OVER-SAMPLING FOR ACCURATE MASKING THRESHOLD CALCULATION IN WAVELET PACKET AUDIO CODERS

Ferdinan Sinaga[#], *Eliathamby Ambikairajah*[#] *and Andrew P. Bradley**
[#]School of Electrical Engineering and Telecommunications
The University of New South Wales
NSW 2052, Australia

*Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP)
School of Information Technology and Electrical Engineering
The University of Queensland,
QLD 4072, Australia

ABSTRACT

Many existing audio coders use a critically sampled discrete wavelet transform (DWT) for the decomposition of audio signals. While the aliasing present in the wavelet coefficients is cancelled in the decoder, these coders normally perform calculation of the simultaneous masking threshold directly on these aliased coefficients. This paper uses over-sampling in the wavelet packet decomposition in order to provide alias-free coefficients for accurate simultaneous masking threshold calculation. The proposed technique is compared with masking threshold calculation based upon the FFT and critically-sampled wavelet coefficients, and the results show that a bit rate saving of up to 16 kbit/s can be achieved using over-sampling.

Keywords: wavelet packet, over-sampling, simultaneous masking.

1. INTRODUCTION

The discrete wavelet transform (DWT) is a powerful technique for audio coding because the DWT coefficients provide a compact, non-redundant representation of the signal [1, 2]. Moreover, using the more general wavelet packet decomposition, the decomposed sub-bands can be arranged to approximate the critical bands of the human auditory system, allowing the calculation of simultaneous masking thresholds directly from the resulting coefficients. In wavelet-based coders, simultaneous masking threshold calculations are normally performed on the critically sampled wavelet packet coefficients, however these are known to be affected by the aliasing inherent in the wavelet decomposition structure.

In this paper, we demonstrate that the use of over-sampling in the decomposition stage allows us to accurately calculate the simultaneous masking threshold, thus enabling a lower signal-to-mask ratio. The critically sampled wavelet coefficients are available as a subset of the over-sampled wavelet coefficients, and hence the decoder is unaffected by this masking calculation technique.

In sections 2 and 3, the critically sampled and over-sampled wavelet decompositions are explained respectively, including their relationship. Simultaneous masking for the removal of perceptually redundant signal components is described in section 4. In section 5, an audio coder is proposed that combines critically sampled wavelet packet decomposition with masking threshold calculations based on the over-sampled wavelet packet coefficients. Computational complexity is discussed in section 6, and performance comparisons on the proposed coder are made in section 7.

2. CRITICALLY SAMPLED DISCRETE WAVELET TRANSFORM

The discrete wavelet transform and discrete wavelet packet decomposition have been instrumental in localizing transient events in the time-frequency domain, and has hence found strong applications in audio coding. The DWT is described as follows [3]

$$w(2^i, 2^i n) = \frac{1}{\sqrt{2^i}} \sum_k \overline{\psi}\left(\frac{k}{2^i} - n\right) s(k), \qquad (1)$$

where ψ is the mother wavelet, *s(k)* is the discrete signal, 2^i is time dilation and $2^i n$ is time translation of wavelet transform indicating decimation.

The filter bank implementation of the critically sampled DWT is performed by down sampling the wavelet coefficients by a factor of two after sub-band filtering using the quadrature mirror filters. For the critically sampled wavelet decomposition, the number of output coefficients is identical to the number of input samples. In this paper, the Daubechies wavelet was selected for the decomposition, with db8 for the first to sixth levels and db2 for seventh and eighth levels.

3. OVER-SAMPLED DISCRETE WAVELET TRANSFORM

The over-sampled DWT is different from the critically sampled DWT in a filter bank implementation, in that it is performed without down sampling. Over-sampled wavelet packet (WP) coefficients can be obtained by performing critically sampled wavelet packet decomposition twice, firstly without shifting and secondly by shifting one sample and then interleaving the resulted wavelet coefficients [4], however this is computationally intensive.

Over-sampled WP decomposition can be performed more efficiently using the A Trous algorithms [3]. The A Trous algorithm is performed by inserting $2^i - 1$ zeroes between filter coefficients, where i is the decomposition level, and with no sub-sampling.

The critically sampled wavelet coefficients and the over-sampled WP coefficients have a close relationship since the critically sampled wavelet coefficients exist in the over-sampled WP coefficients. The critically sampled wavelet coefficients can be obtained by down-sampling the over-sampled WP coefficients by a factor of 2^i [3].

The coefficients of the over-sampled WP decomposition are closer to those of the continuous wavelet transform than those of the critically sampled WP decomposition [5]. Therefore, the over-sampled WP coefficients derived using the A Trous algorithm provides more accurate time-frequency information than the conventional WP decomposition, in addition to shift-invariance, a property that can be important in some applications. The disadvantages are the increased computational complexity and the increased memory required to represent the signal.

4. SIMULTANEOUS MASKING

Auditory masking is a well-known phenomenon in the human auditory system whereby signal components are rendered inaudible by the presence of masking signals that occur within the same critical band. The simultaneous masking model used in this work was obtained from [6].

After obtaining the simultaneous masking threshold (dB), the signal to mask ratio (SMR) is calculated using the maximum power (dB) in the processed frame. The SMR is used in the coder bit allocation algorithm to determine the minimum number of bits needed to represent the input signal in a perceptually lossless fashion. Reducing the number of bits used to represent the audio signal increases the quantisation noise, however the use of the maximum power in each critical band in the SMR calculation ensures that the maximum quantisation noise is still under the masking threshold.

Temporal masking, which has been shown to provide bit rate reductions of up to 20 kbit/s in wavelet packet-based audio coders, can also be combined [7] with the simultaneous masking to further reduce the bit rate. The functional model used to produce this improvement was based upon the following equation [8]:

$$TM_F = a(b - log_{10} t)(L_m - c), \qquad (3)$$

where TM_F is the amount of forward masking threshold in dB in the mth band. t is the time difference between the masker and the maskee in milliseconds. L_m is the masker level in dB obtained by taking the average power of all samples in the mth critical band. a, b, and c, are parameters derived from psychoacoustic data [8].

Temporal masking was not used in this work, since the objective of this paper was to evaluate the effect of critically sampled wavelet coefficients on simultaneous masking threshold calculation.

5. AUDIO CODER

The audio coder developed in this work improves upon the simultaneous masking threshold calculation used in existing critically sampled wavelet packet coders. Following the masking threshold calculation, the over-sampled WP coefficients are discarded, whereas the critically sampled WP coefficients obtained from the over-sampled WP coefficients are retained for the actual coding, as seen in Fig. 1.

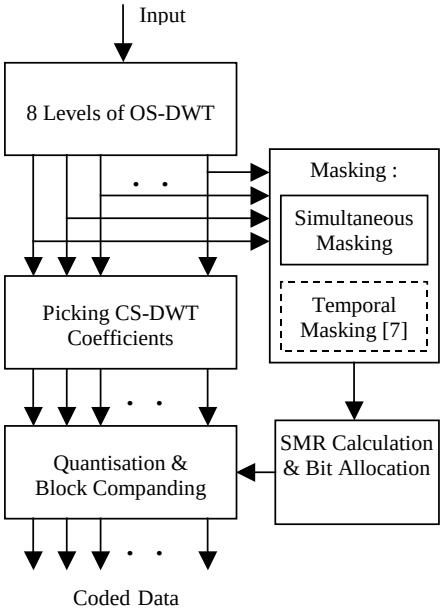

Figure 1 : Wavelet packet-based audio coder.
OS: Over-sampled, CS: Critically sampled.

6. COMPUTATIONAL COMPLEXITY COMPARISON

One major disadvantage of the over-sampled DWT using the A Trous algorithm compared to critically sampled DWT is that the computational complexity increases. Nevertheless the increase in computational complexity occurs in the coding process only, because the decoder simply receives the coded critically sampled data obtained from over-sampled DWT coefficients.

In order to compare computational complexity between critically sampled WP decomposition and over-sampled WP decomposition, the fully decomposed discrete wavelet packet is considered. In this comparison, the number of multiplications is taken as the measure of computational complexity. The results are seen in Table 1, where N is the number of samples in a frame, L is the number of decomposition levels and K is the number of non-zero filter coefficients, applicable in the A Trous algorithm, to avoid multiplication by zero.

Table 1. Computational complexity of fully decomposed critically sampled DWT and over-sampled DWT

Method	*Computational Complexity*
FFT	$N \log N$
Critically sampled DWT	$L\ N\ K$
Over-sampled DWT	$NK\sum_{i=1}^{L} 2^{L-i+1}$

As seen in Table 1, the over-sampled wavelet decomposition is more computationally expensive than the critically sampled wavelet decomposition, however it is still perfectly feasible for applications in which the critically sampled wavelet decomposition is normally used. The FFT also requires less computation than the over-sampled wavelet decomposition, but the small improvement in complexity is easily offset by the reductions in bit rate possible by the more accurate masking threshold calculation resulting from the use of the over-sampled wavelet decomposition. Moreover, in the MPEG standard, FFT is only used for masking threshold calculations, while a separate computation is needed for time frequency mapping. In DWT, the coefficients are directly used for masking threshold calculation.

In this work, the wavelet packet was not fully decomposed. The decomposition was limited to approximate the critical bands of the human auditory system. This simplification of the full decomposition reduces computational complexity by about 25% in our implementation.

7. PERFORMANCE EVALUATION

In evaluating the efficacy of the over-sampled DWT for bit rate reduction, four audio materials with 44.1 kHz sampling frequency were used. Additionally, the FFT algorithm, as used in MPEG 1 layer I, was also included in the comparison.

7.1. SMR and Bit Rate Comparison

In these experiments, DWT coefficients are scaled to have unity gain, while FFT coefficients have been normalized as per MPEG 1 layer I. This normalization equalizes the signal power of DWT and FFT so that the SMR can be compared, in order to make a meaningful bit rate comparison. The results are shown in Figure 2 and Table 2.

The average SMR for each band was calculated, to observe the effect on the bit rate of improved masking threshold calculation from the alias-free wavelet coefficients. Figure 2 shows that the average SMR value of the over-sampled DWT is lower than that of the critically sampled DWT, meaning that the over-sampled DWT can be exploited to reduce the bit rate. The average SMR value of over-sampled DWT is also lower than that calculated using the FFT.

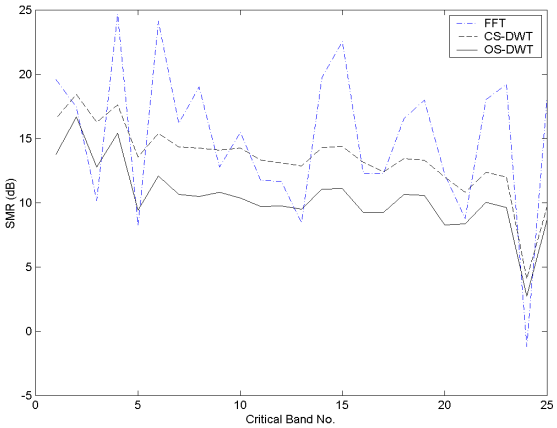

Figure 2. Average SMR for FFT, Critically Sampled (CS) DWT and Over-Sampled (OS) DWT

Table 2. Comparison of bit rate (kbps) for three schemes and four audio materials

Audio Material	*Length (sec)*	*FFT (kbps)*	*CS DWT (kbps)*	*OS DWT (kbps)*
Africa	12	184.8	160.7	144.4
Pop	25	168.0	162.2	149.1
Tracy Chapman	24	174.2	150.1	137.9
Raihan	22	162.4	164.5	150.9

It can be seen from Figure 2 that the aliasing inherent in the critically sampled DWT produces an inter-band *smoothing* effect on the SMR. Therefore, the difference between critical bands with a high and a low SMR becomes less pronounced and hence the difference in bits allocated to these bands becomes less pronounced. This means that the critically sampled DWT is less able to take advantage of any narrow band simultaneous masking. It should be noted that the difference between the average level of the critically sampled and over-sampled SMR graphs is due to the fact that the maximum signal power found in the critically sampled DWT will always be less than or equal to that in the over-sampled DWT.

7.2. Subjective Test

Semi-formal subjective tests, involving 14 subjects, were performed on the decoded audio to produce the result in Table 2. The subjective tests comply with the ITU-R Recommendation BS.1116 [9].

The test procedures were conducted double-blindly using A-B-C triple stimuli with hidden reference. These criteria were achieved by using ABC/HR software [10]. Audio A was the original version as the reference while B and C were the original and the decoded version that were assigned randomly by the software. This is a double blind criterion where neither subject nor test administrator knows which one of B and C is the reference during the test.

Firstly, the original version (A) was presented as the reference to the subjects. Secondly, the randomly assigned original and the decoded version were presented to the subjects. The subjects could listen to A, B or C as many times as they like. Thirdly, the subjects were asked to identify B or C as the decoded version after comparing to A. The grading scale was as follows: 1.0 to 1.9 for Very Annoying quality, 2.0 to 2.9 for Annoying quality, 3.0 to 3.9 for Slightly Annoying quality, 4.0 to 4.9 for Perceptible but Not Annoying quality and 5 as Imperceptible quality with 0.1 resolution. The score was calculated in subjective difference grade (SDG)

$$SDG = Grade_{decoded} - Grade_{original} \qquad (4)$$

where $Grade_{decoded}$ is the score of the audio material that is selected by the subject as the decoded version and $Grade_{original}$ is the score of the original version or the reference, which is 5.0. Correct selection of the decoded version results in negative SDG while the incorrect selection results in positive SDG. The average SDG is then subtracted from 5.0 as the original grade to be mean subjective grade (MSG).

After the subjects graded the audio quality, the original and the decoded audio signals were presented to the subject, and then a signal was randomly selected from these two audio signals and presented to the subjects. The subjects then identified the audio signal as the original or the decoded signal. These steps were repeated five times. The probability that the subject is guessing if the subject identifies correctly all the times is 0.031 and 1.0 if the subject identifies incorrectly all the times.

From the data obtained from the subjective tests, the MSG for critically sampled DWT and the over-sampled DWT is 4.900 and 4.854 respectively. The average probability that the subjects are guessing is 0.45 for critically sampled DWT and 0.44 for over-sampled DWT.

These numbers show that the MSG of over-sampled DWT is very close to the MSG of critically sampled DWT decoded output, which shows almost equal quality. By combining this with the probability that the subjects are guessing, (both probabilities are close to 0.5), the transparent quality has been achieved.

8. CONCLUSIONS

The use of over-sampling in wavelet packet audio coding has been presented in this paper. It has been shown that bit rate reductions of up to 16 kbit/s can be achieved using over-sampling in a variable bit rate scheme, as compared with the conventional critically-sampled DWT or the FFT as used in MPEG 1, layer I. Further, subjective tests have shown that this bit rate reduction can be achieved while maintaining transparent quality in the decoded audio signals. Future research will concentrate on the integration of functional temporal masking models in an over-sampled wavelet packet audio coder, and the performance of fixed bit rate coders based on over-sampled wavelet packet coefficients.

9. REFERENCES

[1] A. P. Bradley, "Shift-invariance in the discrete wavelet transform," in *Proc. of Digital Image Computing : Techniques and Applications (DICTA'03)*, Sydney, Australia, pp. 29-38, 2003.

[2] A. Cohen and J. Kovacevic, "Wavelet : the mathematical background," *Proceedings of the IEEE*, vol. 84, pp. 514-522, 1996.

[3] M. J. Sensa, "The discrete wavelet transform : wedding the a trous and mallat algorithm," *IEEE Transactions on Signal Processing*, vol. 40, pp. 2464-2482, 1992.

[4] Y. Andreopoulos, M. V. Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Complete-to-overcomplete discrete wavelet transforms for scalable video coding with MCTF," *Visual Communication and Image Processing*, vol. 5150, pp. 719-731, 2003.

[5] H. Guo and C. S. Burrus, "Convolution using the undecimated discrete wavelet transform," in *Proc. of Acoustics, Speech and Signal Processing*, pp. 1291-1294, 1996.

[6] M. Black and M. Zeytinoglu, "Computationally efficient wavelet packet coding of wide-band stereo audio signals," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, pp. 3075-3078, 1995.
[7] F. Sinaga, T. S. Gunawan, and E. Ambikairajah, "Wavelet packet based audio coding using temporal masking," in *Proc. of International Conference on Information, Communications and Signal Processing and Pacific-Rim Conference on Multimedia*, Singapore, 2003.
[8] W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *Journal of Acoustic Society of America*, vol. 71, pp. 950-962, 1982.
[9] ITU, "ITU-R BS.1116.1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *International Telecommunication Union, Geneva*, 1997.
[10] ff123, "ABC/Hidden Reference : Tool for comparing multiple audio samples," 2002.