LOW DELAY MULTI-LEVEL DECOMPOSITION AND QUANTISATION TECHNIQUES FOR WI CODING

N. R. Chong, I. S. Burnett, J. F. Chicharo

Whisper Laboratories,
University of Wollongong,
Wollongong, NSW, Australia,

ABSTRACT

For efficient coding of speech, it is desirable to separate the slowly and rapidly evolving spectral components to take advantage of their different perceptual qualities. In this paper, we present a multi-level wavelet decomposition mechanism, using low-delay FIR filters, applied to Waveform Interpolation coding. The technique overcomes the substantial delay problems of [2] and identifies a preferred technique for the quantisation of the decomposed surfaces. Phase is shown to be particularly sensitive to the compounding of quantisation errors within the tree-structured transform. The proposed solution involves the use of VDVQ on separately decomposed magnitude/phase surfaces. This approach provides for coarse or no phase quantisation while maintaining high speech quality. The techniques discussed may also be applied to other transforms and to the quantisation of surfaces in the standard Waveform Interpolation coder.

1. INTRODUCTION

The need to efficiently quantise signals to maximise the transmission channel utilisation, while still maintaining high quality, has led to the exploitation of human speech perception. In particular, for high perceptual quality, it is necessary to preserve the correct degree of periodicity of the speech. This result is employed in coders based on pitch-cycle waveforms (known as characteristic waveforms (CW)) such as the Waveform Interpolation (WI) coder [1]. To quantise the speech more efficiently, it is advantageous to isolate the underlying pulse shape from the signal evolution. This requires some mechanism to decompose the evolutionary signal into frequency subbands. The original decomposition method of WI involves simple filtering of the CW surface in the evolution domain. The outcome is a slowly evolving waveform (SEW) representing the periodic component of speech, and a rapidly evolving waveform (REW) containing the random noise-like component. To further decompose the signal, we proposed a multirate digital filter bank approach using finite impulse response (FIR) wavelet filters [2]. In contrast with [6], for this work the wavelet decomposition was performed on oversampled pitch-cycle waveforms, so as to maintain a fixed transmission rate. This technique integrated well into the WI framework, however, it had the disadvantage of incurring long system delays, making the decomposition impractical for real-time applications. The issue of delay was addressed in [3] with the proposal of infinite impulse response (IIR) quadrature mirror filter banks. Substantial delay reductions result from IIR techniques, however, the decomposition proves to have high sensitivity to low-rate quantisation of the surfaces.

In this paper, we present a multi-level decomposition technique using low-delay FIR filters. We retain the non-recursive nature of FIR filters while having the ability to impose a very low delay. Advantages of the proposed decomposition include scalability in quantisation, multi-scale signal evolution analysis, and low delay which makes the decomposition compatible with real-time speech coding. Quantisation of the decomposed surfaces is a non-trivial issue, and therefore, several quantisation and reconstruction techniques are analysed.

The outline of the paper is as follows. A description of the low-delay FIR filters is given in Section 2. The three decomposition methods investigated are outlined in Section 3, with sensitivities of the parameters in Section 4. Section 5 gives a detailed account of possible quantisation techniques, for both magnitude and phase, and their attributes. Reconstruction of the CW surface is described in Section 6, and the criteria for the technique which will offer good quantised speech quality is stated in Section 7. Finally, the conclusions are summarised in Section 8.

2. LOW DELAY FIR FILTERS

The multi-resolution wavelet decomposition based on Quadrature Mirror Filter (QMF) banks has an inherently long system delay and thus has limited application in the field of real-time speech processing. A possibility for reducing the delay is to use filters with lower delay for the inner layers, since in a tree-structured system, the decomposition levels located further inside the tree contribute more significantly to the overall delay than the outer stages.

Usually, for analysis and synthesis FIR filters of length N, the overall system delay is N-1. However, Nayebi *et al*. [4][5] have presented a time-domain approach for the design of the analysis-synthesis systems, in which the system delay can be considered to be relatively independent of the length of the analysis and synthesis filters. For satisfactory results and faster convergence, minimum phase starting filters [5] are more suitable for low-delay systems. Decreasing the system delay results in the reduction of the overall quality of the system filters - in particular, the stopband attenuation. Also, imposing a very low delay on systems with relatively high order filters may not be as effective as imposing a similar delay on systems with lower order filters.

In our wavelet decomposition, we use eight-tap filters with an imposed system delay of one sample. The frequency responses

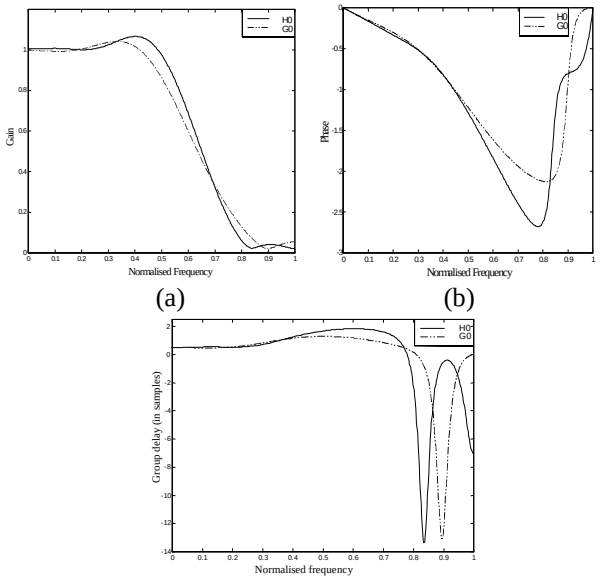

Figure 1. Two-band filter bank with 8-tap filters. System delay is one sample (a) Magnitude Response (b) Phase Response (c) Group Delay

and group delays for $H_0(z)$ and $G_0(z)$, the analysis and synthesis lowpass filters respectively, are given in Figure 1. Note, $G_0(z)$ and $H_1(z)$, the analysis highpass filter, are related by the equation $G_0(z)=H_1(-z)$.

3. DECOMPOSITION OF THE CW SURFACE

The CWs are sampled at a constant rate of 320Hz and are thus oversampled, though not by a constant degree (due to the natural pitch track variation of the speech). Further, since the CWs are not extracted pitch synchronously, or with attention to phase, it is necessary to align the CWs. In this work all CWs were aligned to a common pulse waveform before decomposition so as to remove overall phase offsets. In the time domain such offsets correspond to rotation of the pitch-length waveform. Several methods of decomposition of the aligned CW surface were tested:

1. *Filtering of Time Domain Characteristic Waveforms*
2. *Frequency Domain Techniques:*

   a) *Real/Imaginary*

   The Real and Imaginary components of each CW coefficient are separately decomposed by the filter bank. If required magnitude and phase are calculated on the final surfaces.

   b) *Separate Magnitude/Phase*

   The magnitude and phase of the DFT coefficients of the CWs were separated and filtered individually. To avoid phase wrapping, unity magnitude complex values were used to represent phase.

Note that the surfaces obtained by the two frequency domain techniques have different characteristics. While the highpass surfaces for the Real/Imaginary decomposition exhibit rapid changes in magnitude and phase, the magnitude decomposition of the Separate Magnitude/Phase technique neglects any phase

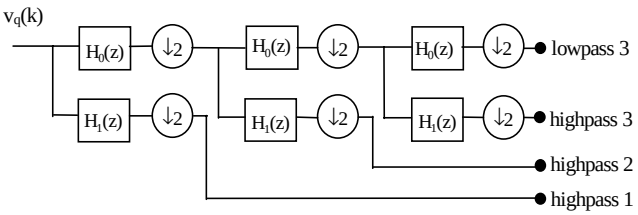

Figure 2. Three-level wavelet decomposition structure where $v_q(k)$ is the evolutionary signal, $H_0(z)$ is lowpass and $H_1(z)$ is highpass.

variations. Thus, a section of speech with slow magnitude evolution but rapid phase evolution will dominate a highpass surface in the Real/Imaginary decomposition, but will be passed by a lowpass filter in the Magnitude/Phase case. Comparison of the decomposed surfaces for the two frequency domain techniques showed that the phase surfaces possessed similar characteristics. The magnitude surfaces were smoother in the case where the magnitude and phase were separated before filtering. When converted to the time-domain, the pulse shape was still visible in the highpass surfaces for the Real/Imaginary decomposition, whereas the surfaces were uncorrelated in the Magnitude/Phase case.

4. PARAMETER SENSITIVITY

Evangelista [6] suggests scalar quantisation schemes for his pitch-synchronous wavelet transform (PSWT) surfaces. Indeed, our results show that very coarse quantisation (2-4 bits) of individual magnitude and phase components maintains high quality output speech. This result suggests that while the exact magnitude and phase values are not significant, the trend of these values is important. This is particularly true with the phase parameter.

To test the sensitivity of phase to quantisation errors, random noise was added to the phase of the surfaces (the original magnitude of the parameters was maintained). The permissible level of Gaussian additive phase noise prior to audible distortion was measured. Of the three innermost phase surfaces (refer Figure 2), highpass 2 was the least sensitive, with added noise of variance 1.0 radians maintaining good performance. Noise with variance 0.6 radians was the maximum allowed before distortion for highpass 3, while noise of variance greater than 0.2 radians added to the phase of the lowpass 3 surface created distortion. Thus the phase of the highpass surfaces can be judged to be significantly less sensitive than that of the lowpass, scale surface. This concurs with the phase assumptions made in the standard WI coder. However, as explained in Section 5.2 the dependence of the Wavelet decomposition on phase and phase surface interrelationships adds a significant dimension to the phase quantisation/representation problem. Such dependencies are particularly relevant when considering the use of VQ schemes for surface quantisation.

5. QUANTISATION

An advantage of the multi-level wavelet decomposition is that the transmission frequency required for the surfaces is defined, due to the decimation. Each surface is sent at a rate

corresponding to its sampling frequency. In addition, the decomposition exhibits potential for higher and variable rate WI coding since bit allocation for the surfaces can be flexible, allowing a more accurate description of perceptually important scales. In this case, the highpass 1 surface did not add significant perceptual detail to the signal and was thus not transmitted. Also, the entropy of the wavelet decomposition is low, with coefficients tightly clustered about zero. This will further increase coding efficiency.

In [6], the PSWT is encoded using Adaptive PCM (APCM) for the highpass surfaces, and an 8-bit uniform quantiser for the lowpass 3 surface. The results given were based on the performance of the PSWT on a single vowel sound. While the paper fails to comment on the performance during an utterance containing voiced and unvoiced speech, the high proposed bit rate of 21 kbits/sec for the transform coefficients alone, or even the reduced rate of 9 kbits/sec, when the two higher rate surfaces were omitted, are impractical for low-rate speech compression. In contrast, here we propose several vector quantisation techniques. These potentially lower the overall coder bit rate to 2.4 kbits/sec. In addition, we add significantly to the work of [6] by discussing the necessary relationships between the decomposed surfaces to provide perceptually accurate reconstruction.

5.1 Magnitude Quantisation

The extraction of pitch-length waveforms results in the need to quantise variable length sequences. A Variable Dimension Vector Quantisation (VDVQ) [7] scheme was applied to quantise the individual surfaces. To quantise the magnitudes, the variable dimension vector is extended to a fixed length and zeros are inserted in frequency bins which do not contain a value. A consequence of the zero-insertion is that in order to correctly train the codebooks, the average value of the incoming vectors must be zero. To achieve this, the mean of the entire training data was subtracted from each value before training. For example, if the variable dimension vector, v1= (1, 2, 3), the chosen fixed length is 6, and v is the entire training data sequence, then, v1→ (1/mean(v), 0, 2/mean(v), 0, 3/mean(v), 0). Codebooks are then trained on these sequences. The resulting fixed-length codebooks are mean-adjusted, sub-sampled, and compared with the variable-length input vector during the codebook search.

Best quantisation results are achieved when the magnitudes are decomposed separately, independent of phase. The decomposed surfaces have reduced dynamic range, enabling smaller codebooks to be used which offer the same quality magnitude-quantised speech as that obtained by using large codebooks with the Real/Imaginary approach. Good results were obtained using an 8-bit codebook for lowpass 3, 4-bits for highpass 3 and 3-bits for the highpass 2 surface.

Application to Standard WI Coder:

The VDVQ techniques described have also been applied to the standard WI coder for the SEW/REW surfaces. This technique enables efficient quantisation of complete SEW magnitudes, removing the requirement for reduced-bandwidth spectral bin approaches. This makes the implementation of scalable and higher rate WI coders more practicable, since the low-pass nature of SEW quantisation in previous coders has been found to be an inherent limitation.

5.2 Phase Quantisation

The original CW and decomposed phase surfaces were compared. The lowpass, scale surfaces have similar phase characteristics to that of the original CW phase. This is particularly true in slowly evolving sections of the speech. In these regions, due to the wavelet decomposition, the highpass surfaces also retain significant phase characteristics of the original surface. Experiments indicate that while the actual CW phase before decomposition takes place is not critical, the phase relationships between the decomposed surfaces is important for crisp output speech. In other words, relative phase is very sensitive within the decomposition. Errors in the phase of the surfaces compound and can produce very rough sounding speech. Hence, to minimise distortion, the phase relationship between the surfaces needs to be preserved and the quantisation method is required to include this constraint.

Phase Quantisation with VDVQ:

Phase sensitivity tests showed that random phase cannot simply be applied to the highpass surfaces without causing considerable distortion. Therefore, VDVQ was employed to quantise the phase of the coefficients. Some adjustments, however, were made to the VDVQ training algorithm to cope with phase wrapping. For the Frequency domain Real/Imaginary decomposition, magnitude and phase were calculated for each coefficient. The absolute phase was then converted to a unity magnitude phase vector for training. This unity magnitude was rectified after each codebook update.

One method of maintaining the phase relationships is to place limits on the phase vector selection for the highpass surfaces once the phase has been chosen for the lowpass surface, rather than quantising all phases independently. As an alternative approach, the phase of the CW surface may be quantised. This, however, results in multiple transmissions of the phase, which takes advantage of neither the lower-rate surfaces of the wavelet decomposition nor human perception of phase.

5.3 Time-Domain Quantisation

VDVQ can also be used to encode the surfaces formed by decomposing the time-domain CWs. However, a difficulty arises since phase information is still contained in the surfaces, and is not able to be controlled explicitly. This results in noisy output speech even with large codebooks.

6. RECONSTRUCTION

The reconstruction techniques for the three decomposition methods are:

1. Reconstructing Time Domain Characteristic Waveforms

Reconstruction of the CW surfaces follows a direct reversal of the decomposition, in which the innermost lowpass and highpass surfaces are upsampled, filtered by the synthesis mirror filters, then combined to form the lowpass surface of the next highest level.

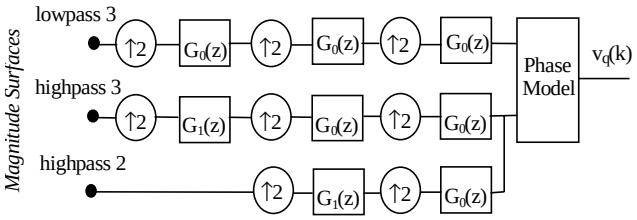

Figure 3. Structure of three-level wavelet reconstruction using the Frequency Domain: Separate Magnitude/Phase method, where $v_q(k)$ is the evolutionary signal, $G_0(z)$ is lowpass and $G_1(z)$ is highpass.

2. *Frequency Domain Techniques:*

   a) Real/Imaginary

   Some phase models were experimented with to represent the phase of the individual surfaces, but with poor results. To improve the quality, transmission of quantised phase is required.

   b) Separate Magnitude/Phase

   Magnitudes and phases are reconstructed separately, with each surface reconstructed and upsampled to the original sampling frequency as shown in Figure 3. By reconstructing all surfaces individually, the phase of one surface may be adjusted, without affecting the phase of other surfaces. This allows, for example, the application of random phase to the third level highpass surface without affecting the phase of higher surfaces. This makes the use of phase models effective, and the need to quantise phase, unnecessary.

In existing coders, such as the waveform interpolation and sinusoidal coders, phase models are used [8]; some requiring a voiced/unvoiced decision. The approach taken in the Separate Magnitude/Phase reconstruction was to apply random phase to the recomposed highpass surfaces, while a linear phase model was used for the lowpass surface. This phase linearity was randomised during periods of unvoiced speech. The measure of the level of disruption of the phase linearity was found by comparing the energy contained in the lowpass surface to that of the corresponding highpass surface. This corresponds to a dynamic voiced/unvoiced mix measurement. In addition, the amount of added noise can be weighted by the inverse of the LPC power spectrum, to ensure that formant frequencies are not distorted.

7. RESULTS

For best results, the magnitude and phase of the characteristic waveform surface were decomposed separately. Using this method, magnitude codebooks could be small and phase did not require quantisation. By choosing a good phase model and adapting this to control the amount of CW correlation required, good quality output speech could be achieved with minimal buzziness or reverberation.

While phase is not perceptually important in speech, the inter-surface phase relationships of the decomposition require accurate transmission for good reconstruction. Due to the conjugate mirror symmetry of the filter banks, quantisation errors result in non-cancellation of aliased components. These errors compound through the tree-structure of the transform, causing any errors due to the encoding of the inner surfaces to be magnified greatly.

8. CONCLUSION

The wavelet decomposition method using low-delay FIR filters allows a multi-resolution analysis of the signal evolution. The complications associated with the quantisation of the decomposed surfaces have been identified. In particular, phase is important in the wavelet decomposition. The key to good quantised speech lies in the ability to separate the magnitude and phase in the decomposition and reconstruction. This removes the need to quantise the phase of the surfaces and hence the compounding phase quantisation errors which lead to perceptually annoying CW phase behaviour. Further through exploiting separation of magnitude and phase it is possible to take advantage of well-established perceptual qualities of voiced and unvoiced sounds.

The wavelet decomposition and quantisation techniques discussed in this paper may also be applied to other transforms, such as those operating pitch-synchronously in [9].

9. ACKNOWLEDGEMENTS

We wish to acknowledge the useful suggestions made by T.P.Barnwell III at ICASSP '98. N.R. Chong is in receipt of an Australian Postgraduate Award (Industry) and a Motorola (Australia) Partnerships in Research Grant. Whisper Laboratories is funded by Motorola, and the Australian Research Council.

10. REFERENCES

[1] W.B. Kleijn, J. Haagen, “Waveform Interpolation for Coding and Synthesis”, in Speech Coding and Synthesis, edited by W. B. Kleijn and K.K. Paliwal, Elsevier 1995.

[2] N.R.Chong, I.S.Burnett, J.F.Chicharo, M.M.Thomson, “Use of the Pitch Synchronous Wavelet Transform as a New Decomposition Method for WI”, *Proc. Int Conf. Acoustics, Speech, Sig. Processing*, Seattle, USA, vol. 1., pp 513-516, May 1998.

[3] N.R.Chong, I.S.Burnett, J.F. Chicharo, "An Improved Decomposition Method For WI Using IIR Filter Banks", to be published in *Proc. 5th Int. Conf. Spoken Language Processing*, Sydney, Australia, Dec. 1998.

[4] K.Nayebi, T.P.Barnwell, III, M.J.T.Smith, "Time Domain Filter Bank Analysis: A New Theory", *IEEE Trans. Sig. Proc.*, vol. 40, no.6, June. 1992, pp. 1412-1429.

[5] K.Nayebi, T.P.Barnwell, III, M.J.T.Smith, "Low Delay FIR Filter Banks: Design and Evaluation", *IEEE Trans. Sig. Proc.*, vol. 42, no.1, Jan. 1994, pp. 24-31.

[6] G.Evangelista, “Pitch-Synchronous Wavelet Representations of Speech and Music Signals”, *IEEE Trans. Sign. Process.*, vol. 41, no. 12, pp. 3313-3329, 1993

[7] Das, A., Rao, V. R., Gersho, A., "Variable-Dimension Vector Quantisation", *IEEE Signal Processing Letters*, vol.3, no.7, pp. 200-202, July 1996.

[8] S.Torres, F.Casajus-Quiros, "Vocal System Phase Coder for Sinusoidal Speech Coders", *Electronics Letters*, vol. 33, no. 20, pp1683-1685, Sept. 1997.

[9] H.Yang, W.Kleijn, "Pitch-Synchronous Subband Representation of the Linear Prediction Residual of Speech", *Proc. Int Conf. Acoustics, Speech, Sig. Processing*, Seattle, USA, May 1998.