July 2000

Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition

N. R. Chong-White
University of Wollongong

I. Burnett
University of Wollongong, ianb@uow.edu.au

Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition

Abstract

A waveform-matched waveform interpolation (WMWI) technique is presented which offers improved signal analysis over standard WI coders and, in the unquantised case, perfect reconstruction. In WMWI, an accurate representation of speech evolution is formed by extracting consecutive pitch periods of a time-warped, constant pitch residual. A pitch track optimisation technique is described which ensures that the critically sampled pitch periods can be effectively decomposed into a slowly evolving and rapidly evolving waveform, allowing efficient quantisation.

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was originally published as: Chong-White, NR & Burnett, I, Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition, Electronics Letters, 6 July 2000, 36(14), 1245-1247.

H_0F_0) be zero. This gives a set of simultaneous constraint equations, which upon solving yields,

$$A = -(2+\alpha)$$
$$B = \frac{4\alpha^3 + 2(2\beta+5)\alpha^2 + (\beta^2+4\beta+8)\alpha + 2}{2\alpha^2 + (\beta+4)\alpha + 2}$$
$$C = -\frac{2\alpha(\alpha+\beta+1)^2}{2\alpha^2 + (\beta+4)\alpha + 2} \quad (3)$$

The parameters α and β can be regarded as the free parameters on which the other parameters A, B and C depend. Setting α = –0.6848 and β = –1.6848 would give back the original '9/7' filter pair.

With rational values for α and β, the filter coefficients will also be rational but not neccesarily binary. For binary coefficients additional constraints need to be imposed on the values of α and β. We shall not present the details of the analysis here for lack of space but the results are as follows:

$$\alpha = -1 \qquad \beta = -\frac{2^c}{m} \quad (4)$$

where c, $m \in \mathbf{Z}$. This condition ensures that the resulting filter coefficients are binary. Although the value α is fixed, β can approximate quite closely almost any value by appropriate choices of c and m.

The 6/10 pair can be obtained from the 9/7 pair by first changing H_0 and F_0 in eqn. 2 around and then by extracting the factor $(1 + z^{-1})$ from the shorter filter (length 7) and inserting it into the longer filter (length 9). This gives

$$H_0 = K_1(1+z^{-1})(Z^2+\alpha Z+\beta)$$
$$F_0 = K_2(1+z)(Z+1)(Z^3+AZ^2+BZ+C) \quad (5)$$

where K_1 and K_2 are normalisation constants.

Table 1: Coefficient values for 9/7 filter pair for various values of parameter β

β	H_0	β	F_0
–1	[1, 0, –4, 8, 22, 8, –4, 0, 1]/32	–1 (–1.000)	[–1, 0, 5, 8, 5, 0, –1]/16
–8/5	[5, 0, –32, 64, 182, 64, –32, 0, 5]/256	8/5 (–1.600)	[–5, 0, 37, 64, 37, 0, –5]/128
–2	[1, 0, –8, 16, 46, 16, –8, 0, 1]/64	–2 (–2.000)	[–1, 0, 9, 16, 9, 0, –1]/32
–8/3	[3, 0, –32, 64, 186, 64, –32, 0, 3]/256	–8/3 (–2.667)	[–3, 0, 35, 64, 35, 0, –3]/128

Table 2: Coefficient values for 6/10 filter pair for various values of parameter β

β	H_0	β	F_0
–1 (–1.000)	[–1, 1, 4, 4, 1, –1]/8	–1	[1, 1, 4, 4, 30, 30, 4, 4, 1, 1]/64
–8/5 (–1.600)	[–5, 5, 32, 32, 5, –5]/64	–8/5	[5, 5, –32, 32, 246, 246, 32, –32, 5, 5]/512
–2 (–2.000)	[–1, 1, 8, 8, 1, –1]/16	–2	[1, 1, –8, 8, 62, 62, 8, –8, 1, 1]/128
–8/3 (–2.667)	[–3, 3, 32, 32, 3, –3]/64	–8/3	[3, 3, –32, 32, 250, 250, 32, –32, 3, 3]/512

Filter pair examples: In the original CDF '9/7' pair [5], β = –1.6848. We therefore choose values of β around –1.6848. Tables 1 and 2 show, respectively, the coefficients of the 9/7 pair and the 6/10 pair for various values of β. The '9/7' (β = –2) pair is the same as the MIT '9/7' pair reported by Strang in [6] (and also independently by Sweldens [7]). This filter is also known as filter no. 102 in the JPEG2000 (ISO/IEC JTC1/SC29/WG1) verification model. The method used in [6] for constructing this filter is via a complementary filter method and is different from our method. In our method, the '9/7' (β = –2) pair is only one out of the many pairs that can be obtained by varying the value of β.

Figs. 1 and 2 show, respectively, the frequency response of the 9/7 filter pairs and the 6/10 filter pairs for two values of β. It can be seen that the filter and scaling function characteristics can be changed by changing β. The corresponding scaling function and wavelet characteristics (not shown) also change by changing β.

The measure of spatial and frequency localisations of wavelet filters is defined in [1] and they characterise the localisation ability of the filter in signal analysis. For the filters presented here, we found that in general (without showing numerical results for lack of space), when β decreases (becomes more negative), the filter is more spatially localised, and when β increases, the filter is more frequency localised.

Conclusion: Two families of binary coefficient wavelet filters parametrised in a simple manner by a free parameter have been presented. The characteristics of the filters can be changed easily by varying the value of the free parameters to suit the application at hand. The main idea behind the technique used to obtain the filters is to allow some degree of freedom in choosing the coefficients by freeing some zeros of the LHBF.

27 April 2000
Electronics Letters Online No: 20000860
DOI: 10.1049/el:20000860

D.B.H. Tay (*Department of Electronic Engineering, LaTrobe University, Bundoora, Victoria 3083, Australia*)

E-mail: d.tay@ee.latrobe.edu.au

References

1 AKANSU, A.N.: 'Multiplierless PR quadrature mirror filters for subband image coding', *IEEE Trans. Image Proc.*, 1996, **5**, (9), pp. 1359–1363
2 REDMILL, D.W., BULL, D.R., and MARTIN, R.R.: 'Design of multiplier free linear phase perfect reconstruction filter banks using transformations and genetic algorithms'. Int. Conf. Image Processing and its Applications (IPA-97), 1997
3 WEI, D., TIAN, J., Jr., WELLS, R.O., and BURRUS, C.S.: 'A new class of biorthogonal wavelet systems for image transform coding', *IEEE Trans. Image Proc.*, 1998, **7**, (7) pp. 1000–1013
4 TAY, D.B.H.: 'Rationalizing the coefficients of popular biorthogonal wavelet filters'. Proc. IEEE Symp. Circuits Syst., 1998
5 COHEN, A., DAUBECHIES, I., and FEAUVEAU, J.C.: 'Biorthogonal bases of compactly supported wavelets', *Comm. Pure Appl. Math.*, 1992, **45**, pp. 485–560
6 STRANG, G. and NGUYEN, T.: 'Wavelets and filter banks' (Prentice-Hall, 1996)
7 SWELDENS, W.: 'The lifting scheme: A custom-design construction of biorthogonal wavelets', *Appl. Comput. Harmonic Anal.*, 1996, **3**, (2), pp. 186–200

Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition

N.R. Chong-White and I.S. Burnett

A waveform-matched waveform interpolation (WMWI) technique is presented which offers improved signal analysis over standard WI coders and, in the unquantised case, perfect reconstruction. In WMWI, an accurate representation of speech evolution is formed by extracting consecutive pitch periods of a time-warped, constant pitch residual. A pitch track optimisation technique is described which ensures that the critically sampled pitch periods can be effectively decomposed into a slowly evolving and rapidly evolving waveform, allowing efficient quantisation.

Introduction: Waveform interpolation (WI) coders are able to achieve high quality speech at low bit rates by using a decomposition motivated by human perception [1]. The decomposition is performed over a surface comprising of extracted, aligned pitch-length segments of the residual signal, called characteristic waveforms (CWs). However, in standard WI coders, relative phase information of the speech signal is destroyed during the rotation of CWs to form a surface suitable for decomposition.

Here, the WI analysis process has been adapted to allow waveform coding of the signal; a base mechanism was described in [2]. The proposed waveform-matched WI (WMWI) technique continuously warps the input linear prediction residual to a constant pitch period. Pitch-length segments of the warped residual are then critically sampled to form the CW surface for decomposition. However, unlike the surfaces of other WI coders [1, 3], an accurate description of the signal evolution is produced, without errors due to cyclic rotation or the repetition or omission of segments due to selective extraction.

In this Letter, we discuss the mapping of a signal to the warped time-domain such that the analysis, decomposition and quantisation of the pitch periods is effective.

Time-domain warping: The warping operation removes the pitch variations of the linear prediction residual signal to enforce a constant pitch period. In WMWI, an unwarping procedure is performed to reconstruct the residual. Perfect reconstruction can be achieved if the pitch track is accurately transmitted. Alternatively, at low rates, standard WI reconstruction can be performed on the warped CWs using a low-resolution pitch track.

Optimising the pitch track: To efficiently quantise the speech residual, the signal is decomposed into a slowly evolving waveform (SEW) and a rapidly evolving waveform (REW) [1]. The decomposition relies on the extracted pitch periods being well-aligned to work effectively. This corresponds to correctly warping the residual.

The pitch track is designed to align all pitch pulse peaks to a fixed position in each warped period. To minimise discontinuities at the period boundaries, this position is chosen to be the central sample of the pitch period.

Definition of terms: For the purpose of correctly warping to align pitch periods, the following terms are interpreted as follows:

(i) Frames which contain sections of high periodicity and exhibit clear pulse peaks in the residual signal are labelled as *voiced*, otherwise they are *unvoiced*

(ii) The *pitch period*, during voiced frames, is the distance between adjacent pulse peaks. Hence, every period has an associated pitch. During unvoiced frames, the pitch has no clear definition -- it is simply assigned a value, to allow continuous time-warping.

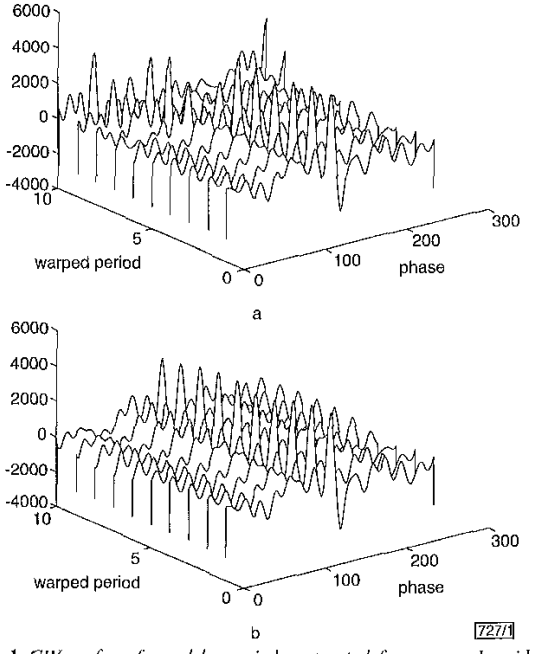

Fig. 1 *CW surface formed by periods extracted from warped residual for the case where pitch track is correct and not correct*

a Pitch track not correct
b Pitch track correct

Effect of non-optimal pitch track: The effect of an incorrect and correct pitch track for a section of voiced speech residual is shown in Fig. 1. If the pitch track is non-optimal (Fig. 1*a*), the poor alignment of pitch periods causes periodic pulses to be decomposed into the REW, making REW quantisation difficult. The well-aligned periods of Fig. 1*b* lead to most of the signal energy being separated into the SEW, as desired. It should also be noted that for effective SEW quantisation, pitch pulses following an unvoiced region must also be aligned with those pulses preceding that section.

Formation of the pitch track is best performed on a pitch period basis, rather on a frame basis. Analysis has shown that it is more important to align the pitch pulse peaks than correlate the pitch period as a whole, since this reduces the possibility of pulse peaks being incorrectly decomposed into the REW.

Locating of pitch pulses: To accurately determine the location of the pitch pulse peaks within the frame, the residual signal is low-pass filtered. A pulse detection algorithm, an extension of the technique described in [1], is then applied. Here, an initial pitch estimate for the frame, τ_{init}, is calculated from the autocorrelations of K segments, combined to form a composite function. For the case where $K = 5$, the composite autocorrelation function, R_C, for each candidate pitch value, d, can be expressed as

$$R_C(d) = R_3(d) + \sum_{\substack{k=1 \\ k\neq 3}}^{5} a_k \max[w(i) \cdot R_k(d-i)] \qquad -l(d) \le i \le l(d) \quad (1)$$

where, for segment k, R_k is the autocorrelation function, a_k is a weighting factor determined by the voicing decision of the previous frame, $w(i)$ is a window function, and $l(d)$ is the window length.

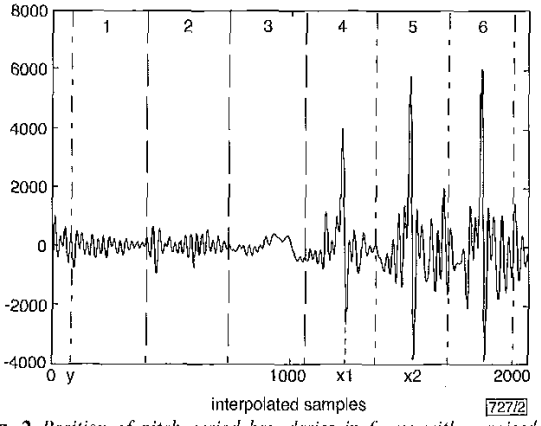

Fig. 2 *Position of pitch period boundaries in frame with unvoiced-to-voiced transition*

Pitch of periods 1–3 is selected to allow the peaks of periods 4–6 to be centred within the pitch period

The composite function is then recalculated (on an interpolated, filtered residual) for a small set of pitch period values surrounding the estimated pitch, τ_{init}, using segments of length equal to that value. If the refined R_C exceeds an adaptive threshold, it is proposed that the period contains a pulse, and the pulse peak location is determined at fractional sample resolution.

Calculating of pitch track for analysis: Given the pitch pulse locations, the pitch track is then formed. We define the pitch track for a set of four possible frame types: continuous voiced, continuous unvoiced, unvoiced-to-voiced, and voiced-to-unvoiced. It should be noted that the true pitch contour, which reflects the nature in which the glottis opens and closes during speech production, may not be the optimum pitch track for good signal analysis and decomposition.

During a continuous voiced section, a simple, yet effective, technique is to simply allow the pitch to remain constant for the duration of the pitch period. During continuous unvoiced frames, the pitch takes on a nominal value.

For unvoiced-to-voiced frame transitions (see Fig. 2), the key requirement is to ensure that pitch cycles surrounding a variable duration unvoiced segment are aligned. Hence, the number of periods, n, and the pitch, τ, of the unvoiced section preceding the period with the first pulse must be chosen such that the first pulse peak is warped to the correct position. To minimise pitch variation, we solve

$$\arg\min_{n} \left|(x_2 - x_1) - M\tau\right| \qquad n = 1, 2, 3, \ldots \quad (2)$$

where

$$M\tau = \frac{x_1 - \frac{x_2 - x_1}{2} - y}{n} \qquad n = 1, 2, 3, \ldots \qquad \tau_{min} < \tau < \tau_{max} \quad (3)$$

where x_i is the position of the ith pulse peak, y is the end boundary of the last period of the previous (unvoiced) frame, and M is

the interpolation constant. For the frame depicted in Fig. 2, $n = 3$. If x_1 is very close to the beginning of the frame, eqn. 2 may be indeterminate due to the constraints on τ. In these cases, y is shifted back to the previous period boundary, and τ is recalculated.

The pitch track calculation places a great deal of importance on the positions of the pitch pulse peaks, and the locations of the pitch period boundaries. This is necessary to achieve the waveform coding objective.

Conclusion: For effective decomposition and quantisation of the speech signal, using WMWI, the CW surface must contain aligned pitch periods. This requires accurate pitch pulse detection and careful derivation of the pitch track. Our technique ensures consistent positioning of the pitch pulses, even after unvoiced segments, while maintaining the waveform coding objective.

3 May 2000
Electronics Letters Online No: 20000871
DOI: 10.1049/el:20000871

N.R. Chong-White and I.S. Burnett (*Whisper Laboratories, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia*)

E-mail: i.burnett@elec.uow.edu.au

I.S. Burnett: Corresponding author

References

1 HAAGEN, J., and KLEIJN, W.B.: 'Waveform interpolation' *in* RAMACHANDRAN, R., and MAMMONE, R. (Eds.) 'Modern methods of speech processing' (Kluwer Academic Publishers, 1995)

2 KLEIJN, W.B., YANG, H., and DEPRETTERE, E.: 'Waveform interpolation coding with pitch-spaced subbands'. Proc. 5th Int. Conf. Spoken Language Processing, December 1998

3 ERIKSSON, T., and KLEIJN, W.B.: 'On waveform-interpolation coding with asymptotically perfect reconstruction'. Proc. IEEE Workshop Speech Coding, June 1999, pp. 93-95

Noise-robust speech recognition based on difference of power spectrum

Jinfu Xu and Gang Wei

A new noise-robust speech recognition method is presented based on the difference in the power spectrum. The idea is to remove the additive noise by filtering in the power spectrum domain. Feature extraction is carried out in two steps: (i) the short-time power spectrum of the speech signal is allowed to pass through a filter bank; and (ii) the differences in the filter outputs are calculated. Theoretical analysis and experimental results show that using the proposed features can significantly improve the recogniser's performance in a noisy environment.

Introduction: The performance of current speech recognisers is degraded to a significant extent in noisy environments by the discrepancy between training and testing conditions. If the speech signal is contaminated by noise, features such as its linear predictive coding cepstral coefficients (LPCC) and mel frequency cepstral coefficients (MFCC) will be changed. Many studies have been carried out into the representation of speech signals to improve the performance of speech recognisers in noisy environments. Some methods focus on the search for noise-resistant features. In this Letter, we propose new noise-robust features based on the difference in the power spectrum. Our idea is based on the work of Hermansky *et al.* [1], You *et al.* [2] and Hirsch *et al.* [3], where the noise effect can be removed by filtering in different domains. Test experiments have shown that the proposed features can vastly improve the recognition rate in noisy environments.

Noise removal based on difference in the power spectrum: Suppose that a clean speech signal is contaminated by additive noise, giving noisy speech, and that the noise is stationary zero-mean white noise and uncorrelated with the clean speech. Since an acoustic speech signal is usually processed in frames, for convenience, we denote $|X_m(k)|^2$, $|S_m(k)|^2$ and $|N(k)|^2$ as the mth frame short-time discrete power spectrum of the noisy speech, clean speech and noise, respectively. We then have

$$|X_m(k)|^2 = |S_m(k)|^2 + |N(k)|^2 \quad \begin{matrix} 0 \le k \le K-1 \\ 0 \le m \le M-1 \end{matrix} \quad (1)$$

where K is the size of the discrete Fourier transform (DFT), m is the frame index, and M is the number of frames within an utterance. Differentiating eqn. 1 on both sides with respect to time, we obtain

$$\frac{\partial |X_m(k)|^2}{\partial t} = \frac{\partial |S_m(k)|^2}{\partial t} \quad (2)$$

Eqn. 2 indicates that $\partial|Xm(k)|^2/\partial t$ is unaffected by noise. Similar to the method used to deduce the delta vector from its original vector, $\partial|X_m(k)|^2/\partial t$ can be calculated approximately by the following polynomial

$$\frac{\partial |X_m(k)|^2}{\partial t} \simeq G_T \sum_{t=-T}^{T} t|X_{m+1}(k)|^2 \quad (3)$$

where $G_T = 1/\Sigma_{t=-T}^{T} t^2$, T being the number of frames used before or after the present frame. We call the result calculated by eqn. 3 the difference in the power spectrum.

Noise-robust speech features: As described above, noise-robust features may be extracted from the difference in the power spectrum, since it is unaffected by noise. To do this, we use a filter bank spread over the frequency range of the speech signal and let the power spectrum of the speech signal pass through it. The output of each filter is the weighted sum of a given discrete power spectral within the passband of this filter, as shown in eqn. 4:

$$Y_m(i) = \sum_{k=k_1}^{k_2} \alpha_k |X_m(k)|^2 \qquad 0 \le i \le FN-1 \quad (4)$$

where i is the filter index, FN the number of filters, k_1 and k_2 are determined by the frequency range of this filter, and α_k non-negative and determined by the shape of the filter.

From eqns. 2 and 4, we obtain

$$\frac{\partial Y_m(i)}{\partial t} = \sum_{k=k_1}^{k_2} \alpha_k \frac{\partial |X_m(k)|^2}{\partial t} = \sum_{k=k_1}^{k_2} \alpha_k \frac{\partial |S_m(k)|^2}{\partial t} \quad (5)$$

$\partial Y_m(i)/\partial t$ is also unaffected by noise, and it can be computed approximately as follows:

$$\frac{\partial Y_m(i)}{\partial t} \simeq G_T \sum_{t=-T}^{T} t Y_{m+t}(i) \quad (6)$$

We call the result computed using eqn. 6 the difference in the output of the filter. The noise-robust speech feature vector can be constituted by the differences in the outputs of all the filters.

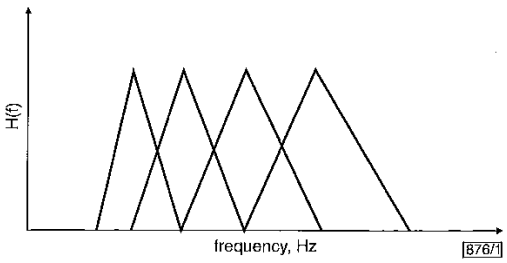

Fig. 1 *General form of filter bank*

Although eqns. 4 and 6 are simple, the following questions have to be solved for practical usage. What is the optimal value of T in eqn. 6? How many filters should be selected? What shape should each filter be, trianglular or rectanglular? Should the passbands of the neighbouring filters be disjoint or overlapping? It is not easy to answer the first two questions theoretically, but their answers can be easily searched experimentally. We only discuss the last two questions here.