Audio Compression using the MLT and SPIHT

Mohammed Raad, Alfred Mertins and Ian Burnett

School of Electrical, Computer and Telecommunications Engineering
University Of Wollongong
Northfields Ave Wollongong NSW 2522, Australia
email: mr10@uow.edu.au

Abstract

This paper discusses the application of the Set Partitioning In Hierarchical Trees (SPIHT) algorithm to the compression of audio signals. Simultaneous masking is used to reduce the number of coefficients required for the representation of the audio signal. The proposed scheme is based on the combination of the Modulated Lapped Transform (MLT) and SPIHT. Comparisons are also made with the Discrete Wavelet Transform (DWT) based scheme. Results presented reveal the compression achieved as well as the scalability of the proposed coding scheme. The MLT based scheme is shown to have compression performance that is superior to the DWT based scheme.

1   Introduction

The compression of audio signals refers to the reduction of the bandwidth required to transmit or store a digitized audio signal. The analogue audio signal is usually digitized using the Compact Disk (CD) standard of 44.1 kHz sampling rate and 16 bit PCM quantization [1]. A number of audio compression techniques are well known. MPEG standards [1] present several techniques of compressing audio signals, as do some commercial coders such as the Dolby AC series of coders [2]. The techniques presented by those standards and products are aimed at constant rate transmission, although MPEG has made some attempts at standardising scalable compression techniques [1][3].

A scalable audio compression technique would relate the quality obtained from the synthesized audio signal to the number of bits used to code the digital audio signal. At the same time acceptable audio quality must be obtained at the lowest rate. A scalable audio compression method would find application in packet based networks such as the Internet where variable bit rates are the norm.

The Set Partitioning In Hierarchical Trees (SPIHT) algorithm sorts the coefficients in terms of relative importance, determined by coefficient amplitude, and transmits the amplitudes partially, refining the transmitted coefficients continuously until the bit limit is reached [4]. The work presented in this paper combines SPIHT with the Modulated Lapped Transform (MLT) and compares the results to those obtained by using the DWT based scheme in [5]. The results presented show clearly the advantage of using the MLT instead of the wavelet transform with SPIHT.

2   Set Partitioning In Hierarchical Trees

The Set Partitioning In Hierarchical Trees algorithm (SPIHT) was introduced by Said and Pearlman [4]. The algorithm is built on the idea that spectral components with more energy content should be transmitted before other components, allowing the most relevant information to be transmitted using the limited bandwidth available. The algorithm sorts the available coefficients and transmits the sorted coefficients as well as the sorting information. The sorting information transmitted modifies a pre-defined order of coefficients. The algorithm tests available coefficients and sets of coefficients to determine if those coefficients are above a given threshold. The coefficients are thus deemed significant or insignificant relative to the current threshold. Significant coefficients are transmitted partially in several stages, bit plane by bit plane.

As SPIHT includes the sorting information as part of the partial transmission of the coefficients, an embedded bit stream is produced, where the most important information is transmitted first. This allows the partial reconstruction of the required coefficients from small sections of the bit stream produced.
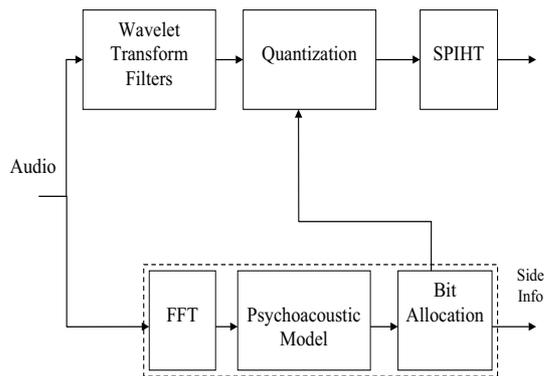
Figure 1: The wavelet based coding scheme

3 The compression schemes used

3.1 The use of wavelets with SPIHT

The wavelet transform has been combined with SPIHT in [5] to compress audio. The attractive property of the wavelet transform is the fact that the transform is implemented in a tree structure and so the sets (or trees) originally developed in [4] could still be used. The filter pairs used in [5] were the 20-length Daubechies filter pairs.The sets that are required for SPIHT can be developed as given in [6].

The scheme based on the wavelet transform is diagrammatically represented by Figure 1. In the scheme shown, the psycho acoustic model determines the bit allocation that should be used in the quantization of the wavelet coefficients. This requires side information to be transmitted. The results presented by Lu and Pearlman indicated that imperceptible distortion in the synthesized signal could be obtained at bit rates between 55-66 kbps [5].

As an indication of how SPIHT reduces the bits required, Table 1 lists initial results for the eight test signals used in this work coded using a maximum of 16 bits per coefficient. The test signals are Sound Quality Assessment Material (SQAM) signals obtained from [7]. The signal content of the files tested is also given in Table 1. The results given are in terms of average bit rates per frame and should be compared to 706 kbps which is the CD rate. Since this set of results is for complete reconstruction combined with bit allocation using the MPEG masking model, the sound quality of the synthesized files were the same as the original. The objective results given are the Segmental Signal to Noise Ratios (SegSNRs) of the synthesised signals.
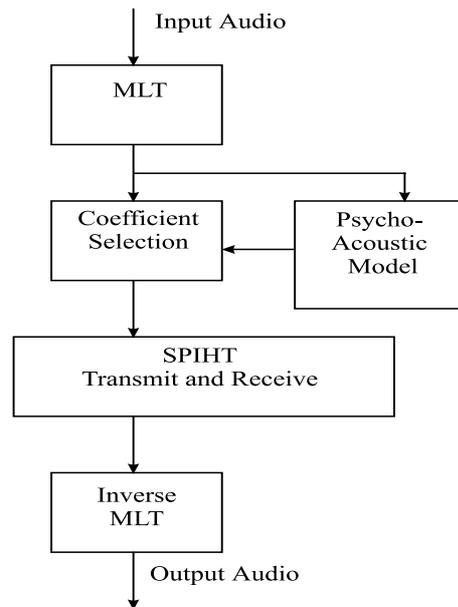

Figure 2: The codec used

The results presented in Table 1 are for complete reconstruction. It was found that the described DWT based scheme may be used to code the SQAM files at lower bit rates than those listed with good results. In fact at bit rates between 42 and 64 kbps, most of the synthesized audio had almost no perceivable distortion which is in agreement with the results presented in [5].

3.2 The MLT combined with SPIHT

The codec based on the combination of the MLT with SPIHT is shown in Figure 2. In Figure 2, the audio signal is divided into overlapping frames and the MLT is applied to each frame. The obtained coefficients are subjected to the Johnston psycho acoustic model [8] and any coefficients that are found to be below the masking threshold are set to zero before scalar quantization is carried out on all of the coefficients. The quantized coefficients are transmitted by the use of SPIHT. At the decoder, SPIHT is used to decode the bit stream received and the inverse transform is used to obtain the synthesized audio.

3.2.1 Setting up the SPIHT sets

In applying the MLT to an SPIHT based codec, the sets that were used for the wavelet based coding scheme no longer describe the relationship between the transform coefficients appropriately. In [4] sets are based on the tree structure organization of the coefficients, whereas the uniform M-band decomposition carried out by the MLT is a parallel operation.

Table 1: Coding Results using the Wavelet Transform.

Signal	Content	SegSNR (dB)	Mean Rate (kbps)
x1	Bass	46.1	167
x2	Electronic Tune	50.9	71
x3	Glockenspiel	46.6	180
x4	Glockenspiel	44.4	201
x5	Harpsichord	31.1	227
x6	Horn	48.0	94
x7	Quartet	43.2	174
x8	Soprano	43.7	162

There has been a reported work that used the tree structure based sets on a non-tree structured transform [9] in image compression with very good results. This indicates that as long as the trees define large sets of insignificant coefficients and small sets of significant coefficients, SPIHT will not use an excessive amount of bits to carry out the sorting.

In the following we define SPIHT sets that link together the frequency domain coefficients for a given frame. The roots of the used sets are at the low frequency end of the spectrum and the outer leaves are at the higher end of the spectrum. Thus, the sets link together coefficients in the frequency domain in an order that fits the expectation that the lower frequency coefficients should contain more energy than the higher frequency coefficients. This ordering is similar to, although not the same as, the sets defined in [4].

In this implementation the sets are developed by assuming that there are N roots. One of the roots is the DC-coefficient and because it is not related to any of the other coefficients in terms of multiples of frequency, it is not given any offspring. Each of the remaining $N - 1$ roots are assigned N offsprings. In the next step each of the offsprings is assigned N offsprings and so on, until the number of the available coefficients is exhausted. The offsprings of any node (i) where (i) varies between 1 and $M - 1$ (M is the total number of coefficients and $i = 0$ is the DC coefficient), are defined as

$$O(i) = iN + \{0, N - 1\}. \tag{1}$$

Any offspring above $M - 1$ are ignored. The descendants of the roots are obtained by linking the offsprings together. For example, if $N = 4$, node number 1 will have offsprings $\{4,5,6,7\}$, node 4 will have offsprings $\{16,17,18,19\}$ and the descendants of node 1 will include $\{4, 5, 6, 7, 16, 17, 18, 19,...\}$.

As part of the development of the M-band transform plus SPIHT coding system, a number of experiments were conducted to determine if the size of N
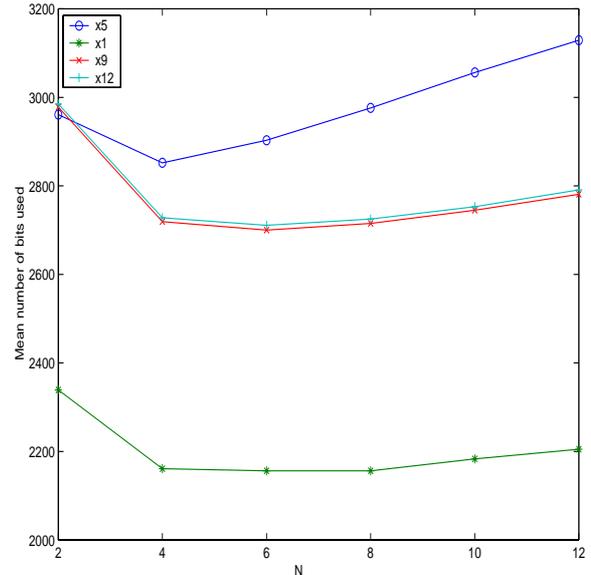

Figure 3: The mean number of bits required as functions of N for various audio files

affects the performance of the coder. Figure 3 shows the results of some of these experiments. Figure 3 indicates that the use of $N = 4$ is better than or equivalent to the use of any other value. This result can be explained by the way in which SPIHT performs the sorting. If a compromise between a few large sets and many smaller sets is obtained one would expect SPIHT to perform better than in either extreme case. This is because SPIHT gains from identifying large insignificant sets as well as having small significant sets. $N = 4$ presents such a compromise.

3.2.2 The MLT

The MLT is a uniform M-channel filter bank. In traditional block transform theory, a signal $x(n)$ is divided into blocks of length M and is transformed by the use of an orthogonal matrix of order M. More general filter banks take a block of length L and transform that block into M coefficients, with the

Table 2: Coding Results using the MLT.

Signal	Full Reconstruction		Partial Reconstruction with Masking	
	SegSNR (dB)	Mean Rate (kbps)	SegSNR (dB)	Mean Rate (kbps)
x1	55.5	**145**	16.7	**53**
x2	64.2	**31**	19.2	**14**
x3	49.4	**60**	17.9	**25**
x4	54.1	**110**	21.8	**47**
x5	45.8	**183**	7.6	**65**
x6	61.1	**68**	23.3	**33**
x7	55.5	180	20.1	**65**
x8	54.2	**140**	21.4	**47**

condition that $L > M$ [10]. In order to perform this operation there must be an overlap between consecutive blocks of $L - M$ samples [10]. This means that the synthesized signal must be obtained by the use of consecutive blocks of transformed coefficients. In the case of the modulated lapped transform L is equal to $2M$ and the overlap is thus M. The basis functions of the MLT are given by:

$$a_{nk} = h(n)\sqrt{\frac{2}{M}}\cos\left[(n + \frac{M+1}{2})(k + \frac{1}{2})\frac{\pi}{M}\right] \tag{2}$$

where $k = 0, \cdots, M - 1$ and $n = 0, \cdots, 2M - 1$. The window chosen is $h(n) = \sin((n + \frac{1}{2})\frac{\pi}{2M})$.

3.2.3 Results of combining the MLT with SPIHT

Table 2 shows the obtained results for complete reconstruction. The results shows that almost all of the SQAM files are coded using a lower mean rate than when the DWT is used, this is indicated by bold font values in the table. Also, note the high SegSNR results which illustrate the resilience of the MLT to quantization noise. The results in Table 2 are obtained with and without the use of the simultaneous masking.

The results presented in Table 2 are for the synthesized signals that are indistinguishable from the original. The reduction in bandwidth is very significant when the masking model is included in the coding, justifying the use of the psycho- acoustic model in the manner described.

The results show that at a rate of 65 kbps almost all of the SQAM signals tested may be reproduced to sound identical to the original. The MLT combined with simultaneous masking produces significant bandwidth savings and the addition of SPIHT also adds the dimension of scalability to the scheme. At the 54 kbps mark almost all of the files had no audible or very little distortion in them.

4   Conclusion

This paper has presented a comparison between two schemes of audio compression based on SPIHT. The results show clearly that significant savings may be obtained if the Modulated Lapped Transform is used in place of the Wavelet transform. The most significant savings are obtained when the Johnston technique of determining masked components is combined with the MLT based scheme. The results presented have also highlighted the usefulness of the SPIHT algorithm, combined with relevant transform coefficient relationships, to scalable audio coding, as the algorithm is designed with the aim of producing an embedded bit stream.

Acknowledgements

References

[1] Peter Noll, "Mpeg digital audio coding," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 59–81, Sept. 1997.

[2] G.A. Davidson, *Digital Signal Processing Handbook*, chapter 41, CRC Press LLC, 1999.

[3] H. Purnhagen and N. Miene, "Hiln - the mpeg-4 parametric audio coding tools," in *Proceedings of ISCAS 2000*, 2000, vol. 3, pp. 201–204.

[4] Amir Said and William A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems For Video Technology*, vol. 6, no. 3, pp. 243–250, June 1996.

[5] Zhitao Lu and William A. Pearlman, "An efficient, low-complexity audio coder delivering multiple levels of quality for interactive applications," in *1998 IEEE Second Workshop on*

Multimedia Signal Processing, 1998, pp. 529–534.

[6] Zhitao Lu, Dong Youn Kim, and William A. Pearlman, "Wavelet compression of ecg signals by the set partitioning in hierarchical trees algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 849–856, July 2000.

[7] "Mpeg web site at http://www.tnt.uni-hannover.de/project/mpeg/audio," .

[8] James D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal On Selected Areas In Communications*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[9] T.D. Tran and T.Q. Nguyen, "A lapped transform progressive image coder," in *Proceedings of ISCAS 1998*, 1998, vol. 4, pp. 1–4.

[10] Henrique S. Malvar, *Signal Processing with Lapped Transforms*, Artec House, Inc., Boston, 1992.