AUDIO CODING BASED ON THE MODULATED LAPPED TRANSFORM (MLT) AND SET PARTITIONING IN HIERARCHICAL TREES

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 presents an audio coder based on the combination of the Modulated Lapped Transform (MLT) with the Set Partitioning In Hierarchical Trees (SPIHT) algorithm. SPIHT allows scalable coding by transmitting more important information first in an efficient manner. The results presented reveal that the Modulated Lapped Transform (MLT) based scheme produces a high compression ratio for little or no loss of quality. A modification is introduced to SPIHT which further improves the performance of the algorithm when it is being used with uniform M-band transforms and masking. Further, the MLT-SPIHT scheme is shown to achieve high quality synthesized audio at 54 kbps through subjective listening tests.

1. INTRODUCTION

Scalable audio compression techniques are of interest for audio transmission over packet based networks such as the Internet. Such compression techniques are also relevant to mobile telephone service providers that aim to deliver different classes of quality to different customers. 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 quality audio must be obtained at the lowest rate.

A number of audio compression techniques are in common usage. MPEG standards [1] contain several techniques and algorithms for the compression of audio signals, as do some proprietary coders such as the Dolby AC series of coders [2]. The techniques presented by those standards and products are aimed at non-scalable transmission rates; that is, most of the techniques standardized are defined for a given bit rate. Although MPEG has made some attempts at standardizing scalable compression algorithms [3][4], the scalability defined in the MPEG standards remains heavily reliant on changing the compression paradigm with varying available bandwidth.

At lower rates, MPEG has adopted the Harmonic and Individual Lines Plus Noise (HILN) coder which is based on the original sinusoidal coders [5][4]. The HILN coder operates at rates ranging from 6 kbps to 24 kbps and has been built into an Internet audio transmission scheme [4]. The HILN coder focuses on signal bandwidth less than 8 kHz and utilizes a perceptual re-ordering scheme of the parameters.

This paper presents an audio compression scheme that utilizes the MLT and a transmission algorithm known as Set Partitioning In Hierarchical Trees (SPIHT) [6]. The algorithm was initially proposed as an image compression solution but it is general enough to have been applied to audio and electrocardiogram (ECG) signal compression, in combination with the Wavelet transform, as well [7][8]. The algorithm aims at performing an ordered bit plane transmission and sorts transform coefficients in an efficient manner allowing more bits to be spent on coefficients that more heavily contribute to the energy of the signal.

The results presented show that the MLT achieves a high compression ratio and the degree of compression obtained is enhanced by the use of a masking model. Further improvements are obtained by using a modified version of the SPIHT algorithm proposed in this paper. The modification tests for absolutely insignificant coefficients (i.e. zeros or coefficients below a given threshold) and removes those coefficients from the sorting and transmission algorithm completely. Subjective test results show that the MLT-SPIHT scheme produces high quality audio at 54 kbps.

2. THE MLT-SPIHT CODER

The codec based on the combination of the MLT transform and SPIHT is shown in Figure 1. In Figure 1, the input audio signal is transformed into the frequency domain by the MLT. The obtained coefficients are applied to a psychoacoustic model, that determines which coefficients are perceptually redundant, before being quantized and 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. The frame length used in this implementation was 20 ms (at a sampling rate of 44.1 kHz). As the MLT is being used, the overlap of the frames must be set to half the frame length.

2.1. The Modulated Lapped Transform (MLT)

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. On the other hand, lapped trans-

This work is supported by Motorola Australia Research Centre.

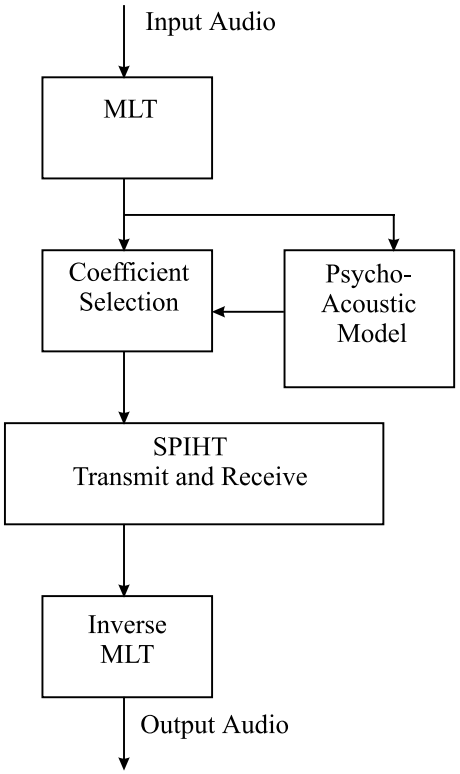

Fig. 1. The MLT-SPIHT codec

forms take a block of length L and transform that block into M coefficients, with the condition that $L > M$ [9]. In order to perform this operation there must be an overlap between consecutive blocks of $L - M$ samples [9]. 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:

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

where $k = 0, \cdots, M-1$ and $n = 0, \cdots, 2M-1$ with $h(n) = \sin(n + \frac{1}{2})\frac{\pi}{2M}$ being the perfect reconstruction window used.

2.2. The Use of Masking

There are two known masking mechanisms; frequency and time domain masking. Frequency domain masking is referred to as simultaneous masking and determines how tones reaching the ear simultaneously mask each other. Time domain masking (or temporal masking) occurs when a signal component masks another signal component before and/or after its onset (known as pre- and post-masking).

Two well known psychoacoustic models for determining the masked and masking components in the frequency domain are the *Johnston model* (first proposed by Johnston in [10]), and the *MPEG model 1* (as described in [11]). Both models allow the development of a masking curve for the entire spectrum of an audio signal. The masking curve defines the perceptual significance of signal components in the frequency domain. The major difference between the two is that the Johnston model specifies a masking value per critical band [10] whereas the MPEG model 1 specifies a masking value for each frequency bin used to describe the signal in the frequency domain (assuming that there are more frequency bins than critical bands).

The traditional way of using the masking curves has been to provide information on how much noise may be allowed in a given frequency band [11][1][2][10], or how accurately a given band needs to be quantized for transmission. For this purpose, a calculation of the mask-to-noise ratio in each critical band is carried out and more bits are allocated to the band with the lowest mask to noise ratio. An iterative procedure is employed where bits are assigned according to some distortion criteria [11]. This technique is used in the MPEG and Dolby AC transform coders [1][2]. Another way of using the masking curves is to ignore all spectral components below the curve. Our informal listening tests showed that if the Johnston technique is used in this manner the audio reconstructed from non-masked components sounds the same as the original audio, which is not the case for the MPEG model 1. The masking curve produced by the MPEG model was found to be too aggressive for this type of use, as the resulting synthesized audio takes on a characteristic similar to low-pass filtering the original audio signal. Hence, in the implementation used to obtain the results presented in this paper, the Johnston model is utilized.

2.3. Set Partitioning In Hierarchical Trees

The Set Partitioning In Hierarchical Trees algorithm (SPIHT) was introduced by Said and Pearlman in [6]. The complete algorithm is listed in [6] as *Algorithm II* and is not presented again here.

SPIHT is built on the principle that spectral components with more energy content should be transmitted before other components; thus the most relevant information is sent first. SPIHT sorts the available transform coefficients and transmits both the sorted coefficients and sorting information in an embedded bit stream. The algorithm is provided with an expected order of the coefficients defined in the form of trees; those coefficients closer to the roots of the trees are expected to be more significant than those at the leaves. The transmitted sorting information is used to modify this pre-defined order. The algorithm tests available coefficients and sets of coefficients to determine if those coefficients are above a given threshold. The result of the test is transmitted and the coefficients are deemed significant or insignificant relative to the current threshold. Significant coefficients are transmitted bit plane by bit plane.

In [6] the SPIHT algorithm used a pre-defined order that linked sub-band coefficients together in trees (with each tree being made up of a number of sets). The trees follow the natural sub-band progression of a dyadic wavelet transform having the lower frequencies located at the base of the trees [6]. In the audio coding work reported in [7], the wavelet transform was used, and so a similar way of organizing the coefficients in sets to that in [6] was used.

In the following we propose a new scheme for defining sets that are more relevant to uniform M-band transforms. The set development is initiated 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 offspring. In the next step each of the offspring is assigned N offspring and so on, until the number of the available coefficients is exhausted. We define the offspring 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), as

$$O(i) = iN + \{0, N-1\}. \quad (2)$$

Any offspring above $M-1$ are ignored. The descendants of the roots are obtained by linking the offspring together. For example, if $N=4$, node number 1 will have offspring $\{4,5,6,7\}$, node 4 will have offspring $\{16,17,18,19\}$ and the descendants of node 1 will include $\{4,5,6,7,16,17,18,19,...\}$. It has been determined experimentally that the use of $N=4$ is better than or equivalent to the use of any other value and so it is the value used in the implementation proposed in this paper.

3. A MODIFIED SPIHT ALGORITHM

SPIHT expects the parameters closer to the roots of the trees to be more significant than those at the leaves. In the frequency domain this translates to the expectation that lower frequencies hold more significant information than higher frequency components. The introduction of the masking creates a representation whereby a number of lower frequency parameters are deemed masked and thus insignificant. This representation in turn leads to a less efficient application of SPIHT. In this Section a modification is introduced into SPIHT to account for such 'unexpected' representations.

In combining the masking model with the MLT, masked coefficients are set to zero. If a masked coefficient is expected to be non-zero (through its position in the SPIHT trees) then SPIHT will test that coefficient a number of times for significance. Since a zero coefficient will never be significant and so will not be transmitted by SPIHT, a number of test bits are wasted on these significance tests. The effect of these wasted bits on the overall bit rate depends on how divergent the transform representation of the signal is from the expected representation. As a remedy, another test was introduced into SPIHT. The new test determines if a given amplitude is significant enough that it may ultimately be included in the transmitted amplitudes. Although this test adds one bit per amplitude to the cost of the algorithm, the savings made by removing insignificant amplitudes from the sorting process are usually greater when the masking model is applied. This saving, however, varies from signal to signal as the properties of the signal vary.

In terms of the algorithm, let the added test be $T_n(k)$ or $T_n(k,l)$ which determines if k (or (k,l) in the case of a matrix input) is above a set threshold. This test is included in the algorithm in step **(2.2.1)** (see [6]) as follows:

output $T_n(k,l)$
if $T_n(k,l) = 1$
- output $S_n(k,l)$
- if $S_n(k,l) = 1$ add (k,l) to the LSP and output the sign of $c_{k,l}$
- if $S_n(k,l) = 0$ add (k,l) to the end of the LIP.

Thus, if (k,l) is below the given threshold, the corresponding coefficient is no longer tested.

Name	Content	Name	Content
x1	Bass	x9	English F Speech
x2	Electronic Tune	x10	French F Speech
x3	Glockenspiel	x11	German F Speech
x4	Glockenspiel	x12	English M Speech
x5	Harpsicord	x13	French M Speech
x6	Horn	x14	German M Speech
x7	Quartet	x15	Trumpet
x8	Soprano	x16	Violoncello

Table 1. The Signal Content

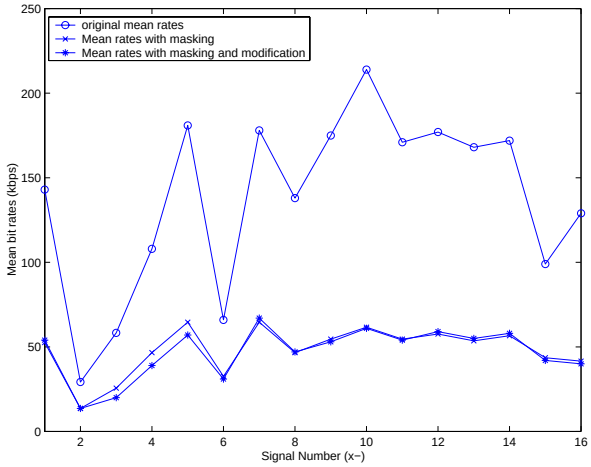

Fig. 2. Mean bit rates using the MLT with SPIHT

4. RESULTS

Figure 2 shows three plots, showing the mean bit rate used to compress the Sound Quality Assessment Material (SQAM [12]) files, of which there are sixteen with content as listed in Table 1. The plots are the results for SPIHT coding of the MLT coefficients for the following three cases:

- without masking or the modification
- with masking and without the modification
- with both masking and the modification applied

It is seen from Figure 2 that the masking reduces the mean bit rate significantly and the modification adds to this improvement for most of the SQAM files.

A set of informal listening tests (pairwise comparison tests) have been conducted to determine the subjective quality of the MLT-SPIHT coding scheme when masking is employed. The

score	*Sound Quality*
1	Very annoying distortion heard
2	Annoying distortion heard
3	Slightly annoying distortion
4	Some perceptible distortion heard, but its not annoying
5	No distortion can be heard

Table 2. Subjective test score guide [13]

Results	54 kbps	64 kbps
Overall mean	4.24	4.44
% No distortion heard	47.4	57.2
% No annoying distortion heard	80.9	88.5

Table 3. Subjective Test Scores for the 64 kbps and 54 kbps codecs

tests consisted of all of the SQAM files listed in Table 1 and nineteen subjects. The subjects varied in gender and age group. The subjects were asked to listen to the original signal and the synthesized signal and judge the similarity of the two signals by allocating a score between 1 and 5 according to Table 2 [13].

The test results obtained showed that in 63.5% of all test cases no distortion could be heard; in other words, the score allocated was a 5. Also, in 90.1% of all test cases any distortion heard was judged to be not annoying, that is, the score allocated was either a 4 or a 5. Finally, the overall mean of the scores given for the MLT-SPIHT coding scheme with masking was 4.52. The results of the subjective test indicate that high quality audio is obtained by the combination of the MLT with masking and SPIHT.

Using the MLT-SPIHT based coder with masking and the modification described in Section 3 a 54 kbps codec and a 64 kbps codec were produced by limiting the bit rate usable by SPIHT. Both of these coders were tested using the same methodology described above. Table 3 lists the results of those subjective tests.

The results listed in Table 3 show that very good quality audio may be obtained by using the MLT-SPIHT based coder with masking at rates between 54 and 64 kbps. More than 80% of all test cases indicated that no annoying distortion can be heard for both the 54 and 64 kbps cases. Table 3 also shows that the test subjects distinguished between the two bit rates, indicating little or no saturation in terms of quality even at relatively high rates as a result of the use of a scalable coding algorithm such as SPIHT.

5. CONCLUSION

This paper has presented a coding scheme built around the MLT and SPIHT. Masking was used to reduce the number of bits required to achieve high quality synthesized audio. A modification was also introduced to SPIHT and described. The modification has been shown to further improve the compression provided by SPIHT. Finally, the subjective test results presented showed that high quality synthesized audio may be achieved using this scheme at 54 kbps.

6. ACKNOWLEDGEMENTS

Mohammed Raad is in receipt of an Australian Postgraduate Award (industry) and a Motorola (Australia) Partnerships in Research Grant.

7. 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] K Brandenburg, O Kunz, and A Sugiyama, "Mpeg-4 natural audio coding," *Signal Processing: Image Communication*, vol. 15, no. 4, pp. 423–444, Jan. 2000.

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

[5] B. Edler and H. Purnhagen, "Parametric audio coding," in *Proceedings of the Fifth International Conference on Signal Processing*, 2000, vol. 1, pp. 21–24.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

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

[13] T. Ryden, "Using listening tests to assess audio codecs," in *Collected Papers on Digital Audio Bit Rate Reduction*, Neil Gilchrist and Christer Grewin, Eds., USA, 1996, pp. 115–125, Audio Engineering Society, Inc.