

Faculty of Informatics

Faculty of Informatics - Papers

University of Wollongong

Year 2000

A new waveform interpolation coding
scheme based on pitch synchronous
wavelet transform decomposition

N. R. Chong*

I. Burnett[†]

J. F. Chicharo[‡]

*University of Wollongong, n.chong@st.elec.uow.edu.au

[†]University of Wollongong, ianb@uow.edu.au

[‡]University of Wollongong, chicharo@uow.edu.au

This article was originally published as: Chong, NR, Burnett, I & Chicharo, J, A new waveform interpolation coding scheme based on pitch synchronous wavelet transform decomposition, IEEE Transactions on Speech and Audio Processing, May 2000, 8(3), 345-348. Copyright IEEE 2000.

This paper is posted at Research Online.

<http://ro.uow.edu.au/infopapers/17>

Correspondence

A New Waveform Interpolation Coding Scheme Based on Pitch Synchronous Wavelet Transform Decomposition

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

Abstract—This correspondence uses a pitch synchronous wavelet transform (PSWT) as an alternative characteristic waveform decomposition method for the waveform interpolation (WI) paradigm. The proposed method has the benefit of providing additional scalability in quantization than the existing WI decomposition to meet desired quality requirements. The PSWT is implemented as a quadrature mirror filter bank and decomposes the characteristic waveform surface into a series of reduced time resolution surfaces. Efficient quantization of these surfaces is achieved by exploiting their perceptual importance and inherent transmission rate requirements. The multiresolution representation has the additional benefit of more flexible parameter quantization, allowing a more accurate description of perceptually important scales, especially at higher coding rates. The proposed PSWT-WI coder is very well suited to high quality speech storage applications.

Index Terms—Multiresolution, waveform interpolation, wavelet transform.

I. INTRODUCTION

This correspondence addresses the issue of speech compression for the storage of voice information. Speech coders based on the waveform interpolation (WI) paradigm allow 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 LP residual, aligned, and then 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. In noisy environments, further frequency subband separation is advantageous in order to isolate undesired noise components. In addition, the ability to improve the speech quality by allowing some components of the signal to be coded more accurately is desirable. This stimulated the motivation for an alternative decomposition method to decompose the CW evolution into multiple subbands.

In this correspondence, we investigate the use of the pitch synchronous wavelet transform (PSWT) [3] as a solution. The proposed technique employs wavelets, dilated and shifted to form a nonuniform filter bank, to create an alternative description of signal evolution. Application of the PSWT to WI offers significant advantages due to its perfect reconstruction properties and multi-scale decomposition of the evolving CW surface. The time-scale analysis provides the potential for improved processing of the input speech and enables scalability to higher and variable bit rates through flexible bit allocation for the frequency subbands.

Manuscript received May 14, 1998; revised June 19, 1999. N. R. Chong was supported by the Australian Postgraduate Award (Industry) and a Motorola (Australia) Partnerships in Research Grant. Whisper Laboratories is funded by Motorola, the Australian Research Council, and ATERB. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peter Kroon.

The authors are with Whisper Laboratories, Telecommunications and Information Technology Research Institute, University of Wollongong, NSW, Australia (e-mail: n.chong@st.elec.uow.edu.au).

Publisher Item Identifier S 1063-6676(00)03451-9.

The outline of the paper is as follows. Section II describes the wavelet transform in detail, extending its application to the evolution domain of the CW surface to take advantage of the quasiperiodicity of speech. The decomposition of voiced and unvoiced sounds is analyzed in Section III, with the method for the quantization of these surfaces outlined in Section IV. Finally, Section V concludes the paper.

II. THE PSWT DECOMPOSITION

The PSWT is similar to the SEW/REW decomposition, involving simple filtering of the CW evolution. This similarity allows the technique to be easily applied to the WI paradigm. However, in contrast to [3], we perform our decomposition on the DFT coefficients, rather than on the time-domain waveforms, so perceptual knowledge may be incorporated in the quantization techniques. We also 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 PSWT can be viewed as performing the discrete wavelet transform in the evolution domain. This involves exploiting the extracted pitch information, stored in $P(k)$, to form a CW surface. The evolutionary waveform, $v_q(k)$, where $q = 0, 1, \dots, P(k) - 1$, is correlated with dilated and translated versions of a unique analysing wavelet, $\psi_{n,m}(k)$

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

where index $n = 1, 2, \dots, N$ represents scale and $m = 0, 1, 2, \dots, M$ represents time shift, and

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

The transfer function of the mother wavelet, $\Psi_{n,0}(\omega)$, is obtained from the lowpass and highpass filter transfer functions, H_0 and H_1 respectively, as follows [3]:

$$\Phi_{n,o}(\omega) = \prod_{k=0}^{n-1} H_0(e^{j2^k \omega}) \quad (3)$$

$$\Psi_{n,0}(\omega) = H_1(e^{j2^{n-1} \omega}) \Phi_{n-1,0}(\omega). \quad (4)$$

Translation of the mother wavelet, given by

$$\Psi_{n,0}(k) = \text{IDFT}[\Psi_{n,0}(\omega)] \quad (5)$$

generates the wavelet sequences

$$\Psi_{n,m}(k) = \psi_{n,0}(k - 2^n m). \quad (6)$$

The objective is to separate the characteristic waveform surface into uncorrelated frequency subbands (in the evolution domain). A diagram of the maximally decimated analysis/synthesis system is shown in Fig. 1 where $H_0(z)$, $G_0(z)$ are scaling sequences (lowpass characteristic)

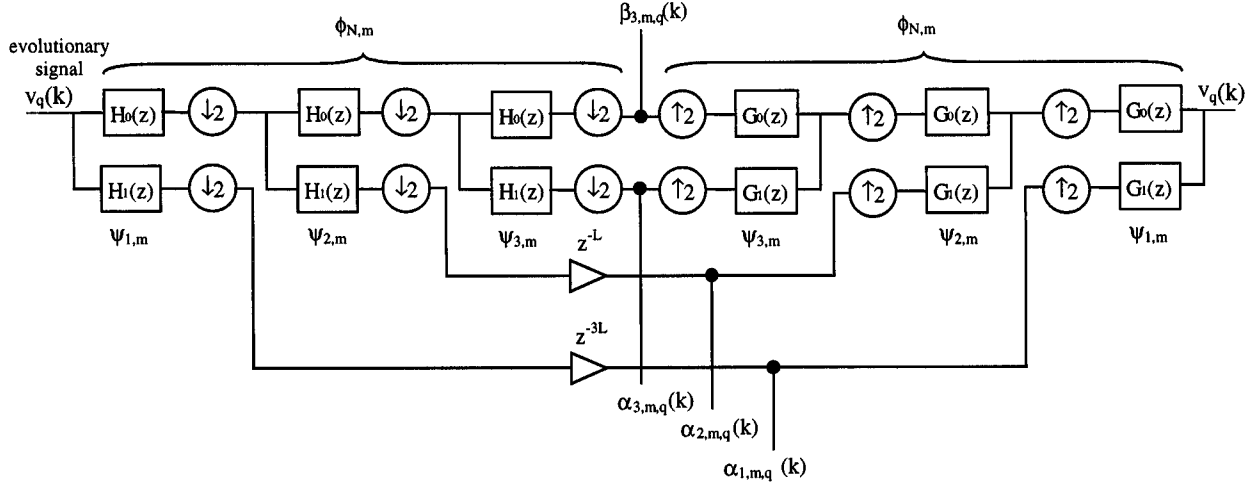


Fig. 1. Realization of the PSWT and its inverse to three decomposition levels.

and $H_1(z)$, $G_1(z)$ are wavelet sequences (highpass characteristic). In order to cancel aliasing, the filters are related as follows:

$$G_0(z) = H_1(-z) \quad (7)$$

$$G_1(z) = -H_0(-z). \quad (8)$$

The signals capturing the periodic nature of speech, called approximation signals, are obtained by correlating the evolutionary signal with scaling sequences, $\phi_{N,m}$. The difference between two approximation signals at the resolutions 2^{n+1} and 2^n is decomposed using a wavelet basis, $\psi_{n,m}$, and the resulting detail signal is transmitted. Thus, for a system comprising n decomposition levels, $n+1$ signals are transmitted, these being n detail signals as well as the approximation signal of the final stage. Higher detail can be achieved by increasing the number of stages in the filter bank implementation. However, this is at the cost of increased system delay. Note that delay is a critical issue for real-time applications, however, in the case of voice storage it is less significant.

For a N -level decomposition, the final approximation (lowpass) surface

$$r_N(k, q) = \beta_{N,m,q}(k) \phi_{N,m}(k)$$

with

$$\beta_{N,m,q}(k) = \sum_k v_q(k) \phi_{N,m}(k) \quad (9)$$

represents the periodic trend, while each detail (highpass) surface,

$$w_n(k, q) = \sum_m \alpha_{n,m,q}(k) \psi_{n,m}(k)$$

with

$$\alpha_{n,m,q}(k) = \sum_k v_q(k) \psi_{n,m}(k) \quad (10)$$

represents the fluctuations at scale 2^n . The sum of these contributions results in the CW surface

$$s(k, q) = \sum_{n=1}^N w_n(k, q) + r_N(k, q). \quad (11)$$

In order to synchronize the signals, extra delays are added in some paths, as indicated in Fig. 1. The total delay incurred for the PSWT decomposition and reconstruction is $z^{(2^N-1)L}$ where L is the combined delay of the analysis/synthesis pair.

In order to obtain perfect reconstruction, i.e., no aliasing, amplitude or phase distortion, either orthogonal or biorthogonal wavelets must be used. The orthogonal solution has the advantage of design simplicity. However, these wavelets lack symmetry and therefore possess nonlinear phase. As a result, we choose finite impulse response (FIR) biorthogonal wavelets derived from the biorthogonal spline wavelet family for the quadrature mirror filter (QMF) bank.

III. ANALYSIS OF THE DECOMPOSED SURFACES

To illustrate the decomposition of the CW into frequency subbands we will use biorthogonal wavelets determined from the filters, $H_0(z)$ and $H_1(z)$, having effective lengths of eight and four, respectively. These filters possess adequate spectral characteristics while incurring a single-level system delay of seven. Further improvement of spectral performance results in increased delay; the next highest order wavelet in the biorthogonal family incurs a system delay of 11. The analysis filter transfer functions are

$$H_0(z) = 0.0663 - 0.1989z^{-1} - 0.1547z^{-2} + 0.9944z^{-3} \\ + 0.9944z^{-4} - 0.1547z^{-5} - 0.1989z^{-6} + 0.0663z^{-7} \quad (12)$$

$$H_1(z) = -0.1768z^{-2} + 0.5303z^{-3} - 0.5303z^{-4} + 0.1768z^{-5}. \quad (13)$$

The resulting wavelet and scaling sequences are shown in Fig. 2.

A three-level PSWT decomposition was performed on both voiced and unvoiced sounds. The decomposed surfaces are shown in Figs. 3 and 4.

Due to the quasiperiodicity of voiced speech, most of the energy appears in the residue, $r_3(k, q)$, producing a smooth surface with a strong correlation between adjacent characteristic waveforms [Fig. 3(e)]. This residue contains the underlying pulse shape evident in the CW surface. In comparison, the detail surfaces, $w_n(k, q)$ are very flat and contain only a very small amount of energy.

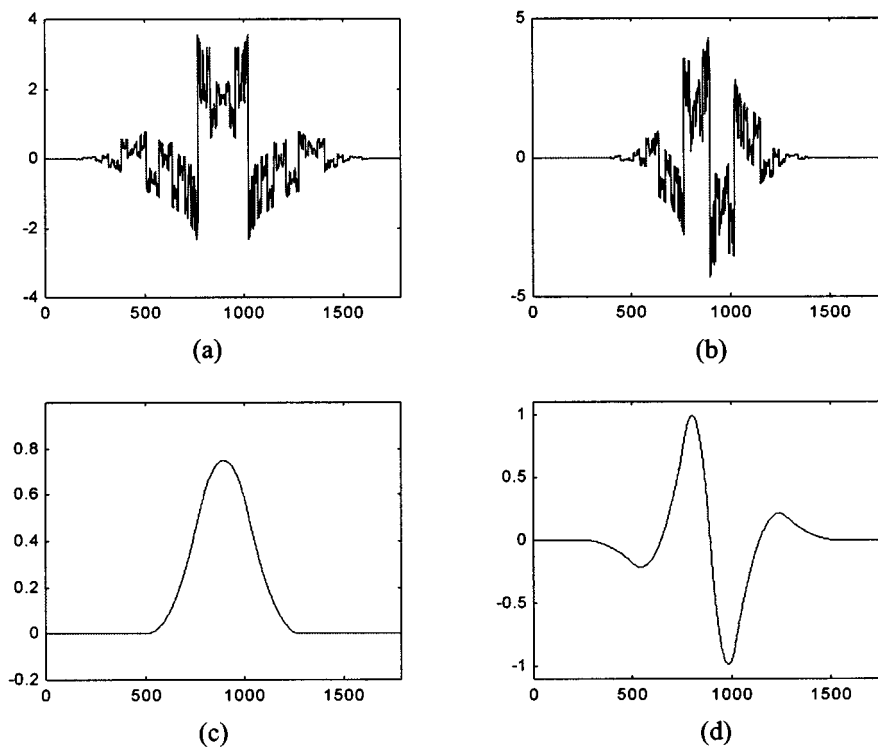


Fig. 2. Scaling and wavelet sequences: (a) analysis scaling function, (b) analysis wavelet function, (c) synthesis scaling function, and (d) synthesis wavelet function.

Although, it may seem less beneficial to have three REW-like surfaces, the advantage of multiple decimated surfaces of rapid evolution lies in the different scales of information available; all surfaces are required for perfect reconstruction. At low rates, these surfaces may be regarded as contributing little perceptual importance, and thus may be coarsely quantized or even eliminated. However, at higher rates, the PSWT offers the flexibility and scalability to define the amount of waveform detail required to achieve the desired quality. In addition, since the faster evolving components have been decomposed into different resolutions, noise suppression techniques may be effectively applied to the surfaces if required.

Subjective analysis has shown that components with lower evolution frequencies possess greater perceptual significance than those of higher evolution frequencies. Hence, the lowest resolution frequency band is awarded the highest quantization accuracy, with each subsequent sub-band receiving less. The lack of energy decomposed into the surface $w_1(k, q)$ [Fig. 3(b)] suggests that these coefficients may be discarded at low rates such as 2.4 kb/s. Note that these surfaces have been upsampled to the original sampling rate.

For the case of the unvoiced sounds, the CW surface is very irregular [Fig. 4(a)]. Energy is decomposed by the PSWT to all frequency sub-bands [Figs. 4(b)–(e)], with no distinctive characteristic existing in any particular surface, as for the voiced case. The highpass decomposition surfaces will be especially important for extracting or enhancing certain features, for example, when the speech is corrupted by background noise.

The delays incurred can be large for multiple decomposition levels. For the filters used above, a delay of 49 (seven frames of 8 CWs/frame) is experienced for three decomposition levels. While this makes the decomposition method impractical for real-time applications, the advantages achieved by the PSWT make the method very beneficial for speech storage applications.

IV. QUANTIZATION

An advantage of the PSWT, arising from the decimation process, is that in order to accurately reconstruct the speech, the data required at each decomposition level is explicitly defined. This contrasts the original SEW/REW decomposition. The information necessary for each surface corresponds to its sampling frequency. Since the filter outputs are decimated at each level, the transmitted information required for the surfaces $w_1(k, q)$, $w_2(k, q)$, $w_3(k, q)$, and $r_3(k, q)$ is in the ratio 4:2:1:1.

To optimize quantization efficiency, we may omit the least perceptually important surface(s), deliberately giving up exact reconstruction. The lack of time-synchrony in WI output speech (or residual), makes objective distortion measures, such as segmental signal-to-noise ratio and mean squared error, unreliable. One solution would be the computation of the distortion measures on a prototype-by-prototype basis. However, the resulting short segment-length also results in an unreliable measure. Thus, for this work, informal pairwise comparison listening tests were the preferred performance measure. Informal listening tests indicated that omission of the first (and highest rate) highpass output, $w_1(k, q)$, did not produce noticeable perceptual distortion of the output speech and thus, in low-rate coding, can be discarded.

For efficient coding, we base the quantization accuracy on the perceptual importance of the decomposed surfaces. Since bit allocation for the surfaces can be flexible, scalability is achievable by allowing a more accurate description of perceptually important scales. While the residue demands a precise description, the detail surfaces require less accuracy. The magnitude spectra of $w_2(k, q)$, $w_3(k, q)$, and $r_3(k, q)$ are quantized in the ratio 3:4:8 using variable dimension vector quantization techniques [5], with a trained codebook for each surface. Best results are obtained when each

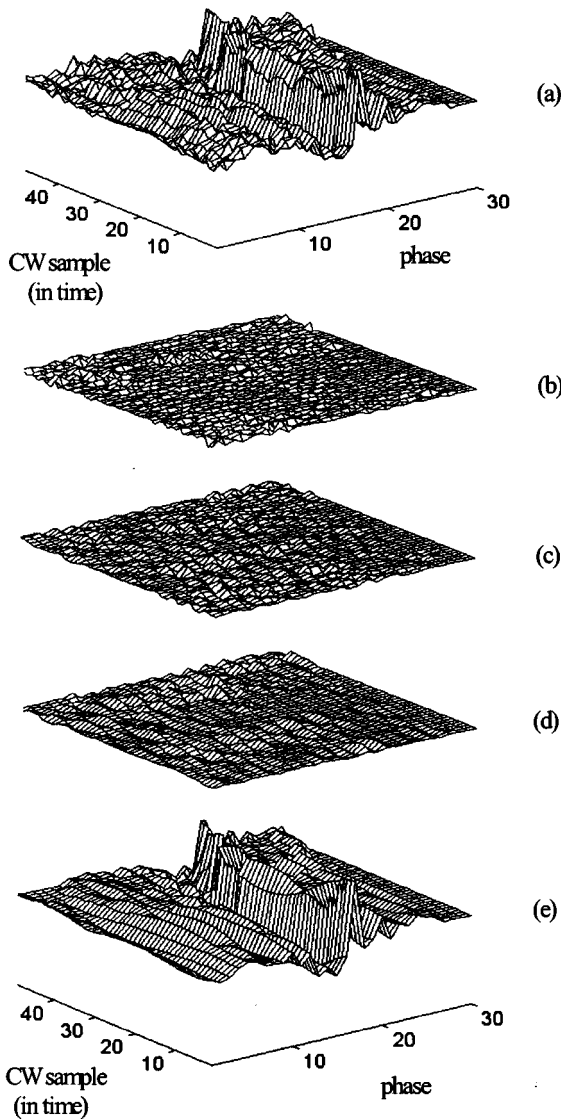


Fig. 3. Decomposition of the voiced sound "oo" taken from the word "foolish." (a) CW surface, $s(k, q)$, (b) $w_1(k, q)$, (c) $w_2(k, q)$, (d) $w_3(k, q)$, and (e) $r_3(k, q)$.

surface is recomposed and upsampled to the original sampling frequency separately, rather than reconstructed up through the tree structure as in Fig. 1 [4]. This allows, for example, the application of random phase to $w_3(k, q)$ without affecting the phase of lower-rate surfaces, such as $r_2(k, q)$. (Note that in the conventional reconstruction method, surface $r_2(k, q)$ is dependent on the surfaces $w_3(k, q)$ and $r_3(k, q)$.) This makes the use of phase models effective, and the need to quantize phase, unnecessary.

V. CONCLUSION

The PSWT provides an alternative description of signal evolution. The similarities between the proposed wavelet decomposition and the existing SEW/REW decomposition method make it easily applicable to the WI paradigm. Its multi-resolution analysis enables more detailed

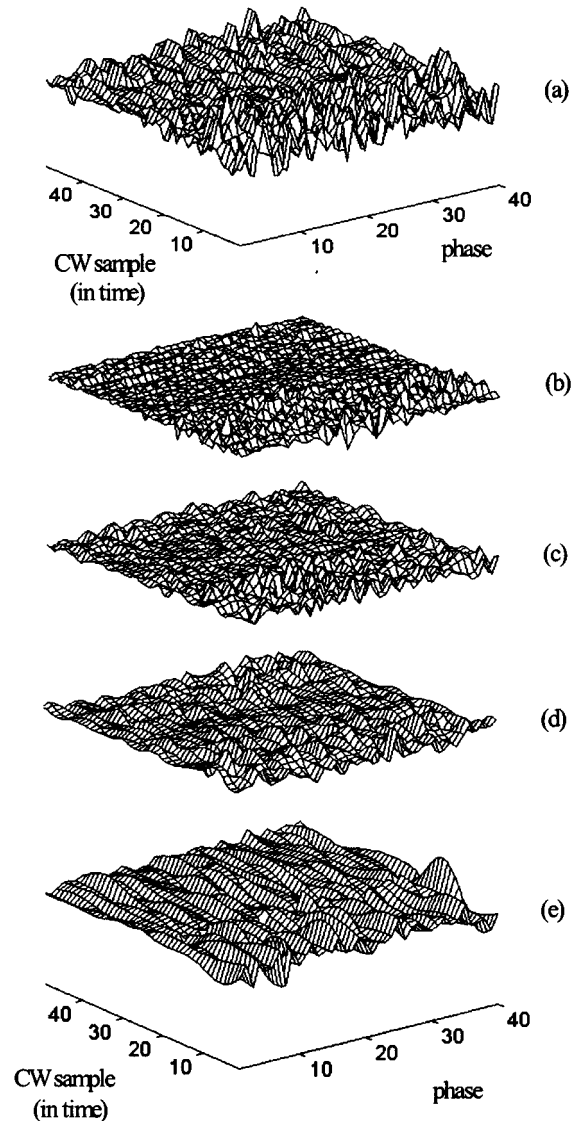


Fig. 4. Decomposition of the unvoiced sound "sh" taken from the word "foolish." (a) CW surface, $s(k, q)$, (b) $w_1(k, q)$, (c) $w_2(k, q)$, (d) $w_3(k, q)$, and (e) $r_3(k, q)$.

characterization of the evolutionary behavior, leading to scalable performance of WI coding.

REFERENCES

- [1] W. B. Kleijn and J. Haagen, "Transformation and decomposition of the speech signal for coding," *IEEE Signal Processing Lett.*, vol. 1, pp. 136–138, Sept. 1994.
- [2] —, "Waveform interpolation for coding and synthesis," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier.
- [3] G. Evangelista, "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Trans. Signal Processing*, vol. 41, pp. 3313–3329, Dec. 1993.
- [4] N. R. Chong, I. S. Burnett, and J. F. Chicharo, "Low-delay multi-level decomposition and quantization techniques for WI coding," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 241–244, 1999.
- [5] A. Das and A. Gersho, "Variable dimension spectral coding of speech at 2400 bps and below with phonetic classification," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 492–495, 1995.