PERCEPTUAL ZEROTREES FOR SCALABLE WAVELET CODING OF WIDEBAND AUDIO

Ashish Aggarwal, Vladimir Cuperman, Kenneth Rose, Allen Gersho
Signal Compression Lab
Department of Electrical and Computer Engineering
University of California, Santa Barbara
[ashish,vladimir,rose,gersho]@laurel.ece.ucsb.edu

ABSTRACT

This paper introduces a new algorithm for scalable coding of wideband audio signals. The technique is based on quantization of bi-orthogonal wavelet transformed coefficients using a perceptual zerotree method. An initial zerotree estimate of the wavelet coefficients is computed, followed by scalar quantization of the coefficients according to perceptual thresholds. The choice of wavelet decomposition and encoding parameters for each frame is adapted to the source characteristics employing a rate distortion criterion. The scalability of the coder is due to the tree structure, which enables graceful degradation with decrease in bit rate. Preliminary subjective tests indicate near-transparent quality for average bit rates in the range of 1.5 to 2.5 bits per sample.

1. INTRODUCTION

Low bit rate, scalable coding of 8 kHz bandwidth audio signals is required in a growing number of applications, including audio over IP and wireless audio transmission. Existing audio coding algorithms use either linear predictive coding (LPC) or the discrete cosine transform (DCT) as their central module, coupled with quantization based on perceptual masking techniques. The former method tends to perform better with speech signals, while the latter offers better music performance. However, owing to the nonadaptive nature of the transform, the compression performance relies heavily on the stationarity of the signal. Furthermore, psychoacoustic studies [4][7] point out several important time-frequency localization properties of the auditory system, as well as the fact that the human ear performs a nontrivial "location-frequency" transformation. Wavelet packets (WP)[2][3] provide a computationally viable framework for exploiting this time-frequency tradeoff, along with flexibility to adapt to signal nonstationarities. These observations motivate investigation of wavelet decomposition as a tool for audio coding. (e.g., [3].)

The starting point of this work is the zerotree (ZT) coding method [5] and in particular, a variant which was proposed by Said and Pearlman [6]. This method exploits well the statistical properties (and hierarchical correlations) of the wavelet coefficients when applied to images. There are three main difficulties in adopting ZT for audio coding: (1) the standard statistical assumptions on the hierarchical correlations of image wavelet coefficients may not be valid for audio signals, (2) the method fails to exploit the known properties of the human ear and, (3) the method fails to take into account the dynamic range of audio signals.

This paper describes a novel algorithm for scalable coding of wideband audio. The algorithm incorporates perceptual considerations into an adaptive rate-distortion (RD) framework for ZT coding of a wavelet-decomposed signal. The transform and coding steps are both (frame) adaptive, where the adaptation involves RD optimization. The resulting coder is hence called the perceptual zerotree wavelet (PZW) coder. The paper is organized as follows: Section 2 describes the coder and details the coding modes. Section 3 provides preliminary experimental results. Section 4 consists of a summary and discussion of directions for future work.

2. PERCEPTUAL ZEROTREE CODING

Figure 1 shows the functional block diagram of the PZW audio coder. The encoder operates on a frame by frame basis. A frame of audio, ***s***, undergoes critically sampled forward wavelet

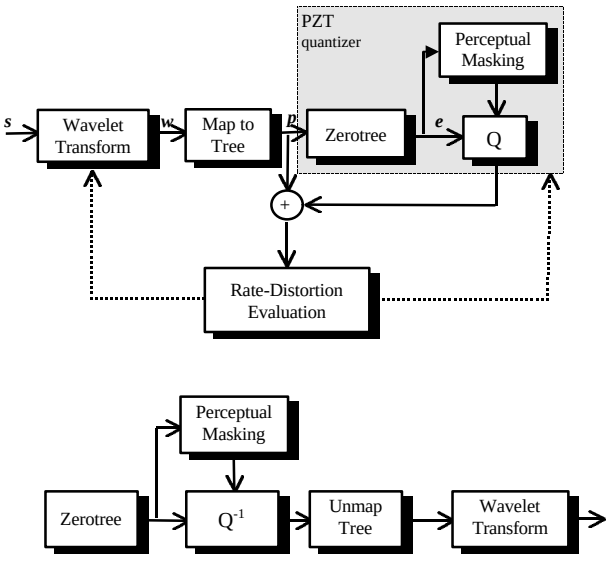

Figure 1. Block diagram for the PZW encoder (top) and decoder (bottom).

transformation (FWT) to produce a vector of wavelet coefficients, ***w***. These coefficients are then mapped to the nodes of a binary tree, denoted by ***p***, and quantized using a perceptual zerotree

This work is supported in part by the NSF under grant MIP-9707764, the University of California MICRO program, ACT Networks Inc., Cisco Systems Inc., Conexant Systems, Dialogic Corp., DSP Group Inc., Fujitsu Laboratories of America Inc., General Electric Corp., Hughes Network Systems, Intel Corp., Lernout & Hauspie Speech Products NV, Lucent Technologies Inc., Nokia Mobile Phones, Panasonic Speech Technology Lab., Qualcomm Inc., Sun Microsystems Inc., and Texas Instruments Inc.

(PZT) coding algorithm. The quantization and decomposition parameters are controlled by an RD module, which adapts the operation to the prescribed level of bit rate or distortion.

The decoder reconstructs $\boldsymbol{p}$, maps the tree back to a standard wavelet coefficient array, and performs the inverse wavelet transform (IWT) to obtain the reconstructed output of $\boldsymbol{s}$. Note that FWT and IWT are lossless, and the reconstruction error is only due to PZT quantization

2.1 The Wavelet Transform

Figure 2 shows the implementation of the wavelet transform using subband filtering, wherein the signal undergoes successive filtering by a bank of quadrature mirror filters (QMF) and down-sampling by a factor of two. The QMF filter bank we used in simulations is the bi-orthogonal 9-7 tap Daubechies type [1]. We further implemented the "symmetric extension" technique at frame boundaries.

We employ a signal-adaptive decomposition that differs from the traditional (logarithmic) wavelet decomposition. The decomposition is made adaptive by varying the decomposition depth and allowing a split at any one of the two terminating nodes. The decomposition is completely characterized by two parameters; the depth of the decomposition and a binary decision at every depth indicating which of the two terminating nodes is split further. Figure 2 shows examples of some possible

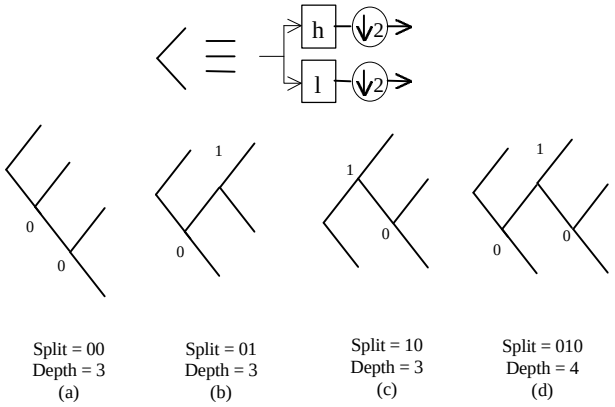

Figure 2. Examples of possible wavelet decompositions.

decompositions, with 2(a) being the traditional decomposition.

2.2 Binary Tree Formation

If we associate a wavelet coefficient with a position in the time-frequency plane containing the most energy of its basis function, we can define a binary tree in the plane. Figure 3(a) shows a mapping for a frame of 16 samples, for the decomposition given in 3(b). Except for the root node, each node grows two branches, which partition into two parts the temporal region of the parent node at a different frequency band. Figure 3(a) also shows the labeling convention used for the tree. The wavelet coefficients of level l in the tree are stored in vector $\boldsymbol{p}$ as the elements p_i where the index i satisfies $2^l-1 < i < 2^{l+1}$.

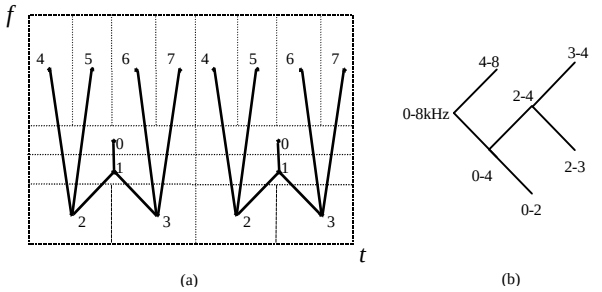

Figure 3. Time-Frequency map and binary tree formation using wavelet transform

2.3 Perceptual Zerotree (PZT) Quantization

After organizing the wavelet coefficients into a binary tree structure, we quantize them using a PZT coding method. The underlying technique is based on the ZT algorithm [5][6] which aims to capture the nonlinear statistical correlation of wavelet coefficients in the time-frequency plane. ZT quantizes the coefficients by transmitting a significance map (SM) using a variable length prefix code. The SM indicates the location of all coefficients that are above a threshold. This procedure is iterated a prescribed (fixed) number of times, each time reducing the threshold by a factor of 2. In this way ZT achieves quantization by transmitting the binary representation of the coefficient.

PZT however, transmits only the most significant bit of the coefficient. We let the iterations run as in the ZT, but once the first threshold is known for which the coefficient becomes significant, no more information about that coefficient is transmitted. Using this information, PZT generates an initial estimate of the coefficient and quantizes it using perceptual considerations. Quantization based on perceptual considerations assigns different error and bit allocation for the wavelet coefficients depending on their position in the frequency scale, the spectrum of the signal, and the initial estimates. The starting threshold of the PZT iteration is adapted to the frame energy thereby accommodating a wide range of signal levels.

The initial estimate, e_i, of the coefficient, p_i, is set equal to the threshold value, given by:

$$\boldsymbol{e_i} = 2^{\boldsymbol{n_i}} \text{ where } \boldsymbol{n_i} = \lfloor \log_2 |\boldsymbol{p_i}| \rfloor \ \forall \boldsymbol{i}$$

Once the initial estimate of a wavelet coefficient e_i is known, the bit allocation, b_i, required to scalar quantize a coefficient according to a simple perceptual model is given by:

$$\boldsymbol{b_i} = \lceil \log_2 |\boldsymbol{K e_i} / \boldsymbol{E_l}| \rceil \text{ where } \boldsymbol{l} = \lfloor \log_2 \boldsymbol{i} \rfloor$$

$$\boldsymbol{E_l} = \frac{\sum_{i=2^l}^{2^{l+1}} \boldsymbol{e_i}^2}{2^l} \text{ where } \boldsymbol{E_l} \text{ is the average energy of level } \boldsymbol{l}$$

The model controls the amount of noise introduced in different frequency bands by quantizing each coefficient to a fixed average SNR, given by the constant K.

2.4 Rate-distortion (RD) tradeoff control

A RD module (see Figure 1) controls the encoding operation. This module adapts the encoding parameters on a frame-by-frame basis to ensure that the available rate is efficiently used. We adapt several coding parameters (as given below) to the current frame statistics, and thereby improve the rate allocation to the frames by achieving better consistency in the level of distortion from frame to frame. The adaptation is based on an operational RD tradeoff, the distortion criterion being the segmental signal-to-noise ratio (SSNR). The RD module adapts the following encoding parameters per frame:

- Number of iterations (threshold steps) for ZT encoding
- Depth of wavelet decomposition
- Binary split decision during wavelet decomposition.

The set of permissible parameter values is restricted to maintain moderate complexity.

2.5 Scalability

ZT transmits the binary representation of a coefficient using a progressive transmission method making it a bit-wise scalable technique. The only bottleneck in making PZT scalable is the transmission of the bit allocation information b_i, which is calculated from the average energy for each level, E_l, after obtaining the complete initial estimate of wavelet coefficients. However, transmitting E_l at beginning of each frame easily eliminates this problem. Further, the bits required for E_l ($\approx$ 0.1 bps) can be saved by transmitting E_l in lieu of one of the wavelet coefficient in that level.

For each frame, the scalable bit stream transmitted to the decoder includes the following: (1) depth of wavelet decomposition, (2) binary split decision, (3) number of iterations, (4) average energy in each level of the tree E_l and, (5) for each iteration, initial estimates and the quantized coefficients.

The decoder can stop after any iteration and still reconstruct the wavelet coefficients, thus making PZW a highly scalable coding scheme. Preliminary subjective tests show a very graceful degradation in quality as the number of iterations is reduced.

3. RESULTS

Subjective evaluation of different coding modes was done using A/B tests on a database of 8 kHz bandwidth music and speech samples. The database included male and female speech, opera, modern rock and trumpet samples. The test was done with 8 untrained listeners and forced-choice A-B comparisons. The frame size in all the experiments was kept to a constant value of 512 samples.

Three coding methods were evaluated: ZT, PZW and G.722. The ZT method used a fixed transform (depth=9, split=0), a fixed number iterations (=6) and no perceptual considerations. The PZW algorithm was run with a fixed target SSNR, adaptive wavelet decomposition and PZT quantization as described in section 2. Tables 1 and 2 give the result of the A/B tests. The bit rate indicated is the average rate obtained over the entire database. For PZW and ZT, the variation in bit rates for individual files was in the range of 24-40kbps.

Test	**PZW @ 30 kbps**	**ZT @ 35 kbps**
Music	0.58	0.42
Speech	0.55	0.45
Total	**0.56**	**0.44**

Table 1. A/B test result for ZT and PZW

Test	**PZW @ 30 kbps**	**G.722 @ 48kbps**
Music	0.70	0.30
Speech	0.66	0.33
Total	**0.69**	**0.31**

Table 2. A/B test result for PZW and G.722

Table 1 shows that the PZW approach outperforms the ZT, 56% to 44%, despite the lower average rate used for the PZW (30 kb/s compared to 35 kb/s for the ZT). When compared to G.722 at 48kbps, the PZW was preferred 70% to 30% despite the lower average bit rate of 30kbps.

4. CONCLUSION

In this paper we presented a scalable method for coding 8 kHz bandwidth audio using wavelet decomposition and perceptual zerotree coding. The coder is successful in exploiting the time-frequency domain properties and a nonlinear frequency mapping of wavelet decomposition using the zerotree coding. Perceptual considerations and a rate-distortion framework were incorporated in the coder. Initial experiments show an average bitrate of 1.5 to 2.5 bits per sample for near-transparent quality. Better masking models and a more suitable distortion measure are expected to yield better results.

REFERENCES

[1] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, "Image Coding Using Wavelet Transform," *IEEE Transactions on Image Processing*, vol. 1, pp. 205-220, April 1992.

[2] K. Ramchandran and M. Vetterli, "Best Wavelet Packet Bases in a Rate-Distortion Sense," *IEEE Transactions on Image Processing*, vol. 2, pp. 160-175, April 1993.

[3] D. Sinha and A. Tewfik, "Low Bit Rate Transparent Audio Compression using Adapted Wavelets," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3463-3479, December 1993

[4] J. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE Journal on Selected Areas in Communication*, vol. 6, pp. 314-323, February 1988.

[5] J. Shapiro, "Embedded Image Coding using Zerotrees of Wavelet Coefficients," *IEEE Trans. On Signal Proc.* vol. 41, pp. 3445-3462, Dec., 1993.

[6] A. Said and W. Pearlman, "Image compression using the spatial-orientation tree," *IEEE Int. Symposium on Circuits and Systems*, pp. 279-282, May 1993.

[7] H. Fletcher, "Auditory Patterns," Rev. Modern Physics, vol. 12, pp. 47-65, 1940.