USE OF THE PITCH SYNCHRONOUS WAVELET TRANSFORM AS A NEW DECOMPOSITION METHOD FOR WI

N. R. Chong [*], I.S. Burnett [*], J.F. Chicharo [*] and M.M. Thomson [†]

[*] Whisper Laboratories, The Institute for Telecommunications Research,
University of Wollongong, NSW, Australia
[†] Motorola Australian Research Centre,
Botany, NSW, Australia

ABSTRACT

A new characteristic waveform decomposition method based on wavelets is proposed for the Waveform Interpolation (WI) paradigm. In WI, pitch-cycle waveforms are filtered in the evolution domain to decompose the signal into two waveform surfaces, one characterising voiced speech and a second representing unvoiced speech. The slow roll-off of FIR filters leads, however, to a significant inter-relationship between the decomposed surfaces. Here we present the Pitch Synchronous Wavelet Transform (PSWT) as an alternative decomposition mechanism. Filtering is again performed in the evolutionary waveform domain, producing characteristic surfaces at several resolutions. This multi-scale characterisation leads to more flexible quantisation of parameters, especially at higher rates than WI's 2.4kb/s. FIR filters are replaced in the Wavelet filter bank by causal, stable IIR filters which achieve significant delay reductions over their FIR counterparts. Furthermore, IIR filters track the dynamic aspects of the evolutionary surfaces faster, overcoming problems existing in the current WI decomposition.

1. INTRODUCTION

Waveform Interpolation (WI) allows the efficient compression of signals by exploiting the perceptual importance of speech characteristics [1][2]. In recent WI coders, pitch-cycle waveforms (characteristic waveforms(CW)) are extracted from the linear prediction residual signal. These CWs are filtered in the evolution domain to decompose the signal into a slowly evolving waveform (SEW) characterising voiced speech and a rapidly evolving waveform (REW) representing noise-like unvoiced speech. This decomposition is motivated by human perception and results in high coding efficiency. However, our previous work [3] showed that the REW magnitude surface contains significant components of the SEW magnitude. This highlighted the necessity for an improved decomposition method.

The Pitch Synchronous Wavelet Transform (PSWT) [4] offers a solution. The similarities between the PSWT and WI are remarkable. In the PSWT, pitch-length segments of speech are extracted to form a 2-D representation of the signal in exactly the same manner as is currently performed in the WI coder.

Deficiencies, however, exist in [4] with the lack of consideration of the extraction and alignment of prototypes when performed in real-time. We adapt the PSWT to maintain a fixed sampling rate and oversample prototypes as in WI, as opposed to the critical sampling proposed. This guarantees a fixed rate of parameters which is appropriate for fixed frame-rate encoding. The prototypes are then filtered in the evolution domain. However, the filtering is not carried out by a 20Hz cut-off FIR filter as in the SEW/REW decomposition, but instead with dilated and translated wavelets. Here, we link this idea to the WI paradigm.

The PSWT thus gives an alternative description of signal evolution and offers significant advantages due to its flexibility in the filter (wavelet) design and time-scale representation. The perfect reconstruction properties of biorthogonal wavelets and the dynamic capture achieved from sharper filters render the wavelet decomposition very appealing. The sharp roll-off achievable with, in particular, IIR filter banks contrasts with the FIR filters used in the SEW/REW approach [1]. Further, the PSWT facilitates scalability to higher and variable bit rates through flexible bit allocation for the frequency subbands.

The outline of this paper is as follows: the underlying motivation for decomposition is given in Section 2. Sections 3 and 4 introduce the PSWT and its application to WI coders. Surfaces resulting from the decomposition of voiced and unvoiced sounds using biorthogonal FIR wavelets are shown in Section 5, with discussions on the spectral quantisation in Section 6. The disadvantage of FIR filters is described in Section 7 and a causal, stable, low-delay IIR solution is presented in Section 8.

2. MOTIVATION FOR DECOMPOSITION OF SPEECH

Voiced and unvoiced speech have fundamentally different quantisation requirements for perceptually accurate reconstruction [2]. Unvoiced speech requires a high time resolution but only low quantisation accuracy. In contrast, voiced speech requires a more precise description, but has a low evolution bandwidth and can be transmitted accurately at low bit rate. By decomposing the signal into voiced and unvoiced components, the various speech sounds can be quantised with a precision which is based on the nature of human speech and perception. This results in high coding efficiency.

3. THE PSWT

The PSWT is an extension to the Multiplexed Wavelet Transform [5] in which a variable pitch period is employed. The pitch information, stored in P(k), is exploited to represent the

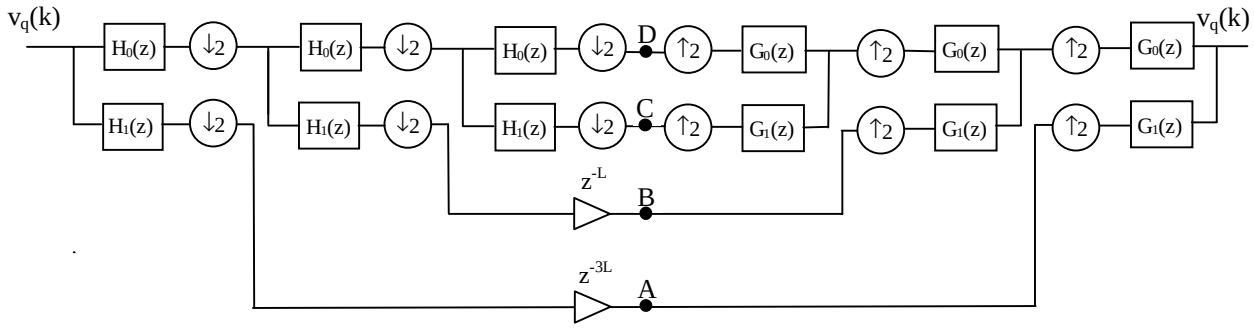

Figure 1. Multirate realisation of the Discrete Wavelet Transform and its inverse for degree L perfect reconstruction FIR QMFs.

speech in terms of a periodic trend and fluctuations over this trend. The Discrete Wavelet Transform is performed, correlating the evolutionary signal, $v_q(k)$, where q=0,1,…..P(k)-1, with dilated and translated versions of a unique analysing wavelet, $\varphi_{n,m}(k)$ where index n=1,2,…,N represents scale and m=0,1,2,…,M represents time shift.

$$v_q(k) = \sum_{n,m} V_{n,m,q}\varphi_{n,m}(k) \quad \ldots(1)$$

where,

$$V_{n,m,q} = \sum_k v_q(k)\varphi_{n,m}(k) \quad \ldots(2)$$

To obtain perfect reconstruction of a signal, one must use either orthogonal or biorthogonal wavelets. Orthogonal wavelets have the advantage of increased algorithm simplicity since the analysis bank is simply inverted by its transpose. However these wavelets lack symmetry. For this reason, we turn to biorthogonal wavelets which possess linear phase and can decompose and reconstruct an input signal with no aliasing or distortion. This is achieved by designing the generating filters to form a quadrature mirror filter (QMF) pair.

4. WAVELET DECOMPOSITION

The PSWT is similar to the operation carried out in the SEW/REW decomposition, involving simple filtering of the DFT coefficients. This similarity enables wavelet decomposition to be easily applied to the WI paradigm. The main difference here is that the filters used are perfect reconstruction wavelets, dilated and shifted to form a non-uniform filter bank. Hence, deviations from periodicity are detected, isolating the periodic trend and at the same time characterising the aperiodic behaviour at several scales.

The original motivation of our work was speech enhancement. While the SEW/REW method allows efficient coding of speech, the decomposition of narrowband and broadband noise [3] showed that although the underlying concept was favourable, it did not produce the desired separation of noise sources. However, the multiresolution wavelet decomposition allows preferential quantisation and coding to achieve high quality performance. It also leads to possible speech enhancement mechanisms through its superior separation of the evolutionary frequency band.

The basic objective is to separate the characteristic waveform surface into uncorrelated frequency subbands (in the evolution domain). Since each subband signal occupies only a fraction of the original frequency bandwidth it can be downsampled to the Nyquist rate without loss of information. A diagram of the maximally decimated analysis/synthesis system is shown in Figure 1 where $H_0(z)$, $G_0(z)$ are scaling sequences (lowpass characteristic) and $H_1(z)$, $G_1(z)$ are wavelet sequences (highpass characteristic). In order to cancel aliasing, the filters are related as follows:

$$G_0(z) = H_1(-z) \quad \ldots(3)$$

$$G_1(z) = -H_0(-z) \quad \ldots(4)$$

The difference between two approximations at the resolutions 2^{n+1} and 2^n is extracted by decomposing the signal using a wavelet basis and the resulting signal is transmitted. Thus, for a system comprising n decomposition levels, n+1 signals are transmitted, these being n detail signals (obtained by highpass filtering) as well as the approximation signal (obtained by lowpass filtering) of the final stage. In order to synchronise the signals, extra delays are added in particular paths.

Initially, finite impulse response (FIR) biorthogonal wavelets with a lowpass decomposition filter effective length of 8 and a highpass decomposition filter effective length of 4 were used. The magnitude response is shown in Figure 2. The analysis/synthesis pair causes a delay of 7 at the coarsest scale. Both voiced and unvoiced sounds were decomposed to three levels. Surfaces at each of the points A to D depicted in Figure 1 are shown in Figures 3 and 4.

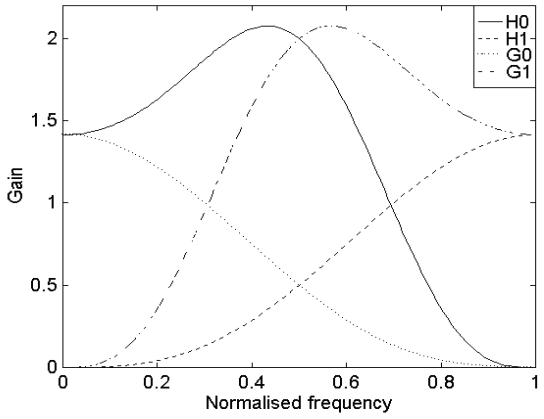

Figure 2. Magnitude Response of Biorthogonal FIR QMF bank used in decomposition.

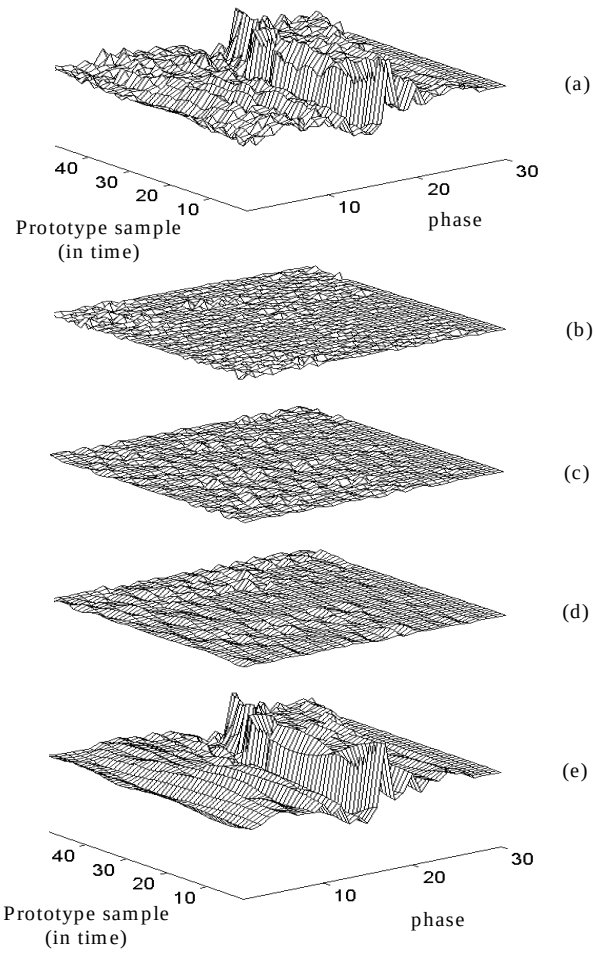

Figure 3. Decomposition of voiced sound "oo" taken from the word "foolish". (a) Prototype waveforms (b) Outputs at point A, (c) B (d) C and (e) D as shown in Figure 1.

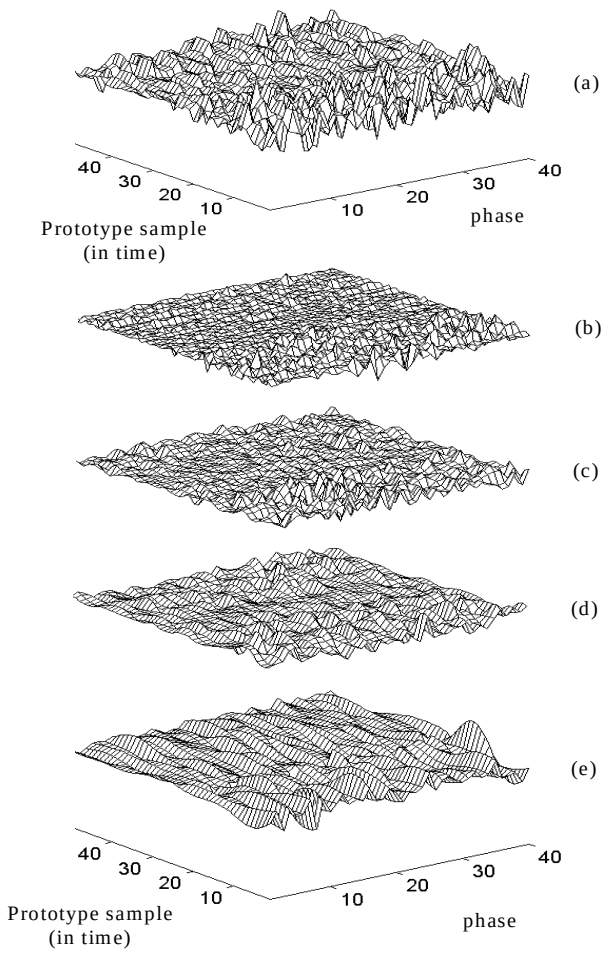

Figure 4. Decomposition of unvoiced sound "sh" taken from the word "foolish". (a) Prototype waveforms (b) Outputs at point A, (c) B (d) C and (e) D as shown in Figure 1.

5. RESULTS

As depicted in Figure 3, during voiced speech, the signal is very slowly evolving. This results in most of the energy passing through the lowpass filters, producing a smooth surface (Fig. 3(e)). Only a very small amount filters into the highpass outputs. The three highpass outputs can be described as three REW surfaces. However, the advantage of these additional surfaces lies in the different scales of information available, caused by the decimation. The first level highpass output (Fig. 3(b)) is very flat, with amplitude approximately 10 times smaller than that in Fig 3e, suggesting that this information is of lesser significance and these coefficients may not need to be quantised and transmitted at low rates such as 2.4kb/s. Note that these surfaces have been upsampled to the CW sampling rate.

For the case of the unvoiced sound “sh”, the prototype surface is very irregular (Fig. 4(a)). The decomposition shows the energy being spread throughout all surfaces, both lowpass and highpass, (Figs. 4(b)-(e)).

6. QUANTISATION OF THE SURFACES

In the same way that the SEW/REW decomposition lead to increased coding efficiency, the wavelet decomposition also offers advantages. Currently in the WI coder, a trained codebook is used for the quantisation of the SEW magnitude information [6] which is transmitted once per frame. The REW magnitude spectrum is quantised using Chebyshev polynomial techniques and is transmitted multiple times per frame. The magnitude spectrum of the lowpass output of the PSWT can be quantised using the techniques for the SEW surface. Likewise, the magnitude spectrum of the highpass highpass outputs can be quantised in the same way as the REW. However, due to the decimation used in the PSWT, the frequency that the surfaces should be transmitted is more defined. Each surface is sent at a rate corresponding to its sampling frequency. The surface with the lowest resolution can be transmitted once per frame, the next lowest twice per frame, and the next 4 times per frame.

The phase spectrum of the lowpass output has similar characteristics to the SEW phase spectrum and is quantised using a phase model derived from natural speech. Random phase is used for the highpass phase spectra.

The PSWT allows scalability in quantisation since bit allocation for the surfaces can be flexible, allowing a more accurate description of perceptually important scales. Also, scales which are relatively insignificant may be upsampled to a common frequency and combined. Further, PSWT decomposition exhibits potential for WI coding at higher and variable bits rates. In addition, by moving to infinite impulse response (IIR) wavelet filters, the dynamic aspects of speech can be tracked faster, overcoming a current problem of the SEW/REW decomposition.

7. FIR AND CAUSAL, STABLE IIR QMF BANKS FOR DECOMPOSITION

A significant disadvantage of the decomposition using FIR filters is the delay. At present, using the biorthogonal FIR filters described above, a delay of 49 (6 frames of 8 prototypes/frame) is incurred for 3 levels.

This inherent disadvantage of FIR filters is of major concern in speech coding applications. The delay is far greater than desired, increasing exponentially with additional decomposition levels, and must be reduced. A solution lies with IIR QMF banks. Compared to FIR filters, IIR filters offer computational and spectral magnitude performance advantages.

Literature on IIR QMF banks give solutions for filter banks possessing causal, unstable synthesis filters [7]. These can be implemented as stable, anti-causal filters which is beneficial for image coding, however inappropriate for real-time applications such as speech coding.

Basu *et al.* [7] described a design for causal, stable IIR filters. Following this design, we derived filter banks from stable IIR prototype filters. These give perfect reconstruction and incur a delay of only 1 sample for the analysis/synthesis pair (7 samples for 3 levels). Bessel filters, out of all the standard filter types, were found to produce the best results. Using the variables defined in [7] and setting $H_0(z)$ to be a 2^{nd} order half-band Bessel filter, we chose $r = 5z^2 + 4z + 1$, and let p and q be 2^{nd} and 3^{rd} order respectively. The surfaces of the final decomposition level are shown in Figure 5, and can be compared to the FIR outputs of Figure 3 (d) and (e). The gradual roll-off characteristic of the Bessel filter ensures smoothing of the surfaces in a similar manner to the FIR biorthogonal filters. This smoothness allows quantisation to be performed using the minimum number of bits.

8. CONCLUSION

The PSWT offers an alternative description of signal evolution. The similarities between the multirate wavelet decomposition and the existing SEW/REW decomposition method make it easily applicable to the WI paradigm. Its multi-resolution analysis enables further characterisation of the approximation and detail signals. However, FIR filters are impractical for use due to the substantial delay incurred. A solution lies with causal, stable IIR QMF banks which incur a very low-delay and produce surfaces with smooth characteristics for efficient quantisation. IIR filters also better follow the dynamic aspects of speech, contrasting the problems of the more slowly reacting SEW/REW decomposition.

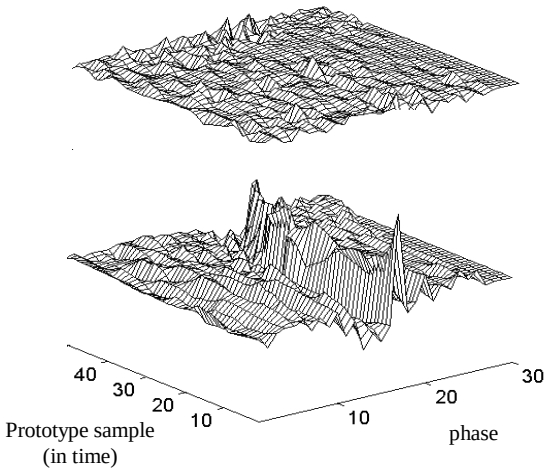

Figure 5. Decomposition of voiced sound "oo" using 2^{nd} order Bessel prototype filter. Outputs at point (a) C and (b) D defined in Figure 1.

9. ACKNOWLEDGEMENTS

Miss N.R. Chong is in receipt of an Australian Postgraduate Award (Industry) and is the recipient of a Motorola (Australia) Partnerships in Research Grant. Whisper Laboratories is funded by Motorola, the Australian Research Council and ATERB.

10. REFERENCES

[1] W.B. Kleijn and J. Haagen "Transformation and Decomposition of the Speech Signal for Coding", *IEEE Signal Processing letters*, vol. 1 no.9, pp. 136-138, 1994.

[2] 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.

[3] N.R.Chong, I.S.Burnett, J.F.Chicharo, M.M.Thomson, "The Effects of Noise on the Waveform Interpolation Speech Coder", *Proceedings of IEEE TENCON'97*, Brisbane, Australia, Dec., 1997.

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

[5] G.Evangelista, "Comb and Multiplexed Wavelet Transforms and their Applications to Signal Processing", *IEEE Trans. Sign. Proc.*, vol. 42, no. 2, pp. 292-303, 1994.

[6] W.B.Kleijn, Y.Shoham, D.Sen, R.Hagen, "A Low-Complexity Waveform Interpolation Coder:, *Proceedings of the Int. Conf. Acoust. Speech Sign. Process*, vol. 1, pp. 212-215, 1993.

[7] S.Basu, C-H.Chiang, H-M.Choi, "Wavelets and Perfect Reconstruction Subband Coding with Causal Stable IIR Filters, *IEEE Trans. Circuits and Systems II*, vol.42, no. 1, pp. 24-38, 1995.