and enhancement layers. It may also allow range extension in wireless communications [22].

The difficulties with achieving scalability are two-fold: 1) all coders are not necessarily scalable and 2) the quality of the decoded signal derived from scalable coding cannot be better than nonscalable coding. Therefore, the goal of designing a scalable system is to generate a layered bit stream and to minimize the difference in signal quality between scalable and nonscalable coding. The difficulty with achieving scalability for speech signals is that speech coding techniques have been investigated and designed by taking advantage of speech production models and human perceptual systems. In recent years, a number of researchers have proposed SNR scalable speech coder designs, and at least one speech coding standard, the MPEG-4 speech coding toolbox, offers SNR scalability as an option. Further details on existing SNR scalable coders are provided in Section II.

With the rapidly increasing installed base of nonscalable speech coders for wireline and wireless telephony, voice over IP, and teleconferencing, it is of interest to determine the possibility of adding an SNR scalable capability to these coders without modifying the core speech coder. This problem has attracted relatively less attention, with one notable effort by Ramprashad [6], [23]. We develop this idea in some generality, which we designate as independent SNR scalability, and consider the rate distortion optimality of such a structure.

In adding the scalability property to an existing coder, it is desirable that the enhancement layer coder introduce as little additional algorithmic coding delay as possible. Furthermore, it is necessary that the enhancement layer coder interoperates with a variety of core layer perceptual distortion measures, and indeed, it is important that the enhancement layer be robust to any background impairments that might appear in various applications. We introduce one independent SNR scalable speech coding solution for the enhancement layer based upon a low-delay tree coder, which due to the choice of the distortion measure in the enhancement layer, is extremely robust to background impairments.

In Section II, we outline previous work on SNR scabbable speech coding, and in Section III, we define the concepts of dependent and independent SNR scalable speech coders. We relate these structures to the concepts of tree structured and multistage vector quantization, and we categorize several of the existing speech coders in terms of dependent and independent SNR scalability. We provide a summary of the relevant rate distortion theory in Section IV, wherein we also discuss the rate distortion theoretic concept of successive refinement of information, with a particular emphasis on the Shannon additive backward
channel, which plays an important role in proving the optimality of the independent SNR scalable designs.

Section V briefly discusses the MPEG-4 SNR scalable coder, a dependent SNR scalable design, which is used for later performance evaluations and comparisons. In Section VI-A, we describe three CELP speech coders that are used as core layer coders for the tree coding enhancement layer, and in Sections IV-B and C, we develop the proposed approach to independent SNR scalability using a low delay tree coder and describe the backward adaptive predictor algorithms, code tree design, and distortion measure. A voice activity detection algorithm is also discussed. Simulation results and performance comparisons for the MPEG-4 scalable speech coder and the CELPTree speech coder are presented in Section VII, where clean and noisy speech inputs are considered. The effects of the perceptual weighting filters in the core and enhancement layers are examined in Section VIII, and conclusions are given in Section IX.

II. PREVIOUS WORK ON SNR SCALABLE SPEECH CODING

SNR or bit rate scalable speech coding has attracted broad research interest only in the last decade. However, the more familiar embedded codes can be utilized as SNR scalable coding schemes. Embedded codes allow the $R'$ least significant bits of an $R$-bit codeword to be stripped off (usually in the network) without informing the encoder. Embedded codes can be adapted to SNR scalability applications by treating the $R'$ least significant bits as the enhancement layer, and the other $R-R'$ bits as the core layer. A prominent example of an embedded code is linear PCM that is inherently embedded [24]. Embedded DPCM [10], [25], [26], which was standardized by the ITU as G.727 [9] to provide a flexible way to alleviate congestion at any point in a network without exchanging control messages, also functions effectively as an SNR scalable coder. However, the operating rates of G.727 are 2 bits per sample or higher, which is above the data rates of interest today. The variable rate speech coding approach presented by Jayant [7] based on explicit coding of the reconstruction noise (coding error) in ADPCM, can also be employed as an SNR scalable coder, but again, only at higher bit rates. An early effort on scalable speech coding that uses only multistage vector quantization and not predictive coding at 8 to 32 kb/s is described in [27].

Recent work in SNR scalable speech coding has included modifying the excitation sequences in CELP, investigating different signal analysis and decomposition methods, and developing a pyramidal structure for the CELP codebook [28]–[37]. Both telephone bandwidth speech and wider band audio signals have been considered.

In 1998, a two stage hybrid embedded speech/audio coding structure was proposed by Ramprashad [6]. His structure uses a CELP standard speech coder in the core layer, but the second stage consists of a transform coder using an Modified Discrete Cosine Transform (MDCT) and perceptual coding principles. This stage is itself embedded both in complexity and bit rate, and provides various levels of enhancement of the core output, particularly for general audio signals like music. Due to its design, the second stage incurs a significant delay and can be sensitive to bit errors.

III. INDEPENDENT AND DEPENDENT SNR SCALABLE CODING

In order to investigate SNR scalable coding structures in more detail, we make a distinction between two classifications of scalable coders: dependent SNR scalable coders and independent SNR scalable coders. We define dependent SNR scalable coders as coders with enhancement layers that are intimately tied to the core layer using internal parameters such as predictor coefficients or the core layer codebook entries. Independent SNR scalable coders, on the other hand, are designed such that the enhancement layer depends minimally on the core layer (only through the core decoder output).

Fig. 1 shows a block diagram of a general dependent SNR scalable coder, where $f_1(\cdot)$ and $g_1(\cdot)$ are the encoder and decoder, respectively, in the core layer, which produce the decoded output denoted as $\hat{X}_c$. The enhancement layer has the encoder and decoder $f_2(\cdot), f_1(\cdot)$ and $g_2(\cdot), g_1(\cdot)$, where we show the explicit dependence of the enhancement layer encoder on the core layer by including the functional dependence of $f_2$ on $f_1$. The outputs of the core and enhancement layer encoders are both used as input to the decoder $g_2(\cdot)$ to produce the enhanced output denoted as $\hat{X}_e$. Note that although not explicitly shown in any of the figures, the encoder is assumed to have full knowledge of the corresponding decoders and can generate the decoder output as needed. This assumption is in agreement with existing speech coding methods.

Fig. 2 shows a block diagram of a general independent SNR scalable coder, where the notation for the core layer encoder and

---

**Figure 1.** Dependent SNR scalable structure where the enhancement layer encoder explicitly depends on the core layer encoder.

A CELP SNR scalable coder along with a CELP bandwidth scalable coder was standardized as a part of the MPEG-4 natural audio coding system in 1998. The MPEG-4 CELP [38] operates at more than fifty bit rates by changing its frame size and coding parameters for both wideband and narrowband speech. SNR scalability in the MPEG-4 CELP coder is achieved by encoding the speech signal using a combination of the core coder and the bit rate scalable tool. The core coder is based on a CELP algorithm and encodes the input speech signal at a predetermined bit rate range between 4 and 12 kb/s for narrowband speech. In the bit rate scalable tool, a residual signal that is produced at the core coder is encoded utilizing multislave vector quantization to enhance the coding quality by an analysis-by-synthesis structure. The bit rate of each enhancement layer is 2 kb/s, and up to 3 enhancement layers may be combined for better quality. In each enhancement layer, the LP filter and weighting filter are the same as those in the core layer.
decoder is the same as in Fig. 1. We explicitly show the decoded output of the core layer as being provided to the enhancement layer encoder, but that is all of the knowledge that the enhancement layer has concerning the core layer encoder/decoder, as is indicated by the notation for the enhancement layer encoder as $f_2(\cdot)$. The output of the enhancement layer is generated using the output of the enhancement layer encoder and the decoded output of the core layer $\hat{X}_c$. Note the significant differences between Figs. 1 and 2. The enhancement layer in Fig. 2 only needs the decoded output of the core layer and the source itself to generate the enhancement layer output.

A less general, but more common SNR scalable structure is shown in Fig. 3. The core layer is the same as in Fig. 2, but the enhancement layer now explicitly uses the difference between the original source and the decoded core layer output as input to the enhancement layer encoder. The enhancement layer decoder output is then summed with the core layer output to produce the enhanced version of the reconstructed source.

Another independent SNR scalable structure that should be mentioned is residual vector quantization (RVQ) [39], where the successive layers of encoding are based only upon the difference between the decoded output of the immediately preceding stage and the input. For one stage RVQ, this is equivalent to the additive dependent SNR scalable structure in Fig. 3, but without the enhancement layer encoder having access to the source directly.

Scalable coding structures have been investigated recently by Tuncel and Rose [40]. They classify SNR scalable coding structures in two ways: tree-structured vector quantization (TSVQ) and multistage vector quantization (MSVQ). Their TSVQ is equivalent to our dependent SNR scalable structure in Fig. 1, while MSVQ fits our additive independent SNR scalable structure in Fig. 3. As noted by Tuncel and Rose [40], TSVQ is the most general and optimal scalable coding strategy, but it may sometimes prove impractical due to its high codebook storage requirements and the necessity of a large training set during the design phase. On the other hand, MSVQ is a special case of TSVQ where the refinement is based on adding a new vector, which depends only on the current layer encoding index. Tuncel and Rose [40] also denote MSVQ as additive successive refinement and show that additive refinement is optimal (achieves the rate distortion function) for any continuous source with tight Shannon lower bound.

Both dependent and independent SNR scalable structures each have advantages in certain applications. For dependent SNR scalable coders, the enhancement layer can make full use of the information of the core so that the bit rate for the enhancement layer can be minimized and still achieve good performance. Independent SNR scalable coders take advantage of existing coders and standards by providing a scalable coding option to existing infrastructure without requiring modifications to the core layer coder. Furthermore, as we shall demonstrate, existing dependent SNR scalable structures use the same distortion measure for the enhancement layers as in the core layer; however, for an independent SNR scalable structure, a different distortion measure can be used for the enhancement layers, thus providing the flexibility to address specific characteristics of the residual distortion in the core layer output.

Based on this classification, linear PCM, embedded DPCM, and the MPEG-4 SNR scalable speech coding tools are all dependent SNR scalable structures, while Jayant's explicit coding of the reconstruction error in ADPCM and Ramprashad's speech coder are examples of independent SNR scalable methods. In the following, we introduce a new independent SNR scalable method and compare and contrast the expected performance and limitations of the two different approaches.

IV. RATE DISTORTION THEORY AND SUCCESSIVE REFINEMENT OF INFORMATION

The rate distortion function of a source with respect to a fidelity criterion is the minimum rate, $R$, at which the source can be transmitted subject to a constraint on the average distortion, $D$. The rate distortion function for a chosen $D$ is defined by

$$ R(D) = \min_{E[d(X, \hat{X})] \leq D} I(X; \hat{X}) $$

where $I(X; \hat{X})$ is the mutual information between the input source $X$ and the reconstructed output $\hat{X}$, $d(X, \hat{X})$ is the distortion measure, and the average distortion constraint determines the admissible set of transition probabilities between the input and the reconstructed output.

One of the most-quoted rate distortion results is the rate distortion function for a mean squared error (MSE) distortion measure and a memoryless Gaussian source with variance $\sigma_x^2$ and arbitrary mean, given by [41], [42]

$$ R_{G}(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma_x^2}{\sigma_r^2}, & \sigma_r^2 > D \\ 0, & \sigma_r^2 \leq D \end{cases} $$

This result, in its distortion rate form, has served as the basis for optimal quantizer design and for bit allocation in transform coding of speech, audio, and still images [2], [24].
Given by \( R(D) \) in most cases, one often resorts to investigating bounds. The rate distortion function for a non-Gaussian memoryless source with respect to the MSE distortion measure is upper bounded by \( R_G(D) \) given in (2) and lower bounded by \( R_L(D) \) given by

\[
R_L(D) = \frac{1}{2} \log \frac{Q_1}{D}
\]

where \( Q_1 \) is the entropy power of the source, given by \( Q_1 = \exp\{1/2\pi \int_\mathbb{R} \log S(\omega) d\mathbb{R} \} \), where \( S(\omega) \) is the power spectral density of the source. An important observation for speech coding applications is that \( Q_1 \) is the one-step mean squared prediction error for Gaussian sequences [41], [43], [44], and hence can be calculated from the autocorrelation matrix of the source as shown in Jayant and Noll [24] and in Gerbino and Gray [45].

The result (3) is a special case of the important Shannon lower bound for difference distortion measures derived by Shannon in 1959 [41], [46]. The Shannon lower bound can be expressed as

\[
R_L(D) = h(p) - \max_{g \in G_D} h(g) \leq R(D)
\]

where \( g_a(x) = e^{sp(x)}/\int dx e^{sp(x)} \), \( p(\cdot) \) represents the distortion measure (to avoid confusion with the differential), \( G_D \) is the set of all admissible probability densities that satisfy \( \int dx p(x) g_a(x) \leq D \) and \( p(\cdot) \) is the source probability density function (pdf).

The conditions where the Shannon lower bound is satisfied with equality, that is, when \( R(D) = R_L(D) \) in (4), yield the Shannon optimum backward channel shown in Fig. 4, where \( Z \) is distributed according to \( g_a(\cdot) \) and is statistically independent of \( Y \). The condition where the Shannon lower bound is tight has been important to deriving conditions for successive refinement of information, and we use the Shannon backward channel in the Appendix to show that the independent SNR scalable speech coding structure satisfies the conditions for successive refinability; that is, no other SNR scalable structure can do better.

SNR scalability, represented in Fig. 5, has been investigated from the rate distortion theory viewpoint as successive refinement of information. A sequence of random variables \( X_1, \ldots, X_n \) is successively refined by a two-stage description that is rate distortion optimal at each stage. The \( X \) sequence is encoded as \( \hat{X} \) at rate \( R_1 \) bits per symbol with average distortion \( D_1 \). Then information is added to the first message at rate \( R_2 = R_2 - R_1 \) bits per symbol so that the resulting two-stage reconstruction \( \hat{X}_r \) now has average distortion \( D_2 \) at rate \( R_2 \geq R_1 \).

Most rate distortion theory research for SNR scalability has been concerned with finding the conditions under which successive refinement is achievable. The successive refinement problem was first introduced by Koshelev [47]–[49] as the problem of divisibility, and he proved the sufficiency of a Markov chain relationship between the source and the refined reconstructions in 1980. Equitz and Cover [16] proved necessity in 1991. They also showed, using the Shannon backward channel formulation, that the Markov chain condition holds for Gaussian sources and squared error distortion, Laplacian sources and the absolute error distortion, and all discrete sources and Hamming distortion measures.

We state the following without the proofs, which can be found in the references.

**Theorem:** References [16], [17], [47]–[49]. Successive refinement with distortion \( D_1 \) and \( D_2 \) (\( D_1 \geq D_2 \)) can be achieved if and only if there exists a conditional distribution \( p(\hat{x}, \hat{x}_r \mid x) \) with \( Ed(X, \hat{X}) \leq D_1 \) and \( Ed(X, \hat{X}_r) \leq D_2 \), such that \( R(D_1) = I(X; \hat{X}) \) and \( R(D_2) = I(X; \hat{X}_r) \) and \( p(\hat{x}_r \mid x) p(\hat{x} \mid \hat{x}_r) \).

The last condition is equivalent to saying that \( X, \hat{X}, \hat{X}_r \) can be written as the Markov chain \( X \rightarrow \hat{X}_r \rightarrow \hat{X} \), or, equivalently, as \( X \rightarrow \hat{X}_r \rightarrow X \). According to the generalization by Rimoldi, we can extend this result for different distortion measures at each layer as follows [17]. Rimoldi’s result substantiates our use of different distortion measures in the core and enhancement layers.

**Theorem:** Successive refinement with distortion \( D_1 \) and \( D_2 \) (\( D_1 \geq D_2 \)) can be achieved if and only if there exists a conditional distribution \( p(\hat{x}_r \mid \hat{x}_r \mid x) \) with \( Ed_1(X, \hat{X}) \leq D_1 \) and \( Ed_2(X, \hat{X}_r) \leq D_2 \), such that \( R_1 \geq I(X; \hat{X}) \) and \( R_2 \geq I(X; \hat{X}_r) \).

V. MPEG-4 CELP SNR SCALABLE SPEECH CODING

The MPEG-4 CELP SNR scalable coder uses a CELP coder for its core, and is a dependent SNR scalable coder. The diagram of the MPEG-4 CELP bit rate scalable codec with the core and one enhancement layer is shown in Fig. 6.

The core coder is a typical CELP coder and the enhancement layer encodes the residual signal produced at the core layer by utilizing multipulse vector quantization. The algebraic-structure codebook at the enhancement layer is obtained by minimizing the perceptually weighted distortion between the reconstruction error signal from the core and the output signal from the enhancement layer. The adaptive codebook, fixed-codebook, and gains of the core layer are used to optimize the algebraic codebook search of the enhancement layer. The same LP synthesis filter, perceptual weighting filter and pitch filter are used for...
Fig. 6. MPEG-4 CELP SNR (bit rate) scalable speech coder block diagram.

Fig. 7. Two layered CELPTree additive independent speech coder block diagram.

Fig. 8. Original voiced speech segment (top) and the reconstruction error signal from G.729 (bottom). (a) Time domain waveform and (b) spectrum.

Table I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>CELP 1</th>
<th>G.723.1</th>
<th>G.729</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>550</td>
<td>800</td>
<td>1800</td>
</tr>
<tr>
<td>Pitch delay</td>
<td>600</td>
<td>600</td>
<td>1400</td>
</tr>
<tr>
<td>Gains</td>
<td>1600</td>
<td>2267</td>
<td>3400</td>
</tr>
<tr>
<td>Fixed-codebook</td>
<td>1700</td>
<td>2567</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3650</td>
<td>5267</td>
<td>8000</td>
</tr>
</tbody>
</table>

Fig. 8. Original voiced speech segment (top) and the reconstruction error signal from G.729 (bottom). (a) Time domain waveform and (b) spectrum.

TABLE I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>CELP 1</th>
<th>G.723.1</th>
<th>G.729</th>
</tr>
</thead>
<tbody>
<tr>
<td>LPC</td>
<td>550</td>
<td>800</td>
<td>1800</td>
</tr>
<tr>
<td>Pitch delay</td>
<td>600</td>
<td>600</td>
<td>1400</td>
</tr>
<tr>
<td>Gains</td>
<td>1600</td>
<td>2267</td>
<td>3400</td>
</tr>
<tr>
<td>Fixed-codebook</td>
<td>1700</td>
<td>2567</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3650</td>
<td>5267</td>
<td>8000</td>
</tr>
</tbody>
</table>

A. CELP Coders for the Core Layer

Although the CELPTree method is applicable to any core CELP coder, three ACELP coders have been investigated here for use as a core: a low bitrate CELP coder at 3.65 kb/s (the MPEG-4 3.85 kb/s coder with 200 bps of control information removed, labeled CELP 1), G.723.1 at 5.3 kb/s and G.729 at 8 kb/s. All three CELP coders use a 10th order LPC filter, but they have different structures for the codebook, pitch filter and weighting filter. They also have different frame sizes, windowing schemes, and quantization methods for the parameters. Table I gives a bit allocation comparison of the three CELP coders. It is important to note that G.723.1 and G.729 do not have SNR scalability options. Therefore, to achieve SNR scalability with an installed base of these coders, an independent SNR scalable solution is attractive. More details on each coder are given in [38], [50], [51].

Fig. 8(a) shows a typical voiced waveform segment for a female speaker and the corresponding core layer reconstruction error signal from G.729 at 8 kb/s. Note that even though G.729 is an excellent speech coder at 8 kb/s, the error signal is speech-like and its amplitude is correlated with the amplitude...
of the original speech. In fact, the coding error from G.729 at 8 kb/s is easily understandable. Fig. 8(b) shows the short-time spectra of this segment and its coding error from G.729. The formants and pitch of the error signal are strongly related to the formants and pitch of the original signal.

A comparison of Figs. 6 and 7 makes more explicit the differences between a dependent SNR scalable coder and an independent SNR scalable coder. Note that in Fig. 6 the synthesis filter $A(z)$ and the weighting function $W(z)$ are the same in the core and enhancement layers. In contrast, for the independent SNR scalable coder in Fig. 7, the synthesis filters, the weighting functions, and even the type of excitation are different in the two layers. The differences are also clearly evident in the decoders where, for the independent SNR scalable coder in Fig. 7, we see that the refined output is the sum of the core layer output and the enhancement layer output. On the other hand, the decoder in Fig. 6 uses the same synthesis filter and applies an additional excitation to the core layer decoder.

The speech coders used in the core and enhancement layers of CELPTree in Fig. 7 are described in more detail in the following sections.

B. Tree Coder for the Enhancement Layer

A functional block diagram of a tree coder is shown in Fig. 9. The code generator in Fig. 9, which is also the decoder, consists of possible excitation sequences into a short-term (or formant) predictor followed by a long-term predictor (or pitch prediction). The excitation sequences are generated from variables placed on the branches of a tree, where the tree can be populated with quantizer output levels or random values from a particular distribution. The excitation sequences used here are based on adaptive nonuniform quantizers as described in [52]. The distortion between the source sequence and each possible reconstructed sequence to some depth $L \leq 16$ in the tree is calculated, and the path through the tree with the smallest weighted distortion is selected as the best path. The $(M, L)$ algorithm that retains the $M = 8$ best paths to depth $L = 16$ is used to search the tree. Path map digits corresponding to this path are then released as encoder output digits and sent to the decoder or receiver for reconstruction. Two different tree codes are considered in this work, the 4–2 multistage that has an average rate of $3/2$ bit/sample and a 2 bit/sample tree [53]. Thus, the two possible data rates for the enhancement layer are 12 kb/s and 16 kb/s. At the corresponding receiver, the source sequence is reconstructed by using the encoded output digits as input to the code-generator.

Backward adaptation of the short-term predictors is attractive because it is low delay (here equal to 2 msec. for $L = 16$) and there is no requirement to transmit side information. Numerous algorithms have been proposed for backward adaptation which usually operate on a sample-by-sample basis, in contrast to the block update of forward adaptation. In the enhancement layer of CELPTree, backward adaptive tree coders are used that incorporate a robust backward coefficient adaptation structure [52], [54] based upon the powerful backward adaptive LSL (least square lattice) algorithm [55]. These algorithms achieve good speech quality while maintaining robustness to transmission errors for bit error rates as high as $10^{-5}$. The sensitivity to channel errors is reduced by using the receiver excitation sequence for adaptation of the short-term predictor coefficients, but the ability to track rapid changes in the speech is retained by shaping the excitation sequence with an all-zero filter.

For more details on the tree coder used here and tree coding, in general, the reader is referred to [52]–[56].

CELP coders may not work well for noisy speech due to the LPC model and the formant related weighting function, so tree coding in the second layer may compensate for this drawback and get natural sounding noisy speech by using an unweighted MSE distortion measure. To keep the bit rate of the enhancement layer as low as possible, a voiced/unvoiced detector is used in the enhancement layer, and only information from voiced segments is encoded to refine the core output (called partial refinement).

C. Voice Activity Detection (VAD)

The overall bit rate of the system is determined by the voiced/unvoiced detector, and a voice activity detection (VAD) algorithm based on zero crossing rates and energy levels is used in the refinement layer of the CELPTree scalable coder. The VAD used by CELPTree is adapted from G.729 Annex B [57], wherein only the energy and zero crossings are employed to detect the voiced portions of clean speech. We do not use the additional parameters available in G.729 Annex B, so if significant noise is present, the enhancement layer bit rate will increase and code the noise. This method is adopted here only to illustrate the available improvements based upon using different distortion measures in the core and enhancement layers. For example, to reconstruct natural sounding speech for noisy inputs, full refinement information (that is, refinement of voiced, unvoiced, and silent segments) and hence a higher bit rate, may be needed, while only partial refinement is required for clean speech.

For the speech segments used in this work, 30% of a clean speech signal is detected as voiced, so the overall bit rate of the scalable coder can be decreased significantly by only transmitting information during the voiced speech. For example, if the G.723.1 coder at 5.3 kb/s is used as the core and the 12 kb/s tree coder is used for the refinement layer, the overall bit rate for this system would be $5.3 + 12 \times 30\% = 8.90$ kb/s. This case is denoted Rf12V to indicate that the 12 kb/s rate in the
enhancement layer is used only to code the voiced segments. If, however, the proportion of voiced speech is 40%, the bit rate for the core and enhancement layer of CELPTree would be $5.3 + 12 \times 40\% = 10.1$ kb/s. If full rate enhancement is used, the transmitted data rate is $5.3 + 12 = 17.3$ kb/s, which is denoted as RF12F.

Since we are very interested in having minimal delay introduced by the enhancement layer, it is important to note that the voice activity detection in the CELPTree structure is performed based upon the frame size of the core layer and while the core layer is being processed. Note that this means the enhancement layer has access to the input signal as indicated in Fig. 2, and thus the CELPTree SNR scalable coder is independent of the core layer but it is not an RVQ coder. As a result, there is no additional delay due to the use of VAD.

VII. PERFORMANCE COMPARISONS

Compared to the MPEG-4 CELP enhancement layers [38], where decoders reproduce the speech signal at different qualities depending on the number of received fixed codebooks, the enhancement layer in the CELPTree of Fig. 7 depends on the core only through the CELP core decoder output. The advantage is that the second layer can be linked to any CELP coder with no modification to either the core or refinement systems. In addition, the second layer can use an entirely different coding paradigm from the core layer. Thus, the CELPTree structure offers a flexible coding scheme for different communications environments.

A. Scalability

When the standards G.723.1 and G.729 and the low rate CELP1 coder were used as core coders, the average bit rates of the CELPTree coder for 8 sentences are as given in Table II. Recall that RF12V means the core is refined by voiced(V) segments only from tree coding at 12 kb/s and RF16F means the core is refined by the full(F) error signal from tree coding at 16 kb/s. Also, on the average, 22% of these 8 speech files are detected to be voiced. Therefore while the MPEG-4 CELP coder covers bit rates from 3850 to 18 200 bits/s, the CELPTree coder covers bit rates from 3650 to 24 000 bits/s.

B. Objective and Subjective Performance

The segmental SNR’s averaged over voiced segments of coded speech from the CELPTree coders (with unweighted MSE in the enhancement layer) and for the MPEG-4 CELP coders are compared in Table III at bit rates 8.82 kb/s and 9.2 kb/s, respectively, for clean speech, where the CELPTree uses G.723.1 at 5.3 kb/s for the core layer and 3.52 kb/s at the enhancement layer (16 kb/s tree coding with 22% partial refinement), and the MPEG-4 CELP is operated at 5.2 kb/s from the core layer and 4 kb/s from the two enhancement layers (2 kb/s for each enhancement layer). It shows that the CELPTree coder performs much better than the MPEG-4 CELP in terms of SNR. As shown in Table III, babble noise with a speech signal power to noise power ratio of 3 dB and car noise at 0 dB decrease the SNR performance of both coders, but the CELPTree coder is more robust to background noise than the MPEG-4 CELP coder.

A comparison of the spectrograms of coded clean speech for the MPEG-4 and CELPTree coders is shown in Fig. 10(a). An inspection of these spectrograms reveals that CELPTree tends to reproduce a slightly more accurate version of the original, while the MPEG-4 coder inserts additional harmonic structure in comparison to the original.

The CELPTree coder can employ either a weighted or unweighted MSE distortion measure in the second layer coding as is appropriate. Therefore, when coding noisy speech, the CELPTree coder has the flexibility to get better natural sounding reconstructed noisy speech, without modifying the core layer, although with the simple VAD algorithm used, at much higher bit rates than the MPEG-4 CELP coder. Fig. 10(b) shows that the spectrogram of the coded noisy speech. CELP-Tree with the core coder at 5.3 kb/s and the enhancement coder at 16 kb/s (RF16F) matches that of the original speech better than that of coded speech from the MPEG-4 CELP with the core coder at 5.2 kb/s and the enhancement layer at 6 kb/s (the maximum rate for the enhancement layer in MPEG-4 CELP coder) in most areas.

Informal listening test results are shown in Table IV. The fraction of trials in which the two systems are preferred by 10 listeners over the MPEG-4 CELP coder is shown. With G.723.1 at 5.3 kb/s as the core, the CELPTree with the RF16V coder refines the voiced segments of the error signal at 2 bits per sample, which results in a total rate of 8.82 kb/s, and the CELPTree with the RF16F coder refines the full error signal at 2 bits per sample, which results in 21.3 kb/s. These coders are compared with the MPEG-4 CELP coder operating at 9.2 kb/s. The clean speech results in Table IV show that the voiced-segment only refinement works well for clean speech. For noisy speech the results show that the CELPTree scalable coder is more robust, especially for

### Table II

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Background</th>
<th>CELPTree</th>
<th>MPEG-4 CELP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male1</td>
<td>clean</td>
<td>19.86</td>
<td>13.29</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>19.33</td>
<td>12.97</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>18.50</td>
<td>12.12</td>
</tr>
<tr>
<td>Male2</td>
<td>clean</td>
<td>17.95</td>
<td>11.90</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>18.03</td>
<td>11.68</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>17.08</td>
<td>11.06</td>
</tr>
<tr>
<td>Female1</td>
<td>clean</td>
<td>21.36</td>
<td>15.41</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>21.00</td>
<td>14.81</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>20.33</td>
<td>14.20</td>
</tr>
<tr>
<td>Female2</td>
<td>clean</td>
<td>20.62</td>
<td>14.95</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>19.92</td>
<td>14.52</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>19.28</td>
<td>13.76</td>
</tr>
</tbody>
</table>

### Table III

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Background</th>
<th>CELPTree</th>
<th>MPEG-4 CELP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male1</td>
<td>clean</td>
<td>19.86</td>
<td>13.29</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>19.33</td>
<td>12.97</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>18.50</td>
<td>12.12</td>
</tr>
<tr>
<td>Male2</td>
<td>clean</td>
<td>17.95</td>
<td>11.90</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>18.03</td>
<td>11.68</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>17.08</td>
<td>11.06</td>
</tr>
<tr>
<td>Female1</td>
<td>clean</td>
<td>21.36</td>
<td>15.41</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>21.00</td>
<td>14.81</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>20.33</td>
<td>14.20</td>
</tr>
<tr>
<td>Female2</td>
<td>clean</td>
<td>20.62</td>
<td>14.95</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>19.92</td>
<td>14.52</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>19.28</td>
<td>13.76</td>
</tr>
</tbody>
</table>
Fig. 10. Spectrograms comparing the SNR scalability performance of the MPEG-4 and CELPTree speech coders: original (left), MPEG-4 CELP (middle), and CELPTree (right). (a) Clean female segment and (b) male segment in car noise background.

Table IV

<table>
<thead>
<tr>
<th>speaker</th>
<th>background</th>
<th>CELPTree w/Rf16V</th>
<th>CELPTree w/Rf16F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male1</td>
<td>clean</td>
<td>5/10</td>
<td>5/10</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>6/10</td>
<td>9/10</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>7/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Male2</td>
<td>clean</td>
<td>6/10</td>
<td>6/10</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>5/10</td>
<td>9/10</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>6/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Female1</td>
<td>clean</td>
<td>5/10</td>
<td>5/10</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>6/10</td>
<td>8/10</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>5/10</td>
<td>10/10</td>
</tr>
<tr>
<td>Female2</td>
<td>clean</td>
<td>5/10</td>
<td>5/10</td>
</tr>
<tr>
<td></td>
<td>babble noise</td>
<td>6/10</td>
<td>8/10</td>
</tr>
<tr>
<td></td>
<td>car noise</td>
<td>6/10</td>
<td>10/10</td>
</tr>
</tbody>
</table>

background car noise speech, and that the unvoiced segments are important for natural sounding noisy speech.

To provide a further evaluation of the CELPTree independent scalable coder performance for noisy speech, we also compare CELPTree performance to the highest rate possible for the MPEG-4 narrowband, scalable speech coding tool, which is a core rate of 12.2 kb/s plus three enhancement layers of 2 kb/s each for a total rate of 18.2 kb/s. Note that this option does not fit our scenario of starting with a fixed core layer rate and then adding enhancement layers, but in those applications where some kind of feedback mechanism is available to the encoder, it may be possible to switch to a higher rate core layer coder as considered here for MPEG-4. We do not consider this option for the independent scalable coder, CELPTree.

For clean speech, both the MPEG-4 coder at 18.2 kb/s and the CELPTree coder at 21.3 kb/s reproduce the original speech very well, with little to discern between them. For noisy speech, the CELPTree coder and the MPEG-4 coder both do well on the voiced portions of the sentence, with the CELPTree reproducing the original speech plus noise slightly better but with the MPEG-4 coders providing a very slight reduction in noise during the voiced speech. For the unvoiced segments and the noise during “silence” intervals, the CELPTree coder produces a more natural sounding reproduction, whereas the MPEG-4 coder sounds shaky, or equivalently, has a varying background noise amplitude. The spectrograms for an original noisy female speech segment, the MPEG-4 coder at 9.2 kb/s, the MPEG-4 coder at 18.2 kb/s, and the CELPTree coder at 21.3 kb/s are shown in Fig. 11. In analyzing these spectrograms, it is clear that the MPEG-4 coders reproduce the harmonic structure of the
voiced segments better than CELPTree and that CELPTree has reduced granular noise between voice segments. It is also evident that the MPEG-4 coder at 9.2 kb/s inserts more harmonic structure than is in the original, as can be seen around 2 kHz at 1–1.2 s and between 2 and 3 kHz at 2.5–2.7 s, for example, and also, changes the frequency content of the noise during non-voiced speech as a function of time, as can be seen from 0 to 1 s and around 4 s, for example. At 18.2 kb/s, the MPEG-4 coder does reduce the noise levels compared to the MPEG-4 coder at 9.2 kb/s and adds less harmonic structure, except in certain areas, such as around 1000 to 1500 Hz at 3.2 s, it inserts additional harmonics. The variation in the background noise frequency content is reduced but still evident.

The CELPTree coder does not add harmonic structure, reduces the noise level, and does not modulate the frequency content of the noise between voiced segments, but it also does a
poorer job of reproducing the higher frequency harmonics than the MPEG-4 coders at both rates. As a result, the CELPtree speech sounds slightly more natural and with less noise, but perhaps less “crisp.” The MPEG-4 coders have audible artifacts due to modulating the background noise frequency content as a function of time. The primary reasons behind these differences are likely that tree coding with the unweighted MSE in the enhancement layer is more like waveform coding, while CELP coding of all layers is more model-based. These results are examined in more depth in Section VIII.

VIII. WEIGHTING FUNCTIONS AND ERROR SPECTRA

A. Clean Speech

Fig. 12(a) shows the speech and reconstructed error spectra of a typical frame of male speech coded by CELPtree at 8.82 kb/s. The solid line shows the original speech spectrum, and the dashed line shows the error spectrum from layer 1 where a standard weighted distortion measure is used. The dashed-dotted line shows the refined error spectrum from two layers with the standard weighted distortion measure in layer 1 and unweighted
MSE for layer 2. The dotted line shows the error spectrum from two layers with the same weighted distortion measure used in both layers. We can see that the speech-like error spectrum from the core is below the speech spectrum across the band 200–3400 Hz. At the higher rate with two layers, the refined error spectrum with a weighted distortion measure only in the core is whiter than the refined error spectrum using a weighted distortion measure in each enhancement layer, which demonstrates the well-known whitening property of the MSE distortion measure.

Fig. 12(b) shows the speech and reconstruction error spectra of a typical frame of male speech coded by the MPEG-4 CELP scalable coder at 9.2 kb/s. The solid line shows the speech spectrum, and the dashed line shows the error spectrum from the core. The dashed-dotted line shows the refined error spectrum with one refinement layer. The dotted line shows the error spectrum with three refinement layers. It is evident that all error spectra are below the speech spectrum across the main band because of the weighted distortion measures; however, with multiple enhancement layers, the shaping of the reconstruction error spectrum becomes sharper at the formants and the nulls become deeper.

The spectra in these figures imply that both MPEG-4 CELP and CELPTree scalable coders work well for clean speech.

### B. Noisy Speech

Fig. 13(a) shows original noisy speech, coded noisy speech, and coding error spectra for a voiced frame of noisy speech, including the MPEG-4 coder at 9.2 kb/s, the MPEG-4 coder at 18.2 kb/s, and the CELPTree coder at 21.3 kb/s (because of the simple voice activity detector and the relatively high noise level). The solid lines in both plots show the original noisy speech spectrum. In the top plot, the dashed line shows the error spectrum for MPEG-4 CELP at 9.2 kb/s (denoted MP4CELP at 9 kb/s), the dotted line shows the error spectrum for MPEG-4 CELP at 18.2 kb/s (denoted MP4CELP at 18 kb/s), and the dash-dot line shows the error spectrum for the CELPTree coder with the MSE distortion measure in the enhancement layer. The same notation is used for the coded spectra in the bottom plot. It is evident that for all of the coders, the coding error spectrum is below the speech spectrum across the band of interest (below 3400 Hz) and that the CELPTree error spectrum is flatter than the MPEG-4 coders. Additionally, the MPEG-4 coder at 18.2 kb/s begins to lose the basic shape of the speech spectrum, with the first peak getting very close to the speech spectrum around 300 Hz. The coded error spectra in the bottom plot show that the CELPTree coder more nearly approximates the original speech spectrum than the MPEG-4 coders below about 3 kHz. Fig. 13(b) shows the corresponding spectra for an unvoiced speech frame using the same notation. Of particular interest is the shape of the coded speech spectra in the bottom plot, wherein the CELPTree is seen to reproduce the spectrum better below about 3 kHz, but also it is evident that the MPEG-4 coders have created spectral peaks which were not in the original noisy speech spectrum. Thus, for noisy speech, the CELPTree coded speech more closely approximates the original and sounds more natural, while the creation of formant structure by the MPEG-4 coders causes audible artifacts in the reconstructed output.

### IX. DISCUSSION AND CONCLUSION

We classify SNR scalable speech coders as dependent or independent as a function of whether the enhancement layer coders rely on parameters within the core layer or only on the core layer output, respectively. It is shown that when the Shannon lower bound is tight, independent SNR scalable coders can be rate distortion optimal. We introduced an independent SNR scalable coder, designated CELPTree, which has an CELP-based core and uses a robust, backward adaptive tree coder with voice activity detection as the enhancement layer coder. Thus, CELPTree can be used to provide a scalable coding option for the existing installed base of nonscalable coders such as G.729, without introducing significant additional coding delay beyond that of the core layer. The CELPTree coder is shown to be competitive with the MPEG-4 scalable coders for clean speech and to offer improved performance for noisy speech, although at a higher bit rate. The effects of frequency-weighted distortion measures are investigated, and it is demonstrated that perceptual weighting in the enhancement layers of an SNR scalable speech coder can produce artifacts in the coded speech if there is background noise.

**APPENDIX**

**INDEPENDENT SNR SCALABILITY AND SUCCESSIVE REFINEMENT**

We show that the general independent SNR scalable structure in Fig. 14(a) can satisfy the conditions for rate distortion theoretic successive refinability as discussed in Section IV when the Shannon lower bound in (4) is tight (achieved with equality). The most conspicuous example of a source and distortion measure for which the Shannon lower bound is tight is a memoryless Gaussian source subject to the weighted or unweighted MSE distortion measure [41]. From Fig. 14(a), we can write $X = \hat{X} + E$ and $\hat{X} = \hat{X}_r + \hat{E}$. But, when the Shannon lower bound is achieved with equality, we can also write that $E = \hat{E} + \Delta$, where $\Delta$ is statistically independent of $\hat{E}$, so we can write $X = X_r + \Delta$. Furthermore, when the Shannon lower bound is tight in layer 1, $\hat{E}$ is statistically independent of $\hat{X}$, and $\Delta$ is statistically independent of $X_r$. As a result we can draw the diagram in Fig. 14(b), from which we can conclude that $\hat{X} \rightarrow X_r \rightarrow X$ is a Markov chain.

---

**Authorized licensed use limited to: Univ of Calif Santa Barbara. Downloaded on November 7, 2008 at 01:20 from IEEE Xplore. Restrictions apply.**
Further, from $p(\hat{x}|\hat{e}, x)$, we use the Shannon lower bound relationships to write

$$p(\hat{x}|\hat{e}, x) = p(\hat{x}|\hat{e}, \hat{v} + \epsilon) = p(\hat{x}|\hat{v} + \delta + \hat{e}).$$

Which implies that $X \rightarrow \hat{X} \rightarrow \hat{X}$ is a Markov chain and the conditions for successive refinability in Section IV are satisfied. We are thus able to conclude that when the Shannon lower bound is satisfied with equality, the independent SNR scalable structure achieves rate distortion optimal performance, and no dependent SNR scalable coder can perform better. Tuncel and Rose also have shown that additive successive refinement using an MSQV structure can achieve rate distortion optimality when the Shannon lower bound is tight.

**ACKNOWLEDGMENT**

The authors gratefully acknowledge the constructive comments of the reviewers that significantly improved the presentation of these results.

**REFERENCES**


Hui Dong received the B.S. and M.S. degrees from Wuhan University, China, in 1987 and 1990, respectively, and the M.S. and Ph.D. degrees from Texas A&M University, College Station, and Southern Methodist University, Dallas, TX, respectively, in 1998 and 2002, all in electrical engineering.

She worked as a Lecturer/Researcher in data processing at Tongji Medical University, China, from 1990 to 1996, and as a post-doc in speech coding at the University of California at Santa Barbara from 2003 to 2004. She currently works for Ditech Communications Corp., Mountain View, CA, as a DSP Research Engineer. Her research interests include speech coding/processing, digital signal processing, information theory, and digital communication.


Dr. Gibson was Associate Editor for Speech Processing for the IEEE TRANSACTIONS ON COMMUNICATIONS from 1981 to 1985 and Associate Editor for Communications for the IEEE TRANSACTIONS ON INFORMATION THEORY from 1988 to 1991. He was President of the IEEE Information Theory Society in 1996. He served as Technical Program Chair of the 1999 IEEE Wireless Communications and Networking Conference, Technical Program Chair of the 1997 Asilomar Conference on Signals, Systems, and Computers, and General Co-Chair of the 1993 IEEE International Symposium on Information Theory. Currently, he serves on the Steering Committee for the Wireless Communications and Networking Conference. He is an elected Member-at-Large on the Communications Society Board of Governors for 2005–2007, and he serves on the Editorial Board of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. In 1990, he received The Fredrick Emmons Terman Award from the American Society for Engineering Education, and in 1992, was elected Fellow of the IEEE “for contributions to the theory and practice of adaptive prediction and speech waveform coding.” He was co-recipient of the 1993 IEEE Signal Processing Society Senior Paper Award for the Speech Processing area.