PHASE MODELLING OF SPEECH EXCITATION FOR LOW BIT-RATE SINUSOIDAL TRANSFORM CODING

Xiaoqin Sun, Fabrice Plante*, Barry M.G. Cheetham* & Kenneth W. T. Wong#*

*Dept.of Electrical Engineering and Electronics, The University of Liverpool, Liverpool, UK.
e-mails: xqsun@liv.ac.uk, fabrice@liv.ac.uk & barry@liv.ac.uk,
BT Laboratories, Martlesham Heath, Ipswich, IP5 7RE, UK.

ABSTRACT

Sinusoidal transform coding (STC) techniques model speech as the sum of sine-waves whose frequencies, amplitudes and phases are specified at regular intervals. To achieve a low-bit rate representation, only the spectral envelope is encoded and the phases are regenerated according to a minimum phase assumption. In this paper, the inaccuracy of the minimum phase model is demonstrated. It is shown that the phase spectra of decoded speech segments may be corrected using either the parameters of a Rosenberg pulse model or a second order all-pass filter. Experiments have shown that by applying this correction, the phase accuracy increases and the speech quality improves.

1. INTRODUCTION

Good quality speech may be synthesised from a sinusoidal model with frequencies, amplitudes and phases specified at regular intervals [1]. This model has proved effective for low bit-rate speech coding since assumptions about the characteristics of speech and perception may be made to allow highly efficient representations of the model's parameters.

The Sinusoidal Transform Coding (STC) technique proposed by McAulay and Quatieri [2] applies FFT analysis to segments of speech extracted at suitable intervals and determines the amplitudes and frequencies of peaks in the resulting short-term magnitude spectra. For voiced speech, the frequencies of the peaks are assumed to be harmonics of an extracted pitch-frequency. The amplitudes of the peaks are efficiently represented by fitting an envelope to the magnitude spectrum and characterising this envelope by a set of coefficients (e.g. LSP coefficients).

Phase information required to synthesise speech at the decoder is obtained from the envelope assuming that it is the magnitude spectrum of a minimum phase transfer function modelling the effect of the human vocal apparatus when impulsively excited. This assumption has proved to be reasonably effective, though experiments have shown that the speech quality may be improved by providing more accurate phase information [3]. In this paper we examine the minimum phase assumption and propose modification of STC to improve the accuracy of phase derivation.

2. MINIMUM PHASE MODEL

Voiced speech may be modelled by a pseudo-periodic sequence of impulses e[n] driving a glottal filter G(z), a vocal tract filter V(z) and a lip-radiation filter L(z) as shown in figure 1.

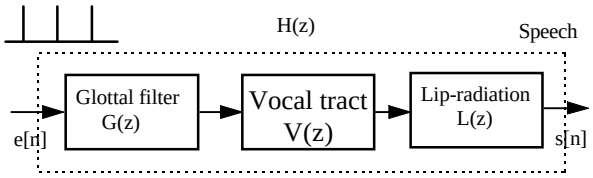

Figure 1: Model of speech production.

The combination of these three filters is defined to be the "vocal system" filter H(z). Assuming the spectrum of e[n] to be a series of lines at the pitch-frequency harmonics, it follows that at the frequency of each harmonic, the speech spectrum satisfies:

$$\left|S(e^{j\omega})\right| = \left|E(e^{j\omega})\right|\left|H(e^{j\omega})\right| = A\left|H(e^{j\omega})\right|$$

$$\begin{aligned}\arg(S(e^{j\omega})) &= \arg(E(e^{j\omega})) + \arg(H(e^{j\omega})) \\ &= -n_0\omega + 2k\pi + \arg(H(e^{j\omega}))\end{aligned}$$

where A is a constant determined by the amplitude of e[n], n_0 is the position of the excitation pulse and k is any integer.

The STC coder [2] encodes parameters which represent $\left|H(e^{j\omega})\right|$ as closely as possible. The phase spectrum $\theta(\omega)$ is derived at the decoder from the received version of $\left|H(e^{j\omega})\right|$ assuming that H(z) is minimum phase [4]. This assumption is not entirely true since the glottal filter G(z) is not minimum phase. The derived phase spectra will therefore be to some degree in error.

3. MODIFICATION USING ROSENBERG PULSES

One way of reducing this phase error is to assume that the impulse response of the glottal filter G(z) may be approximated by a Rosenberg pulse [5] as illustrated in Figure 2 (dashed line). Such pulses are commonly used to model the vocal tract excitation and are governed by three parameters which are the pitch-period, P, the "opening time" ,T_P, and the "closing time", T_N. Taking the opening and closing times to be fixed at 33% and 10% of P respectively, the magnitude and phase spectra, $R(\omega)$ and $\varphi(\omega)$ say, of a unity amplitude Rosenberg pulse may be readily calculated for a given pitch-period. Hence the assumed effect of G(z) on the spectrum of H(z) may be removed by dividing the magnitude of $H(e^{j\omega})$ by $R(\omega)$ and subtracting $\varphi(\omega)$ from its phase at each frequency ω. The resulting magnitude and phase spectra should then, in principle, correspond to the minimum phase transfer function KV(z)L(z) for some constant K.

This technique may be applied to an STC decoder by dividing the received magnitude spectrum of H(z) by $R(\omega)$, deriving a phase spectrum via a discrete Hilbert transform and then adding $\varphi(\omega)$ to obtain the required phase spectrum [4]. It was found [3] that the quality of speech obtained from STC could thus be improved without modifying the encoder. The Rosenberg pulse technique is, in principle, more applicable to versions of STC which represent the short term spectral envelope by Fourier transform or cepstral coefficients [2] rather than the parameters of an all-pole model [6]. This is because the Rosenberg pulse may contain the effect of zeros which are not well represented by the all-pole model. The all-pass filtering approach now proposed is applicable to both representations.

4. MODIFICATION USING ALL-PASS FILTER MODEL

The magnitude response of G(z), i.e. $|G(e^{j\omega})|$, is often assumed [7] to be close to the magnitude response of a second order all-pole integrator with transfer function

$$G_I(z) = \frac{1}{(1-\alpha z^{-1})\,(1-\beta z^{-1})}$$

The constants α and β are close to unity within the unit circle. One of the poles, α say, is often assumed to coincide with the single zero of the lip-radiation filter L(z), so that the low-pass effect of one of the poles of $G_I(z)$ on the magnitude spectrum of the speech is to some degree cancelled out by the high-pass filtering of the lip-radiation model.

However the impulse response of the minimum phase transfer function $G_I(z)$ can never resemble a typical glottal pulse which is clearly not minimum phase. A minimum phase signal, in comparison to all possible causal signals with exactly the same magnitude spectrum, will have maximum energy concentrated at the beginning of the waveform. This is clearly not the case with glottal pulses which tend to have rather slowly rising leading edges caused by the vocal cords opening relatively slowly and are terminated by much sharper trailing edges caused by sudden reduction in volume velocity as the vocal cords snap together.

The shapes of glottal excitation pulses may be better modelled as impulse responses of $G_I(z)$ time-reversed and appropriately delayed. The shapes of such time-reversed impulse responses can be made very close to those of typical Rosenberg pulses. The dotted line in figure 2 is the impulse response of a two pole model with $\alpha = 0.8$ and $\beta = 0.79$. The solid line is this impulse-response delayed and time-reversed, and the dashed line is a Rosenberg pulse with $T_P \approx 17$, $T_N \approx 5$ and $P \approx 50$ samples.

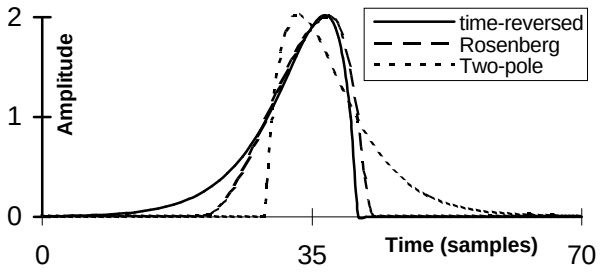

Figure 2: Waveforms of two-pole model, Rosenberg pulse & time-reversed pulse

The magnitude response of $G_I(z)$ is unaffected by the time-reversal whereas the phase response is changed from $\varphi(\omega)$ say to $-\varphi(\omega)$ disregarding a linear phase component.

Since $G_I(z)$ is minimum phase, applying a Hilbert transform to the spectral envelope of a segment of voiced speech will produce the phase spectrum of $G_I(z)V(z)L(z)$ rather than $G(z)V(z)L(z)$. The phase spectrum produced by STC will therefore not reflect the time-reversal referred to above. However, it can be corrected by subtracting the phase response of the "all-pass" filter:

$$G(z) / G_I(z) = \frac{(z-(1/\alpha))(z-(1/\beta))}{(z-\alpha)(z-\beta)}$$

which is

$$\phi(\omega) = 2\arctan\left(\frac{\alpha\sin\omega}{\alpha\cos\omega - 1}\right) + 2\arctan\left(\frac{\beta\sin\omega}{\beta\cos\omega - 1}\right) - 2\omega$$

Figure 3 (dotted line) shows the difference between the minimum phase spectrum obtained from the envelope of a 160-sample frame of male voiced speech and the true phases measured at pitch-frequency harmonics. This may be compared with the solid line which is the phase response $\phi(\omega)$ of an all-pass filter with optimised parameters α and β. There is clear similarity, indicating that the all-pass filter is capable of compensating, to some degree, for the inaccuracy of the minimum phase assumption

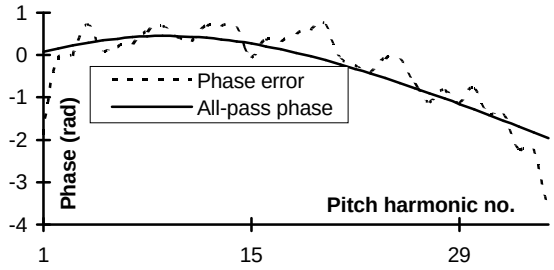

Figure 3: All-pass phase response & error between measured and minimum phase response

The values of α and β for the all-pass filter were calculated by minimising a "phase error":

$$\varepsilon = \sum_{\ell=1}^{L} \left[\Phi_o(\omega_\ell) - \Phi_d(\omega_\ell)\right]^2$$

where $\Phi_o(\omega_\ell)$ and $\Phi_d(\omega_\ell)$ denote the original and derived phases respectively at the pitch-frequency harmonic ω_ℓ, and L is the number of harmonics. The derived phase $\Phi_d(\omega_\ell)$ is the sum of the minimum phase $\theta(\omega_\ell)$, the all-pass phase $\phi(\omega_\ell)$ and a linear phase component $n_0\omega_\ell$. The value of n_0 is found as part of the optimisation procedure which requires the phase error and its gradients with respect to α, β and n_0 to determine the best combination.

Experiments were carried out to optimise the phase error for 412 frames of voiced speech from both male and female speakers. Histograms showing the distribution of α and β obtained are shown in figures 4 and 5.

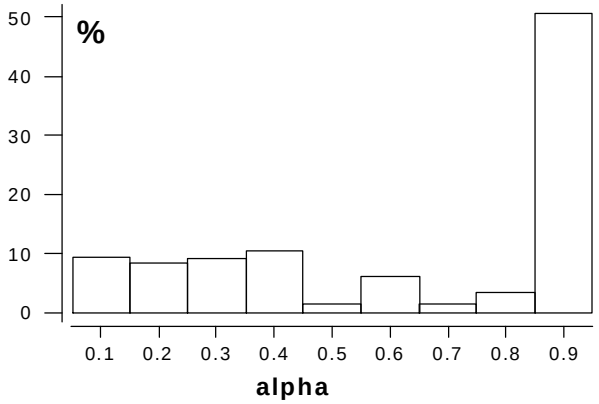

Figure 4: Percentage distribution of α

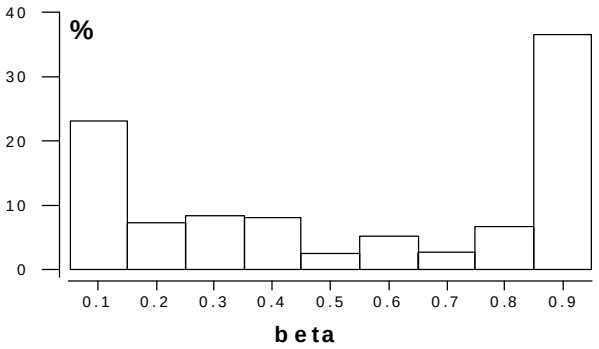

Figure 5: Percentage distribution of β

As expected, many values α and β occurred around 0.9. However, other values of α and β often occurred. This may be due to innacuracies in the linear phase component, the definition of the error measure (which may be better defined in terms of delay rather than phase) and the effect of IRS filtering. If α and β are to be encoded, several bits will be required. As this may not be possible for very low bit-rate STC, fixed values for α and β (=0.9) were tried and found to give significant improvement in derived phase accuracy.

5. RESULTS

Results obtained from the two modified forms of STC are summarised in figures 6 and 7, and compared with results from the original (minimum phase) method. Figure 6 shows the average phase error over 240 frames of voiced speech. It may be seen that both the Rosenberg pulse method and the all-pass filter method with fixed α and β can give significant improvement in the accuracy of phase derivation.

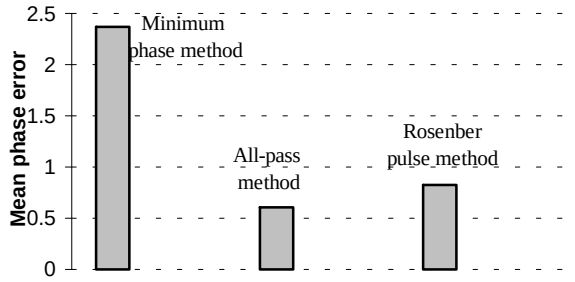

Figure 6: Mean phase error obtain with the three models.

Figure 7 summarises the results of an informal listening test where eleven subjects were asked to compare four samples of STC synthesised speech with different phase regeneration techniques:

(i) Synthesised with true phase
(ii) Synthesised with standard minimum phase assumption
(iii) Rosenberg pulse approximation (with T_P= 33% and T_N = 10% of pitch-period)
(iv) All-pass filter compensation (with fixed $\alpha = \beta = 0.9$)

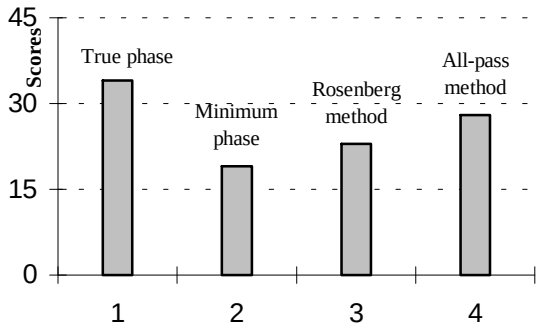

Figure 7: Listening scores obtained for the four synthetic speech

The speech segments included male and female voices. Each subject gave an order of preference for the synthesised speech quality and a score was recorded as follows: first = 4; second = 3; third = 2 and fourth = 1. Figure 7 shows the total score for each of the four samples. It was concluded that both the Rosenberg pulse method and the all-pass filter method have potential for improving synthetic speech quality as well as improving the accuracy of phase derivation without increasing the bit-rate of an STC coder.

6. CONCLUSION

The quality of STC voiced speech can be improved by reducing phase error due to the minimum phase assumption. This may be achieved by modifications based on assumptions about the shapes of the vocal tract excitation. For very low bit-rates it may not be possible to encode additional information about the excitation and the methods proposed have been found to be effective with fixed parameters. At higher bit-rates, the parameters can be optimised and encoded to further decrease the phase error.

7. REFERENCES

[1] R. J. McAulay & T. F. Quatieri, “Speech Analysis/Synthesis Based on a Sinusoidal Representation”. IEEE Trans. ASSP-34, 744-754, 1986.

[2] R. J. McAulay & T.F.Quatieri. "Low-Rate Speech Coding Based on the Sinusoidal Model", In: "Advances in Speech Signal Processing", ed. S. Furui & M. M. Sondhi, Marcel Dekker Inc., 1992.

[3] X. Q. Sun, B. M. G. Cheetham & W. T. K. Wong, “Spectral envelope and phase optimisation for sinusoidal speech coding”, Proc. IEEE Workshop on Speech Coding for Telecommunications, pp. 75-76, Annapolis, USA, September, 1995.

[4] A.V. Oppenheim & R. W. Schafer, “Digital Signal Processing”, Prentice-Hall, Englewood Cliffs, N.J., 1975.

[5] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", Journal of the Acoustical Society of America, Vol.49, No.2, pp.583-590, 1971.

[6] R. J. McAulay, T. Champion & T. F. Quatieri, “Sinewave amplitude coding using line spectral frequencies”, Proc. IEEE Workshop on Speech Coding for Telecommunications Speech Coding for the Network of the future, Canada, pp. 53 - 54, Oct. 1993.

[7] L. R. Rabiner & R. W. Schafer, “Digital processing of speech signals”, Prentice- Hall, Inc., 1978.