An Efficient, Low-Complexity Audio Coder Delivering Multiple Levels of Quality for Interactive Applications

Zhitao Lu and *William A. Pearlman*

Electrical,Computer and Systems Engineering Department
Rensselaer Polytechnic Institute
Troy, NY 12180

July 6, 2000

Abstract

This paper proposes an efficient, rate-scalable, low complexity audio coder based on the SPIHT (set partitioning in hierarchical trees) coding algorithm [5], which has achieved notable success in still image coding. A wavelet packet transform is used to decompose the audio signal into 29 frequency subbands corresponding roughly to the critical subbands of the human auditory system. A psychoacustic model ,which, for simplicity, is based on MPEG model I, is used to calculate the signal to mask ratio, and then calculate the bit rate allocation among subbands. We distinguish the subbands into two groups: the low frequency group which contains the first 17 subbands corresponding to 0-3.4 KHz, and the high frequency group which contains the remaining high frequency subbands. The SPIHT algorithm is used to encoder and decode the low frequency group and a reverse sorting process plus arithmetic coding algorithm is used to encode and decode the high frequency group. The experiment shows that this coder yields nearly transparent quality at bit rates 55-66 Kbits/second, and and degrades only gradually at lower rates. The low complexity of this coding system shows its potential for interactive applications with levels of quality from good to perceptually transparent.

1 Introduction

Source coding of wideband audio signals for storage and/or transmission application over bandlimited channels is currently a research topic receiving considerable attention. Its applications are in the fields of audio production, program distribution and exchange, digital audio broadcasting, digital storage, video conference and multimedia applications. The industrial standard for wideband audio signal with sampling rate at 44.1 KHz which covers the entire audible frequency range of the human hearing system, each sample is quantized into 16 bits, without compression, the bit rate will be 705.6 Kb/sec for one channel. The goal of audio data compression is to get the bit rate as low as possible without perceptible distortion.

Most proposed audio coders are transform coders or subband coders. They mainly include three parts: subband decomposition or transform, dynamic bit allocation and the coding algorithm. First the original audio data is transformed into subband signals; the target bit rate is dynamically allocated among the subbands through a psychoacustic model; and then each subband signal is encoded to a bit stream.

Wavelet packet decomposition is widely used, since by changing the time resolution and frequency resolution, it represents the audio signal more efficiently. The subbands are usually similar

to the critical subbands of the human hearing system. The algorithms proposed in [6] and [4] find the optimal basis for each data frame. In [3] the audio signal is decomposed into harmonic components and noise-like residue.
MPEG layer III also uses 6 points and 18 points DCT in subbands to get better high-frequency resolution.

For dynamic bit allocation and quantization, the audio data is usually input into a psychoacustic model to calculate the signal to mask ratio in each subband and to determine how many bits should be used for each coefficient in every subband [4]. In [6], a scalar quantizer is used and in [1] a hybrid quantizer–low frequency band using scalar quantizer and high frequency band using vector quantizer–is used to quantize the coefficients in each subband.

In this paper, we present the coder based on the set partitioning in hierarchical trees (SPIHT) coding algorithm [5] and reverse sorting process plus arithmetic coding algorithm. The SPIHT algorithm, first used for still image compression, combines quantizer and coder together, and exploits the dependence of the coefficients in different subbands by setting up similarity trees. Using a wavelet packet transform, dynamic bit allocation, and the SPIHT coding algorithm, which also features fast encoding and decoding, this coder achieves nearly transparent quality at 55-66Kbits/second. The system is also capable of delivering lower rate service from the same bitstream. The degradation of quality at lower bit rates appears to be quite gradual, making the system attractive for delivery of different qualities of service requiring near real-time to real-time execution, such as interactive applications.

2 Wavelet Packet Transform and Dynamic Bit Allocation

We use the tree structure of filter banks proposed in [6], where 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 [2], which provide sufficiently good bandpass characteristics.

In parallel with execution of the wavelet packet transform (WPT), the audio source data is fed into a psychoacoustic model, which is based on the ISO/MPEG model 1. After calculating the signal-to-mask ratio in each subband SMR_m and combining with the coefficients in each subband, we do the dynamic bit allocation [7]. The rate-distortion relation is modeled as:

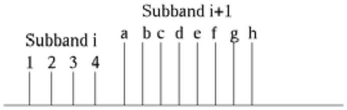
$$d_m(r_m) = w_m g 2^{-2r_m}$$

where:
d_m is distortion in each subband
r_m is bit rate in each subband,
w_m is the subband weight calculated by SMR_m,
g is a constant.

For each subband with n_m samples, the signal-to-noise ratio, $SNR_m > SMR_m$, is satisfied in order to get transparent quality coding. The whole problem of allocating bit rate to achieve the above constraint is a constrained optimization problem, formally solved by use of the Kuhn-Tucker theorem. However, the problem can be solved iteratively by calculating the bit rate for each subband by the formula:

$$r_m = \begin{cases} \frac{1}{2}\log_2[\frac{w_m \sigma_m^2}{\theta n_m}] & , w_m \sigma_m^2 / n_m > \theta \\ 0 & , w_m \sigma_m^2 / n_m \leq \theta \end{cases}$$

where:
n_m is the number of samples in m'th subband.



When the depth of subband i+1 in the wavelet packet tree $l_{i+1} = l_i + 1$, then
a,b will be the offspring of 1.
c,d will be the offspring of 2.
e,f will be the offspring of 3.
g,h will be the offspring of 4.

Figure 1: Figure 3. Definition of The Similarity Tree

σ_m^2 is the variance of the coefficients in m'th subband.
The parameter θ is continually adjusted until the overall target bit rate is met with nonnegative r_m's. After we get the bit rate for each subband, we determine the final minimum quantization intervals or stepsizes for each subband.

3 Coding Algorithm

In this paper, after getting the quantizer stepsize for each subband, we seperate all the subbands into two groups – the low frequency group which contains the first 17 subbands (frequency range is 0-3.4KHz) and the high frequency group which contains other higher frequency subbands. The SPIHT algorithm is used to encode the coefficients in low the frequency group. The coefficients in high frequency group are quantized by the uniform quantizer with stepsizes determined in the rate allocation procedure, and then encoded losslessly by the reverse sorting process plus arithmetic coding.

3.1 SPIHT Algorithm

The principles of the SPIHT algorithm are partial ordering of the transform coefficients by magnitude with a set partitioning sorting algorithm, ordered bit plane transmission and exploitation of self-similarity accross different scales of an audio wavelet packet transform

3.1.1 Similarity Trees

In a typical audio signal, most of the energy is concentrated in low frequency bands, but there are basic frequency and harmonic components in different subbands. It has been observed that there is a temporal self-similarity among different subbands analogous to the spatial self-similarity trees in the 2D wavelet tramsform of an image. The coefficients are expected to be better magnitude-ordered if we move down following the same similarity tree. We define the similarity tree in Figure 3 , so that every point in subband i corresponds to $2^{l_{i+1}-l_i}$ points in the next subband $i+1$, where l_i is the depth of the ith subband in the filter bank tree in Figure 1. This definition is analogous to that of spatial orientation trees in image coding with SPIHT. By defining the similarity trees, we organize the coefficients into subsets as follows:
O(i)— set of all offsprings of point i.
D(i)— set of all descendants of point i.

H(i)— set of all roots of similarity trees.
L(i)— D(i)-O(i)

Each point i and its descendant and offspring sets represent the same time period in different frequency subbands.

3.1.2 Set Partitioning Sorting Algorithm

The same set partitioning rule is defined in the encoder and decoder. The subset T_m is determined if it is significant or insignificant by the magnitude test

$$\max_{i \in T_m} \{|c_i|\} > 2^n$$

on the subband coefficients c_i in T_m. 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 similarity tree until all the significant sets are a single significant point. After this sorting, the indices of the coefficients are put onto 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 LSP entries and binary outcomes of the magnitude tests are transmitted to the decoder. The partitioning proceeds as follows in this so-called sorting pass:

1) The initial partition is formed with point i and its
   descendents D(i).
2) if D(i) is significant, then it is partitioned into L(i)
   and the offspring of i, O(i).
3) if L(i) is significant, then it is partitioned into
   D(k), for every k in O(i).

3.1.3 Coding Algorithm

After each sorting pass, we get the ordered significant points for the threshold 2^n, and then we send to the decoder the nth most significant bit of every coefficient found significant at a higher threshold. The binary representation of the magnitude-ordered coefficient is shown as Table 1, where the arrows indicate the order of the transmission. By transmitting the bit stream in this ordered bit plane fashion, we always transmit the most valuable remaining bits to decoder. The outline of the full coding algorithm is as follows:

1) Initialization.

Set the list of significant points (LSP) as empty. Set the roots of similarity trees in the insignificant points (LIP) and insignificant sets (LIS). Set the significance threshold 2^n with

$$n = \lfloor \log_2(max_{(i)}(|c_i|)) \rfloor$$

2) Sorting pass.

Using the set partitioning algorithm distribute the appropriate indices of the coefficients to the LIP, LIS,and LSP.

3) Refinement pass:
   For each entry in the LSP significant for higher n,
   send the nth most significant bit to the decoder.
4) decrement n by one and return to step 2
   until the specified bit rate is reached.

Table 1: Binary representation of the magnitude ordered coefficients

sign	s	s	s	s	s	s	s	s	s	s	s
5	1	1	0	0	0	0	0	0	0	0	0
4	-	→	1	1	0	0	0	0	0	0	0
3	-	-	-	→	1	1	1	1	0	0	0
2	-	-	-	-	-	-	-	→	1	1	1
1	-	-	-	-	-	-	-	-	-	-	→
0	-	-	-	-	-	-	-	-	-	-	→

3.2 Reverse Sorting Process and Arithmetic Coding

We use reverse sorting process and arithmetic coding algorithm to encode the quantized coefficients in the high frequency group.

After the uniform quantization, the quantized coefficients in high frequency group have little magnitude with high probability. This is the motivation to use reverse sorting. In this algorithm, we have a list of significant points whose magnitude is greater than the threshold. The sorting process is as follows:

1. Initialization : put all coefficients in high frequency group into significant set. Set the threshold as 1.
2. Find the maximum value of the coefficients in the significant set, and transmit it.
3. Pass through the whole significant set; for a coefficient whose magnitude is greater than threshold, send 1, otherwise, send 0, and remove it from the significant set.
4. Pass through the whole significant set; for a positive coefficient, send 1, otherwise send 0.
5. Increase the threshold by 1, and pass through the significant set; for a coefficient whose magnitude is greater than threshold, send 1, otherwise send 0 and remove it from significant set.
6. Repeat last step until the significant set is empty.

After we do the reverse sorting, Arithmetic encoding is used to encode the bitstream. This will improve the compression ratio by about 10-30 percents.

The decoder can reproduce the lists and sets produced at the encoder from the received maximum values and decision and sign bits, so is able to reconstruct the same quantized coefficients using the received refinement bits without additional overhead information.

4 The Presented Audio Coder

The diagram of the proposed audio coder is shown in Figure 4. The audio frame with frame size of 1024 is input into the psychoacustic model to calculate the SMR. In parallel with this, the signal is transformed into 29 subband signals. Then for each subband, according to the desired total bit rate, SMR, and coefficents, we calculate the stepsize of the quantizers used in the SPIHT coding algorithm and reverse sorting process. These stepsizes are also sent as side information to the decoder. The transform coefficients of low frequency group are encoded by SPIHT algorithm

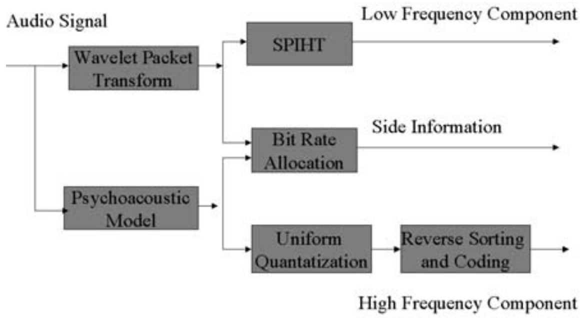


Figure 4: Diagram for SPIHT Audio Encoder

Table 2: Test Result of the Proposed Encoder

signal	total	spiht	arc	SNR(dB)
sting	1.28	.5	.8	13.08
fm	1.32	.5	.8	13.13
stp	1.39	.5	.9	10.90
pj	1.30	.6	.7	11.40
mozart	1.48	.5	1.	11.18
mahler	1.49	.5	1.	14.23

- total is total bitrate (bits/sample)
- spiht is bitrate used for low frequency group (bits/sample)
- arc is the bitrate used for high frequency group (bits/sample)
- SNR is (signal variance/mean square error) in dB.
- all of the reconstructed signals achieve nearly transparent quality at corresponding bitrate.

and high frequency group signal are encoded by reverse sorting process plus arithmetic coding algorithm.

5 Results

Several pieces of music are used to test the performance of this coder. Mozart and Mahler are classical music, FM and Sting are soft rock songs , PJ and Stp are rock songs. All of them are sampled at the rate of 44.1 KHz with 16 bits of precision. Informal subjective testing reveals that the coder can get nearly transparent quality at bit rates 55-66 Kbits/second. The results are shown in Table 2.

In this paper, we investigated the performance of an audio encoder based on SPIHT algorithm. The experiment shows that it is efficient and has low complexity in implementation. We will compare the performance of our encoder with the MPEG2 and MPEG4 standard encoders. We shall also present results for lower coding rates for non-transparent quality applications, where current indications are that the SPIHT audio coder performs very well. We believe this coder to a strong candidate for interactive applications requiring good to perceptually transparent quality.

References

[1] S. Boland, M. Deriche, "Audio Coding Using the Wavelet Packet Transform and a Combined Scalar-Vector Quantization,"*Proc. IEEE Intern. Conf. Acoust., Speech, and Sig. Processing (ICASSP)*, Vol. 2, pp. 1041-1044, 1996.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] T. R. Trinkaus, *Wideband Audio Compression Using Subband Coding and Entropy-Constrained Scalar Quantization*, Master Thesis, Rensselaer Polytechnic Institute, 1995.