HIGH QUALITY SCALABLE STEREO AUDIO CODING

Zhitao Lu and William A. Pearlman

Electrical Computer and System Engineering Department
Rensselaer Polytechnic Institute, Troy NY 12180
luz2, pearlman@rpi.edu

ABSTRACT

This paper proposes an efficient, low complexity, scalable audio coder based on a combination of two embedded coding algorithms: the SPIHT (set partitioning in hierarchical trees) coding algorithm [1] and an embedded, nested binary set partitioning (NBSP) algorithm. The SPIHT algorithm, considered to be the premier state-of-the-art algorithm in still image compression, is used for the low frequency subbands in a wavelet packet audio signal decomposition, while the NBSP algorithm encodes the high frequency audio subbands. Both left and right channels are encoded together to form a single embedded stereo audio bitstream, that can be truncated at any point to produce an optimal lower rate and quality bitstream for delivery to lower quality user services. Using standard MPEG test materials, we evaluate the performance of the proposed encoder compared to the MPEG II standard audio coder through informal listening tests at bit rates of 48Kbs/sec and 64Kbs/sec per channel. We conclude that our coder is comparable with MPEG II at 48Kbs/sec and better at 64 Kbs/sec per channel. The algorithm also features exact bit rate control, progressive transmission and low complexity for both the encoder and decoder. These features show its potential for interactive audio transmission over networks.

1. INTRODUCTION

Applications for transmitting multimedia information over networks is increasing rapidly. Since the volume of multimedia signals containing audio and video is large, the available bandwidth of the network is also limited and changes dynamically. It requires that the compression technique achieves high compression ratio, progressive transmission and low complexity in both the encoder and decoder. Wavelet and wavelet packet transform have both good time and frequency localization. This characteristic is very similar to the characteristic of the human hearing system, so it is widly used in audio coding [3] [4] [5]. In [3] and [5] an optimal wavelet packet transform is used to represent the audio signal and almost transparent quality is achieved at bitrates of 48 to 66 kbits/sec. In [4] where the audio signal is decomposed into harmonic components and noise-like residue, the bitrate for transparent quality is claimed to be as low as 44kbits/sec. But these algorithms do not achieve progressive transmission and the complexity in the encoder side is very high.

In this paper, we propose a new embedded, wavelet stereo audio coder that utilizes the SPIHT (set partitioning in hierarchical trees) image compression algorithm in a one-dimensional mode for the low frequency subbands and an embedded, nested binary set partitioning (NBSP) algorithm for the high frequency subbands. The proposed coder is an improved version of the monaural coder presented in [2]. The diagram of proposed encoder is shown as in figure 1. It integrates the psychoacoustic model, dynamic bit rate allocation and the coding process of the channels together. The paper is organized as follows: Section 2 describes briefly the psychoacoustic model and bit rate allocation. Section 3 presents the coding algorithms based on 1-D SPIHT and embedded NBSP. The tests of the SPIHT-NBSP coder compared to the MPEG II standard audio coder with the standard MPEG test clips and their results are presented in Section 4. Section 5 states the conclusions of the paper.

2. WAVELET PACKET TRANSFORM AND BIT RATE ALLOCATION

We use a fixed wavelet packet transform, according to the wavelet tree proposed in [3]. There are 29 subbands which mimic the critical subbands of the human hearing system. The lowpass and highpass filters are the length-20 Daubechies filter pair[6], which provide sufficiently good bandpass and bandstop characteristics. We distinguish subbands into two groups: low frequency subbands and high frequency subbands. The low frequency subband group comprises subbands 0-16, corresponding to the frequency range 0-3.4 KHz. The rest of the subbands, numbered 17-29, belong to the high frequency group.

The psychoacoustic model is very important for audio coding. It represents the perceptual characteristics of the human hearing system and determines the inaudible noise en-

[1]This work is based on material supported by the National Science Foundation Industry/University Co-operative Research Program under Grant No. EEC-9812706 and the Center for Digital Video and Media Research. The government has certain rights in this material

ergy level in each subband. The MPEG psychoacoustic model II is used here to calculate the signal-mask-ratio SMR_m representing the masking effect for each subband m.

We design a uniform quantizer for each subband, where the quantizer step size is determined through the bit rate allocation process. We use a Laplacian rate-distortion model, so that each bit rate corresponds to a certain distortion and quantizer step. Each time, we allocate the available bits to the subband having the highest noise-to-mask ratio until the bit budget is exhausted. Instead of explicitly quantizing the coefficients, the calculated quantizer step is integrated into the coding process, as will be explained in the next section.

3. CODING ALGORITHM

In typical audio signals, most of the energy is concentrated in the low frequency subbands. In these subbands the mutual correlation is expected to be strong. It has also been observed that there is a temporal self-similarity among different subbands analogous to the spatial self-similarity trees in the 2D wavelet transform of an image in [1]. The coefficients are expected to be decreasing in magnitude as we move down toward the leaves of the similarity tree. The coding algorithm proposed for these subbands is therefore based on the SPIHT coding algorithm using a binary similarity tree instead of the quaternary spatial-orientation tree (quadtree) for two dimensions. The algorithm consists of repeated sorting passes and refinement passes. In sorting pass, the sets are groups of coefficient nodes in a subtree designated by the subtree root and are partitioned according to specific rules. The same sets and partitioning rules are defined in the encoder and decoder. The subset of subband coefficients c_{ij} in $\mathcal{T}$ is said to be significant for bit depth n if

$$\max_{i \in \mathcal{T}} \{|c_{ij}|\} \geq 2^n$$

(where j represents left or right channel), otherwise it is said to be insignificant. If the subset is insignificant, a 0 is sent to the decoder. If it is significant, a 1 is sent to the decoder and then the subset is further split according to the set partitioning rule until all the significant sets are a single significant point. In SPIHT, the subset is partitioned into the two child coefficients (nodes) and the subtrees rooted at these nodes. After the sorting pass, the indices of the coefficients are put into three lists, the list of insignificant points (LIP), the list of insignificant sets (LIS), and the list of significant points (LSP). Only bits related to the lower bit planes of the LSP entries (refinement pass) and binary outcomes of the magnitude tests are transmitted to the decoder. The LIP and LIS are tested again at lower thresholds . Once a set or pixel becomes significant, it is not tested again at a lower threshold.

Compared with the original SPIHT algorithm, we make some modification to accomodate to signal characteristics of the wavelet packet transform of wideband audio signals. Further details about the original SPIHT algorithm and modified 1-D SPIHT algorithm can be found in [1] and [2]. We explain the modification briefly in the following subsections.

3.1. Parent-Children Relation in Similarity Tree

In the original SPIHT algorithm, the transform used is the conventional dyadic wavelet transform, where only lowpass subband is further split. The parent-children relation is shown as in figure 2 (a). For a wavelet packet transform, the highpass subband is also further split. The corresponding parent-children relation is shown as in figure 2 (b). In both cases, each node in the figure representing parent and children corresponds to different subband components of the same temporal period of signal. Each node represents two coefficients, for left and right channel, respectively.

3.2. Set Partitioning rule

The coefficients are set partitioned by the similarity tree analogous to the original SPIHT algorithm which is shown as (a) in figure 3. The whole subset corresponding to a similarity tree is tested. If the subset is significant, the children c_1 and c_2 and descendant set d_1 and d_2 are further tested. The significant subset is further split along the similarity tree until all the significant points are found.

In the high frequency subbands, the correlation among the coefficients in different subbands is not strong and there is likely to be some coefficients of higher magnitudes. This means that the SPIHT algorithm, if used for these subbands, would have to partition the tree deeply at almost every threshold during the sorting pass, spending too many bits in the process. Therefore we truncate the similarity trees and set the leaves before reaching these higher frequency subbands. We employ another procedure for coding these subbands, called nested binary set partitioning (NBSP). The set of coefficients within each highpass subband is recursively partitioned into two equal subsets as shown in (b) of figure 3. In the stage 1, the whole set is tested. If the set is found significant, it is further partitioned into two subsets as in stage 2. Both of them are tested and the significant subset is split again. This process is repeated until all the significant points are found and put into the LSP. The insignificant subsets discovered in the recursive binary splitting process are put into the LIS and are tested again at lower thresholds. As in SPIHT, a significant subset or point is never tested again at a lower threshold.

3.3. Integrated Quantization

The SPIHT and NBSP algorithms employ progressive bit plane transmission. The threshold used in a test of subset significance is reduced by half after each sorting and refinement

pass. From the psychoacoustic model and bit rate allocation, we obtain the step size of the uniform quantizer for each subband. Instead of quantizing the subband coefficients explicitly using the determined step sizes, we integrate it into the coding process. When the threshold becomes less than the quantization step of a certain subband, we move the coefficients in this subband from the LSP to the LIP in both the encoder and decoder. So no more information bits for such subband components are transmitted to the decoder. Experiments show that this so-called implicit quantization procedure is more efficient than explicit quantization.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed codec, we encode and decode the MPEG audio test clips and then compare its performance to that of the MPEG II. The signals used are bass47_1, quar48_1, harp40_1, trpt23_2 and vioo10_2, which are opera, song and instrumental music. Since objective measures, such as signal-to-noise ratio (SNR) between the original and reconstructed signal are not perceptually meaningful for audio coding evaluation, we conducted some informal listening tests in a quiet environment with high quality headphones driven by a DELL multimedia desktop system. We encoded and decoded the audio test signals at both 48 Kbs/sec and 64 Kbs/sec per channel with the MPEG II coder and the proposed SPIHT-NBSP coder. The lower rate bitstream for the latter coder was obtained by truncation of the higher rate bit stream. The results are summarized in Table 1. Generally, our embedded SPIHT-NBSP coder is comparable or better than MPEG II at 64 kbps and comparable at 48 kbps. The quality of both signals is very high at these rates, practically transparent at the higher rate, with the differences being very subtle. SPIHT-NBSP has better fidelity to the original than MPEG II, but at the lower bit rate exhibits some low level aliasing noise absent in MPEG II. Although the MPEG II signal shows some defects not present in the SPIHT-NBSP signal at the lower rate, these defects are less bothersome and noticeable than the aliasing noise. Reducing the rate further can eliminate this noise and create a more pleasing listening experience, because the coefficients producing this noise then become lower than the threshold. We believe we can greatly reduce or eliminate this noise with sharper high frequency filters. Further investigations will be conducted to verify this supposition.

(Note: the aliasing noise mentioned above has been eliminated by use of biorthogonal (10, 18) filters with symmetric extension. The table below therefore is to be modified. Generally, perceptual results are comparable to MPEG II, with some cases of SPIHT-NBSP being slightly better.)

Table 1: SPIHT-NBSP compared to MPEG II in Listening Tests

SIGNAL	48Kbs/s/ch	64Kbs/s/ch
bass47_1	comparable	comparable
quar48_1	comparable	better
harp40_1	better	better
trpt21_2	little worse	comparable
vioo10_2	little worse	comparable

5. CONCLUSION

In this paper, we investigated the performance of an audio coder based on the SPIHT and NBSP embedded coding algorithms. The experiment shows that this coder achieves high coding efficiency. It also has features of exact bit rate control and progressive stereo transmission. So the user can get a different quality reconstructed signal by decoding part of the bit stream according to the available bandwidth. The coding procedure is mainly implemented by comparison and bit operations, so it has low complexity in both the encoder and decoder. We believe that these features make the SPIHT-NBSP coder to be a good candidate for transmission and delivery of compressed stereo and multi-channel audio signals over networks.

6. REFERENCES

[1] A. Said and W. A. Pearlman, "A New, Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees,", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 6 No. 3, pp. 243-250, June 1996.

[2] Zhitao Lu and W. A. Pearlman, "An Efficient, Low-Complexity Audio Coder Delivering Multiple Levels of Quality for Interactive Applications", *Proceedings of 1998 IEEE Second Workshop on Multimedia Signal Processing*, Dec. 7-9,1998 Redondo Beach, CA, pp. 529–534.

[3] D. Sinha and A. H. Tewfik, "Low Bit Rate Transparent Audio Compression Using Adapted Wavelets," *IEEE Trans. on Signal Processing*, Vol 41, No. 12, pp. 3463-3479 Dec. 1993.

[4] K. N. Hamdy, M. Ali and A. H. Tewfik, "Low Bit Rate High Quality Audio Coding with Combined Harmonic and Wavelet Representations,"*Proc. IEEE Intern. Conf. Acoust., Speech, and Sig. Processing (ICASSP)* , , Vol. 2, pp. 1045-1048, 1996.

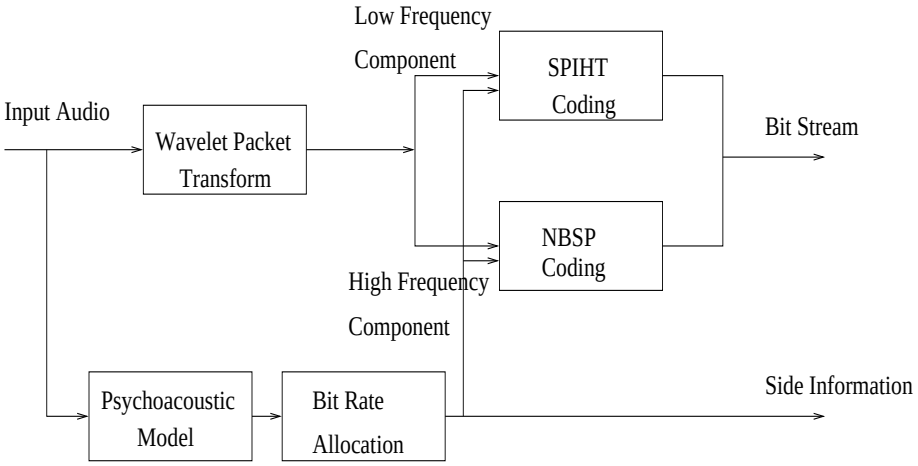

Figure 1: The Diagram of Proposed Audio Coder

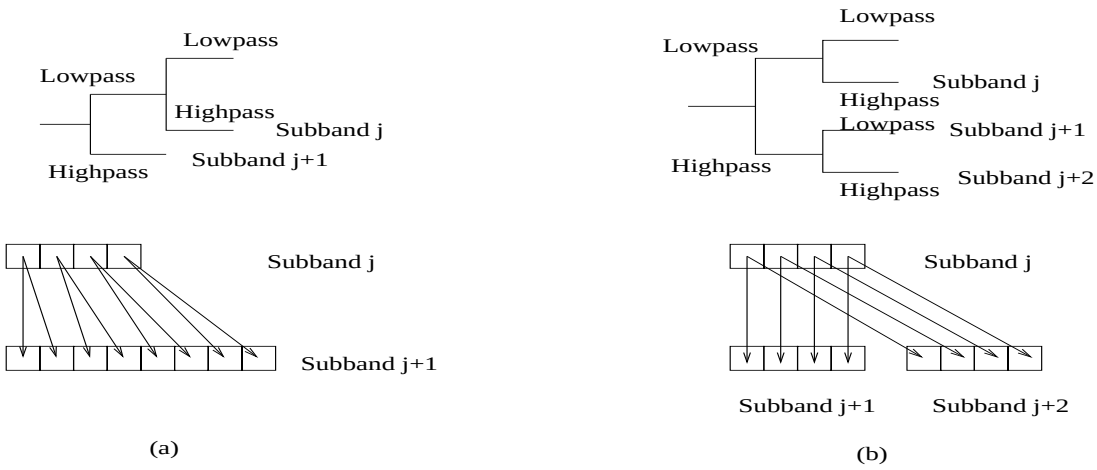

Figure 2: The Parent-Children relation within Similarity Tree

[5] M. Purat and P. Noll, “Audio Coding with a Dynamic Wavelet Packet Decomposition Based on Frequency-Varying Modulated Lapped Transforms,”*Proc. IEEE Intern. Conf. Acoust., Speech, and Sig. Processing (ICASSP)*, Vol. 2, pp. 1021-1024, 1996.

[6] I. Daubechies, “Orthonormal Bases of Compactly Supported Wavelets,” *Commun. Pure Appl. Math.*, vol 41. pp. 909-996, Nov. 1988.

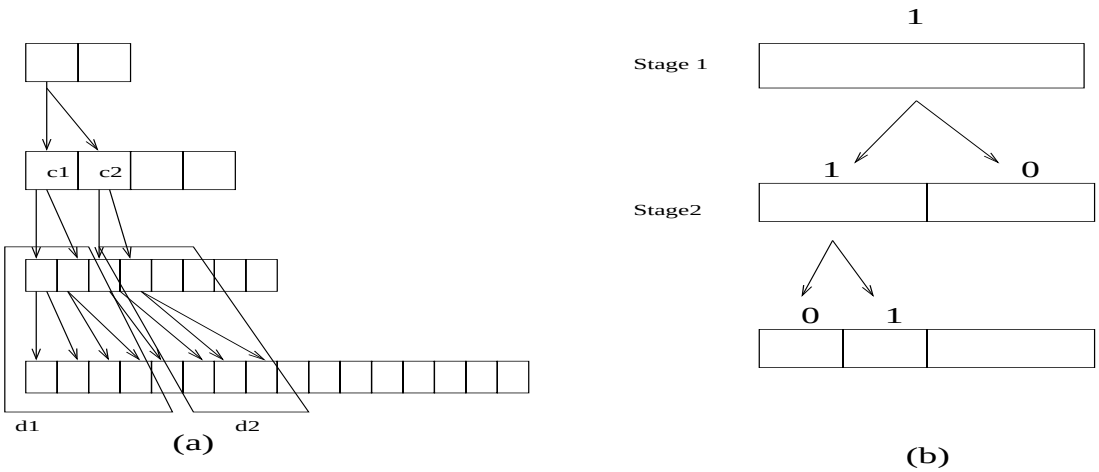

Figure 3: Set Partitioning by Similarity Tree and Nested Binary Set Partitioning