AN IMPROVED RESIDUAL-DOMAIN PHASE/AMPLITUDE MODEL FOR SINUSOIDAL CODING OF SPEECH AT VERY LOW BIT RATES: A VARIABLE RATE SCHEME

Sassan Ahmadi
Nokia Mobile Phones, Inc.
San Diego, CA 92121 USA
Sassan.Ahmadi@nmp.nokia.com

ABSTRACT

An improved harmonic sinusoidal model is presented, where the underlying sine wave amplitudes and phases are efficiently represented using a combination of linear prediction, linear phase alignment, all-pass filtering, and spectral sampling in the residual-domain. The analysis and synthesis systems are introduced and the derivation and encoding of each model parameter is discussed. Performance analysis on a large database indicates effective modeling of the sinusoidal parameters. A variable-rate sinusoidal coder operating at an average bit rate of 1.75 kbps, based on the proposed model, has been developed, yielding reproduced speech of good quality, intelligibility, and naturalness. The proposed model may find applications in low bit rate speech coding in high capacity wireless communication systems.

1. INTRODUCTION

Sinusoidal coders are among the most prominent speech coders capable of reproducing speech of good quality and intelligibility at low bit rates. The use of linear predictive (LP) analysis along with all-pass phase correction and delay compensation for simultaneous representation of sinusoidal amplitudes and phases was recently proposed in [2]. The inclusion of the all-pass phase correction scheme had been inspired by improvements in source-system LPC reported in the literature (e.g., [7]). It was later realized that representation of the residual signal is indeed easier than a direct attempt to model the speech signal using a sinusoidal model. In fact, only perceptually important features of the residual signal have to be preserved. This observation has been the basis for LPC-based speech coding systems for many years. Although sinusoidal representation in the residual-domain has been earlier reported by other researchers [4], the successful use of LPC spectral envelope to model the sinusoidal amplitudes as well as improvements in phase prediction obtained by utilizing an all-pass phase correction scheme motivated the design and development of this improved sinusoidal model for speech coding at very low bit rates.

In the baseline model, the residual signal obtained from a low-order LPC analysis is approximately represented with a harmonic series, where constant sine wave amplitudes and measured phases are employed. It can be shown that the use of properly-scaled constant amplitudes in the residual expansion is equivalent to representing the sinusoidal amplitudes, corresponding to the short-time spectrum of speech, by samples of the LPC envelope at integer multiples of the fundamental frequency [8]. In practice, a two-stage phase prediction algorithm consisting of linear phase alignment and all-pass filtering is used to appropriately model the phases. Speech and silence internals are detected using an open-loop multi-feature classifier. Silence frames are encoded at a lower bit rate without phase information.

The performance of the proposed model was evaluated based on a statistical analysis using speech data taken from TIMIT database. The results of this analysis indicate very small phase prediction error and reasonably small amplitude prediction error for all classes of speech (i.e., voiced, unvoiced, and transition). A variable rate sinusoidal coder at average bit rate of 1.75 kbps has been developed based on the proposed model which yields reproduced speech of good quality, intelligibility, and naturalness.

This paper is organized as follows. In the next section, a detailed description of the proposed analysis and synthesis system is given. The quantization issues are discussed in section 3. In section 4, two meaningful distortion measures are defined and the results of the statistical analysis are presented. Concluding remarks are given in section 5.

2. DESCRIPTION OF THE ALGORITHM

In this algorithm, a harmonic sinusoidal model is used to represent the short-time segments of speech and residual signal [9],[10]. While the parameters of the model remain the same for all classes of speech, silence intervals are treated differently. This is due to the fact that silence intervals are generally perceptually insignificant. Furthermore, there is roughly a high percentage of silence frames in a real two-way communication. The results of classification and voicing activity detection (VAD) are reflected in two bits, selecting between three possible modes of operation; i.e., voiced, unvoiced, and silence. Although there is not any difference in the nature and the number of parameters used for voiced and unvoiced speech, different codebooks are used to quantize transformed LP coefficients corresponding to unvoiced frames, making the system more robust to classification errors. The pitch frequency and classification data are obtained using an open-loop classifier based on the analysis of certain features of speech such as short-time energy, cepstral peak, zero crossing rate, average magnitude difference function, and energy ratio [1]. A 10 ms look-ahead has been included in the classification process. The following subsections provide the detailed description of the analysis and synthesis systems.

2.1. Analysis System

The block diagram of the analysis system is shown in Fig. 1. The input speech, $s(n)$, is first analyzed by a Hanning window of length N. A harmonic model is used to represent the windowed speech as follows:

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

where α_l and θ_l denote the time-varying amplitude and phase of the lth harmonic, ω_{ss} is the spectral sampling frequency, and L is the total number of spectral samples over the entire signal bandwidth. The spectral sampling frequency corresponds to the fundamental frequency during

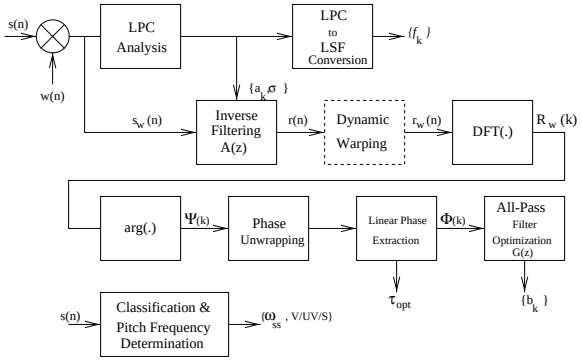

Figure 1. Block diagram of the analysis system

voiced speech segments. For the unvoiced frames a random sampling frequency uniformly distributed over the interval (50-100 Hz) is utilized. The use of a random sampling frequency in the unvoiced intervals was found very effective in significant reduction of the buzziness caused by imposition of a harmonic structure in the noise-like spectrum corresponding to the unvoiced speech. A 10th order LPC analysis, based on the autocorrelation method, is performed. In order to avoid sharp spectral peaks in the LPC spectrum, a fixed 30 Hz bandwidth expansion ($\beta = 0.988$) is applied to the poles of the all-pole transfer function, which is given as follows:

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

where P denotes the order of the LPC analysis. The gain σ and the LP coefficients $\{a_k\}$ are computed over a frame of 20 ms (i.e., $N = 160$ given an 8000 Hz sampling rate) and updated every 20 ms. The inverse filter $A(z)$ is then used to derive the LPC residual signal $r(n)$, whose harmonic representation is given as follows:

$$r(n) = \sum_{l=1}^{L} A_l \cos(l\omega_{ss} n + \phi_l) \qquad n = 0, 1, ..., N-1 \quad (3)$$

where A_l and ϕ_l denote the amplitude and phase corresponding to the lth harmonic in the short-time spectrum of $r(n)$.

The use of an invertible warping function, $\zeta(n)$, to improve the periodicity of the residual signal during voiced intervals is under investigation . In [3], it was shown that the use of a dynamic warping function whose derivative was inversely proportional to the instantaneous pitch variation function improved the performance of the harmonic model by adjusting the location of the harmonic peaks in the spectrum of voiced frames. The dynamically warped residual $r_w(n) = \zeta[r(n)]$ is referred to as the modified residual throughout this paper. In current embodiment of the algorithm no warping function is used.

It can be shown that the use of constant amplitudes along with measured phases in the harmonic representation of the modified residual signal results in reproduced speech of good quality. This important observation suggests that good quality speech can be synthesized provided that a sufficiently accurate estimation for the underlying sine wave phases corresponding to the residual signal is given. Based on the successful phase model described in [2], the analysis system shown in Fig. 1 was derived. In fact, the use of constant amplitude scheme is another interpretation and implementation of an earlier observation reported in [8], where underlying sine wave amplitudes were obtained by sampling the LPC spectral envelope at integer multiples of the fundamental frequency during voiced intervals and a constant spectral sampling frequency in unvoiced sections.

The DFT of the modified residual signal is taken and its short-time phase spectrum is extracted and unwrapped. If $R_w(k)$ represents the short-time Fourier transform corresponding to $r_w(n)$, then the target phase function $\Psi(k)$ can be defined as follows:

$$\Psi(k) = \arg\left[R_w(k)\right] \qquad (4)$$

A general all-pass filter is introduced as follows:

$$T(z) = G(z) z^{-\tau} = \frac{\sum_{m=0}^{M-1} b_m z^m}{\sum_{m=0}^{M-1} b_m z^{-m}} z^{-\tau} \qquad (5)$$

where M denotes the order of the IIR all-pass filter $G(z)$. The phase response of $T(e^{j\omega})$ ideally represents the phase spectrum of the modified residual signal. To compute the parameters of the all-pass filter, the phase estimation procedure is divided into two separate stages. In the first stage, a linear phase component is computed and subtracted from the target phase to further reduce the entropy of the target phase function, simplifying the approximation of the remaining phase information with a lower order IIR all-pass filter. The optimum delay τ can be computed using a least squares method as follows:

$$\tau_{opt} = \arg\left\{ \min_{\tau} \left[\sum_{k=0}^{K-1} \left(\Psi(k) - \frac{2\pi}{K} k\tau \right)^2 \right] \right\} \qquad (6)$$

where K is the number of DFT points. A closed-form expression for τ_{opt} is obtained by setting the derivative of the squared error in (6) to zero.

After computing the linear phase component, the target phase for the IIR all-pass filter can be written as follows:

$$\Phi(k) = \Psi(k) - \frac{2\pi}{K} k\tau_{opt} \qquad k = 0, 1, ..., \frac{K}{2} \qquad (7)$$

The objective for the design of the IIR all-pass filter $G(e^{j\omega})$ is to find the set of coefficients $\mathbf{b} = \{b_0, b_1, ..., b_{M-1}\}$ and the parameter M such that the following error is minimized in a meaningful sense; i.e.,

$$\min_{M,\mathbf{b}} \left\| W(k) \left[\Phi(k) - \hat{\Phi}(k) \right] \right\| \qquad k = 0, 1, ..., \frac{K}{2} \qquad (8)$$

where $\hat{\Phi}(k) = \arg[G(e^{j\omega})]$ and $W(k)$ is an appropriate spectral/perceptual weighting function. For example, in [5], an algorithm based on the minimization of the phase error in a Chebyshev sense is described, where it is shown that the resulting minimax optimization problem can be approximated using a weighted least squares (WLS) approach [1],[5]. In fact, the proposed model is not restricted to any particular algorithm for this purpose and any efficient all-pass filter design scheme may be exploited [7].

The last parameter to be determined is a gain factor γ which is found through an analysis-by-synthesis procedure. It can be shown that γ satisfies the following equation:

$$\gamma = \frac{\sum_{n=0}^{N-1} r_w(n) e(n - \tau_{opt})}{\sum_{n=0}^{N-1} [e(n - \tau_{opt})]^2} \qquad (9)$$

where

$$e(n) = \sum_{l=1}^{L} A \cos(l\omega_{ss} n + \hat{\phi}_l) \qquad n = 0, 1, ..., N-1 \quad (10)$$

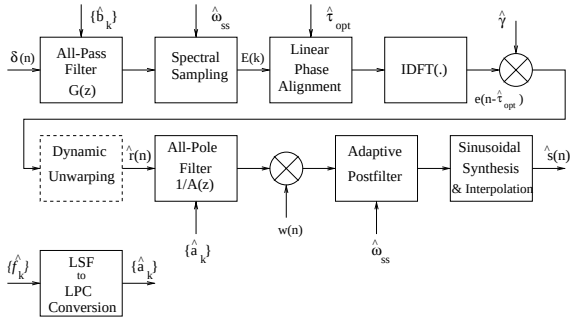

Figure 2. Block diagram of the synthesis system

denotes the constant amplitude approximation to the modified residual signal using the estimated phase $\hat{\phi}_l$ obtained from the phase response of the IIR all-pass filter $G(e^{j\omega})$.

The silence frames, on the other hand, are simply represented by a set of LP coefficients and a scaling factor used to adjust the reconstruction energy. These quantities are quantized and encoded along with two mode bits and transmitted to the receiver.

2.2. Synthesis System

The synthesis system is depicted in Fig. 2. The mode bits are examined first in order to determine how to interpret the information received at the receiver. The unit-sample sequence, $\delta(n)$, at the input of all-pass filter $G(z)$ is used as the excitation to obtain the impulse response or correspondingly the frequency response of the filter. The frequency response of the all-pass filter is then sampled at integer multiples of the quantized spectral sampling frequency. A linear phase component is added and the resulting complex-valued spectrum is inverse Fourier transformed to obtain the best approximation to $e(n-\tau_{opt})$. This estimate is further refined using the quantized scaling factor $\hat{\gamma}$ and used as an approximation to the residual signal, assuming that no warping function is used. The short-time speech segment is reconstructed by passing the residual signal through the all-pole synthesis filter, $H(z)$. The output of the synthesis filter is then weighted using a Hanning window of length N.

An adaptive postfiltering scheme proposed by Chen and Gersho [6] is used to enhance the perceptual quality of the coded speech. The postfilter consists of a long-term postfilter, cascaded with a short-term postfilter, along with a spectral tilt-compensation filter. The short-term postfilter emphasizes speech formants and attenuates the quantization noise in the spectral valleys between the formants. On the other hand, the long-term postfilter is used to emphasize the pitch harmonics during voiced speech segments and to deemphasize the spectral valleys between harmonics.

Two different approaches may be taken for synthesis. One possibility is the use of a mid-frame interpolation algorithm, effectively yielding a frame update rate of 10 ms [9],[10]. However, the other alternative, which is known as the variable frequency synthesis scheme [8],[9], has been found to be more effective [12]. In this method, the speech is synthesized directly in the time-domain by a superposition of sinusoids with continuously varying amplitudes and phases; i.e.,

$$\hat{s}(n) = \sum_{l=1}^{L} \hat{\alpha}_l^k(n) \cos\left(\hat{\theta}_l^k(n)\right) \qquad (11)$$

where the superscript k denotes the frame number. In order to ensure amplitude continuity at the frame boundaries, a linear interpolation scheme is adopted as follows:

$$\hat{\alpha}_l^k(n) = \hat{\alpha}_l^k + (\hat{\alpha}_l^{k+1} - \hat{\alpha}_l^k)(\frac{n}{N}) \qquad (12)$$

where $\{\hat{\alpha}_l^k\}$ and $\{\hat{\alpha}_l^{k+1}\}$ denote the sinusoidal amplitudes corresponding to the kth and $(k+1)$th frames, respectively. The phase and frequency continuity, on the other hand, can be ensured by taking a cubic evolution approach for the phase, or equivalently, a quadratic evolution for its first derivative which is the instantaneous frequency [9]; i.e.,

$$\hat{\theta}_l^k(n) = \eta_{0k}^l + \eta_{1k}^l n + \eta_{2k}^l n^2 + \eta_{3k}^l n^3 + 2\pi\lambda \qquad (13)$$

where λ is an integer-valued parameter used to unwrap the phases and the parameters η_{0k}^l, η_{1k}^l, η_{2k}^l, η_{3k}^l can be determined in terms of frame length N and the sinusoidal parameters $\{\hat{\omega}_{ss}^k, \hat{\theta}_l^k\}$ and $\{\hat{\omega}_{ss}^{k+1}, \hat{\theta}_l^{k+1}\}$ corresponding to the kth and $(k+1)$th frames, respectively [9]. At the bit rates we are targeting, this is probably the most eligible synthesis method.

The silence frames, on the other hand, are reconstructed by passing a white noise through the all-pole synthesis filter and adjusting the output energy using the quantized scaling factor. The resulting signal is weighted by the synthesis window and postfiltered.

3. QUANTIZATION OF THE PARAMETERS

For robust and efficient transmission, the LPC parameters and the coefficients of the stable IIR all-pass filter are transformed into line spectral frequency (LSF) parameters [1],[11].

In order to achieve reasonably small spectral distortions when a limited number of bits are available, the differences in the spectral structures of voiced and unvoiced frames were exploited to encode the LSF parameters. Therefore, separate codebooks were developed and optimized for voiced and unvoiced frames. A weighted Euclidean distance $d(\mathbf{f}, \hat{\mathbf{f}})$, between the original LSF vector $\mathbf{f} = (f_1, f_2, ..., f_P)$ and the quantized LSF vector $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, ..., \hat{f}_P)$, is used to select a codeword in the codebook [11]. The distance is given as follows:

$$d(\mathbf{f}, \hat{\mathbf{f}}) = \sum_{i=1}^{P} W(f_i)\mu_i(f_i - \hat{f}_i)^2 \qquad (14)$$

where $W(f)$ is an appropriate perceptual weighting function. In the above formula, $\{\mu_i\}_{i=1}^P$ is a constant weight vector which is used to give more weight to the lower LSFs than to the higher LSFs.

The gain, spectral sampling frequency, and linear phase alignment parameters are encoded separately using scalar quantizers. The bit allocation for each parameter at the average rate of 1.75 kbps is shown in Table 1. This is under the assumption that in a two-way speech communication, silence intervals approximately make up 60% of the conversation time, whereas active speech makes up 40% of the overall conversation.

Since the human auditory perception is uniformly sensitive to frequency errors in logarithmic scale, the logarithm of the spectral sampling frequency is uniformly quantized on [50-400 Hz] region and encoded in 7 bits for active speech.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed model, a comprehensive statistical analysis was carried out. The clean speech data used in the experiments was taken from TIMIT

Table 1. Bit allocation in the proposed 1.75 kbps variable rate sinusoidal coder

Parameter	Speech	Silence
Mode (V/UV/S)	2	2
Spectral Sampling Frequency	7	-
LSF Parameters (P=10)	18	18
All-Pass Filter Coefficients (M=12)	12	-
τ_{opt}	5	-
Gain	6	5
Total	50 bits	25 bits
Bit Rate	2500 bps	1250 bps

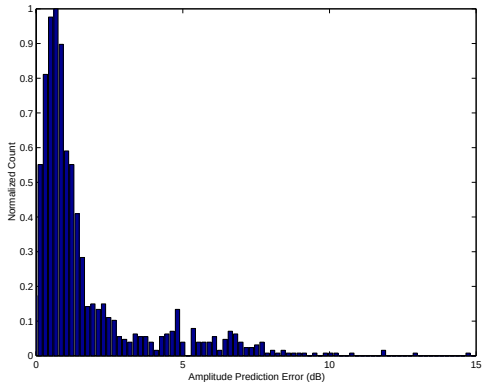

Figure 3. Statistical distribution of ϵ_A^k

database [1]. The following distortion measures were defined and computed for a large number of speech segments. Let α_l^k and $\hat{\alpha}_l^k$ denote the sine wave amplitudes corresponding to the original and the synthesized speech at the kth frame, respectively, the amplitude prediction error (in dB) for the kth frame is defined as follows:

$$\epsilon_A^k = \left[\frac{1}{L_k}\sum_{l=1}^{L_k}\left(20\log\alpha_l^k - 20\log\hat{\alpha}_l^k\right)^2\right]^{\frac{1}{2}} \qquad (15)$$

where L_k is the total number of harmonics in the spectrum of the kth speech segment. The statistical distribution of the amplitude prediction error computed for all classes of speech (i.e., voiced, unvoiced, and transition), with $P = 10$ and $M = 12$, is shown in Fig. 3. The reasonably small amplitude distortions obtained indicate that the modeling of the sinusoidal amplitudes is indeed efficient. The phase prediction error for the kth frame is defined as follows:

$$\epsilon_P^k = \frac{1}{L_k}\sum_{l=1}^{L_k}\left[(\phi_l^k - \hat{\phi}_l^k) \bmod 2\pi\right] \qquad (16)$$

where ϕ_l^k and $\hat{\phi}_l^k$ denote the sinusoidal phases corresponding to the original and the reconstructed modified residual signal, respectively. The statistical distribution of the phase prediction error computed over the entire signal bandwidth including all classes of speech is illustrated in Fig. 4. The small phase prediction errors obtained reaffirm the effectiveness of the proposed model. Note that, unquantized values of the parameters were used in the above expressions. Informal listening tests indicate that speech of good subjective quality, intelligibility, and naturalness can be obtained using the proposed model.

5. CONCLUSION

An improved harmonic sinusoidal coding system was presented, where the underlying sine wave amplitudes and

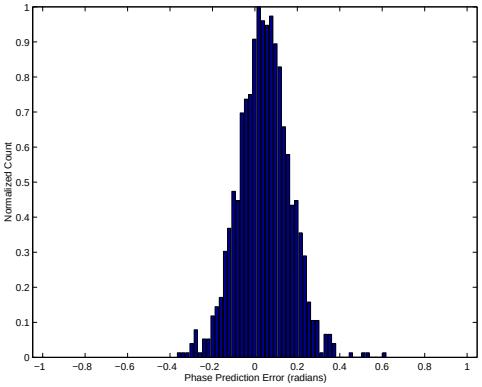

Figure 4. Statistical distribution of ϵ_P^k

phases are efficiently represented using a combination of linear prediction, linear phase alignment, all-pass filtering, and spectral sampling in the residual-domain. The elements of the analysis and synthesis systems were introduced and the extraction and encoding of the model parameters were discussed. A variable-rate sinusoidal coder operating at an average bit rate of 1.75 kbps has been developed based on the proposed model, yielding reproduced speech of good quality, intelligibility, and naturalness.

REFERENCES

[1] S. Ahmadi, "Low bit rate speech coding based on the sinusoidal model", *Ph.D Dissertation*, Arizona State University, Tempe, Arizona, June 1997.

[2] S. Ahmadi, and A. S. Spanias, "A New Phase Model for Sinusoidal Transform Coding of Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 6, No. 5, pp. 495-501, Sept. 1998.

[3] L. B. Almeida, and J. M. Tribolet, "A model for short-time phase prediction of speech", *in Proc. IEEE ICASSP-81*, pp. 213-216, 1981.

[4] B. S. Atal, and N. David, "On synthesizing natural-sounding speech by linear prediction", *in Proc. IEEE ICASSP-79*, pp. 44-47, 1979.

[5] 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.

[6] C. Chen, and A. Gersho, "Adaptive post-filtering for quality enhancement of coded speech", *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 59-71, Jan. 1995.

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

[8] 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.

[9] 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.

[10] 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.

[11] K. K. Paliwal, and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 3-14, Jan. 1993.

[12] I. M. Trancoso, J. S. Rodrigues, and L. B. Almeida, "Quantization issues in harmonic coding" *in Proc. IEEE ICASSP-88*, pp. 382-385, 1988.