LOW BIT RATE WIDEBAND WI SPEECH CODING

C.H. Ritz, I.S. Burnett, J. Lukasiak
Whisper Labs, University of Wollongong
Wollongong, Australia.
chritz@st.elec.uow.edu.au, i.burnett@elec.uow.edu.au, jasonl@elec.uow.edu.au

ABSTRACT

This paper investigates Waveform Interpolation (WI) applied low bit rate wideband speech coding. An analysis of the evolutionary behaviour of wideband Characteristic Waveforms (CWs) shows that direct application of the classical WI algorithm may not be appropriate for wideband speech. We propose a modification whereby CW quantisation is performed using classical WI decomposition for the low frequency region and noise modelling for the high frequency region. Wideband WI coders incorporating this modification and operating at 4 kbps and 6 kbps are described. Subjective testing of these coders shows that WI is a promising approach to low bit rate wideband speech compression.

1. INTRODUCTION

Waveform Interpolation (WI) is a method for coding narrowband speech with high perceptual quality at low bit rates [1]. It is based on the description of speech (in practice the LP residual) as a surface formed from evolving pitch-length Characteristic Waveforms (CWs). To date, most research into WI has focused on narrowband (300 Hz to 3.4 kHz bandwidth) speech [1]. The work described in this paper focuses on WI applied to wideband (50 Hz to 7 kHz bandwidth) speech [2]. The motivation is for a low bit rate (below 6 kbps) solution to wideband speech coding.

The key to quantisation of the CWs at low bit rates is decomposition of that surface into the Slowly Evolving Waveform (SEW) and the Rapidly Evolving Waveform (REW), which represent the voiced and unvoiced speech components, respectively. High speech quality relies on an appropriate decomposition strategy that does not result in excessive REW (causing noisiness) or excessive SEW (causing buzziness) [1].

This paper presents an investigation into the evolutionary behaviour and appropriate decomposition strategy for wideband CWs (Sections 2 to 4). Also presented is a description of appropriate quantisation techniques for 4 kbps and 6 kbps wideband WI (Sections 5 and 6). The quality of these coders was evaluated using subjective tests, the results of which are described in Section 7. Section 8 presents a discussion and conclusion of the major results of this paper.

2. EXTRACTION OF WIDEBAND CWS

Wideband speech was generated by re-sampling speech extracted from the ANDOSL database [3] to 16 kHz and band-limiting to 50 Hz to 7 kHz. For comparison, narrowband speech was also generated by re-sampling the same speech to 8 kHz and band-limiting to the range 300 Hz to 3.4 kHz.

To minimise reconstruction errors and ensure a smooth evolution of CWs [1], extraction was performed in the residual domain following Linear Predictive (LP) analysis of the speech signal. For wideband WI, an appropriate LP order is 20, based on our results in [2]. LP analysis was performed using 30 ms analysis intervals and 25 ms frames.

CW extraction is performed at a rate of 400 Hz; this ensured that CWs are extracted at least once per pitch period for a maximum pitch frequency of 400 Hz [1]. Following extraction, the CWs are aligned by circular rotation to maximally correlate with the previous CW. Alignment is necessary to ensure a smooth evolution of the CW surface [1].

3. EVOLUTION OF WIDEBAND CWS

In classical WI applied to narrowband speech, decomposition is motivated by the differences in evolution of voiced and unvoiced speech. Here, the evolution is analysed through both the correlation decay and the evolution spectrum of the CWs.

3.1 Correlation Decay

The correlation decay refers to the change in correlation between CWs as their separation (in evolutionary time) increases. To investigate this decay, the cross correlation was measured between CWs separated by n extraction points. To facilitate investigation of the CW evolution in different frequency bands we performed the analysis in the DFT domain. The real and imaginary DFT coefficients corresponding to frequency *f*, of the CW extracted at time *n* are defined as:

$$CW_n(f) = \left[a_n(f) \quad b_n(f) \right] \qquad (1)$$

The cross correlation between CWs extracted at time 0 and n was measured using the equation:

$$c_n = \frac{\left| \sum_f CW_0(f) \times CW_n(f)^T \right|}{\sqrt{\sum_f CW_0(f) \times CW_0(f)^T \sum_f CW_n(f) \times CW_n(f)^T}} \qquad (2)$$

where, $f \in [f1, f2]$ defines the frequency band of interest.

Equation (2) was used to measure the CW correlation decay for $0 \le n \le 20$ and two frequency bands covering the 0 to 4kHz and 4 to 8 kHz ranges for voiced and unvoiced speech. For comparison, the correlation decay for CWs extracted from narrowband speech was also measured. The results of these experiments are shown in Figure 1.

Figure 1 shows that the correlation decay in the 0 to 4 kHz range during voiced speech is similar for both wideband and narrowband CWs. In contrast, the 4 to 8 kHz range of wideband CWs during voiced speech demonstrates more rapid correlation decay. For unvoiced speech, the correlation decays rapidly for both narrowband CWs and both frequency bands of wideband CWs. These results indicate that compared with the 0 to 4kHz region, the 4 kHz to 8 kHz region of the wideband CWs does not exhibit slow evolution during voiced speech.

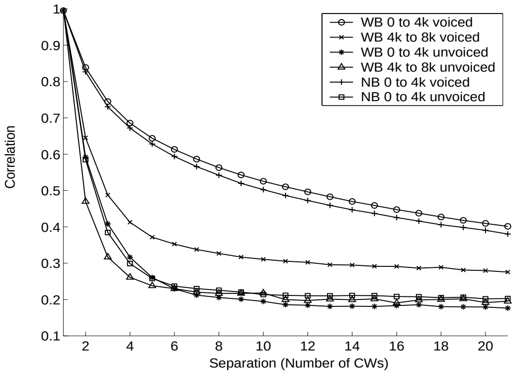

Figure 1. CW Correlation Decay.

3.2 Evolution Spectrum

The correlation results above suggest that the evolution spectrum of the CWs for the 0 to 4 kHz frequency region will differ to the 4 to 8 kHz frequency region. The evolution spectrum can be described by taking the DFT of a sequence of frequency domain CWs [4]. This results in an evolutionary magnitude spectrum for each coefficient of the CWs in the frequency domain. Figure 2 shows an example of the 3-D evolutionary spectrum for a sequence of 20 voiced wideband CWs. The magnitude spectra of the evolutionary coefficients were averaged for eight 1 kHz frequency bands, covering the entire CW bandwidth. For comparison, Figure 3 shows the evolutionary spectrum of narrowband CWs (averaged for four 1 kHz frequency bands) extracted from the same section of speech as for Figure 2

Figure 2 shows that the evolution of the low frequencies of the CW has energy concentrated at low evolution frequencies, while the evolution spectrum of the high frequencies is more evenly distributed. Comparison of the surfaces of Figures 2 and 3 indicates that the evolutionary behaviour of narrowband CWs are similar to the 0 to 4 kHz region of the wideband CWs, with energy concentrated at low frequencies. These results for narrowband CWs agree with the suggestions in [1], that the energy in the evolution spectrum is concentrated at low frequencies.

4. DECOMPOSITION OF WIDEBAND CWS

In narrowband WI, the decomposition of the CWs is usually achieved by low pass filtering the evolution of the CW surface. The low pass component forms the SEW surface while the high pass component forms the complimentary REW surface. For narrowband speech, the decomposition filter typically has a cutoff frequency of 20 Hz [1]. Results from Section 3, suggest that the decomposition requirements for the lower half band of wideband CWs should be similar to that for narrowband CWs; however, the decomposition strategy for high frequencies should differ. Here we analyse the decomposition using the SEW-to-REW (STR) energy ratio, defined as:

$$STR = \frac{\sum_{f} |S(f)|^2}{\sum_{f} |R(f)|^2} \quad f \in [f1, f2] \qquad (3)$$

where, S(f) and R(f) are the DFT magnitudes of the SEW and REW at frequency *f*, respectively and [*f*1,*f*2] defines the bandwidth of interest.

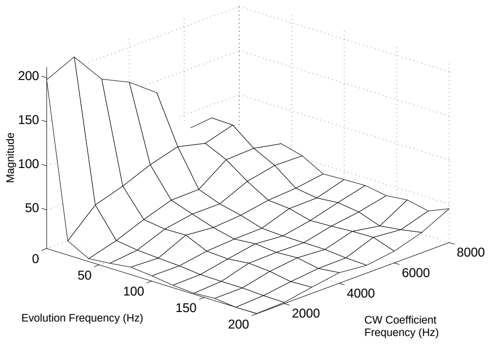

Figure 2. Evolution spectrum of wideband CWs.

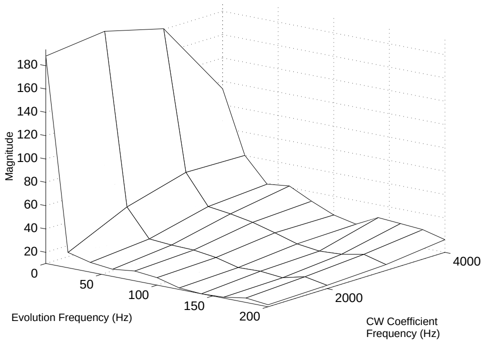

Figure 3. Evolution spectrum of narrowband CWs.

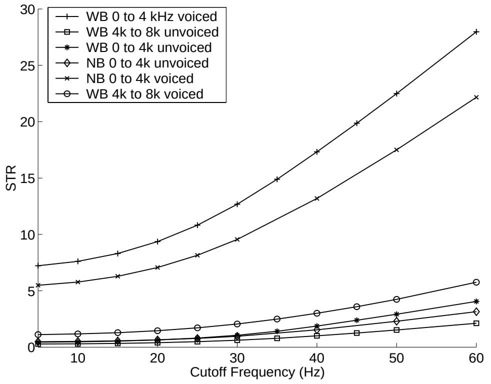

Figure 4. STR versus cutoff frequency.

We decomposed the wideband CWs using cutoff frequencies ranging from 5 Hz to 70 Hz and the STR was measured for both the 0 to 4 kHz and 4 to 8 kHz frequency bands. Figure 4 shows results for both wideband and narrowband CWs generated from identical speech. For comparison, results for voiced and unvoiced speech are shown.

For the 0 to 4kHz frequency band of wideband and narrowband CWs derived for voiced speech, Figure 4 shows similar shaped curves. The small differences in STR can be explained by the band-limiting of the narrowband CWs to 300 Hz to 3.4 kHz as well as the differing LP filters for wideband and narrowband. An appropriate choice of cutoff frequency would thus be the frequency corresponding to the knee of the curve, since at this point most of the slowly evolving energy has been captured. It can be seen that in both cases the knee points correspond to a frequency of around 20 to 30 Hz. This confirms the choice of a 20 Hz cutoff for the decomposition filter in narrowband WI and for the 0 to 4 kHz region of wideband CWs.

In contrast, for the 4 to 8 kHz region of wideband CWs derived for voiced speech, the STR curve does not show a clear knee point. This indicates that there is no clear SEW component at the high frequencies during voiced speech. In addition, this curve has a similar shape to the STR curves for CWs derived for unvoiced speech. This indicates that the 4 to 8 kHz region of wideband CWs derived for voiced speech have similar characteristics to CWs derived for unvoiced speech.

4.1 Alignment of wideband CWs

In classical WI, alignment is performed by maximising the correlation between adjacent CWs. The correlation is measured over the entire CW bandwidth (although a smaller bandwidth has been used for reduced complexity [5]). If only the low frequencies of the CWs are decomposed, it may be appropriate to only align using these frequencies to ensure that the CWs are maximally smooth in this region. An increase in the smooth evolution is indicated by increased SEW energy following decomposition.

To investigate the effects of different alignment bandwidths on decomposition, the STR (defined in Equation (3)) was measured for voiced CWs extracted from two different sentences. A high STR indicates that the CW contains more SEW energy. Alignment bandwidths of 0 to 8 kHz and 0 to 4 kHz were used and the STR was measured for both the 0 to 4 kHz and 0 to 8 kHz frequency ranges. The STR results for these experiments are shown in Table 1.

Alignment Bandwidth	STR	
	0 to 8 kHz	0 to 4 kHz
0 to 8 kHz	1.48	2.38
0 to 4 kHz	1.49	2.58

Table 1. STR measured for voiced CWs with different alignment bandwidths.

Results from Table 1 show an approximate 8 % increase in the STR corresponding to the 0 to 4 kHz frequency range when the alignment bandwidth corresponds to 0 to 4 kHz rather than 0 to 8 kHz. These results indicate that when only decomposing the low frequencies of the CWs, only these frequencies should be considered in the alignment.

5. WIDEBAND CW QUANTISATION

The previous section showed that only low frequencies of the CWs show significant SEW content, hence only low frequencies should be analysed during decomposition. For the coders developed here, we chose a maximum CW frequency for decomposition of 6 kHz. During informal listening tests, we found that decomposition above 6 kHz causes undesirable buzzy distortions in the reconstructed speech. From the results of Sections 3 and 4, we concluded that this was due to an increase in the SEW energy, due to the capturing of REW-like energy rather than slowly evolving components. A similar frequency has also been used for the separation of low and high frequencies in wideband speech coding [6] as well as for the transition frequency of periodic to random excitation in bandwidth extension of wideband speech [7]. The preferred decomposition method is low pass filtering to 20 Hz.

5.1 SEW Quantisation

Since the SEW is obtained by low pass filtering to 20 Hz, and assuming ideal filters, the SEW can be down-sampled to a rate of 40 Hz. For the wideband WI coders described here, this corresponds to one SEW per frame. As with narrowband WI [1], only the SEW magnitude is quantised. We chose Variable Dimension Vector Quantisation (VDVQ) [8] for SEW magnitude quantisation. The SEW phase spectra is replaced in the decoder by a dispersion phase vector taken from a section of male speech, similar to narrowband WI [1].

5.2 REW Quantisation

For quantisation of the REW, a similar method was adopted as used in narrowband WI coding [1], whereby the REW was replaced by white Gaussian noise scaled to the correct power. This method has also been used for coding the high frequencies of unvoiced speech above 4 kHz [9]. Here, only the REW magnitude was quantised and the REW phase spectrum was replaced by uniform random values. The REW magnitude is quantised using VDVQ and updated five times per frame.

5.3 Quantisation of the 6 to 8 kHz region of the CWs

Two noise modelling approaches were investigated for quantisation of the 6 to 8 kHz region of the CWs. Noise modelling for high frequencies of the excitation has previously been used for wideband speech coding [6,10]. The first method replaces the CW with white Gaussian noise, scaled to the correct magnitude. The second method uses modulated noise for the high frequencies. In this method, Gaussian noise is modulated by the time domain envelope of the low frequency components of the CW.

Modulated noise is the preferred method for representing the high frequencies. Through informal listening tests, it was found to provide more natural speech than purely Gaussian noise. This is similar to the results of the authors in [10]. To minimise bit rate in the 4 kbps coder, the magnitude spectrum is not quantised and noise is directly modulated by the time domain envelope corresponding to the 0 to 6 kHz frequency region of the CW. For the 6 kbps coder, the modulated noise is scaled using the magnitude spectrum of the 6 to 8 kHz frequency region that was quantised using VDVQ and updated five times per frame.

6. BIT ALLOCATION OF WIDEBAND CODERS

The bit allocations for the 4 kbps and 6 kbps coders are shown in Table 2. These bit allocations were chosen to ensure that the most perceptually important parameters (LSFs, pitch and gain) were quantised with minimal distortion. The remaining bits were used for CW quantisation.

Parameter	**Bits per frame**
Pitch	8
Gain	10
LSFs	48
SEW	19 (4 kbps) 24 (6 kbps)
REW	15 (4kbps) 30 (6 kbps)
CW (6 to 7 kHz)	0 (4kbps) 30 (6 kbps)

Table 2. Bit rate for the 4 kbps and 6 kbps WI coders.

The LSF vectors were coded using split VQ and 48 bits per frame. This was based on our results in [11], which found that 48 bits gave an average log spectral distortion of approximately

1.3 dB, which is far less than the transparency requirement of 1.6 dB, suggested for wideband LSF coding in [12].

The pitch period is measured in samples and the ranges from 40 to 295. To ensure no errors due to quantisation, uniform scalar quantisation of the pitch period requires 8 bits. The gain was quantised in the log domain twice per frame using differential scalar quantisation. Informal listening tests found a bit allocation of 5 bits per gain (10 bits per frame) resulted in minimal distortions.

For the 4 kbps coder, the above bit allocations left 34 bits available for CW quantisation. An appropriate bit allocation was found to be 19 bits for the SEW and 15 bits for the REW with no bits allocated to the region above 6 kHz. These bit allocations ensured that the SEW magnitude was quantised more accurately than the REW magnitude, which is critical for WI speech coding [1]. For the 6 kbps, 84 bits were available for the quantisation of the CW. An appropriate bit allocation was found to be 24 bits for the SEW, 30 bits for the REW and 30 bits for the CW frequency region above 6 kHz.

6.1 Postfiltering

To reduce quantisation distortions, postfiltering was used. The postfilter adopted can be described by [13]:

$$H(z) = \frac{\hat{A}(z/\alpha)}{\hat{A}(z/\beta)}(1-\mu z^{-1}) \quad 0<\alpha<\beta<1 \qquad (4)$$

where, $\hat{A}(z)$ represents the LP synthesis filter.

Postfiltering for wideband WI was implemented by spectrally shaping the speech domain CWs using the spectrum of the postfilter prior to speech synthesis. Informal listening tests found values for α, β and μ of 0.65, 0.75 and 0.7, respectively, to provide the most improvement in speech quality.

7. SUBJECTIVE TESTING

To evaluate the subjective quality of the wideband WI coders, Mean Opinion Score (MOS) tests were conducted. For these tests, 10 sentences were coded using the 4 kbps and 6 kbps wideband WI coders. These sentences were also coded using the 6.6 kbps wideband 3GPP/ETSI AMR coder [6]. Twenty-eight listeners participated in this test, with results shown in Table 3.

Wideband Coder	MOS (male sentences)	MOS (female sentences)	MOS (total)
WI @ 4 kbps	3.28	2.63	2.95
WI @ 6 kbps	3.25	2.88	3.07
AMR @ 6.6 kbps	3.92	3.34	3.63

Table 3. MOS results for wideband WI coders.

In Table 3, the 6 kbps coder scored about 0.1 of a MOS higher than the 4 kbps coder. However, the AMR coder at 6.6 kbps scored around 0.6 of a MOS higher than the wideband WI coder at 6 kbps.

8. DISCUSSION AND CONCLUSIONS

This paper has presented an investigation into wideband WI speech coding at low bit rates. It was found that the evolution of CWs derived for wideband speech differs across frequency bands. It was found that decomposition by low pass filtering to 20 Hz (as is commonly used in narrowband WI) is justified for the low frequencies of the wideband CWs, but not necessarily for high frequencies. Based on these results, the wideband WI coders described here only decompose the 0 to 6 kHz frequency regions of the CWs. For the remaining frequencies, the CWs were quantised using modulated noise.

Subjective MOS tests showed that the performance of the wideband WI coder operating at 6 kbps is only slightly better than the 4 kbps version. In addition, results showed that the AMR coder at 6.6 kbps had significantly higher quality. Some of the improved quality may be explained by the 10 % higher bit rate of the 6.6 kbps AMR coder. We believe that most of the difference is due to the saturation in quality of wideband WI above 4 kbps. In narrowband speech coding, it is well known that WI produces high quality speech at 4 kbps and below [1], but the performance of WI above 4 kbps is unclear. Being a parametric coder, it is reasonable to assume that WI will not approach transparent quality with increasing bit rate due to the limitations of the model. We believe that the classical WI model also limits the quality of the wideband coders described here, but remains a promising technique for wideband speech coding at low rates.

A distinctive feature of all coders is the significantly higher MOS results for female sentences compared with male sentences. Improving the quality of female speech is an area for future research.

REFERENCES

[1] 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., eds., Elsevier Science B.V., 1995.

[2] Ritz, C.H., Burnett, I.S. and Lukasiak, J., “Extending Waveform Interpolation to Wideband Speech Coding”, *Proc. 2002 IEEE Workshop on Speech Coding*, Japan, October, 2002.

[3] Australian National Database of Spoken Language (ANDOSL), CD ROM.

[4] Tanaka, Y., and Kimura, H., “Low-bit-rate speech coding using a two dimensional transform of residual signals and waveform interpolation”, *Proc. ICASSP’94*, Vol. I, pp. 173-176, April, 1994.

[5] Kleijn, W. B., Shoham, Y., Sen, D. and Hagen, R., “A low-complexity waveform interpolation coder”, *Proc. ICASSP’96*, pp. 212-215, 1996.

[6] Rotola-Pukkila, *et. al*., “AMR wideband codec – leap in mobile communication voice quality”, *Proc. EUROSPEECH’2001*, Vol. 4, pp. 2303-2306, Sep., 2001.

[7] Nilsson, M. and Kleijn, W. B., “Avoiding over-estimation in bandwidth extension of telephony speech”, *Proc. ICASSP’2001*, Vol. 2, pp. 869-872, 2001.

[8] 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.

[9] Epps, J. R. and Holmes, W. H., “A new very low bit rate wideband speech coder with a sinusoidal highband model”, *Proc. ISCAS’01*, Vol. 2, pp. 349-352, 2001.

[10] McCree, A., Unno, T., Anandakumar, A., Bernard, A. and Paksoy, E., “An embedded adaptive multi-rate wideband speech coder”, *Proc. ICASSP’2001*, Vol. 2, pp. 761-764, 2001.

[11] C.H. Ritz and I.S. Burnett, “Temporal Decomposition: A Promising Approach to Wideband Speech Compression”, *Proc., EUROSPEECH 2001*, Vol. 4, pp. 2315-2318, Aalborg, Denmark, September, 2001.

[12] Ferhaoui, M., Van Gerven, S., “LSP Quantization in Wideband Speech Coders”, *Proc. 1999 IEEE Workshop on Speech Coding*, pp.25-27, June 1999.

[13] Chen, J.-H. and Gersho, A., “Adaptive Postfiltering for Quality Enhancement of Coded Speech”, *IEEE Trans. on Speech and Audio Processing*, vol.3, no. 1, pp. 59-71, January, 1995.