A COMPARISON OF TWO HYBRID AUDIO CODING STRUCTURES INCORPORATING DISCRETE WAVELET TRANSFORMS

Michael Mason Sridha Sridharan Vinod Chandran
Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
m.mason@qut.edu.au s.sridharan@qut.edu.au
v.chandran@qut.edu.au

ABSTRACT: This paper compares the performance of two related audio coding structures. Both structures are hybridisations of parametric and subband transform coding schemes. The first coder exploits linear predictive (LP) analysis to model the spectral shape of the audio signal, and uses the LPC analysis filter to extract the residue. The residue is decomposed into non-uniform subbands using a multiband discrete wavelet transform (MDWT), and these subband coefficients are quantised in accordance to a perceptually determined dynamic bit allocation scheme. The second coder directly models the audio signal using a sinusoidal model and the residual is the difference between this model and the original signal. The residual is quantised in the manner as in the first coder. The quality of the decoded audio and the complexity of the coders is compared.

INTRODUCTION

The high fidelity coding of digitised audio signals has been of interest to researchers and commercial enterprises alike for many years now. The need to reduce storage and transmission bandwidth requirements have been the driving forces behind the development of a range of different algorithms and applications. In recent years the International Organisation for Standardisation (ISO) and International Electrotechnical Commission (IEC) have published a series of standards which address the coding of moving pictures and associated audio(MPEG, 1993, 1996, 1999). The Moving Pictures Expert Group (MPEG) is responsible for the ongoing development of these standards.

As part of the most recently published standard (commonly referred to as MPEG-4(MPEG, 1999)) the need to incorporate parametric coding of audio has been recognised. In recent years codecs which incorporate parametric coding components have been the subject of several proposed approaches to audio coding and in particular a range of codecs which combine parametric coding with subband transform methods have emerged(Lin and Steele, 1993; Boland and Deriche, 1995; Hamdy et al., 1996; Boland and Deriche, 1997; Moriya et al., 1997; Rampreshad, 1998). These and similar approaches have become collectively known as hybrid coders.

Previously we have presented work related to two hybrid coding structures, investigating issues relating to scalability(Mason et al., 1997) and the use of vector quantisation(Mason and Sridharan, 1998). In this paper we consolidate these two structures and compare their performance. Both coding structures analyse the audio signals and extract from it a noise- like residue which is quantised in a subband transform domain. The Linear Predictive, Discrete Wavelet Transform (LP-DWT) coder exploits LP analysis to model the spectral shape of each frame of audio and then filters the frame with an all-zero LPC analysis filter to determine the residual. The Sinusoidal Model, Discrete Wavelet Transform (Sin-DWT) coder models the sinusoidal components of the signal directly, and calculates the residual as the difference between the model and original signal. The residual signal is transformed into 10 non-uniform subbands using a 3-level 4-Band MDWT and each of these subbands are quantised directly. The resolution of the quantisers for each subband is controlled by a psychoacoustic model and dynamic bit allocation technique.

Section 2 describes the two coders, first giving a general overview of the common parameters, then the details of the parametric portion of each coder. The common transform domain portion of the coders and how the different parametric models interact with it is then described. Finally a summary of the bit

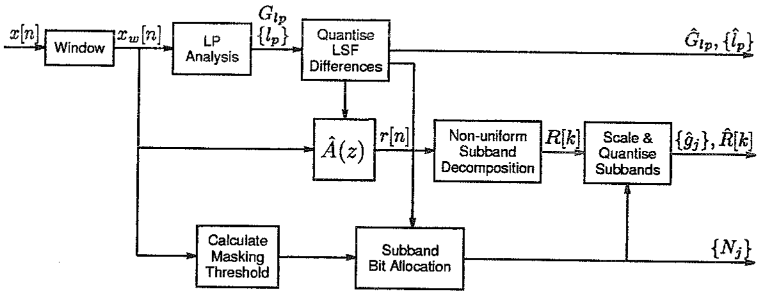
$x[n]$
Window
$x_w[n]$
LP Analysis
G_{lp}
$\{l_p\}$
Quantise LSF Differences
$\hat{G}_{lp}, \{\hat{l}_p\}$
$\hat{A}(z)$
$r[n]$
Non-uniform Subband Decomposition
$R[k]$
Scale & Quantise Subbands
$\{\hat{g}_j\}, \hat{R}[k]$
Calculate Masking Threshold
Subband Bit Allocation
$\{N_j\}$

Figure 1: Block Diagram of LP-DWT Coder

allocation for each coder is presented. Section 3 details the methods and results of the comparisons of computational complexity and quality and section 4 discusses these results and offers future research directions.

CODER DESCRIPTIONS

Both the LP-DWT and Sin-DWT coders use a 512 sample frame which overlaps 1/16 of the previous frame. When coding CD quality audio sampled at 44.1 kHz this represents a frame length of 11.6 ms and approximately 92 frames/sec. To avoid spectral splatter a windowing function is applied to each frame, for the Sin-DWT coder this is done after the sinusoidal model is fitted. The window function used for these coders is a flat topped window with Hamming roll-off in the overlap regions. No window switching is used in either of these coders. While both coders can be configured to code at a range of bandwidths down to 16 kHz, for this comparative study only their performance on 44.1 kHz CD quality audio was considered.

Linear Predictive Analysis (LP-DWT Coder)

Figure 1 provides the block diagram of the LP-DWT coder. Each incoming frame, $x[n]$, is windowed and then the spectral shape of the signal is modeled by a standard linear predictor. The predictor error can be expressed the output of a minimum phase analysis filter $A(z)$,

$$A(z) = G_{lp}\left(1 + \sum_{p=1}^{N} a_p z^{-p}\right) \tag{1}$$

The frequency response of $1/A(z)$ models the spectrum of $x_w[n]$ as shown if figure 2. The LP-DWT coder uses a 16th order LP model. The 16 Linear Predictive Coding(LPC) coefficients, $a_1 \ldots a_{16}$, are transformed into Line Spectral Frequencies(LSF), $l_1 \ldots l_{16}$, and the differences between successive LSFs quantised rather then absolute LPCs to localise the spectral distortion caused by quantisation(Soong and Juang, 1993). The LPC gain, G_{lp}, is quantised using a non-linear scalar quantised designed to match an estimate of the gain factor statistics based on a large set of training data.

Sinusoidal Model (Sin-DWT Coder)

The block diagram for the Sin-DWT Coder is presented in figure 3. Each input frame is fitted to a simplified version of the sinusoidal model first proposed byMcAulay and Quatieri (1986). By restricting the model to contain only undamped sinusoids and to be purely real the complex exponetial model is reduced to (2),

$$x[n] = \sum_{p=1}^{N/2} A_p \sin\left(n\omega_p + \theta_p\right) \tag{2}$$

where N is the model order and $\{A_p,\omega_p,\theta_p\}$ defines the amplitude, frequency (in radians) and fundamental phase of the pth sinusoidal component. A 16th order model was fitted, and the Total-Least

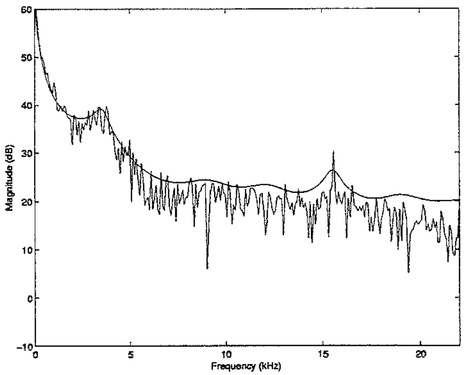

Figure 2: Frequency response of 16th order LPC filter ($1/A(z)$) superimposed on magnitude spectrum of audio frame

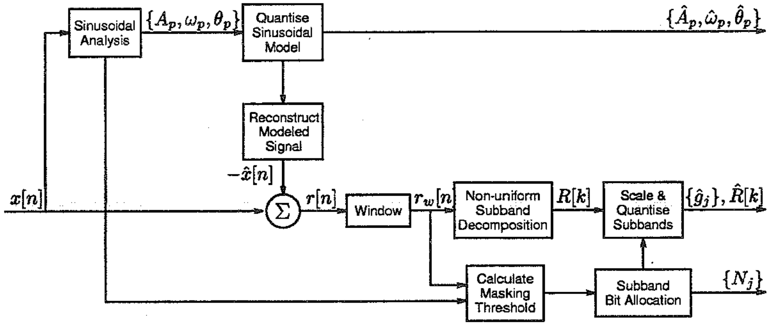
Sinusoidal Analysis
$\{A_p, \omega_p, \theta_p\}$
Quantise Sinusoidal Model
$\{\hat{A}_p, \hat{\omega}_p, \hat{\theta}_p\}$
Reconstruct Modeled Signal
$-\hat{x}[n]$
$x[n]$
Σ
$r[n]$
Window
$r_w[n]$
Non-uniform Subband Decomposition
$R[k]$
Scale & Quantise Subbands
$\{\hat{g}_j\}, \hat{R}[k]$
Calculate Masking Threshold
Subband Bit Allocation
$\{N_j\}$

Figure 3: Block Diagram of Sin-DWT Coder

Squares Prony (TLS-Prony)(Rahman and Yu, 1987) method was used to solve for the individual parameters. Figure 4 shows an example of a single frame with its sinusoidal model overlaid.

Each of the parameters were then individually quantised to provide $\{\hat{A}_p, \hat{\omega}_p, \hat{\theta}_p\}$. To reduce the perceptual effects of quantising the frequency parameters, $\omega_1 \ldots \omega_{16}$, these were first mapped to the critical band scale (Bark) using an approximation of the Bark-Hertz mapping 3, before being quantised with a linear quantiser.

$$\Omega = 6\log\sqrt{\frac{\omega}{1200\pi} + \left(\frac{\omega}{1200\pi}\right)^2 + 1} \qquad (3)$$

Non-uniform Subband Decomposition

The non-linear subband decomposition was achieved using a 3 level MDWT, with each level containing 4 bands. The near perfect reconstruction 4-band filterbank was designed using the method proposed by Ikehara and Nguyen (1997), and each filter had 32 taps. Figure 5 shows the normalised response of these filters and figure 6 depicts the effective power spectral density, Υ_{H_n}, of the 3 level transform. Each residual was symmetrically extended at its boundaries to avoid aberrations caused by the discontinuities that occur with zero padding or periodic extension.

Masking Threshold, Subband Bit Allocation and Quantisation

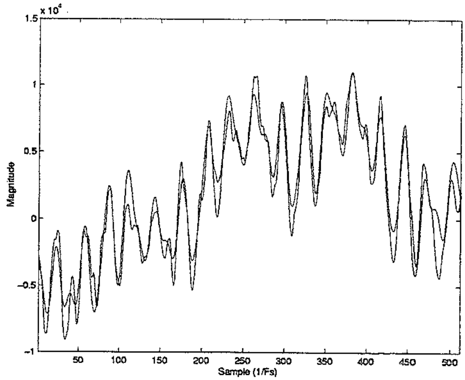

Figure 4: A Single Audio Frame and its 16th Order Sinusoidal Model

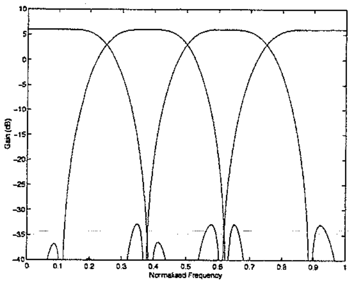

Figure 5: Normalised Frequency Response of 4-Band Filterbank

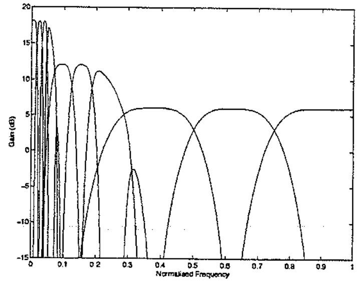

Figure 6: Power Spectral Density of 3-Level MDWT

Each coder calculates an estimate of the simultaneous frequency masking threshold based on the psychoacoustic model 2 proposed in MPEG (1993). As part of the masking threshold calculations for the Sin-DWT coder the frequency parameters of the model are used to simplify the calculation.

The masking threshold was then used to calculate the Signal-to-Mask Ratio (SMR_n) of each subband. For the Sin-DWT this was done directly using (4) and for the LP-DWT the masking threshold, Φ, and power spectral density of the nth subband filter, $\Upsilon_{H_n(z)}$, was modified by the effective power spectral density of the LPC analysis filter, $\Upsilon_{\hat{A}(z)}$, (5).

$$SMR_{n_{sin}} = 10 \log \left(\frac{\sum \Phi * \Upsilon_{H_n(z)}}{\sum r_w[n] * \Upsilon_{H_n(z)}} \right) \tag{4}$$

$$SMR_{n_{lp}} = 10 \log \left(\frac{\sum \Upsilon_{\hat{A}(z)} * \Phi * \Upsilon_{H_n(z)}}{\sum \Upsilon_{\hat{A}(z)} * r[n] * \Upsilon_{H_n(z)}} \right) \tag{5}$$

The Mask-to-Noise Ratio (MNR_n) can then be defined as,

$$MNR_n = SNR_n - SMR_n \tag{6}$$

where the Signal-to-Noise Ratios for each subband, SNR_n, is dependent on the resolution of the quantiser associated with the subband. Mason and Sridharan (1998) describes the the dynamic bit allocation method employed when using D_4 Lattice Vector Quantisers (LVQ).

Parameter	(bits)
$\hat{G}_{lp}$	8
$\hat{l}_1$	6
$15 \times \Delta\hat{l}_{1 \ldots 16}$	75
$10 \times N_j$	50
Total	139

Table 1: Side Information Bit Allocation for the LP-DWT

Parameter	(bits)
$16 \times \hat{A}_p$	80
$16 \times \hat{\omega}_p$	160
$16 \times \hat{\theta}_p$	96
$10 \times \hat{N}_j$	50
Total	386

Table 2: Side Information Bit Allocation for the Sin-DWT

Summary of Bit Allocation

A summary of the side information bit allocation is presented in tables 1 and 2. The side information is coded for each frame and the rest of the bits allocated to quantising the wavelet coefficients. The total number of bits per frame varies from 1044 bits per frame at 96 kbps to 1393 bits per frame when coding at 128 kbps.

PERFORMANCE ANALYSIS

Complexity

When comparing the computational complexity of the two coders it is reasonable to focus simply on the functional blocks which are different between them. With this in mind the blocks of interest from the LP-DWT coder are the LP Analysis, the LSF quantisation and the filtering by $\hat{A}(z)$. For the Sin-DWT coder the complexity of the TLS-Prony analysis, the quantisation of the model and the reconstruction and extraction of the residual, $r[n]$, are of interest. The last functional block which differs between the two coders is the masking threshold and SMR calculations.

Grouping the blocks associated with the parametric modeling and residual extraction together for each coder we found that the Sin-DWT coder was 50% more computationally intense then the LP-DWT coders modeling process. This massive increase was counter slightly when considering the masking threshold and SMR calculations where the direct use of the sinusoidal models parameters provided an 5% improvement in the total computational load. The total effect of these computational differences was found to be that the Sin-DWT coder was 27% more computationally intense then the LP-DWT.

Subjective Quality Assessment

A series of informal comparative listening tests with 11 trained listeners were performed on both coders using a range of different audio sources. In each test the listeners were presented with the original signal, and then three coded versions. In the first series of tests the three coded signals were, the MPEG-1 layer 2 codec at 128 kbps, the LP-DWT at 104 kbps and the Sin-DWT at 104 kbps. For the second series of tests the LP-DWT and Sin-DWT coders were used at 96 kbps, again with the MPEG level 2 coder at 128 kbps. For each audio sample the three coder signals were ordered randomly and the listener was asked to rate the level of distortion introduced by the coding process on a scale of one (1.0) to five (5.0), where five represented imperceptible distortion and one represented destructive and unacceptable levels of distortion. Table 3 summarises the scores produced from these tests.

From the scores gathered in the first test we can say that the Sin-DWT, at 96 kbps, offers quality comparable to MPEG-1 with the average score difference being less than 0.1. The LP-DWT at 96 kbps provides slightly lower quality than MPEG-1, however with an average score still above 4.0 and with only one score significantly lower then 4.0, this coder still offers generally acceptable quality. The second set of tests indicates that the Sin-DWT coder still performs well at 96 kbps, still out performing the LP-DWT at 104 kbps, by virtue of a slightly higher average score and still no score being significantly lower then 4.0. At 96 kbps the LP-DWT coder performs poorly, with scores averaging only 3.6 and few scores approaching the acceptable quality mark of 4.0.

CONCLUSION

The results of the quality comparisons clearly indicates that the Sin-DWT coder offers superior fidelity to the LP-DWT coder at the cost of coding complexity. The primary reason that the Sin-DWT coder

	MPEG layer 2 128 kbps	LP-DWT 104 kbps	Sin-DWT 104 kbps	LP-DWT 96 kbps	Sin-DWT 96 kbps
Female	4.81	4.63	4.76	3.68	4.51
Male	4.84	4.71	4.80	3.91	4.34
Choir	4.73	4.69	4.56	3.57	3.98
Flute	4.79	3.31	4.68	2.98	4.42
Vibraphone	4.80	3.91	4.82	3.05	4.36
Castanets	4.77	4.53	4.68	3.94	4.22
Jazz	4.73	4.50	4.71	4.01	4.51
Pop Music	4.84	4.09	4.81	3.62	4.62
Mean	4.79	4.30	4.73	3.60	4.37

Table 3: Listening Test Results

performs significantly better then LP-DWT coder is due to the constructive nature of the sinusoidal model versus the treatment of the LP-DWT coders residual as an excitation for a filter. Due to this disparity the proportion of residual bits to model bits in the LP-DWT coder must be maintained at a much higher level then in the Sin-DWT coder.

Future work based on these coding structures may focus on two distinct avenues. The reduction of the computational expense of the sinusoidal model, and more effective methods of quantising the residual in the LP-DWT.

REFERENCES

Boland, S. and Deriche, M. (1995), High quality audio coding using mulitpulse lpc and wavelet decomposition, *Proc. of ICASSP*.

Boland, S. and Deriche, M. (1997), A new hybrid lpc-dwt algorithm for high quality audio coding, *Proc. of TENCON*, Vol. 2, pp. 751–754.

Hamdy, K. N., Ali, M. and Tewfik, A. H. (1996), Low bit rate high quality audio coding with combined harmonic and wavelet representations, *Proc. of ICASSP*, pp. 1045–1048.

Ikehara, M. and Nguyen, T. Q. (1997), Time-domain design of linear-phase pr filter banks, *Proc. of ICASSP*, pp. 2077–2080.

Lin, X. and Steele, R. (1993), Subband coding with modified multipulse lpc for high quality audio, *Proc. of ICASSP*, Vol. 1, pp. 201–204.

Mason, M., Boland, S., Sridharan, S. and Deriche, M. (1997), Combined coding of audio and speech signals using lpc and the discrete wavelet transfom, *Proc. of TENCON*, Vol. 2, pp. 747–750.

Mason, M. and Sridharan, S. (1998), Hybrid audio coding using the discrete wavelet transform and vector quantised residuals, *Proc. of ISPACS*, Vol. 2, pp. 606–610.

McAulay, R. J. and Quatieri, T. F. (1986), Speech analysis/synthesis based on a sinusoidal model, *IEEE Trans. Acoust., Speech, Signal Processing* **34**, 744–754.

Moriya, T., Iwakami, N., Jin, A., Ikeda, K. and Miki, S. (1997), A design of transform coder for both speech and audio signals at 1 bit/sample, *Proc. of ICASSP*, Vol. 2, pp. 1371–1374.

MPEG, I. J. (1993), *Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s - Part 3: Audio*, International standard; ISO/IEC 11172, Geneva, Switzerland : ISO/IEC.

MPEG, I. J. (1996), *Generic coding of moving pictures and associated audio*, International standard; ISO/IEC 13818, Geneva, Switzerland : ISO/IEC.

MPEG, I. J. (1999), *Coding of audio-visual objects*, International standard; ISO/IEC 14496, Geneva, Switzerland : ISO/IEC.

Rahman, M. A. and Yu, K.-B. (1987), Total least squares approach for frequency estimation using linear prediction, *IEEE Trans. Acoust., Speech , Signal Processing* **35**(10), 1440–1454.

Rampreshad, S. A. (1998), A two stage hybrid ambedded speech/audio coding structure, *Proc. of ICASSP*, Vol. 1, pp. 337–340.

Soong, F. K. and Juang, B.-H. (1993), Optimal quantisation of lsp parameters, *IEEE Trans. on Speech and Audio Processing*, Vol. 1, pp. 15–24.