A New Phase Model for Sinusoidal Coding of Speech Signals

Sassan Ahmadi and Andreas S. Spanias

Department of Electrical Engineering
Telecommunications Research Center
Arizona State University
Tempe, AZ 85287-7206 USA

Abstract

A new phase modeling algorithm for sinusoidal analysis/synthesis of speech is presented. Short-time sinusoidal phases are approximated using a combination of linear prediction, spectral sampling, delay compensation, and phase correction techniques. The algorithm is different than phase compensation methods proposed for source-system LPC in that it has been optimized for sinusoidal representation of speech. Performance analysis on a large speech database indicates considerable improvement in temporal and spectral signal matching as well as improved subjective quality of the reconstructed speech. The extra parameters used for representation of the sine wave phases require a small number of bits. The method can be applied to enhance phase matching in low-bit rate sinusoidal coders, where underlying sine wave amplitudes are extracted from an all-pole model.

1 Introduction

The sinusoidal model represents speech by a linear combination of underlying sinusoids with time-varying amplitudes, phases, and frequencies [7],[8],[9]. Although successful techniques have been developed for the quantization of the sinusoidal amplitudes and frequencies, there is still a demand for more improvements in the performance of the sinusoidal phase models. The basic motivation for an efficient phase model lies in the fact that coarse quantization of sine wave phases usually results in a performance degradation. In fact, improper modeling and quantization of the sinusoidal phases may lead to strong reverberance in reconstructed speech.

A number of approaches to sine wave phase estimation have been proposed by Almeida *et al.* [1],[6] and McAulay and Quatieri [7],[8]. The approach taken by Almeida *et al.* exploits the correlation between harmonic phases of consecutive voiced segments. McAulay and Quatieri used a mixed-voicing sinusoidal representation where phases for the voiced portion were extracted from a spectral envelope under a minimum phase assumption, while phases for the unvoiced portion were randomized. In their recent work, the spectral envelope was parameterized in terms of all-pole polynomial coefficients [8].

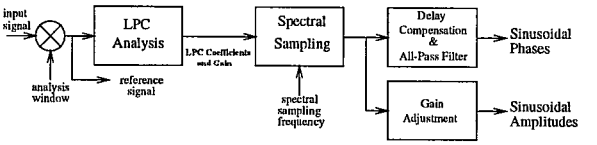

Figure 1: General block diagram of the proposed method.

In this paper, the use of an LPC analysis along with all-pass phase correction and delay compensation for simultaneous representation of the sinusoidal amplitude and phase parameters is proposed. The motivation for phase correction stems from the well-known fact that the LPC analysis does not usually provide correct phase information. The inclusion of the all-pass phase correction scheme in the proposed sinusoidal phase model was inspired by improvements in source-system LPC reported by Hedelin [3],[4], Trancoso *et al.* [10], and Honda [5]. It must be noted that the proposed algorithm is different than phase-compensated LPC-based methods in that: a) spectral sampling and delay compensation are also integrated into the phase correction process, b) the proposed method is applied to sinusoidal analysis/synthesis which is distinctly different than source-system LPC. In addition, the proposed method can be viewed as an alternative sinusoidal phase model relative to [8] in the sense that it captures not only the vocal tract phase effects but also glottal effects through the use of an all-pass phase compensation stage that is shown, through statistical analysis, to work well with a variety of speech segments (i.e., voiced, unvoiced, onset, and transition).

Figure 1 shows the general block diagram of the proposed method. A low-order LPC analysis is performed on a windowed speech segment and the LPC coefficients and gain are used to form an all-pole transfer function (i.e., $H(z)$ in Fig. 2), which is sampled at integer multiples of a predetermined spectral sampling frequency. The spectral sampling frequency corresponds to the fundamental frequency during voiced speech while for unvoiced segments a constant frequency of 70 Hz is used. A delay compensation algorithm is used to compensate for the time-varying delay, *jitter*, and to achieve maximum alignment between the reference and the in-

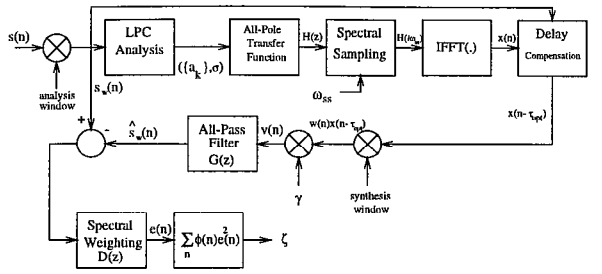

Figure 2: All-pass filter optimization procedure.

put to the all-pass filter. Improvement in temporal and spectral matching is achieved by introducing an all-pass filter [2],[4]. The phase response of the all-pass filter approximates the phase difference between the reference signal and the input to the all-pass filter.

To evaluate the performance of the proposed algorithm, a comprehensive statistical analysis was performed using temporal and spectral distortion measures and speech data taken from the TIMIT database. The results of this analysis reveal small phase estimation error at low frequencies and improved performance over other sinusoidal phase modeling techniques.

A simplified version of the phase model was also developed, where the phase parameters were reduced to an optimal delay, a coefficient, and an integer denoting the all-pass filter order. Experimental results show that considerable improvement in temporal and spectral matching can still be achieved using the simplified algorithm. The simplified representation is intended for low-bit rate coding of sinusoidal phases.

The rest of the paper is organized as follows. In the next section, a detailed description of the algorithm is given. In section 3, the distortion measures used to evaluate the performance of the proposed algorithm are defined and experimental results are presented. Concluding remarks are given in section 4.

2 Description of the Algorithm

In this algorithm, the harmonic sinusoidal model [7] is used to represent the speech signal. If $s_w(n)$ represents a windowed speech segment, then

$$s_w(n) = \sum_{l=1}^{L} A_l \cos(l\omega_{ss}n + \psi_l) \qquad n = 0, ..., N-1 \quad (1)$$

where A_l and ψ_l denote the time-varying amplitudes and phases of the underlying sinusoids, respectively, ω_{ss} is the spectral sampling frequency, and L is the total number of spectral samples over the signal bandwidth. The spectral sampling frequency corresponds to the fundamental frequency during voiced speech segments and to a constant frequency of 70 Hz for unvoiced segments. The input speech is first analyzed by a Hanning window of length N. A Pth order LPC analysis, based on the autocorrelation method, is performed. In order to avoid sharp spectral peaks in the LPC spectrum a fixed 10 Hz bandwidth expansion is applied to the poles of the minimum-phase all-pole transfer function, which represents the time-varying characteristics of the vocal tract. This transfer function is given as follows:

$$H(z) = \frac{\sigma}{1 + \sum_{k=1}^{P} a_k z^{-k}} \quad (2)$$

where the gain, σ, and the coefficients, $\{a_k\}_{k=1}^{P}$, are computed over a frame of 30 ms long, and updated every 15 ms. The complex-valued all-pole transfer function, $H(z)$, is sampled at integer multiples of ω_{ss} as shown in Fig. 2. The time-domain signal corresponding to these samples does not match the original waveform due to the lack of correct short-term phase components. Improved temporal and spectral matching are achieved by introducing an all-pass filter, $G(z)$. The input to the all-pass filter, $v(n)$, has to be preprocessed to ensure maximum alignment between $s_w(n)$ and $v(n)$. This is done by a delay compensation algorithm which minimizes the following weighted squared error in the time-domain:

$$\xi(\tau) = \left\{ \sum_{n=0}^{N-1} [s_w(n) - w(n)x(n-\tau)]^2 \right\}^{\frac{1}{2}} \quad (3)$$

where $x(n)$ denotes the inverse Fourier transform of the output of the spectral sampling unit, and $w(n)$ is the synthesis window, usually chosen to be the same as the analysis window. The optimum delay is found as:

$$\tau_{opt} = \arg\min_{\tau} [\xi(\tau)] \quad (4)$$

In practice, τ is an integer and the interval $[\tau_{min}, \tau_{max}]$ is searched for the value that minimizes the error in (3). The delay compensated signal, $x(n-\tau_{opt})$, is windowed and scaled by a gain, γ, to obtain the closest signal to $s_w(n)$. The gain is given by:

$$\gamma = \frac{\sum_{n=0}^{N-1} s_w(n)w(n)x(n-\tau_{opt})}{\sum_{n=0}^{N-1} [w(n)x(n-\tau_{opt})]^2} \quad (5)$$

Let us define $\mathbf{s}_w = (s_w(0), s_w(1), ..., s_w(N-1))^T$ and $\mathbf{v} = (v(0), v(1), ..., v(N-1))^T$, the weighted squared error at the input of the all-pass filter can be expressed as follows:

$$\epsilon_1 = (\mathbf{s}_w - \mathbf{v})^T \mathbf{\Phi} (\mathbf{s}_w - \mathbf{v}) \quad (6)$$

where $\mathbf{\Phi} = \text{diag}(\phi(0), \phi(1), ..., \phi(N-1))$ is an appropriate $N \times N$ diagonal weighting matrix in the time-domain, and the superscript T denotes the transpose operation. Since our primary focus is to minimize the

phase difference between $\mathbf{s}_w$ and $\mathbf{v}$, a linear all-pass filter of order M is introduced as follows:

$$G(z) = \frac{D(z^{-1})}{D(z)} = \frac{\theta_0 + \sum_{k=1}^{M} \theta_k z^k}{\theta_0 + \sum_{k=1}^{M} \theta_k z^{-k}} \tag{7}$$

where $\mathbf{\Theta} = (\theta_0, \theta_1, ..., \theta_M)^T$ can be computed using a weighted least squares (WLS) method. Without loss of generality, we can assume that $\theta_0 = 1$. Since $G(z)$ is assumed to be all-pass, the phase response of $G(e^{j\omega})$ ideally represents the phase difference between $\mathbf{s}_w$ and $\mathbf{v}$. The objective for the all-pass filter is to choose $\{\theta_k\}$ such that $\hat{\mathbf{s}}_w = (\hat{s}_w(0), \hat{s}_w(1), ..., \hat{s}_w(N-1))^T$ becomes as close as possible to $\mathbf{s}_w$. In that case, the weighted squared error

$$\epsilon_2 = (\mathbf{s}_w - \hat{\mathbf{s}}_w)^T \mathbf{\Phi} (\mathbf{s}_w - \hat{\mathbf{s}}_w) \tag{8}$$

after minimization would be less than ϵ_1.

A spectral weighting function is applied to shape the error between $\mathbf{s}_w$ and $\hat{\mathbf{s}}_w$. There are some advantages associated with the application of a spectral weighting function, such as improvement of the subjective quality of the output (i.e., when a perceptual type of weighting is used), and reduction in computational complexity. In fact, if the spectral weighting function is chosen as $D(z)$, a linear minimization problem is involved [4], in that case, the instantaneous error vector is obtained as:

$$\mathbf{\Delta} = \mathbf{X}\mathbf{\Theta} \tag{9}$$

where $\mathbf{X}$ is an $N \times M+1$ matrix defined as follows:

$$\begin{aligned} \mathbf{X} \;=\; & [\mathbf{s}_w(n) - \mathbf{v}(n) \mid \mathbf{s}_w(n-1) - \mathbf{v}(n+1) \mid ... \\ & ... \mid \mathbf{s}_w(n-M) - \mathbf{v}(n+M)] \end{aligned} \tag{10}$$

It is desirable to find the parameters $\mathbf{\Theta}$ and M such that the following weighted squared error is minimized:

$$\min_{\mathbf{\Theta}, M} [\ \zeta = (\mathbf{X}\mathbf{\Theta})^T \mathbf{\Phi} (\mathbf{X}\mathbf{\Theta})\ ] \tag{11}$$

For convenience, the above equation is written as $\zeta = \mathbf{\Theta}^T \mathbf{R} \mathbf{\Theta}$, where $\mathbf{R}$ is a symmetric and non-negative definite $M+1 \times M+1$ matrix defined as:

$$\mathbf{R} \triangleq \mathbf{X}^T \mathbf{\Phi} \mathbf{X} = \begin{pmatrix} r_{00} & r_{01} & ... & r_{0M} \\ r_{10} & r_{11} & ... & r_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M0} & r_{M1} & ... & r_{MM} \end{pmatrix} \tag{12}$$

To obtain the coefficients $\tilde{\mathbf{\Theta}} = (\theta_1, \theta_2, ..., \theta_M)^T$, equation (12) is rewritten as:

$$\mathbf{R} = \left(\begin{array}{c|c} r_{00} & \tilde{\mathbf{r}}^T \\ \hline \tilde{\mathbf{r}} & \tilde{\mathbf{R}} \end{array} \right) \tag{13}$$

The parameters of the all-pass filter can be obtained by setting the gradient of the weighted squared error, ζ, to zero, which yields:

$$\tilde{\mathbf{\Theta}} = -\tilde{\mathbf{R}}^{-1} \tilde{\mathbf{r}} \tag{14}$$

where $\tilde{\mathbf{r}} = (r_{10}, r_{20}, ..., r_{M0})^T$.

The improvements brought by the introduction of the all-pass filter over previous implementations of the sinusoidal model depend on the order of the all-pass filter. Experiments have shown that good results are obtained using all-pass filters of orders 12 to 18.

The stability of the all-pass filter is important for quantization purposes. On the other hand, steady-state sinusoidal analysis/synthesis requires that the all-pass filter to be stable so that the Fourier transform and consequently sampling on the unit circle can be defined. If no constraints are imposed, the minimization of (11) can lead to an unstable all-pass filter. An alternative approach is to directly minimize ϵ_2, that would result in a non-linear minimization problem, where an iterative method may provide an approximation to the analytical solution. The stability can be checked within the iteration and the algorithm is terminated at the last stable stage [4]. The use of FIR all-pass filters might be another approach to guarantee the stability of the all-pass filter. Here, we opted to use the linearized model (i.e., using $D(z)$ as the spectral weighting function) to reduce in the computational complexity of the algorithm.

The short-time spectrum is synthesized based on the estimated sine wave amplitudes and phases and then inverse Fourier transformed, windowed, and overlap-added to reproduce the output speech.

For lower rate speech coding applications, a simplified version of the all-pass filter is introduced as follows:

$$G(z) = \frac{1 + \alpha z^M}{1 + \alpha z^{-M}} \tag{15}$$

Therefore, the weighted squared error in (11) is reduced to:

$$\zeta(M) = \sum_{n=0}^{N-1} \phi(n)[s_w(n) + \alpha s_w(n-M) - v(n) - \alpha v(n+M)]^2 \tag{16}$$

where

$$\alpha = -\frac{\sum_{n=0}^{N-1} \phi(n)[s_w(n) - v(n)][s_w(n-M) - v(n+M)]}{\sum_{n=0}^{N-1} \phi(n)[s_w(n-M) - v(n+M)]^2} \tag{17}$$

The order of the all-pass filter is calculated as:

$$M_{opt} = \arg\min_{M} [\zeta(M)] \tag{18}$$

In practice, M is found through a search process in the interval $[1, M_{max}]$. The order and single coefficient of the simplified all-pass filter can be encoded using a scalar quantization scheme. Preliminary results with

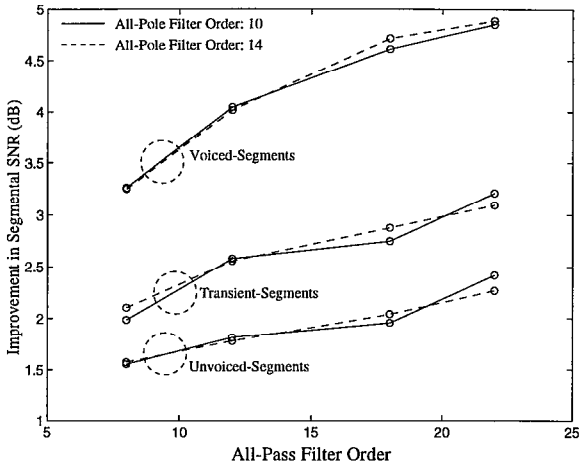

Figure 3: Average improvement in signal to noise ratio (SNRI) as a function of all-pass filter order using 20 ms analysis/synthesis windows.

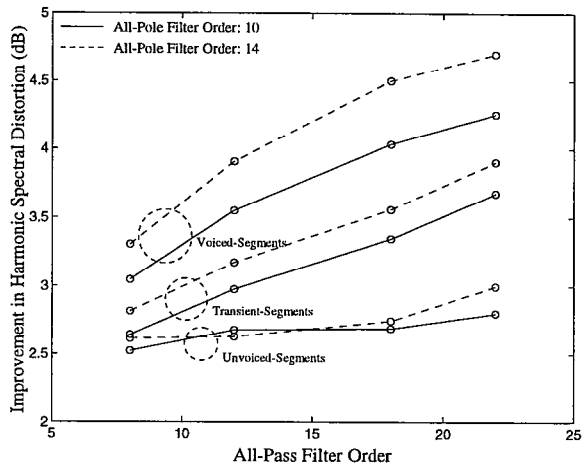

Figure 4: Average improvement in harmonic spectral distortion (HSDI) as a function of all-pass filter order using 20 ms analysis/synthesis windows.

this scheme are reported in the next section. Application of the simplified model in low rate sinusoidal coding is part of our ongoing research.

3 Experimental Results

To evaluate the performance of the proposed algorithm, a comprehensive statistical analysis was carried out. Two distortion measures were defined and computed for 50,000 speech segments taken from the TIMIT database. The first measure is the average improvement in signal to noise ratio (SNRI), which is defined as follows:

$$SNRI = \frac{1}{K}\sum_{k=1}^{K} 10\log\left(\frac{\epsilon_{1k}}{\epsilon_{2k}}\right) \tag{19}$$

where ϵ_{1k} and ϵ_{2k} are the weighted squared errors before and after all-pass filtering for the kth frame, respectively, and K denotes the total number of frames used in the experiment. This measure is used as an indication of the degree of improvement in temporal waveform matching. Figure 3 shows the average SNRI as a function of the all-pass filter order with the all-pole filter order as a parameter. The second measure reflects the improvement in harmonic spectral distortion (HSDI), which is defined as follows:

$$HSDI = \frac{1}{K}\sum_{k=1}^{K} 10\log\left(\frac{\sum_{l=1}^{L_k}|S_w(l\omega_{ss}) - V(l\omega_{ss})|^2}{\sum_{l=1}^{L_k}|S_w(l\omega_{ss}) - \hat{S}_w(l\omega_{ss})|^2}\right) \tag{20}$$

where $S_w(\omega)$, $V(\omega)$, and $\hat{S}_w(\omega)$ are the Fourier transforms of the signals $s_w(n)$, $v(n)$, and $\hat{s}_w(n)$, respectively, and L_k denotes the number of spectral samples in the spectrum of the kth segment. This measure is used as an indication of the degree of improvement in spectral matching at integer multiples of the spectral sampling frequency. The average HSDI as a function of the all-pass filter order with the all-pole filter order as a parameter is illustrated in Fig. 4. It can be concluded, that elaborate design of the all-pass filter (i.e., the way in which the phase difference is approximated) results in improved reproduction of the original signal both in the time and frequency domains. The performance of the simplified model (15) has also been evaluated and shown in Fig. 5. It can be seen, that considerable improvement in temporal and spectral matching is achieved using a small number of parameters to represent the sinusoidal phases.

The statistical distribution of the phase residuals (i.e., the difference between the original and estimated sinusoidal phases) in different frequency bands, obtained from the proposed algorithm, are shown in Fig. 6. The data was extracted based on the analysis performed on a large number of speech segments of different categories (i.e., voiced, unvoiced, onset, and transition). It can be seen that the proposed method yields small phase residuals in low frequencies and the variance of the phase residuals gradually increases with increasing frequency. The significance of the statistical distributions, shown in Fig. 6, lies in the fact, that the human ear is more sensitive to low frequency phase information. Hence, the phase approximations have to be more accurate in low frequencies (i.e., small phase estimation error in low frequencies).

4 Conclusions

A new sinusoidal phase model was presented. The method may find applications to low-bit rate sinusoidal coders, in which LPC envelope is used to represent the sine wave amplitudes. Efficient representation for sinu-

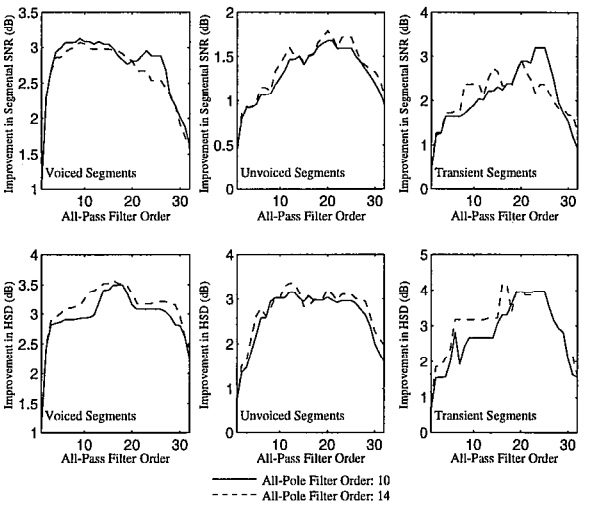

Figure 5: Average improvements in signal to noise ratio (SNRI) and harmonic spectral distortion (HSDI) obtained from the simplified model using 20 ms analysis/synthesis windows.

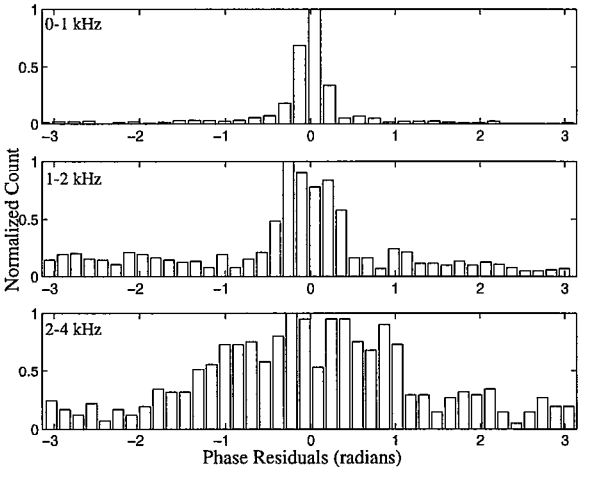

Figure 6: Statistical distribution of the phase residuals obtained from the proposed algorithm in different frequency bands using 30 ms analysis/synthesis windows and LPC analysis of order 14. The general all-pass filter of order 12 is used.

soidal phases is obtained by cascading an all-pass filter to the all-pole filter, where the all-pass filter is used for phase correction. Performance analysis on a large database reveals considerable improvement in matching between the original and reconstructed signals both in the time and frequency domains. Informal listening tests revealed that the use of the proposed phase model results in improved subjective quality of the output speech. A simplified version of the phase model that is well suited for low-bit rate speech coding applications was also presented.

References

[1] L. B. Almeida, and J. M. Tribolet, "Non-stationary spectral modeling of voiced speech", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-31, pp. 664-678, June 1983.

[2] C. K. Chen, and J. H. Lee, "Design of digital all-pass filters using a weighted least squares approach" *IEEE Trans. on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 41, pp. 346-350, May 1994.

[3] P. Hedelin, "High quality glottal LPC-vocoding" *in Proc. IEEE ICASSP-86*, pp. 465-468, 1986.

[4] P. Hedelin, "Phase compensation in all-pole speech analysis" *in Proc. IEEE ICASSP-88*, pp. 339-342, 1988.

[5] M. Honda, "Speech coding using waveform matching based on LPC residual phase equalization" *in Proc. IEEE ICASSP-90*, pp. 213-216, 1990.

[6] J. S. Marques, L. B. Almeida, and J. M. Tribolet, "Harmonic coding at 4.8 kb/s" *in Proc. IEEE ICASSP-90*, pp. 17-20, 1990.

[7] R. J. McAulay, and T. F. Quatieri, "Low-rate speech coding based on the sinusoidal model", *Advances in Speech Signal Processing*, Chapter 6, S. Furui, and M. M. Sondhi Eds., Marcel Dekker, Inc., New York, 1992.

[8] R. J. McAulay, and T. F. Quatieri, "Sinusoidal coding", *Speech Coding and Synthesis*, Chapter 4, W. B. Kleijn, and K. K. Paliwal Eds., Elsevier, 1995.

[9] A. S. Spanias, "Speech coding: a tutorial review" *Proc. IEEE*, vol. 82, pp. 1541-1582, Oct. 1994.

[10] I. M. Trancoso, R. Garcia-Gomez, and J. M. Tribolet, "A study on short-time phase and multi-pulse LPC" *in Proc. IEEE ICASSP-84*, pp. 10.3.1-10.3.4, 1984.