LOW BIT RATE SPEECH AND MUSIC CODING USING THE WAVELET TRANSFORM

S. Boland, M. Deriche, and S. Sridharan

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology.

ABSTRACT - This paper investigates low bit rate coding of high quality speech and music signals using the discrete wavelet transform (DWT). Compression is first exploited by eliminating wavelet coefficients that are zero or near zero in magnitude. The remaining coefficients are then uniformly quantized using 8 bits. Identical experiments are carried out with the DCT and the FFT. In each case, subjective quality and the segmental SNR are recorded to enable accurate comparison with the DWT. Also the effects of frame size, number of vanishing moments, and levels decomposed in the DWT, on the subjective quality and seg. SNR are examined. More efficient techniques for the quantization of wavelet coefficients are currently considered with the aim of reducing the bit rate, while maintaining near transparent quality.

INTRODUCTION

Digital coding of high quality speech and music signals is becoming a very important issue in the fields of telecommunications and signal processing. Several strategic applications which require high quality digital speech and music at low bit rates are emerging. These include digital audio broadcasting, multimedia/ hypermedia applications, satellite TV and teleconferencing. Monophonic compact disk (CD) quality audio is characterized by the high bit rate of 705 kb/s (44.1 kHz sampling frequency and 16 bit PCM). Similarly, uncompressed wideband speech is characterized by the high bit rate of 256 kb/s (16 kHz sampling frequency and 16 bit PCM). Efficient transmission or storage in the above applications requires a compression scheme that is equally suited to coding high quality speech and music signals.

Several attempts have been investigated for "transparent" coding of high quality digital audio at low bit rates. Transparent means that no audible difference exists between the original signal and the coded and reconstructed signal for all possible ears. Most previous methods use either subband coding (Brandenburg and Stoll, 1992) or transform coding (Johnston, 1989). The majority of schemes are perceptually based, exploiting the masking properties of the human hearing process. This is the phenomenon whereby a weak (noise) signal is made inaudible by a simultaneously occuring stronger (audio) signal. For each audio frame, the results of a masking threshold calculation are used to adaptively allocate bits to either the subband signals in the case subband coders, or the discrete cosine transform (DCT) coefficients in the case transform coders.

It is important to notice that the compression schemes for wideband speech are different in several respects to audio coders. One difference is the lower sampling frequency of 16 kHz which is used in the CCITT G.722 wideband speech coding algorithm (Noll, 1993). Higher sampling rates result in oversampling and are not necessary in wideband speech applications. Also wideband speech coding algorithms are not perceptual coders. The CCITT G.722 coder splits the speech signal into only two subbands by a pair of quadrature mirror filters. It does not use a masking threshold calculation to allocate bits to the two subbands. Instead the lower and higher subbands are ADPCM encoded with fixed bit rates of 48 kb/s and 16 kb/s respectively.

Recently, the wavelet transform has been investigated for compression of CD quality audio (Sinha and Tewfik, 1993). Unlike the Fourier transform, the wavelet transform is characterized by a constant relative bandwidth or "constant-Q". This is attractive for coding audio signals since audiological research has shown similar properties in the human ear. The ear integrates sounds over regions of frequencies called "critical bands". Below 500 Hz, the critical bands are 100 Hz wide, while above 500 Hz, the critical bands are approximately a third octave. In subband coding, the frequency ranges of the decomposed subbands cannot be made the same as the critical bands. On the other hand, the discrete wavelet transform (DWT) and in particular, the wavelet packet transform (WPT), can be used to obtain a subband decomposition very close to the critical band divisions. The advantage here is that a more accurate masking model can be calculated, which in turn leads to a reduction in bit rates.

This paper presents the results of preliminary research into the compression of both CD quality music and wideband speech signals using the DWT. Firstly a background of the DWT is provided. A description of simple compression tests is performed and results with the DCT and FFT are then given. Finally, conclusions are made on how further reductions in bit rates can be achieved using the DWT.

BACKGROUND-THE DISCRETE WAVELET TRANSFORM

Wavelets are a new family of basis functions for the space of square integrable signals (Daubechies, 1990). A signal f(t) can be representated by translates and dilations of a single wavelet W(t) as

$$f(t) = \sum_{j=J}^{\infty} \sum_{k=-\infty}^{\infty} \sqrt{2^j} b(j,k) W(2^j t - k) + \sum_{k=-\infty}^{\infty} \sqrt{2^J} a(J,k) g(2^J t - k) \tag{1}$$

where $b(j,k) = \int f(t) W(2^j t - k) dt$ and $a(J,k) = \int f(t) g(2^j t - k) dt$. This expansion provides a multiresolution decomposition of the signal f(t). The coefficients $b(j,k)$ represent details of the original signal at different levels of resolution j and coefficients $a(J,k)$ represent an approximation of the original signal f(t) at resolution J. The wavelet W(t) is obtained from a scaling function g(t) as

$$W(t) = \sum_{k=0}^{K-1} (-1)^k c_{1-k} g(2t - k) \tag{2}$$

where c_k are the coefficients that define the scaling function, g(t), which obeys the dilation equation given by:

$$g(t) = \sum_k c_k g(2t - k) \tag{3}$$

To construct the wavelet W(t), the coefficients c_k must satisfy certain conditions. In most cases, attention is restricted to wavelets with compact support i.e. c_k is nonzero only for $0 \leq k \leq K-1$. Since g(t) is a low-pass function normalized such that $\int g(t) dt = 1$,

$$\sum_{k=0}^{K-1} c_k = 2 \tag{4}$$

Similarly for orthogonality of the translates and dilates of g(t) we require

$$\sum_{k=0}^{K-1} c_k c_{k+2m} = 2\delta_{0,m} \tag{5}$$

where $\delta_{0,m}$ is the Kronecker delta function. To obtain wavelet coefficients, $f_{j,k}$, that decay quickly to zero, another condition is placed on the coefficients c_k's

$$\sum_k (-1)^k k^m c_k = 0 \qquad m = 0, 1, 2, ..., p-1 \tag{6}$$

A wavelet W(t) obeying (6) has p vanishing moments.

Mallat (1989) has shown that the discrete wavelet transform can be implemented by a recursive algorithm. This is done by using the coefficients c_k as the filter coefficients of a pair of quadrature mirror filters, G and H (see figure 1). The output of the low pass filter, G, is the approximation of the input signal for that level. While the output of the high pass filter, H, is the detail for that level. The impulse response of the low pass and high pass filters are related by the equation,

$$g(n) = (-1)^n h(1-n) \tag{7}$$

Here the coefficients $a(J,k)$ and $b(j,k)$ are computed recursively using the efficient pyramid algorithm proposed by Mallat (1989).

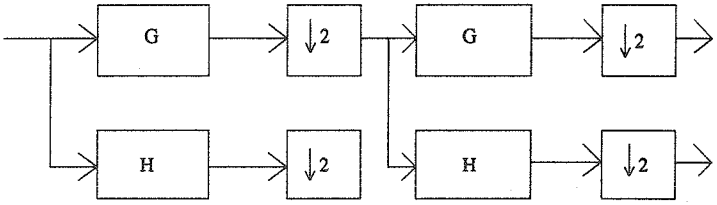

Figure 1. Discrete Wavelet Transform Decomposition

COMPRESSION AND QUANTIZATION OF WAVELET COEFFICIENTS

To assess the usefulness of the discrete wavelet transform in audio compression, some simple tests were performed on three test signals (two music and one speech). The music signals were sampled at 44.1 kHz and 16 bit PCM. The speech signal was sampled at 16 kHz and 16 bit PCM. Each signal was divided into small frames and the DWT was computed. Thresholding was then used to remove the wavelet coefficients that were zero or near zero in magnitude. To be consistent, in each case compression ratios of 2:1 and 4:1 were used for each signal. For 2:1 compression, half the wavelet coefficients are kept and quantized, while for 4:1 compression only one-quarter of the original wavelet coefficients are kept and quantized.

In our initial experiments, frame sizes of 512 and 1024 samples were used with 8-bit uniform quantization. After quantizing the non-zero coefficients, the audio frames were then reconstructed from the quantized coefficients. This same procedure was carried out for both the DCT and the FFT. As with the DWT, this was done for compression ratios of 2:1 and 4:1. Efficient techniques for quantization of coefficients, such as entropy coding, are currently being investigated and implemented. Also further reductions in bit rates can be achieved by shaping the quantization noise below a computed masking threshold. This is presently being designed and implemented also.

COMPARISON OF SEG. SNR AND SUBJECTIVE QUALITY

It is well known that traditional SNR is not an accurate measure in assessing the performance of audio coders. This is since the SNR does not take into account masking effects. However to compare the performance of the DWT with the DCT and FFT, we chose to use listening experiments as well as a segmental SNR calculation for each test signal. The segmental SNR is

determined by calculating the SNR for each frame and averaging it over the entire signal. Table 1 lists the recorded segmental SNR for the Daubechies wavelet with 3, 10 and 20 vanishing moments. The segmental SNR for the DCT and FFT are also recorded in the table. For each case of the DWT, DCT and FFT the size of the frame used was 512 samples.

Listening tests were carried out on each of the audio segments using headphones. For the two music segments the quality of the reconstructed signal was much higher for the DWT than that of the DCT. For the wideband speech segment, the quality was very similar for both the DWT and the DCT. The subjective quality of the reconstructed signal was considerably worse, however, for the each audio segment using the FFT. This is even though the average segmental SNR is very close to the DWT and DCT. The DCT gave near transparent quality for the wideband speech segment with 2:1 compression. Transparent or near transparent reconstruction was achieved for the mozart segment using the DWT and 2:1 compression. This was apparent for 10 and 20 vanishing moments only. As the number of vanishing moments was increased from 3 to 10 vanishing moments, the reconstructed signal was of higher quality.

Method	Compression Ratio	SNR-Guitar	SNR-Mozart	SNR-Speech
DWT - order 3	2:1	30.7	35.1	26.2
	4:1	27.6	28.7	18.9
DWT - order 10	2:1	32.2	35.7	27.4
	4:1	28.5	32.1	19.9
DWT - order 20	2:1	32.1	35.9	27.8
	4:1	28.3	32.0	20.0
DCT	2:1	32.0	34.7	28.3
	4:1	29.9	33.9	21.5
FFT	2:1	30.9	31.9	27.2
	4:1	27.5	29.4	20.3

Table 1. Average Segmental SNR (dB) for the DWT, DCT and FFT.

EFFECT OF DIFFERENT LEVELS OF RESOLUTION AND FRAME SIZE

For a frame size of 512 samples, the DWT decomposes the original signal into 9 levels of resolution and 1 approximation level. The effect of only decomposing the original signal into fewer resolution levels was examined. Table 2 lists the average segmental SNR for 5, 2 and 1 resolution levels. This was done using the Daubechies wavelet with 10 vanishing moments and 2:1 compression. Compared to the complete DWT in table 1, the SNR is several dB higher for the two music segments. Decomposing the audio signal into only 2 levels of resolution means that the decomposed subbands are 0-5 kHz, 5-10 kHz and 10-20 kHz. After taking the wavelet transform, the coefficients for the two detail levels are very small. Hence the removal of wavelet coefficients that are near zero occurs in the two detail signals. The remaining coefficients, most of which lie in the approximation level or normal speech bandwidth (0-5 kHz), are then quantized. The quality of the reconstructed music segments in each case was also higher compared to the complete DWT. Conversely, very little increase in quality in the reconstructed speech segment was apparent for different levels of resolution. Given the lower bandwidth of the speech signal, a dramatic increase in speech quality is not expected.

Different frame sizes and their effect on subjective quality and SNR was also investigated. For each audio signal, frame sizes of 256, 512, 1024 and 2048 samples were used. The segmental SNR in each case is shown in table 3. Only small changes are apparent in the SNR for each test signal. The quality of the reconstructed signals was also very similar for each frame size used.

However it is expected that with signals containing a large variation in energy or containing transients, such as castenets, special consideration should be given to the frame size. This is currently being investigated also.

Levels of Resolution	SNR-Guitar	SNR-Mozart	SNR-Speech
5	33.9	36.4	27.4
2	39.0	39.8	27.5
1	38.9	40.7	25.9

Table 2. Average Segmental SNR (in dB) for the DWT and different levels of resolutions.

Frame Size	SNR-Guitar	SNR-Mozart	SNR-Speech
256	39.0	40.2	27.1
512	39.0	39.8	27.5
1024	38.8	39.3	27.9
2048	38.5	38.8	28.5

Table 3. Average Segmental SNR (in dB) for the DWT and different frame sizes in samples.

CONCLUSION

The work performed here is only preliminary research into low bit rate speech and music coding using the wavelet transform. The results from listening tests and seg. SNR, however, show that the DWT is better suited than the DCT and FFT. Also since the DWT is more computationally efficient than the DCT and FFT, it is very applicable to the task of audio compression. Computation of a masking model and entropy coding of the wavelet coefficients are the two further areas currently being investigated and implemented. This is being done with the aim of reducing the total bit rate. At the conference, we will present more results on this ongoing research.

ACKNOWLEDGEMENT

This work was partially supported by a grant from the Australian Telecommunications and Electronics Research Board (ATERB).

REFERENCES

K. Brandenburg and G. Stoll, *The ISO/MPEG-Audio Codec: A generic standard for coding of high quality digital audio*, Presented at 92nd AES Convention, Vienna, Austria, March 1992.

I. Daubechies, *Orthonormal bases of compactly supported wavelets*, IEEE Transactions on Information Theory, Vol. 36, pp. 961-1005, September, 1990.

J. Johnston, *Perceptual transform coding of wideband stereo signals*, Proceedings ICASSP, 1989, pp.1993-1995, vol.3.

S. Mallat, *A theory for multiresolution signal decomposition : The wavelet representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, 1989.

P. Noll, *Wideband Speech and Audio Coding*, IEEE Communications Magazine, pp. 34-44, November, 1993.

D. Sinha and A. Tewfik, *Low bit rate transparent audio compression using adapted wavelets*, IEEE Transactions on Signal Processing, pp.3464-3479, vol. 41, no. 12, 1993.