# Wideband Speech Coding at 4 kbps using Waveform Interpolation

*C.H. Ritz, I.S. Burnett*
Whisper Laboratories, TITR,
University of Wollongong, NSW, Australia
chritz@st.elec.uow.edu.au, i.burnett@elec.uow.edu.au

## ABSTRACT

In this paper we present a new low rate, wideband speech coder operating at 4 kbps and based on Waveform Interpolation (WI). An outline of WI speech coding is provided together with a description of its adaptation to wideband speech. Particular emphasis is placed on the quantisation of the WI parameters. Included is a detailed analysis of the quantisation requirements for the Line Spectral Frequencies (LSFs) and the Characteristic Waveforms (CWs). We conclude that WI is a feasible technique for high quality low rate wideband speech coding.

## 1. INTRODUCTION

Wideband speech generally has significantly better subjective quality than narrowband speech and it has been proposed that wideband speech coded at a similar rate to narrowband speech may offer significantly better subjective quality [1]. Most current research into wideband speech compression has targeted 8 kbps and above [1-3] but there are clear applications for wideband coding at lower rates (around 4 kbps); examples of such areas are: high quality voicemail and paging services, internet telephony, mobile internet applications such as streaming media and speech storage for online teaching material. To date, most low rate speech coding research has focused on narrowband speech coding, primarily due to the bandwidth limitations of existing infrastructure; these new and emerging applications are not limited by such constraints. The work described in this paper seeks to extend a current, successful narrowband coding technique to the low rate 4kbps wideband scenario.

A number of methods have been proposed for narrowband speech coding at low bit rates [4-6]. Waveform Interpolation (WI) [6,7] is one of these methods and has been shown to outperform many other types of coders at these bit rates [6]. In this work, we describe an adaptation of WI to the coding of wideband speech at 4 kbps. Specifically we focus on quantisation requirements for the wideband WI parameters.

Section 2 presents an overview of WI while Section 3 describes the adaptation of the coding architecture to the wideband scenario. Section 4 describes methods and results for quantisation of the WI parameters, with conclusions presented in Section 5.

## 2. OVERVIEW OF WI

In the WI algorithm, described in [6] speech is viewed as a sequence of evolving pitch-length Characteristic Waveforms (CWs). These waveforms are sampled at a constant rate and described by Fourier series coefficients. By aligning successive CWs, speech can be reconstructed at the decoder by linearly interpolating the CW Fourier series coefficients and then transforming the spectral CW representations back to the time domain for every signal point. The quantisation of the spectrum represented by these coefficients is the key to the performance of WI at low rates.

The WI architecture used in this work is similar to that outlined in [6]. The first step is to filter the speech signal using the Linear Prediction (LP) filter and generate a residual signal. To minimise distortions at frame boundaries, LP filtering is performed using linearly interpolated filter coefficients. The pitch is then estimated from the residual and a series of uniformly spaced CWs are extracted. The CWs are normalised by their power, which is quantised separately to ensure it is preserved. The CWs are then used to form an evolving surface and decomposed into two 'independent' components: the Slowly Evolving Waveform (SEWs) and Rapidly Evolving Waveform (REWs). These resulting surfaces represent the quasi-periodic and non-periodic components of speech, respectively. Since they have different and distinct evolutionary and spectral properties they can be quantised effectively using appropriate algorithms. For each WI frame, the remaining quantised parameters are the Linear Prediction Coefficients (LPCs), the pitch and a representation of the power of the CW surface.

# 3. WI FOR WIDEBAND SPEECH

All speech in this work was taken from the ANDOSL 20 kHz sampled database [8]. Wideband speech was obtained by down-sampling the database speech to 16 kHz with further band-limiting to the range 50 Hz to 7 kHz. In this work, $20^{th}$ order LP analysis was performed using 30 ms analysis windows and 25 ms frames. The choice of $20^{th}$ order LP analysis is based on the 7 kHz spectral range and can be justified by the suggestions in [9] that approximately one LPC is needed per kHz of sampling rate plus a few extra poles to accurately represent the speech spectrum. The frame size and window size are consistent with LP analysis theory for speech [9]. To allow efficient quantisation and interpolation, the resulting LP coefficients were converted to LSFs.

The minimum pitch length used in this work is 40 samples, which equates to a maximum pitch frequency of 400 Hz. To correctly sample the CWs, the minimum extraction rate should also be 400 Hz, corresponding to 10 CWs being extracted per frame. This accords with [6] which states that for narrowband WI operating without quantisation a 400 Hz extraction rate leads to near transparent speech, while lower extraction rates result in speech of less perceptual quality. Informal listening tests confirmed that this is also the case for wideband WI.

# 4. WI PARAMETER QUANTISATION RESULTS

## 4.1 LSF Quantisation

The LSFs were quantised using split Vector Quantisation (VQ). Split VQ was chosen over standard VQ to reduce complexity. Note that other methods such as multistage VQ could also be used, but split VQ has the added benefit of robustness to channel errors. To determine the best bit rate for LSF quantisation, we performed split VQ using both 5 and 6 splits with split vector sizes of (3,3,4,4,6) and (3,3,3,3,4,4). Codebooks of various sizes were trained using LSF vectors derived from approximately 31 minutes of speech. The AverageLog Spectral Distortion (AvLSD) [10] introduced by quantising the LSFs derived for a separate set of 12 male and female sentences was measured for various bit rates, and is shown in Table 1. Also included are percentages of outliers, to indicate perceived distortion for wideband LSF quantisation [11].

In reference [11] it was suggested that an average spectral distortion of 1.6 dB with no more than 4 % of outliers in the range of 3 to 5 dB and no outliers above 5 dB results in transparent quantisation of wideband LSFs. In Table 1, these results are almost met by 40 bits and easily met at 48 bits. Informal listening tests found that quantising the LSFs using 48 bits introduced very few distortions and so was chosen in this work.

## 4.2 Pitch and Power Quantisation

The pitch period varies between 40 samples and 292 samples; hence, this was scalar quantised once per frame using 8 bits. We also experimented with using non-uniform quantisation of the pitch period at 7 bits, allowing more quantisation levels for shorter pitch periods over longer pitch periods. Informal listening tests found that this generally led to transparent pitch quantisation. However, we chose to scalar quantise at 8 bits to ensure no errors were introduced by the pitch quantisation scheme.

The quantise the CW power, it was first converted to the speech domain so that it is not influenced by quantisation errors generated for the LP synthesis filter. Speech domain gain was then quantised twice per frame (80 Hz) using log differential scalar quantisation and 5 bits, similar to that done in narrowband WI [6]. Informal listening tests found that using more bits for the power resulted in little improvement in perceptual speech quality. We propose that speech power is independent of the bandwidth, and so should require a similar bit rate regardless of whether it is wideband or narrowband.

| No. of Bits | Splits | bits/split | AvLSD (dB) | 3<LSD≤5 (%) | LSD>5 (%) |
|---|---|---|---|---|---|
| 25 | 3-3-4-4-6 | (5,5,5,5,5) | 2.5 | 20.4 | 1.7 |
| 30 | 3-3-4-4-6 | (6,6,6,6,6) | 2.2 | 9.3 | 1.3 |
| 40 | 3-3-4-4-6 | (8,8,8,8,8) | 1.6 | 2.8 | 0.5 |
| 48 | 3-3-3-3-4-4 | (8,8,8,8,8,8) | 1.3 | 0.2 | 0 |

Table 1. Average log spectral distortion results and percentages of outliers for quantisation of wideband LSFs at various bit rates.

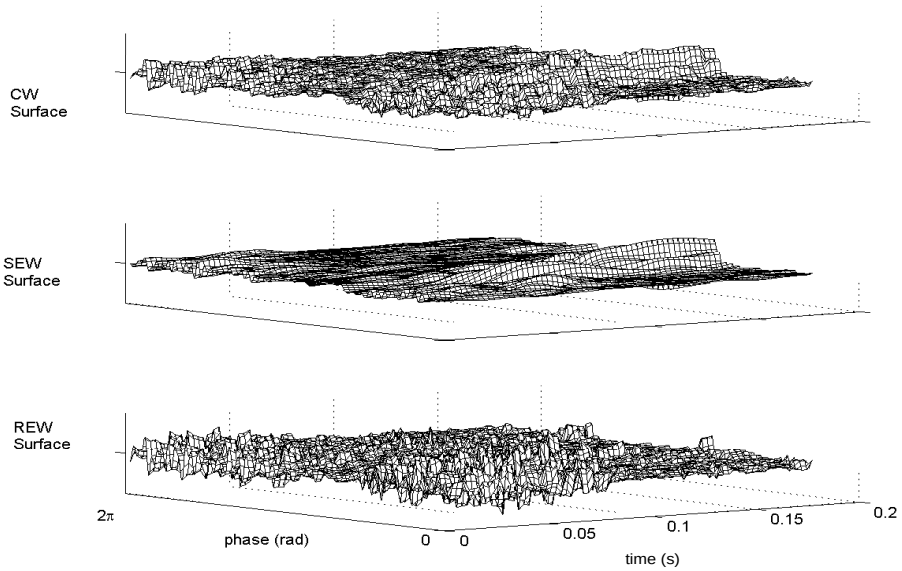


Figure 1. CW, SEW and REW surfaces.

## 4.3 CW Quantisation

As outlined in Section 3, the CW surface is decomposed into a SEW and REW surface; these are quantised separately. Since the SEW represents the slowly evolving part of the waveform, it can be obtained by low pass filtering the sequence of aligned CWs. The REW is obtained by subtracting the SEW from the CW. Kleijn [6] suggests that low pass filtering at 20 Hzgives satisfactory decomposition for narrowband CWs. This is in part based on the work in [12], which concluded that speech intelligibility begins to decrease as the variation of the speech amplitude envelope falls below 16 Hz. Since the CW evolution represents this amplitude variation, and the SEW aims to represent the mainly slow varying components, it is reasonable to conclude that 16 Hz should be the minimum cut-off frequency for the low pass decomposition filter. This work was conducted on speech sampled at approximately 16 kHz and so it is also relevant to wideband speech; hence we propose that low pass filtering at 20 Hz is also appropriate for the wideband CW decomposition.

Figure 1 shows example CW, SEW and REW surfaces derived for a section of wideband speech using a 20 Hz low pass decomposition filter. These were obtained by spectrally padding the aligned waveforms to a common length, converting back to the speech domain and plotting the time domain waveforms at each extraction point. From these surfaces it can be seen that the SEW surface contains mostly slow varying components of the CW while the REW surface contains most of the fast varying components of the CW. Hence, the decomposition method outlined in this work was deemed satisfactory for wideband speech.

### 4.3.1 REW Quantisation

The REW represents the non-periodic and noisy components of the residual waveform. In narrowband WI, it is suggested that the REW requires a high update rate, but can be quantised with few bits [6]. Informal listening tests found that in wideband WI if too few REWs are sent and interpolated in the decoder, the synthesised speech sounds too buzzy. This is because interpolation introduces false periodicity into regions of the CW, which are supposed to represent non-periodic components. Further, in this wideband coder, only the REW magnitude is quantised since it is more perceptually important than the REW phase [6]. We confirmed this through informal listening tests, where replacing the REW phase spectrum with random values resulted in negligible loss of perceptual quality.

In narrowband WI a common REW quantisation method is to model magnitude spectra using polynomials with trained coefficients [6]. An alternative adopted in this work is to quantise the magnitude spectrums using VQ. Since the length of the REW magnitude vectors vary with pitch, we used Variable Dimension Vector Quantisation (VDVQ) [13], using codebooks trained on approximately 68000 vectors. Informal listening tests indicated that coding the REW magnitudes with as few as 3 bits achieved similar quality to that obtained using up to 6 bits. Examination of REW magnitude spectrums found that they are mostly flat above 4 kHz, with most of the variation occurring below 4 kHz. Hence, while coding with only 3 bits leads to a rough description of the REW magnitude spectrum, especially above 4 kHz, informal listening tests found that the perceptual quality loss was not great.

### 4.3.2 SEW Quantisation

Since the SEWs were obtained by low pass filtering at 20 Hz, they can be down-sampled to a minimum of 40 Hz and recovered in the decoder by linear interpolation. This assumes an ideal low pass filter. However, to minimise delay, a low pass filter of only $20^{th}$ order is used in this work (which equates to 1 frame of CWs look ahead). Hence, the filter does not have a sharp cut-off, which leads to some aliasing. To account for this effect, more SEWs can be transmitted per frame. However, informal listening tests found that sending 1 SEW was preferable as it generally leads to smoother sounding speech.

For efficient quantisation, only the SEW magnitude is quantised. In the decoder, a quantised SEW is generated by combining a quantised SEW magnitude vector with a model phase vector. Modelling of the SEW phase, rather than quantisation, has been shown in narrowband WI to provide high quality speech. Since the wideband SEW represents the mainly periodic and voiced components, it is reasonable to assume that a SEW phase model is also relevant to wideband WI.

In narrowband WI a common technique for low rate SEW magnitude quantisation is to only quantise the magnitude below 800 Hz. The remaining magnitude spectrum is calculated as one minus the REW magnitude, based on the assumption of mostly flat magnitude spectrums above 800 Hz. Informal listening tests found that using this method for wideband SEW magnitude quantisation led to a significant degradation of the perceptual quality. We also experimented with quantising a greater proportion of the REW magnitude and inferring less from the REW, however we found that perceptual quality was still lost. Hence, to maintain quality, we quantised the entire SEW magnitude spectrum using VDVQ [13].

In narrowband WI, the low frequency sections of the SEW magnitude are more perceptually significant [6] than the high frequency sections. Since the SEW represents the underlying pulse shape, we propose that this is also true for wideband SEWs; hence we chose to use split VDVQ in this work; we used three splits covering frequency ranges of 0-1000 Hz, 1000 Hz-4000 Hz, and 4000 Hz-8000 Hz. The codebooks were trained on SEWs derived from the same speech as used for REW magnitude quantisation. A variety of codebooks, with sizes ranging from 3 to 8 bits, was trained for each split. Informal listening tests found that using too few bits for the lower split resulted in poor subjective quality, while smaller codebooks (as small as 3 bits) could be used for the highest split without decreasing subjective quality significantly. Based on these observations, and taking into account the 19 bits available after REW quantisation, we found that a suitable compromise was to use split codebooks of sizes 8, 6 and 5 bits, respectively, for each split.

The SEW phase in this work was represented using a linear SEW phase model. In [6] it was suggested that a fixed SEW phase vector can be used and chosen based on the REW magnitude spectrum; this was found to also be applicable to wideband WI. Hence we set the upper 4 kHz phases to be random whenever the SEW/REW energy ratio in this region fell below a threshold; based on informal listening tests, a threshold of 1.5 was chosen. Adaptively modelling SEW phase in this way ensured that pulses were not generated in unvoiced speech, which can lead to buzziness.

## 5. CONCLUSIONS

This paper has presented an investigation into WI applied to wideband speech coding at 4 kbps. We have considered the derivation and quantisation of the fundamental WI parameters when coding Wideband rather than narrowband speech. It was found that the general analysis procedure used in narrowband WI is also appropriate for wideband WI. It was also found that the general principles for SEW and REW quantisation in narrowband WI also apply to wideband WI. One difference is the quantisation of the wideband SEW surface, which requires a higher bit rate than a comparable narrowband SEW surface to obtain high perceptual quality. We propose that WI is a suitable technique for high quality low bit rate wideband speech coding.

## 6. ACKNOWLEDGEMENTS

C.H. Ritz is in receipt of an Australian Postgraduate Award and a Motorola (Australia) Partnerships in Research Grant.

## 7. REFERENCES

[1] McCree, A., Takahiro, U, Anandakumar, A., Bernard, A. and Paksoy, E., " An Embedded Adaptive Multi-Rate Wideband Speech Coder", *ICASSP'2001*, May 7-11, Salt Lake City, Utah.

[2] Salami, R., Laflamme, C, and Adoul, J.-P., "Real-Time Implementation of a 9.6 kbit/s ACELP Wideband Speech Coder", *Proc. GLOBECOM '92*, pp. 447-451, Vol. 1, 1992.

[3] Lin, W., Ho, S.N., Lin. X., "Mixed Excitation Linear Prediction Coding of Wideband Speech at 8 kbps", *Proc. ICASSP'2000*, vol. 2, pp. 1137-1140, 2000.

[4] Cambell, J., Welch, V. and Tremain, T., "An Expandable Error Protected 4800 BPS CELP Coder (U.S. Federal Standard 4800 BPS Voice coder)", *Proc. ICASSP'89*, pp. 735-738.
[5] McCree, A.V., Barnwell, T.P., "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding", *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 242-249, July 1995.
[6] Kleijn, W. B. and Haagen, J., "Waveform Interpolation for Coding and Synthesis", in *Speech Coding and Synthesis*, pp. 175-207, Kleijn, W.B., and Paliwal, K.K., editors, Elsevier Science B.V., 1995.
[7] Gottesman, O. and Gersho, A., "Enhanced waveform interpolative coding at 4 kbps", *Proc. IEEE Workshop on Speech Coding*, pp. 90-92, 20-23 June, 1999.
[8] Australian National Database of Spoken Language (ANDOSL), CD ROM.
[9] Rabiner, L.R. and Schafer, R.W., "Digital Processing of Speech Signals", pp. 419-420, Prentice Hall, 1978.
[10] Paliwal, K.K. and Kleijn, W.B., "Quantization of LPC Parameters", in Speech Coding and Synthesis, pp. 443-444, edited by Kleijn, W.B. and Paliwal, K.K., Elsevier, 1995.
[11] Ferhaoui, M. and Van Gerven, S., "LSP Quantization in Wideband Speech Coders", *Proc. IEEE Workshop on Speech Coding*, pp. 25-27, June, 1999.
[12] Drullman, R., Festen, J.M. and Plomp, R., "Effect of temporal envelope smearing on speech reception", *J. Acoust. Soc. Am.*, 95 (2), Feb., 1994.
[13] Das, A., Rao, A.V. and Gersho, A., "Variable-Dimension Vector Quantization", *IEEE Signal Processing Letters*, pp. 200-202, Vol. 3, No. 7, July, 1996.