# A NEW SINUSOIDAL PHASE MODELING ALGORITHM

*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 and synthesis of speech signals is presented. Short-time sinusoidal phases are efficiently approximated by incorporating linear prediction, spectral sampling, delay compensation, and phase correction techniques. The algorithm is different than phase compensation methods proposed for multi-pulse LPC in that it has been tailored to sinusoidal transform coding of speech signals. Performance analysis on a large speech database indicates considerable improvement in temporal and spectral matching between the original and reconstructed signals as compared to other sinusoidal phase models as well as improved subjective quality of the reproduced speech.

## 1. INTRODUCTION

The sinusoidal model represents speech by a linear combination of sinusoids with time-varying amplitudes, frequencies, and phases [4]-[10]. Although successful techniques have been developed for quantization of the sinusoidal amplitudes and frequencies [5],[9], there is still a demand for more improvements in the sinusoidal phase models. The basic motivation for an efficient phase model is the fact that quantization of phase is usually a major source of degradation in the performance of the sinusoidal coders. Figure 1 illustrates the basic idea of the method presented in this paper. The speech signal at the input is analyzed with a finite duration analysis window. A $P$th order LPC analysis is performed and the estimated complex-valued transfer function of the all-pole filter is sampled at integer multiples of a predetermined spectral sampling frequency, which corresponds to the fundamental frequency during voiced speech segments. The LPC order is constrained due to the fact that the autocorrelation sequence of voiced segments becomes periodic if the maximum lag exceeds the pitch period of the corresponding segment, which results in resolving the fine structure, thereby causing distortion in the spectral envelope [7]. A delay compensation based on the minimization of a time-domain weighted squared error is introduced to compensate for the time-varying delay (jitter) and to achieve maximum alignment between the reference and the input to the all-pass filter. Improvement in temporal and spectral matching is achieved by introducing an all-pass filter. The phase response of the all-pass filter approximates the phase difference between the reference signal and the input to the all-pass filter [1]-[3].

A number of different approaches to sine wave phase estimation have been proposed in the literature. The sinusoidal phase model developed by McAulay and Quatieri [6]-[9] is essentially based on a minimum phase assumption for the vocal tract and it is restricted to voiced speech segments. Another approach taken by Almeida et al. [4],[5] is based on the correlation between harmonic phases of consecutive voiced segments. In contrast to the aforementioned approaches, the proposed algorithm assumes no special condition on the short-time segments of the speech signal and results in improved phase matching for all categories of speech (i.e., voiced, unvoiced, onset, and transition). Furthermore, the method yields improved reproduction of nasalities and vowel sounds. The proposed algorithm is also different than other approaches taken by Hedelin [1], Tribolet et al. [2], and Honda [3] in that spectral sampling and delay compensation are also integrated into phase matching. In addition, those techniques were proposed and utilized in LPC-based algorithms, whereas our algorithm has been tailored to sinusoidal transform coding.

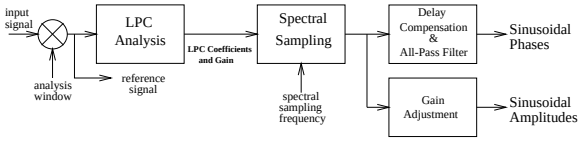


**Figure 1. Simultaneous representation of the sinusoidal amplitudes and phases**

To evaluate the performance of the proposed algorithm, a comprehensive statistical analysis was carried out. Two distortion measures were used to evaluate the performance of the proposed algorithm on 50,000 uncorrelated short-time speech segments taken from TIMIT database. The experimental results indicate superior performance over other sinusoidal phase modeling techniques.

A simplified version of the phase model was also developed, in which the phase parameters were reduced to an optimum delay, a coefficient, and an integer denoting the all-pass filter order. The experimental results show that considerable improvement in temporal and spectral matching can still be achieved using the simplified algorithm. The simplified representation is well suited 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, some experimental results are presented and compared to those obtained from other sinusoidal phase modeling algorithms. Concluding remarks are given in section 4.

## 2. DESCRIPTION OF THE ALGORITHM

In this algorithm, the harmonic sinusoidal model is used to represent finite duration segments of a speech signal [6]-[9]. If $s_w(n)$ represents a windowed segment of the speech

*Portions of this work have been supported by a grant from Intel Corporation

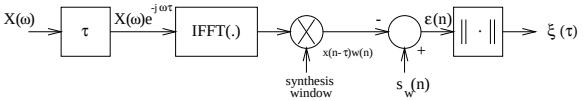


**Figure 2. Delay compensation procedure**

waveform $s(t)$, sampled at $f_s$ samples/sec, 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 $\psi_l$ denote the time-varying amplitudes and phases of the underlying sinusoids, respectively, $\omega_{ss}$ is the frequency at which the corresponding spectrum is sampled, and $L$ is the total number of spectral samples over the $[0, f_s/2]$ region. The spectral sampling frequency corresponds to the fundamental frequency during voiced speech segments. As depicted in Fig. 1, the speech signal at the input is analyzed by a Hanning window of length $N$. A $P$th order LPC analysis, based on the autocorrelation method, is performed. In order to avoid sharp spectral peaks in the LPC spectrum which may result in unnatural synthesized speech, a fixed 10 Hz bandwidth expansion is applied to the poles of the minimum-phase all-pole filter. The transfer function of the all-pole filter, which characterizes the time-varying characteristics of the vocal tract, is given by:

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

where the gain $G$ and the predictor coefficients $\{a_k\}_{k=1}^{P}$ are computed on a short-term basis over a frame of about 20-30 ms long, during which the speech signal can be considered to be approximately stationary. The complex-valued all-pole transfer function, $H(z)$, is sampled at integer multiples of $\omega_{ss}$ as shown in Fig. 3. The time-domain signal corresponding to these samples does not match the original waveform due to the lack of correct short-term phase components. Temporal and spectral matching are achieved by introducing an all-pass filter, $G(z)$, to model a portion of the phase characteristics of the speech signal $s_w(n)$. The input signal to the all-pass filter, $v(n)$, has to be pre-processed to ensure maximum alignment and correlation between $s_w(n)$ and $v(n)$. This is provided by a delay compensation procedure (Fig. 2), that minimizes the following weighted squared error in the time-domain:

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

where $w(n)$ is the synthesis window, usually chosen to be the same as the analysis window. The optimum delay is found as:

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

In practice, $\tau$ is an integer and the interval $[\tau_{min}, \tau_{max}]$ is searched for the value that minimizes the error in (3). Once the delay is computed, the complex-valued signal $\mathbf{X}(\omega)e^{-j\omega\tau_{opt}}$ is inverse Fourier transformed, windowed, and multiplied by a gain, $\gamma$, to obtain the closest signal to $s_w(n)$. The gain is given by:

$$\gamma = \frac{\max |s_w(n)|}{\max |y(n)|} \quad (5)$$

An arbitrary rational function of the form $G(z) = \frac{N(z)}{D(z)}$ is an all-pass filter if $N(z) = \gamma D(z^{-1}) z^{-\tau_{opt}}$. Note that

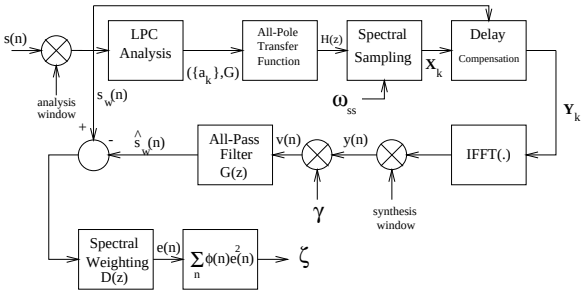


**Figure 3. All-pass filter optimization procedure**

the values of $\gamma$ and $\tau_{opt}$ have already been found. To derive the parameters of the all-pass filter, a weighted least squares (WLS) algorithm is used. If we define $\mathbf{S}_w = (s_w(0), s_w(1), ..., s_w(N-1))^T$ and $\mathbf{V} = (v(0), v(1), ..., v(N-1))^T$, then the weighted squared error at the input of the all-pass filter can be expressed as follows:

$$\rho_1 = (\mathbf{S}_w - \mathbf{V})^T \mathbf{\Phi} (\mathbf{S}_w - \mathbf{V}) \quad (6)$$

where $\mathbf{\Phi} = \mathrm{diag}(\phi(0), \phi(1), ..., \phi(N-1))$ is an appropriate $N \times N$ diagonal weighting matrix in the time-domain. The superscript $T$ will be used throughout the paper to denote 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:

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

where $\mathbf{\Theta} = (\theta_0, \theta_1, ..., \theta_M)^T$ can be found using a 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 $G(e^{j\omega})$ can be interpreted as a model of 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 after minimization

$$\rho_2 = (\mathbf{S}_w - \hat{\mathbf{S}}_w)^T \mathbf{\Phi} (\mathbf{S}_w - \hat{\mathbf{S}}_w) \quad (8)$$

would be less than $\rho_1$. 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, if the weighting function is chosen the same as $D(z)$ [1]. A direct attempt to minimize $\rho_2$ would result in a nonlinear minimization problem. If the spectral weighting function is chosen as $D(z)$, then a linear minimization problem is involved. The instantaneous error vector is obtained as:

$$\mathbf{\Delta} = \mathbf{X}\mathbf{\Theta} \quad (9)$$

where $\mathbf{X} \triangleq [\mathbf{S}_w(n) - \mathbf{V}(n) \mid \mathbf{S}_w(n-1) - \mathbf{V}(n+1) \mid ... \mid \mathbf{S}_w(n-M) - \mathbf{V}(n+M)]$ is an $N \times M+1$ matrix. Therefore, one would like to find the parameters $\mathbf{\Theta}$ and $M$ so as to minimize the following weighted squared error:

$$\min_{\mathbf{\Theta}, M} \left[ \zeta = (\mathbf{X}\mathbf{\Theta})^T \mathbf{\Phi} (\mathbf{X}\mathbf{\Theta}) \right] \quad (10)$$

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:

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

To obtain the coefficients $\tilde{\mathbf{\Theta}} = (\theta_1, \theta_2, ..., \theta_M)^T$, equation (11) is rewritten as:

$$\mathbf{R} = \left( \begin{array}{c|c} r_{00} & \tilde{\mathbf{r}}^T \\ \hline \tilde{\mathbf{r}} & \tilde{\mathbf{R}} \end{array} \right) \quad (12)$$

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

$$\tilde{\mathbf{\Theta}} = -\tilde{\mathbf{R}}^{-1} \tilde{\mathbf{r}} \quad (13)$$

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. Good results are obtained using all-pass filters of orders 12 to 18.

We observe from (9) that values of $s_w(n)$ and $v(n)$ are required outside the interval $0 \leq n \leq N-1$. If we choose not to supply the values outside this interval, and not to extend the signals with zero samples, then we are restricted to a smaller interval for minimization (i.e., $M \leq n \leq N-M-1$). The only drawback with this method is the fact that less data is used to estimate the elements of $\mathbf{R}$. This method is called a *covariance type of approach*. If we choose to extend values of the signals outside the interval $0 \leq n \leq N-1$ with zero samples, then we must resort to using a finite duration weighting sequence, $\phi(n)$, to reduce the end effects. In vector notation, $\phi(n)$ is represented by the diagonal weighting matrix $\mathbf{\Phi}$. To minimize the error, a weighting sequence is applied which smoothly tapers the signals to zero at the ends of the window. This is called an *autocorrelation type of approach*. For practical purposes the covariance type of approach is preferred, because an equal number of samples are used for the computation of the elements of $\mathbf{R}$, and indeed no weighting is required.

If no constraints are imposed, the minimization of (10) can lead to an unstable all-pass filter. The instability of the all-pass filter can be avoided if iterative methods are used. The stability has to be checked within the iteration and the algorithm must be terminated at the last stable stage [1]. Use of FIR all-pass filters might be another approach to guarantee the stability of the all-pass filter.

For efficient transmission and storage, a simplified version of the all-pass filter was developed, in which the transfer function of the all-pass filter was simplified as follows:

$$G(z) = \frac{1+\alpha z^M}{1+\alpha z^{-M}} \quad (14)$$

Therefore, the weighted squared error in (10) is reduced to:

$$\zeta(M) = \sum_{n=0}^{N-1} \phi(n)[s_w(n) + \alpha s_w(n-M) - v(n) - \alpha v(n+M)]^2 \quad (15)$$

where

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

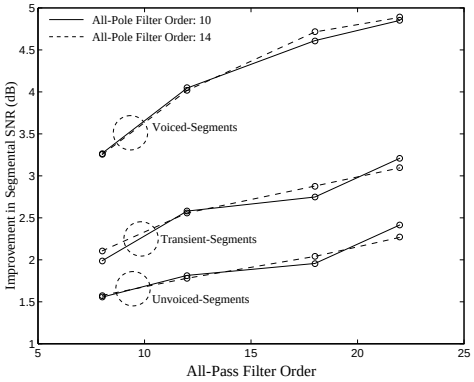


**Figure 4. Improvement in segmental SNR as a function of all-pass filter order**

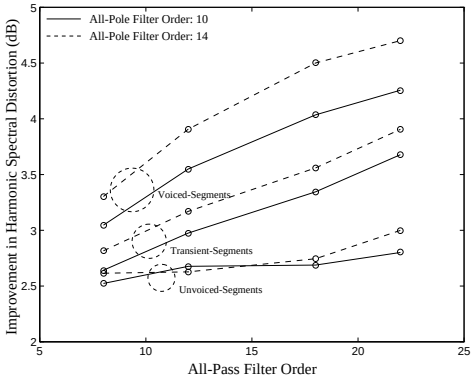


**Figure 5. Improvement in harmonic spectral distortion as a function of all-pass filter order**

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

$$M_{opt} = \arg\min_M [\zeta(M)] \quad (17)$$

$M$ is found through a search process in the interval $[1, M_{max}]$.

## 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 uncorrelated segments of real speech data taken from TIMIT database. The first measure is the average improvement in segmental signal to noise ratio (SNR), which is defined as follows:

$$SNRI = \frac{1}{K} \sum_{k=1}^{K} 10 \log \left( \frac{\rho_{1_k}}{\rho_{2_k}} \right) \quad (18)$$

where $\rho_{1_k}$ and $\rho_{2_k}$ are the weighted squared errors before and after all-pass filtering for the $k$th frame respectively, and $K$ denotes the number of frames used in the experiment. Figure 4 shows the average improvement in SNR 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 (HSD), which is defined as follows:

$$HSDI = \frac{1}{K} \sum_{k=1}^{K} 10 \log \left( \frac{\sum_{l=1}^{L_k} |S_w(l\omega_{ss}) - V(l\omega_{ss})|^2}{\sum_{l=1}^{L_k} |S_w(l\omega_{ss}) - \hat{S}_w(l\omega_{ss})|^2} \right) \quad (19)$$

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 $L_k$

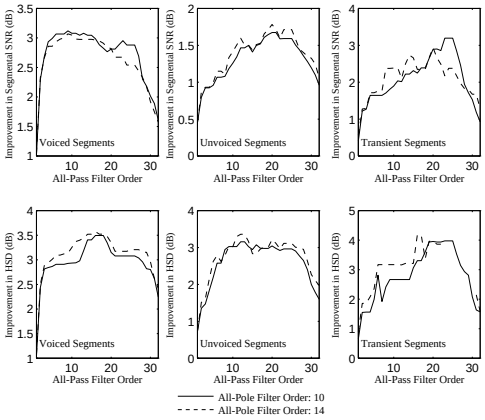


**Figure 6. Average improvements in SNR and HSD obtained from the simplified model**

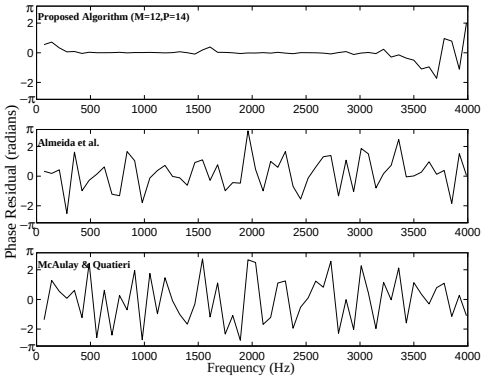


**Figure 7. Phase residuals as a function of frequency obtained from different methods for a typical speech segment**

denotes the number of spectral samples in the spectrum of the $k$th segment. The average improvement in HSD as a function of the all-pass filter order with the all-pole filter order as a parameter is illustrated in Fig. 5. It can be concluded that sophisticated design of the all-pass filter (i.e., the way in which the phase difference is approximated) results in perfect reproduction of the original signal both in the time and frequency domains. The performance of the simplified model (14) also has been evaluated and shown in Fig. 6. It can be seen that considerable improvement in temporal and spectral matching is still achieved using a small number of parameters to represent the sinusoidal phases.

The performance of the proposed algorithm has been compared to other sinusoidal phase modeling techniques. The first method developed by McAulay and Quatieri [6],[8] uses a minimum-phase assumption for the vocal tract and approximates the $k$th short-time phase in terms of an on-set parameter associated with the vocal excitation, a system phase that is derived from the minimum-phase vocal tract system function, and a voicing dependent part that depends on the probability that the frame is voiced. The second method developed by Almeida et al. [4],[5] predicts the phase of the current voiced frame in terms of the phase of the previous frame and a frequency dependent phase increment. The phase residual vector (i.e., the difference between the original and reconstructed phase functions) obtained from different methods for a typical speech segment is shown in Fig. 7. It is clear that very good phase approximation has been achieved over a wide range of frequencies using the proposed method. Finally, typical reproduction of various signals is illustrated in Fig. 8.

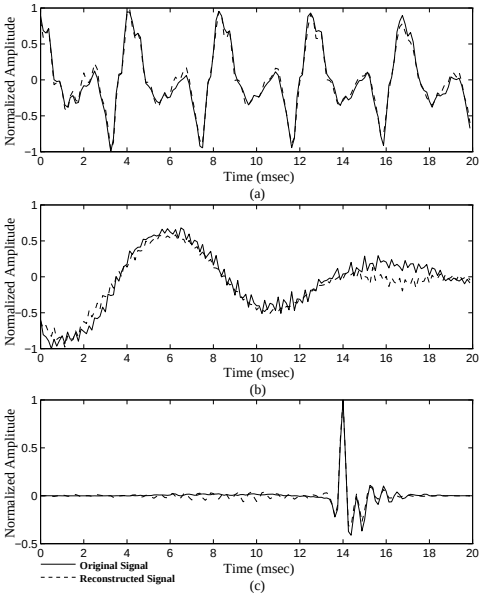


**Figure 8. Typical reproduction of signals of different types, (a) a quasi-periodic signal (b) a non-periodic signal (c) a burst signal**

## 4. CONCLUSION

A new sinusoidal phase model was presented. Efficient simultaneous representation of the sinusoidal amplitudes and phases is obtained by cascading an all-pass filter to the LPC filter, where the all-pass filter is used for phase correction. Performance study on a large database indicates considerable improvement in matching between the original and reconstructed signals both in the time and frequency domains. It was also shown that the proposed method compares favorably against other sinusoidal phase modeling techniques. A simplified version of the phase model, that is well suited for low-bit rate speech coding applications, was also presented.

## REFERENCES


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

[2] 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.

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

[4] 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.

[5] 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.

[6] R. J. McAulay, and T. F. Quatieri, "Sine wave phase coding at low data rates" *in Proc. IEEE ICASSP-91*, pp. 577-580, 1991.

[7] T. G. Champion, R. J. McAulay, and T. F. Quatieri, "High-order all-pole modeling of the spectral envelope" *in Proc. IEEE ICASSP-94*, pp. I.529-I.532, 1994.

[8] 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.

[9] 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.

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