EFFICIENT SCALABLE CODING OF STEREOPHONIC AUDIO BY CONDITIONAL QUANTIZATION AND ESTIMATION-THEORETIC PREDICTION

Ashish Aggarwal, Sang-Uk Ryu and Kenneth Rose

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106-9560
Email:[ashish,sang,rose]@ece.ucsb.edu

ABSTRACT

The standard scalable coding of stereophonic audio suffers from significant performance loss because of (1) poor prediction gain at the enhancement-layer and (2) direct requantization of the reconstruction error, which is suboptimal for the noise-mask ratio (NMR) criterion. To mitigate such performance loss, this paper proposes an integrated approach which employs two complementary techniques, namely, the estimation theoretic (ET) predictor and the conditional enhancement-layer quantizer (CELQ). The ET predictor has been shown to combine information from various sources for efficient enhancement-layer prediction, while CELQ efficiently handles scalable quantization to minimize NMR. We demonstrate that the proposed combined approach can achieve major performance gains in terms of bit rate reduction and reconstruction quality enhancement. For example, the proposed 2x16kbps two layer coder achieves considerably improved reconstruction quality compared to that of the conventional 4x16kbps four layer coder, despite expending only 50% bit of the standard scalable coder bit rate.

1. INTRODUCTION

Bit rate scalability, or embedded coding, has become increasingly important. A scalable bit stream allows the decoder to produce a coarse reconstruction if only a portion of the bit stream (i.e., the base layer) is received, and to improve the quality as more of the total stream (i.e., enhancement-layer) is made available. Scalability is important in applications, such as digital audio/video broadcasting and multicast audio, which require simultaneous transmission over multiple channels of differing capacity. Further, scalability enables error robustness in a lossy transmission scenario. An important example of a scalable audio compression system is the MPEG-4 general audio coder [1, 2] which performs multi-layer coding using AAC modules [3]. Another important requirement of audio compression algorithms, in addition to scalability, is the multi-channel encoding. Of particular interest is stereo encoding which finds wide application in multimedia services. Current coders such as MPEG-AAC and AC3 provide bit stream formats and tools towards this end. These algorithms achieve efficient compression by exploiting two types of redundancies, namely, intra-channel [4], and inter-channel [5] redundancies. The major objection to incorporating scalability in current audio coders is the resulting loss in compression performance compared to the non-scalable coder, especially when low rate coding modules are employed.

In this paper, we attack the problem of efficient scalable coding for low bit rate predictive coding of audio. A conventional scalable predictive audio coder, such as a scalable stereo audio coder, incurs performance penalty because it fails to adequately utilize all the available information available. Further, in sharp contrast to the classical mean square error (MSE) criterion, this loss is particulary pronounced when optimizing a perceptually motivated distortion criterion such as the noise-mask ratio (NMR). An efficient scalable quantization for *memoryless* sources aimed at optimizing the NMR metric was previously proposed in [6, 7]. The proposed conditional (enhancement-layer) quantization (CELQ) scheme was shown to substantially improve the scalable performance of the audio coders by quantizing the base-layer reconstruction error in the compandor's compressed domain using two switchable (conditional) quantizers. For scalable quantization of sources with *memory*, an estimation-theoretic (ET) prediction framework was proposed in [8], which derives an optimal estimate, in the mean-squared prediction error sense, of the signal at the enhancement-layer. Direct application of ET to scalable coding of stereophonic audio was presented in [9]. In this paper, we combine ET with CELQ to derive a superior scalable predictive coder and demonstrate its performance in scalable stereo audio coding. The proposed approach when implemented with multi-layer stereo coder leads to substantial bit rate saving over the conventional scalable approach, while maintaining the same or better reproduction quality.

The paper is organized as follows: Section 2 formulates the main problem. Section 3 derives the ET and CELQ schemes. Section 4 integrates ET and CELQ within a stereo audio coder. Section 5 presents simulation results that substantiate the performance advantage of the new scheme over conventional scalable coding of stereophonic audio.

2. PROBLEM FORMULATION

Let us consider a typical two-layer scalable stereo coder. The input audio signal first undergoes transformation, for example, with the MDCT. At the base layer, the transform coefficients of the left channel are directly quantized. To exploit inter-channel redundancy, the right channel coefficients are predicted from the left channel reconstructed coefficients and the prediction error is quantized. The enhancement-layer for each channel operates independently on the corresponding base-layer reconstruction error. In other words, the signal at the enhancement-layer is *predicted* from the corresponding base-layer reconstruction, and the prediction error is quantized. We refer to this enhancement-layer pre-

This work is supported in part by the NSF under grants no. EIA-9986057 and EIA-0080134, the University of California MICRO Program, Dolby Laboratories, Inc., Lucent Technologies, Inc., Mindspeed Technologies, Inc., and Qualcomm, Inc.

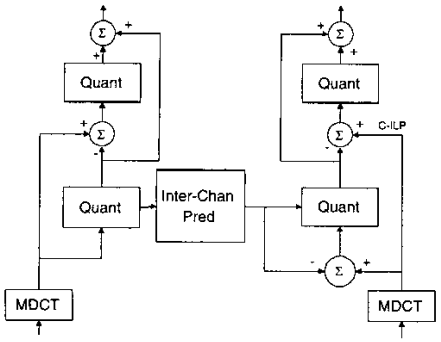

Fig. 1. *Block diagram of a conventional scalable stereo scheme*

diction scheme as the conventional inter-layer prediction (C-ILP). The conventional scalable stereo scheme with C-ILP is shown in Fig. 1. Quantization is performed by first grouping the signal into non-uniform bands to mimic the human auditory system's critical band model. All coefficients within a given band are then quantized using the same quantizer. The step-size of the quantizer is adjusted to match the masking profile, and thus, to minimize the average NMR of the frame for the given bit rate. The quantized coefficients are entropy coded and transmitted to the decoder, while the quantizer step-size for each band is transmitted as side information. This quantization scheme, for example, is employed by MPEG AAC.

Let us focus on the right channel at the base-layer. A non-uniform quantizer is employed in order to effectively handle the weights of the distortion metric (see design of entropy coded scalar quantizer [10]). For example, AAC uses a compressor function of $|x|^{3/4}$. However, the weight of the distortion metric (i.e., NMR) cannot be accurately expressed as a direct function of prediction error at the right channel. Thus, use of a non-uniform quantizer (or equivalently a compressor function) to quantize the prediction error fails to successfully optimize the weighted distortion metric. Conventional systems alleviate the problem by transmitting side information in the form of quantizer step-size which, at low rates, may consume 30%-40% of the total bit rate.

Let us now consider the enhancement-layer of right channel. It has two distinct sources of information: the base-layer reconstruction and the left channel's enhancement-layer reconstruction. Note that the latter has better quality than the left channel's base-layer reconstruction and therefore provides additional useful information. Nevertheless, most standard coders derive their prediction solely from the base-layer so as to ensure that the base-layer compressed residual is fully utilized. The ET prediction makes use of both sources of information and derives an optimal estimate of the right channel signal at the enhancement-layer. However, a direct quantization of the resulting prediction error, as done in [9], has the same drawbacks as those mentioned above for the base-layer, i.e., the problem of quantizer mismatch exists, and in fact is exacerbated at the enhancement-layer.

3. ET PREDICTION AND CELQ

Let x be the signal to be quantized. We use the subscripts $_l$ and $_r$ to denote the left and right channel, superscripts ^{b} and ^{e} to denote the base- and enhancement-layer, and finally, $\hat{}$ and $\tilde{}$ to denote the quantized and predicted values, respectively. For example, $\tilde{x}_r^b$ denotes the right channel predicted value at the base-layer. Similarly, $\hat{x}_l^e$ denotes the left channel quantized value at the enhancement-layer. Further, we use the notation (a, b) to specify a quantization interval. The relation between x_r and x_l, the corresponding transform coefficients at the right and left channels is modelled as,

$$x_r = \rho x_l + z \tag{1}$$

where ρ is the correlation coefficient, and z is a zero-mean, stationary process that is independent of x_l. Further, we model the marginal density of x_l and x_r by a Laplacian distribution. The probability density function (pdf) of z is then given by [11]:

$$p_z(z) = \rho^2 \delta(z) + (1-\rho^2)\tfrac{2}{\lambda} e^{-\lambda|z|}. \tag{2}$$

where the value of λ is estimated from a training set.

3.1. ET Prediction

The base quantizer at the right channel quantizes the prediction error, $\tilde{r}_r^b = x_r - \tilde{x}_r^b$, and transmits index i_r^b. The quantization interval associated with index i^b is denoted as (a_r^b, b_r^b), i.e., $\tilde{r}_r^b \in (a_r^b, b_r^b)$. ET prediction focuses on prediction at the right channel enhancement-layer, which is given as:

$$\begin{aligned} \tilde{x}_r^e &= E[x_r \mid \hat{x}_l^e, x_r \in (\tilde{x}_r^b + a_r^b, \tilde{x}_r^b + b_r^b)], \\ &= \rho\hat{x}_l^e + E[z \mid z \in (\tilde{x}_r^b + a_r^b - \rho\hat{x}_l^e, \tilde{x}_r^b + b_r^b - \rho\hat{x}_l^e)] \end{aligned}$$

where E denotes the expectation operator with respect to the pdf of x_r.

For the first order Laplace-Markov process, a closed form solution for this centroid calculation can be derived in terms of the interval limits as [8],

$$E_Z[z|z \in (s,t)] = \begin{cases} \frac{se^{-s\lambda} - te^{-t\lambda}}{e^{-s\lambda} - e^{-t\lambda}} + \frac{1}{\lambda} & \text{if } s > 0 \\ \frac{te^{t\lambda} - se^{s\lambda}}{e^{t\lambda} - e^{s\lambda}} - \frac{1}{\lambda} & \text{if } t < 0 \\ \frac{e^{s\lambda}(1-\lambda s) - e^{-t\lambda}(1+\lambda t)}{\lambda(2 - e^{s\lambda} - e^{-t\lambda} + \frac{2\rho^2}{1-\rho^2})} & \text{otherwise.} \end{cases} \tag{3}$$

Note that ET predictor can be extended in a straightforward manner to the multi-layer coding. For complete details of ET prediction see [8].

3.2. CELQ

CELQ is based on three observations:

1. Direct quantization the base-layer reconstruction error yields asymptotically optimal scalability for MSE.
2. An optimally designed compressor function converts a *weighted* squared-error (WSE) measure, such as NMR, to MSE.
3. The conditional pdf of the signal seen at the enhancement-layer varies substantially with the base-layer reconstruction value and parameters.

Based on the first two observations CELQ achieves asymptotic optimal quantization by quantizing the base-layer reconstruction error in the compandor's compressed domain. The third observation necessitates the design and use of different quantizers depending on the base-layer reconstruction value. However, given

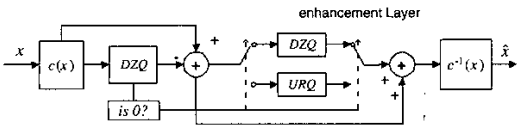

Fig. 2. *Block diagram of CELQ scheme*

	$\rho\hat{x}_l^e \in (\tilde{x}_n^b + a_r^b, \tilde{x}_n^b + b_r^b)$	otherwise
$r_r^e =$	$c(x_r) - c(\tilde{x}_r^e)$	$c(x_r) - c(\tilde{x}_r^e)$
$i_r^e =$	$s(r_r^e)\,\lfloor \lvert r_r^e\rvert/\Delta_r^e \rfloor$	$s(r_r^e)\,\lfloor \lvert r_r^e\rvert/\Delta_r^e + 0.5 \rfloor$
$a' =$	$e\big(\Delta_r^e(i_r^e + \frac{s(i_r^e)-1}{2})\big)$	$e\big(\Delta_r^e(i_r^e - 0.5)\big)$
$b' =$	$e\big(\Delta_r^e(i_r^e + \frac{s(i_r^e)+1}{2})\big)$	$e\big(\Delta_r^e(i_r^e + 0.5)\big)$

Table 1. *Quantization at right channel enhancement-layer using conditional quantizer.*

the assumption that the source can be modelled as Laplacian, the conditional quantizer can be implemented using two switchable distinct quantizers: one for use when the base-layer reconstruction is zero and one when it is non-zero. When the base-layer reconstruction is zero, the conditional distribution of the signal at enhancement-layer is similar to that of the base-layer's. In this case the enhancement-layer continues to employ a scaled version of the base-layer quantizer. When the base-layer is not zero, the enhancement-layer signal is quantized using a simple uniform scalar quantizer (USQ). The CELQ scheme for memoryless quantization of the audio signal using a dead-zone quantizer (DZQ) [6] and a USQ is shown in Fig. 2.

4. SCALABLE STEREO CODING WITH ET-CELQ

In this section we outline the implementation of the integrated ET-CELQ algorithm for scalable coding of stereophonic audio. The block diagram of the proposed scheme is shown in Fig. 3.

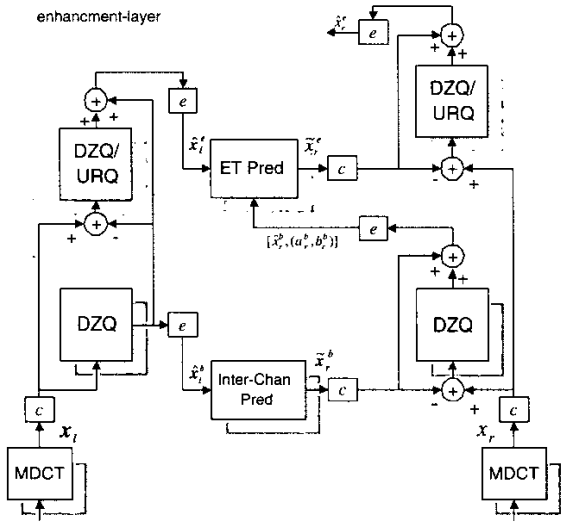

Fig. 3. *Block diagram of the proposed scheme (ET-CELQ)*

4.1. Base-Layer

Quantization of the left channel signal, x_l, at the base-layer is straightforward. It simply performs a non-uniform quantization of the transform coefficients. This operation is shown by the compressor function, $c(x)$, and the DZQ. The output at base-layer of the left channel is the quantized signal $\hat{x}_l^b$. Function $e()$ denotes the expandor (inverse compressor).

The signal at the right channel is predicted from the left channel reconstruction in a manner identical to the conventional scheme. The main difference in the proposed method scheme is the formation of the prediction error in the compressed domain. A DZQ is then employed to quantize the compressed domain prediction error, i.e.,

$$\begin{aligned} \tilde{x}_r^b &= \rho\hat{x}_l^b, \quad r_r^b = c(x_r) - c(\tilde{x}_r^b), \\ \hat{x}_r^b &= \tilde{x}_r^b + \hat{r}_l^b \quad \text{s.t.} \quad r_r^b \in \big(c(a_r^b), c(a_r^b)\big). \end{aligned}$$

4.2. Enhancement-layer

For quantization of the left channel signal at the enhancement-layer, we directly employ CELQ. The base-layer reconstruction error at the left channel is quantized in the compressed domain using either DZQ or USQ.

Let us next focus on the right channel enhancement-layer. ET prediction is employed to drive the estimate of the right channel signal, i.e.,

$$\tilde{x}_r^e = \rho\hat{x}_l^e + E[z \mid z \in (\tilde{x}_r^b + a_r^b - \rho\hat{x}_l^e, \tilde{x}_r^b + b_r^b - \rho\hat{x}_l^e)]. \quad (4)$$

The expectation in (4) is calculated using (3). The prediction error is then formed in the compressed domain as,

$$r_r^e = c(x_r) - c(\tilde{x}_r^e). \quad (5)$$

The next step is the quantization of the prediction error in (5) using conditional quantization. This is shown by the equation Table 1, where $s()$ denote the signum operation and Δ_r^e gives the quantizer stepsize in use.

The final step is the formation of the quantization interval, (a_r^e, b_r^e), which is the union of the quantization intervals from the base-layer and the enhancement-layer. This quantization interval is used only by subsequent encoding layers.

$$a_r^e = \max\big(\tilde{x}_r^b + a_r^b - \tilde{x}_r^e, a'\big), b_r^e = \min\big(\tilde{x}_r^b + b_r^b - \tilde{x}_r^e, b'\big) \quad (6)$$

5. SIMULATION RESULTS

In this section, the performance of the combined scheme (ET-CELQ) is compared experimentally to that of C-ILP (Fig. 1) and ET in conjunction with conventional quantization (ET-CS) [9] for a four-layer coder. Also shown for reference is the performance of the non-scalable coding scheme. We observe the bit rate savings provided by ET-CELQ, relative to C-ILP and ET-CS, at the same quality of reconstruction (measured by objective and subjective criteria). The average NMR and bit rate used for *right channel* encoding are measured at each layer. The bit rate used to encode the left channel is nearly identical in all methods. Eight critical

test database of 44.1kHz sampled music files from the MPEG-4 SQAM were used in the simulation. Fig. 4 depicts the rate-distortion curve of four-layer coder with each layer operating at 16kbps. The solid curve represents the operational rate-distortion bound or the non-scalable performance of the coder.

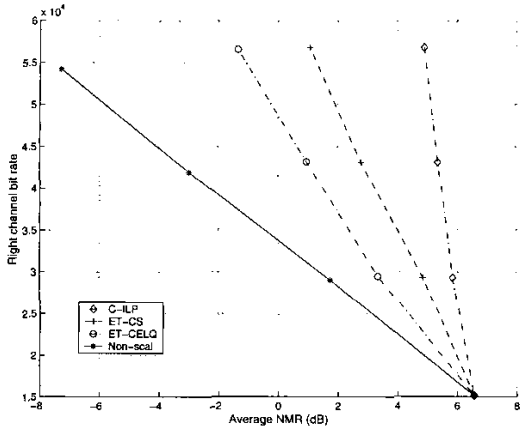

Fig. 4. *Rate distortion curve of four-layer coder for ET-CELQ (proposed), ET-CS and C-ILP*

ET-CELQ yielded considerable consistent bit rate savings over the ET-CS, which itself has substantial bit rate saving over the conventional C-ILP. For example, the 3x16kbps three layer ET-CELQ coder achieves similar reconstruction quality to ET-CS's 4x16kbps four layer coder, whose reconstruction quality is superior to the C-ILP 4x16kbps four layer coder. Furthermore, the ET-CELQ 2x16kbps two layer coder achieves better reconstruction quality than the conventional C-ILP 4x16kbps four layer coder – saving 50% in bit rate. Note that the results of ET-CS are similar to those given in [9].

We also performed an informal "AB" comparison test to subjectively evaluate the quality of the proposed scheme. Two tests, each with eight experienced listeners, were carried out using a 3-choise AB test over the test database. In the first test, we compared the ET-CELQ 2x16kbps two layer coder with the C-ILP 4x16kbps four layer coder. Table 2 gives the test results. It is evident that the subjective quality of the audio signal produced by ET-CELQ's 2x16kbps two layer coder is superior to that of the C-ILP's 4x16kbps four layer coder, despite the fact that ET-CELQ uses 50% fewer bit than the C-ILP.

preferred ET-CELQ @2x16kbps	preferred C-ILP @4x16kbps	no preference
59.30%	25.00%	15.70%

Table 2. Subjective performance of a two-layer ET-CELQ (2x16kbps), and four-layer C-ILP (4x16kbps) coder.

In the second test, we compared the ET-CELQ 3x16kbps three layer coder with ET-CS 4x16kbps four layer coder. The conditional coding via CELQ achieves major performance gains. This can be seen by Table 3, which compares ET-CELQ with ET-CS. The former operating at 3x16kbps achieves better subjective and objective quality than the latter running at 4x16kbps, a saving of 25% in bit rate. From the objective and subjective tests, we conclude that the proposed scheme leads to substantial bit rate reduction over the conventional method, while maintaining the same or better quality.

preferred ET-CELQ @3x16kbps	preferred ET-CS @4x16kbps	no preference
32.81%	29.69%	37.50%

Table 3. Subjective performance of a three-layer ET-CELQ (3x16kbps), and four-layer ET-CS (4x16kbps) coder.

6. CONCLUSION

In this paper, we presented a combined approach which integrates ET prediction and CELQ for improved scalable coding of stereophonic audio. It has been shown in earlier work that ET predictor provides advantages for scalable predictive coding of stereophonic audio. However, the approach still suffered from the performance penalty under the NMR criteria. As a remedy for the problem, we incorporated CELQ in conjunction with ET. The scheme leads to significant bit rate savings over the scheme employing ET with conventional quantization, which itself considerably outperforms the standard technique.

7. REFERENCES

[1] ISO/IEC 14496-3:2001(E) (Part 3: Audio), "Information technology - very low bitrate audio-visual coding."

[2] B. Grill, "A bit rate scalable perceptual coder for MPEG-4 audio," *103rd AES Conv.*, 1997. preprint 4620.

[3] M. Bosi, *et al.*, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. AES.*, vol. 45, pp. 789–814, October 1997.

[4] G. Schuller , *et al.*, "Perceptual audio coding using adaptive pre-and post-filters and lossless compression," *IEEE Trans. SAP*, vol. 10, pp. 379–90, September 2002.

[5] J. Johnston, *et al.*, "MPEG-2 NBC audio-stereo and multichannel coding methods," *101st AES Conv.*, 1996. preprint 4383.

[6] A. Aggarwal and K. Rose, "A conditional enhancement-layer quantizer for the scalable MPEG Advanced Audio Coder," *Proc. IEEE ICASSP*, vol. 2, pp. 1833–6, May 2002.

[7] A. Aggarwal and K. Rose, "Approaches to improve quantization performance over the Advanced Audio Coder," *112th AES Conv.*, 2002. preprint 5557.

[8] K. Rose and S. L. Regunathan, "Towards optimal scalability in predictive coding," *IEEE Trans. IP*, vol. 10, pp. 965–76, July 2001.

[9] A. Aggarwal, S. L. Regunathan, and K. Rose, "Optimal prediction in scalable coding of stereophonic audio," *109th AES Conv.*, 2000. preprint 5273.

[10] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. IT*, vol. IT-25, pp. 373–380, July 1979.

[11] N. Farvardin and J. W. Modestino, "Rate-distortion performance of DPCM schemes for autoregressive sources," *IEEE Trans. IT*, vol. IT-31, no. 3, pp. 402–18, 1985.