SEW REPRESENTATION FOR LOW RATE WI CODING

J. Lukasiak, I.S. Burnett
Whisper Laboratories, TITR
University of Wollongong
Wollongong, NSW, Australia, 2522

ABSTRACT

This paper considers low-rate Waveform Interpolation (WI) coding. It compares the existing, common Slowly Evolving Waveform (SEW) quantisation scheme with two new schemes for representing and quantising the SEW. The first scheme uses a minimum phase estimate to reconstruct the SEW whilst the second scheme uses a pulse model whose parameters are implicitly transmitted in the quantised rapidly evolving waveform (REW). These new schemes maintain or reduce the bit rate required for transmission of the SEW. Results indicate that, for low rate WI coding, necessarily coarse SEW magnitude spectrum quantisation limits the contribution of the SEW to perceptual quality. Perceptual tests indicate that avoiding coarse spectral shape quantisation and using a fixed shape model that lends itself to smooth interpolation, maintains the perceptual quality of the synthesized speech. The proposed fixed shape model requires no bits for transmission, allowing a 12 percent reduction in the overall coder bit rate.

1. INTRODUCTION

The waveform interpolation (WI) paradigm proposed by Kleijn [1] is the focus of much current research in speech coding circles. The WI paradigm involves firstly linear predictive (LP) filtering the input speech. The residual signal is then separated into pitch cycles (known as characteristic waveforms (CW) [1]) and these are used to form a two dimensional waveform which evolves on a pitch synchronous nature. To maximize the smoothness of the two dimensional surface the individual pitch length segments are aligned when constructing the surface. This two dimensional waveform is then decomposed into Slowly evolving and Rapidly evolving waveforms (SEW/REW). The SEW and REW are down sampled and quantised separately in the encoder. The decoder reconstructs the SEW and REW via interpolation before recombining them. Synthesised speech is produced by converting the reconstructed two dimensional surface back to a one dimensional signal and passing this signal through the linear predictive synthesis filter.

The current research involving WI can be broadly grouped into distinct categories these being; a)low rate speech coding; b) perfect reconstruction allowing waveform coding. A commonality in much of the present research involves identifying better means of representing the SEW to achieve improved perceptual quality [2][3]. The original WI coder [1] operating at 2.4kbps down samples the SEW by 10, leaving only a single Sew vector per frame. Only the DFT magnitude values of this vector are then quantised and transmitted. A phase model is used in the decoder to reconstruct the SEW vector. New methods proposed for better representation of the SEW waveform include directly quantising the SEW DFT phase in an analysis by synthesis structure (AbyS) [2] and critically sampling and warping to a constant length to achieve perfect reconstruction [3]. These methods report improved perceptual quality but at the expense of increased complexity and bit rate.

This paper compares two new alternate schemes for quantising and reconstructing the SEW for low rate WI coders, to the well known existing scheme given in [1]. The new methods are:

a) Quantising only the SEW DFT magnitudes and reconstructing the SEW using a minimum phase estimation technique.

b) The use of a new pulse model to represent the SEW.

The comparison involves contrasting both the waveform shape of the quantised SEW's and the perceptual quality of the synthesized speech produced using each SEW method.

The paper is organized as follows. In section 2 the SEW reconstruction techniques shown above as well as the well known existing technique [1] are detailed. Section 3 reports and analyses the characteristics of the SEW shapes for each of the above techniques. In section 4 experimental results obtained for the techniques are reported. Finally the major points are summarised in section 5.

2. SEW QUANTISATION SCHEMES

2.1 Existing WI SEW Quantisation [1]

The DFT magnitude coefficients for the lowest 800Hz are quantised using a 7 bit Vector quantisor. The remaining DFT magnitude coefficients are calculated from the reconstructed REW using the fact that above 800Hz the overall residual signal magnitude spectrum may be considered flat. The SEW phase spectrum is not transmitted but is set equal to that of a predetermined model. The bit rate required for this method is 280bps for the frame size used.

2.2 Minimum Phase model SEW

A minimum phase model for a given magnitude spectrum can be calculated via the use of either the Hilbert transform or via the Complex Cepstrum as used in many low rate sinusoidal coders such as [4]. The minimum phase component is calculated via the Cepstrum as [5]:

$$C(n) = FFT^{-1}(\log(FFT|F(n)|))$$
$$\Phi(n) = FFT^{-1}(e^{FFT(w(n)C(n))}) \quad (1)$$
$$\text{where: } w(n) = \begin{cases} 0 & n < 0 \\ 1 & n = 0 \\ 2 & n > 0 \end{cases}$$

Where $|F(n)|$ is the input magnitude spectrum, C(n) is the cepstrum and $\Phi(n)$ is the minimum phase spectrum.

Using the phase generated in equation (1) assumes that the input signal is minimum phase which is not strictly true for speech. However, this method of estimating the phase is widely used in low rate sinusoidal speech coders such as [4] and is reported to improve the perceptual quality of these coders when compared to using only linear phase or phase models.

The minimum phase model SEW continues to use the existing quantisation scheme to transmit the SEW magnitude spectrum and then calculates the minimum phase estimate in the decoding stage. This requires that in common with the existing scheme [1], the minimum phase scheme requires 7 bits per frame to be used for transmission of the SEW magnitude spectrum.

2.3 New SEW pulse model

A new scheme to quantise and reconstruct the SEW was first proposed by these authors in [6]. This method uses a new pulse modeling mechanism for reconstruction of the SEW waveform.

The use of a pulse model for the SEW results in no bits being used for transmission of this parameter. The model used is based on the Zinc function which is defined in the discrete time domain as [7]:

$$z(n-\lambda) = A\text{sinc}(n-\lambda) + B\text{cosc}(n-\lambda) = ...$$
$$..... = \begin{cases} A & n-\lambda = 0 \\ \dfrac{2B}{(n-\lambda)\pi} & n-\lambda = odd \\ 0 & n-\lambda = even \end{cases} \quad (2)$$

The zinc model has been found to be superior in modeling the LP residual and is widely used in time domain Analysis by Synthesis schemes where the parameters A,B and λ are selected to minimize the error between the pulse and the residual signal [7]. To allow the zinc pulse to be conveniently used in the WI structure, the pulse was translated to the frequency domain via the DFT. This allows straight forward interpolation between adjacent pulses of different lengths via zero padding. For low rate coding, the parameters A,B and λ cannot be transmitted due to bit rate constraints. It was found that speech of high perceptual quality could be produced by setting λ to zero. This forces the zinc pulse to be wholly positive and also places the pulse peak at the beginning of the frame. Positioning the pulse at the beginning of each frame is equivalent to removing the SEW's linear phase component, which is acceptable in WI as the linear phase is already modified by aligning the waveforms in the encoding stage prior to down sampling. The values of A and B are set according to:

$$A = B = 1 - \frac{1}{N}\sum_{i=0}^{N-1}|REW(i)| \quad (3)$$

Equation (3) uses the assumptions that the LP residual spectrum is flat and that spectrum has been normalized to unity value previously in the coding process by the removal of the gain term. Using these assumptions and the properties of the Inverse Discrete Fourier Transform it can be shown that the resultant value of equation (3) is equal to the height of a single sample time domain pulse that would exhibit the required flat magnitude spectrum. Whilst the zinc pulse consists of a large initial pulse followed by further impulses whose amplitudes decrease rapidly with time, informal perceptual testing has shown that setting the height of the initial pulse equal to the height of a single impulse as calculated by equation (3), produces good results.

The method of deriving an implicit SEW pulse from the REW allows the pulse height to be dynamically varied according to the magnitude of the REW. Thus for sections with a high noise content the pulse is small and vice versa. Forcing the pulse to be positive and of fixed position may appear to be sub optimal in modeling the SEW. However, initial perceptual testing indicated a preference for this configuration over a pulse of variable position and polarity.

3. CHARATERISTICS OF THE SEW QUANTISATION SCHEMES

3.1 Waveform Shape

Figure 1 shows examples of quantised SEW for each of the methods detailed in Section 2.

Comparing the unquantised and existing quantised waveforms in Figure 1 indicates that the existing method models the main impulse quite well. However, the fine detail information away from the main pulse (phase information) is lost. In contrast the minimum phase estimate models the initial impulse less well than the existing method but better reproduces the detail away from the main pulse. This characteristic indicates that minimum phase estimate causes the pulse energy to be dispersed across the waveform. This has the effect of reducing the pulse height and thus de-emphasises the contribution of the pulse. The zinc pulse model reproduces the height of the main impulse quite well. It also adds some further impulses away from the main pulse, these impulses decrease rapidly in value as the distance from the main pulse increases. The shape of the zinc pulse model is also constant with only the magnitude varying. This characteristic makes the zinc model very slowly evolving and thus ideal for producing smoothly evolving surfaces via interpolation.

3.2 Recombination of the SEW and REW

In [1] the synthesized CW is generated by adding the Fourier series coefficients representing the SEW and REW. However, due to the prior discarding of both the SEW and REW phase information, this method of recombination does not ensure constructive rather than destructive recombination. Also it is possible that the REW may introduce an extra pulse into the reconstructed section. One option for improving the recombination procedure is to time align the entire SEW and

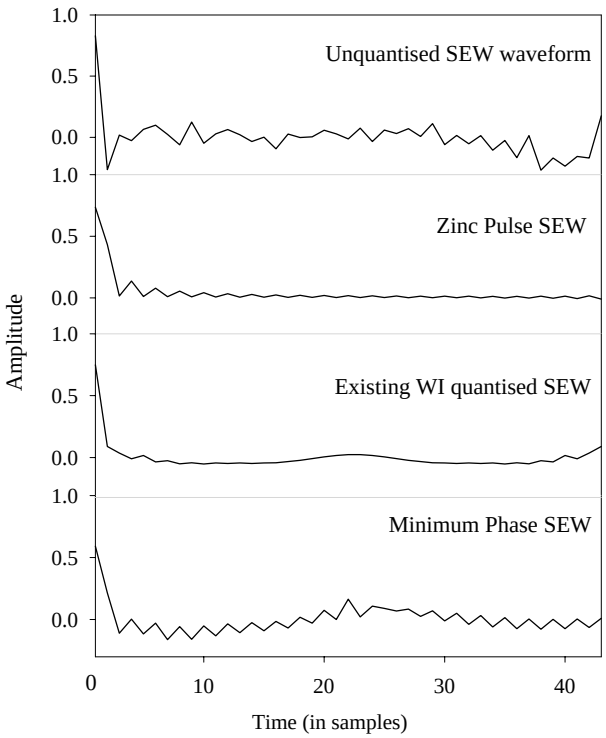

Figure 1- Comparison of Quantised SEW

	Standard SEW	Minimum Phase	Pulse Model
MOS Score	3.5	3.42	3.49
95% Conf. level	0.13	0.15	0.13

Table 1- MOS test Results

REW waveforms. This method works well for the SEW quantised using the existing scheme as the reconstructed waveform is generally dominated by a single impulse as seen in Figure 1. However, for the new zinc pulse and minimum phase schemes time aligning before recombination does not work well and produces hiss in the synthesized speech. This is due to the fact that, as seen in Figure 1 these waveform exhibit additional information or impulses away from the main pulse. This characteristic makes the maximum correlation criteria for time alignment unreliable as the correlation due to the subsequent impulses can tend to subtract from the correlation for the initial impulse, thus causing the waveforms to align incorrectly. To achieve good reconstruction with the alternate methods, a method that matches the REW phase to that of the SEW below 800Hz is used. Matching the low frequency phase where the SEW magnitude is usually dominant, ensures that these low frequencies are aligned in the time domain and thus produce a degree of temporal masking around the SEW pulse. This masking removes the hiss from the reconstructed signal.

To determine the effect of the phase matching method on the existing SEW quantised waveforms, informal listening tests were used. These tests indicated that when compared to time alignment, the phase matching method resulted in neither an improvement nor degradation in the perceptual quality of the reconstructed speech.

4. EXPERIMENTAL RESULTS

4.1 Coder Configuration

The structure of the coder is as detailed in [1]. The coder allocates 26 bits for the LSF parameters, 10 bits for the power, 6 bits for the pitch and 8 bits for the REW waveform per frame, with a frame size of 25ms. The 8 bits allocated to the REW are used to represent the REW magnitude spectrum with random phase used in the decoder to reconstruct the REW. The REW quantisation scheme used is the same as that detailed [1].

The overall bit rate of the base coder is 2kbps. In the case of the existing and minimum phase SEW quantisation methods, 280bps must be added to this value.

4.2 Synthesised Speech Waveforms

The coder detailed in 4.1 was used to generate synthesized speech for each of the SEW quantisation schemes. Figure 2 contains an example of unquantised speech and the synthesized versions of this speech for each of the SEW methods.

Figure 2 shows that the zinc pulse method has produced synthesized speech that most closely resembles the original unquantised speech. This is most clearly evident in the reproduction of the pitch pulses. The zinc model produces clearly defined pulses similar to those seen in the unquantised speech. Whist for the other methods the pulses, in particular the positive going pulses are not as clearly visible. This is due to poor matching of the SEW from the limited size codebooks available in a low rate coder. The minimum phase method may have produced better reconstruction of the fine detail between pitch pulses however, this has been lost due to the use of the same codebook selection for reconstructing the magnitude spectrum as the existing method.

The problem of selecting poorly matched pulse shapes due to the limited codebook size available for low rate coding does not occur for the zinc pulse method, as the shape is a consistent pulse and only the magnitude is varied. This fixed shape does not produce adverse effects in unvoiced sections as the magnitude of the pulse is the unitary compliment of the average REW magnitude. In these unvoiced sections the average REW magnitude is close to unity and thus the zinc pulse SEW has little effect.

4.3 Subjective Listening Test Results

For testing purposes, the coder detailed in section 4.1 was used to code 8 input speech sentences (4 male,4 female) of flat speech from the TIMIT database using each of the SEW representations detailed in 2.1-2.3. MOS testing comprising 20 untrained listeners was carried out on the coded speech files. The listeners used Sony headphones and the files were played via Turtle Beach Montego sound cards. The results are shown in Table 1.

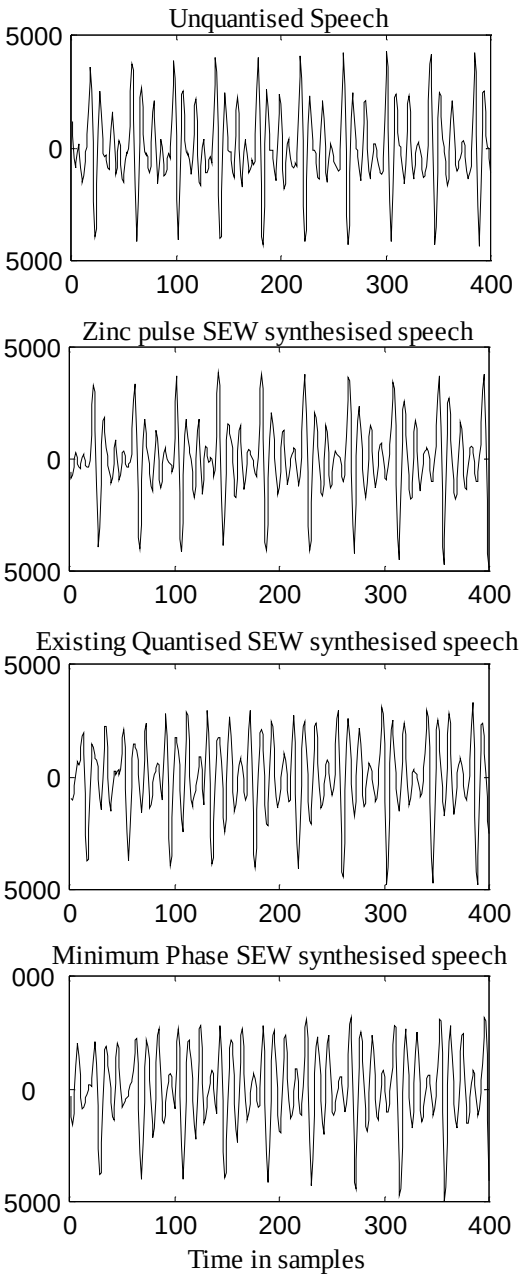

Figure 2- Synthesised speech comparison

The results obtained indicate that there is no statistical difference in perceptual quality between each of the SEW configurations used. For the minimum phase SEW this infers that any improvement over the existing method in modeling the phase is overwhelmed by using the same coarsely quantised magnitude spectrum as the existing method. The retention in perceptual quality for the pulse model SEW representation is despite the fact that the pulse model requires no bits for transmission, compared to 280bps for the other methods. This result indicates that attempting to retain the spectral shape with the limited bit count available at low rates offers no perceptual advantage over using a fixed shape zinc model that lends itself to smooth interpolation.

5. CONCLUSIONS

The results obtained for the minimum phase SEW indicate that altering the representation of the SEW phase used for reconstruction, whilst maintaining a coarsely quantised magnitude spectrum offers little perceptual modification. This agrees with the well documented belief that phase is less perceptually important than spectral shape. It also infers that for low rate WI coding, the perceptual quality due directly to the SEW is restricted by the limited bit count available to quantise the SEW spectral shape. Abandoning the attempt to quantise the SEW spectral shape with limited bit count and instead using a fixed shape pulse model that is smoothly evolving, was found to maintain the perceptual quality of the synthesized speech. The retention of quality is despite the pulse model requiring no bits for transmission. This results in a saving of 280bps and constitutes a 12 percent reduction in the overall coder bit rate.

In conclusion it can be deduced that to achieve the best trade off for bit rate and perceptual quality, WI should either quantise the SEW very accurately requiring a higher bit rate as in [2] and [3] or opt for a parametric representation that is smoothly evolving and thus easily interpolated. Attempting to coarsely maintain the SEW shape with very limited bits appears to offer little benefit.

6. ACKNOWLEDGEMENTS

J. Lukasiak is in receipt of an Australian Postgraduate Award (Industry) and a Motorola (Australia) Partnerships in research Grant. Whisper Laboratories is funded by Motorola and the Australian Research Council.

7. REFERENCES

[1] W.B. Kleijn and J. Haagen, “A speech coder based on decomposition of characteristic waveforms”, Proc. ICASSP 95, Vol. 1, pp.508-511, 1995.

[2] O. Gottesman, “Dispersion phase vector quantisation for enhancement of waveform interpolative coder”, Proc. ICASSP 99,

[3] N.R. Chong, I.S. Burnett, J.F. Chicharo, “Adapting waveform interpolation (with pitch spaced subbands) for quantisation”, Proc. of IEEE workshop on speech coding, pp. 96-98, 1999.

[4] R.J. McAulay and T.F. Quatieri, “The Sinusoidal transform coder at 2400b/s”, Conf. Rec of MILCOM ’92. Communications- Fusing command, control and intelligence, Vol.1, pp.378-380, 1992.

[5] F.M. Tesche, “On the use of the Hilbert transform for processing measured CW data”, IEEE trans. On Electromagnetic Compatability, Vol.34, part1, pp259-266, August 1992.

[6] J. Lukasiak and I.S. Burnett, “Exploiting simultaneously masked linear prediction in a WI speech coder”, Proc. of IEEE Workshop on Speech Coding 2000, pp.11-13, 2000.

[7] D.J. Hiotakakos and C.S. Xydeas, “Low bit rate coding using an interpolated zinc excitation model”, Conf. Proc. ICCS 94, Vol.3, pp.884-997, 1994.