PERCEPTUAL CONSIDERATIONS IN A LOW BIT RATE SINUSOIDAL Vocoder

E. Bryan George and Mark J. T. Smith
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT
The concept of error spectrum shaping has recently been applied to a number of low bit rate speech coding techniques based on linear prediction models, dramatically improving the subjective quality of the coded speech. At the same time, models which represent speech as a sum of amplitude- and frequency-modulated sinusoids have been reported whose parameters may be coded at low bit rates. This paper presents an approach to coding the parameters of a harmonic sinusoidal model which incorporates error spectrum shaping in order to improve the subjective quality of the speech coded at low bit rates.

1 INTRODUCTION
In recent years there has been a great deal of interest in techniques for digital transmission of speech at medium to low bit rates. A number of approaches to low bit rate speech coding based on linear prediction models have been investigated and reported. Several of these, such as Multi-pulse-Excited-, Code-Excited- and Self-Excited-LPC have been shown to yield performance superior to conventional pitch-excited LPC at bit rates in the range of 4800 bits per second [1,2,3]. Much of the performance gain achieved by these approaches may be attributed to the technique of shaping the error spectrum to exploit the noise masking property of aural perception, which has the effect of improving the perceptual or subjective quality of the coded speech.

As an alternative to linear prediction based speech modeling, representations of speech based on sums of sinusoidal signals have been developed and reported. These representations are particularly well-suited to speech modeling since they are inherently time-varying and quasi-periodic in nature. McAulay and Quatieri have introduced a sinusoidal model formulation which sums sinusoids with continuous linear amplitude and cubic phase parameters [4]. They have demonstrated that the parameters of this model may be efficiently coded at a bit rate of 4800 bits per second [5]. This paper presents a sinusoidal model formulation which differs slightly from the formulation of McAulay and Quatieri. We then present an approach to coding the parameters of this model at a bit rate of 4800 bits per second which incorporates the concept of error spectrum shaping in an attempt to improve the resultant subjective quality of the coded speech.

2 THE MODEL
In its most general form a sinusoidal model is described by the equation
\[ s(n) = \sum_{k=1}^{K} A_k(n) \cos(\Phi_k(n)), \]
where \( s(n) \) is the original speech signal. Unfortunately, there is no simple way to determine appropriate values of \( A_k(n) \) and \( \Phi_k(n) \). The approach to finding parameter values proposed by McAulay and Quatieri is to choose amplitude and phase values at specific time intervals by peak-picking the Discrete Fourier Transform (DFT) of the Hamming-windowed signal and then interpolating these amplitude and phase values to obtain a linear \( A_k(n) \) and a cubic \( \Phi_k(n) \). The authors have investigated an alternate approach to the problem whereby the technique of analysis-by-synthesis is applied to a special form of the general sinusoidal model to determine appropriate parameter values [6].

The version of the sinusoidal model we use has the form
\[ \tilde{s}(n) = \sigma(n) \sum_{i=-\infty}^{\infty} w(n-iN) \tilde{a}(n-iN), \]
where
\[ w(n) = \begin{cases} \cos^2(n\pi/2N), & -N \leq n \leq N \\ 0, & \text{otherwise} \end{cases} \]
and
\[ \tilde{a}(n) = \sum_{k=1}^{K} A'_k(\cos(\omega_k n + \phi'_k)). \]
N determines the frame length, \( \sigma(n) \) is a time-varying gain signal or "envelope", and \( \{ A_i \}, \{ \omega_k \} \) and \( \{ \phi_k \} \) are the sets of amplitudes, frequencies and phases, respectively, associated with frame \( i \). Note that for each frame the signal \( \tilde{a}_i(n) \) is produced by a particularly simple form of the general sinusoidal model and that each of these signals contributes to the overall approximation \( \tilde{s}(n) \) via an overlap-add technique.

2.1 Analysis-by-Synthesis Technique

The parameter set which must be determined in order to approximate a given \( s(n) \) consists of the time-varying gain \( u(n) \) and the amplitudes, frequencies and phases of the sinusoidal components for each frame. The value of \( \sigma(n) \), which is a very slowly varying signal, can be reasonably estimated by filtering the full-wave rectified signal \( |s(n)| \) using a recursive quasi-Gaussian lowpass filter. Given \( u(n) \), we would like to find amplitudes, frequencies and phases for each frame such that \( \tilde{s}(n) \) is closest to \( s(n) \) in some sense. One possibility would be to minimize the mean-square error

\[
E = \sum_{n}(s(n) - \tilde{s}(n))^2
\]

in terms of the parameters \( \{ A_i \}, \{ \omega_k \} \) and \( \{ \phi_k \} \) given in Equation 4. However, it is not generally feasible to solve for these parameters simultaneously due to the complexity involved. A way around this difficulty is to solve for the amplitude, frequency and phase parameters one frame at a time. For each frame \( i \), then, the goal is to minimize

\[
E_i = \sum_{n=-N}^{N}(s(n) - \tilde{s}(n))^2
\]

Weighting the error by \( w(n) \) in Equation 6 accounts for the fact that it is more important for \( \tilde{a}_i(n) \) to be accurate at the frame center, where its contribution to \( \tilde{s}(n) \) is dominant. If we now define

\[
x(n) \triangleq w(n)s(n + iN)
\]

\[
g(n) \triangleq w(n)\sigma(n + iN)
\]

and use the definition of \( \tilde{a}_i(n) \) given in Equation 4, the equation for \( E_i \) simplifies to

\[
E = \sum_{n=-N}^{N}(x(n) - g(n)\sum_{k=1}^{K}A_k \cos(\omega_n + \phi_k))^2,
\]

where the subscript and superscripts \( i \) have been omitted for clarity. Attempting to minimize \( E \) simultaneously in terms of the amplitude, frequency and phase parameters leads to a complex nonlinear optimization problem which is difficult to solve. As an alternative, a slightly suboptimal but computationally simple analysis-by-synthesis algorithm is used to determine the amplitude, frequency and phase of each sinusoidal component separately.

This algorithm works as follows: Suppose we have determined the parameters of \( k - 1 \) sinusoidal components,
The values for $a_k$ and $b_k$ are given by

$$a_k = \frac{(\gamma_{12} \psi_1 - \gamma_{13} \psi_2)}{\Delta}$$
$$b_k = \frac{(\gamma_{11} \psi_2 - \gamma_{13} \psi_1)}{\Delta},$$

where

$$\Delta = \gamma_{11}^2 \gamma_{12}^2 - \gamma_{13}^2.$$  

By the Principle of Orthogonality, given $a_k$ and $b_k$ we can calculate $E_k$ by

$$E_k = \sum_{n=-N}^{N} \epsilon_{k-1}^2(n) - a_k \psi_1 - b_k \psi_2$$
$$= \epsilon_{k-1}^2 - a_k \psi_1 - b_k \psi_2.$$  

Having determined $a_k$ and $b_k$, $A_k$ and $\phi_k$ are given by the relations

$$A_k = \sqrt{(a_k^2 + b_k^2)}$$
$$\phi_k = -\tan^{-1}(b_k/a_k).$$

This establishes a method for determining the optimal amplitude and phase parameters for a single sinusoidal component at a given frequency. To determine an appropriate frequency for this component we can perform an ensemble search procedure in which $\omega_k$ is varied over a set of candidate values given by $\Omega_j = j\pi/512$ for $0 \leq j \leq 512$. For each $\Omega_j$ we calculate the corresponding value of $E_k$, choose $\omega_k$ as that value of $\Omega_j$ which yields the minimum error and choose $A_k$ and $\phi_k$ as the amplitude and phase parameters associated with that frequency.

Having determined the parameters of the $k$-th component, the approximation and error sequences are updated as

$$\epsilon_k(n) = \epsilon_{k-1}(n) + g(n)A_k \cos(\omega_k n + \phi_k)$$
$$\epsilon_k(n) = \epsilon_{k-1}(n) + g(n)A_k \cos(\omega_k n + \phi_k).$$

At this point the procedure may be repeated for the next component. From this we see that, starting with the initial conditions

$$\epsilon_0(n) = 0$$
$$\epsilon_0(n) = \pi(n),$$

it is possible to determine $A_k$, $\omega_k$ and $\phi_k$ for $k = 1, \ldots, K$ in a recursive fashion using the above equations. Figure 1 shows a block diagram of this closed-loop analysis-by-synthesis procedure.

### 2.2 Fundamental Frequency Tracking

The previous section described a method of determining the amplitude, frequency and phase parameters necessary to accurately model an arbitrary sequence using an overlap-add formulation of a sinusoidal model. If accurate modeling were the overall goal, what has been described would be sufficient. However, since our goal is to code speech at low bit rates, further information is needed. Quasi-periodic signals are by definition associated with a fundamental frequency or "pitch". For the purposes of coding, knowledge of this information is useful since it provides a bound on the number of sinusoidal components necessary to model the given signal and reduces the amount of information necessary to code the frequency parameters. We now describe a robust and computationally efficient algorithm for estimating and tracking the fundamental frequency of speech signals.

Before discussing algorithms for estimating and tracking pitch, some theoretical basis for the notion of fundamental frequency is necessary. Recall from Equation 4 that for frame $i$

$$\tilde{\epsilon}_i(n) = \sum_{k=1}^{K} A_k \cos(\omega_k n + \phi_k), \quad -N \leq n \leq N.$$  

Suppose that

$$\omega_k = k\omega_0 + \Delta_k \approx k\omega_0$$

\[\text{(41)}\]
The sinusoidal model just described has a special form which is referred to as "quasi-harmonic". This means that associated with \( \omega_k \) there is a frequency \( \omega'_k \) given by

\[
\omega'_k = 2\pi f_s / F_s,
\]

where \( f_s \) is the fundamental frequency in Hz and \( F_s \) is the sampling frequency in samples per second, and that there is one and only one sinusoidal component associated with each harmonic. For speech signals we can reasonably expect that \( f_s > 40 \) Hz.

We now offer the following definition of fundamental frequency: The fundamental frequency associated with a quasi-harmonic signal is that value of \( \omega_o \) which minimizes

\[
E(\omega_o) = \sum_{n=-N}^{N} \left[ \sum_{k=1}^{K} A_k \cos(\omega_k n + \phi_k) - \cos(\omega_o n + \phi_o) \right]^2
\]

where again the superscripts \( i \) have been omitted. Since the term \( \omega_o \) is embedded in the cosine functions in this expression, it is not possible to derive a closed-form expression for \( \omega_o \) based on this definition. However, it is possible to derive an expression for \( \omega_o \) based on an approximation to \( E(\omega_o) \).

We define

\[
F(\omega) = \sum_{n=-N}^{N} \cos \omega n = \frac{\sin \omega (N + 1/2)}{\sin \omega / 2} = F(-\omega)
\]

and note the relation

\[
\sum_{n=-N}^{N} \cos(\omega_1 n + \phi_1) \cos(\omega_2 n + \phi_2) = \frac{1}{2} \cos(\phi_1 + \phi_2) F(\omega_1 + \omega_2) + \frac{1}{2} \cos(\phi_1 - \phi_2) F(\omega_1 - \omega_2).
\]

From Equation 48, it is reasonable to assume that \( F(\omega) \approx 0 \) when \( \omega > 2\pi/(2N + 1) \). Substituting this into Equation 49 yields

\[
\sum_{n=-N}^{N} \cos(\omega_1 n + \phi_1) \cos(\omega_2 n + \phi_2) \approx \frac{1}{2} \cos(\phi_1 - \phi_2) F(\omega_1 - \omega_2)
\]

when \( \omega_1 + \omega_2 > 2\pi/(2N + 1) \). For a frame length of 20 msec this approximation holds when the corresponding Hertzian frequencies obey the relation \( f_1 + f_2 > 25 \) Hz.

From Equation 47, recalling the assumptions that \( \omega_k \approx k \omega_o \) for all \( k \) and that \( f_s > 40 \) Hz, the approximation holds in Equation 47 for all \( k \) and \( \ell \). Substituting Equation 50 into Equation 47 yields

\[
E(\omega_o) \approx \frac{1}{2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} A_k A_\ell \cos(\phi_k - \phi_\ell) \left( \frac{F(\omega_k - \omega_o) - F(\omega_k - \omega_o)}{F(\omega_k - \omega_o)} \right)
\]

Further simplification of this expression is possible, since in this equation the arguments of the functions \( F \) are differential frequencies, the absolute value of which are approximately \( k - \ell \omega_o \). Given this and the assumption that \( f_s > 40 \) Hz, the summand terms in Equation 51 are approximately zero when \( k \neq \ell \), resulting in the simplified expression

\[
E(\omega_o) \approx \sum_{k=1}^{K} A_k^2 (2N + 1 - F(\omega_k - \Delta k))
\]

This expression for \( E(\omega_o) \) still cannot be solved in a closed-form fashion, but since \( \omega_k \approx k \omega_o \), \( \Delta k \approx 0 \) hence the second-order Taylor approximation for \( F(\omega) \), given by

\[
F(\omega) \approx \sum_{n=-N}^{N} \left[ 1 - (\omega n)^2/2 \right]
\]

\[
= 2N + 1 - \omega^2 \sum_{n=1}^{N} n^2
\]

\[
= 2N + 1 - \omega^2
\]

may be substituted into Equation 53, yielding

\[
E(\omega_o) \approx E'(\omega_o) = \epsilon \sum_{k=1}^{K} A_k^2 \Delta k
\]

\[
= \epsilon \sum_{k=1}^{K} A_k^2 (\omega_k - k \omega_o)^2.
\]
\( E'(\omega) \) is easily minimized in terms of \( \omega \), resulting in

\[
\omega = \frac{\sum_{k=1}^{K} A_k^2 \omega_k}{\sum_{k=1}^{K} A_k^2}. \quad (59)
\]

Note that this expression depends only on the amplitudes and frequencies of the component sinusoids which make up \( \hat{z}(n) \) and may be thought of as the average of \( \omega_k/k \) weighted by \( (kA_k)^2 \).

Given a rough estimate of the fundamental frequency, the parameters of the sinusoidal model for a given frame of speech can be arranged in a quasi-harmonic form and a refined estimate of \( \omega \) calculated by Equation 59. As there is no a priori knowledge of the fundamental frequency of a given frame of speech, this rough estimate must be determined from the speech frame itself. We now discuss an algorithm for obtaining this estimate.

Recalling the expression for \( E'(\omega) \) given in Equation 58,

\[
E'(\omega) = e \sum_{k=1}^{K} A_k^2 (\omega_k - \omega) / \sum_{k=1}^{K} A_k^2, \quad (60)
\]

we now define a related measure by

\[
\sigma^2 = \frac{E'(\omega)}{\sum_{k=1}^{K} A_k^2} = \frac{\sum_{k=1}^{K} A_k^2 (\omega_k - \omega)^2}{\sum_{k=1}^{K} A_k^2}. \quad (61)
\]

This measure gives the weighted mean-square frequency error of the set \( \{\omega_k\} \) about the set \( \{k\omega_k\} \).

A qualitative result of the analysis-by-synthesis algorithm described in Section 2.1 is that the first component found using the algorithm typically has the highest energy, meaning that this first component is in fact a signal component and not associated with background noise or model inaccuracy. This, combined with the observation that the weighting terms in Equation 58, \( A_k^2 \), correspond approximately to the energies of the components, leads to the following assumption: The frequency of the first component found is approximately equal to some multiple of the fundamental frequency.

With this in mind, given the frequency of the first component we can postulate an initial set of candidate fundamental frequency estimates given by \( \omega_j/j \) for \( j_{\text{min}} \leq j \leq j_{\text{max}} \), where \( j_{\text{min}} \) and \( j_{\text{max}} \) are chosen such that the initial estimates are within a desired range, typically 40 to 400 Hz for speech. Given these initial estimates, for each new component found using the analysis-by-synthesis algorithm we can perform the fundamental frequency refinement algorithm described above in parallel for each of the candidate frequencies. At the same time \( \sigma \omega \), the weighted RMS frequency error, is calculated for each of the candidate frequencies. We can then determine which candidate frequencies are invalid by disqualifying those whose associated value of \( \sigma \omega \) exceeds a threshold value, which by empirical observation is set to 10 Hz. Experiments indicate that after only five components are found the highest frequency of the remaining valid candidates is approximately the expected fundamental frequency.

If it were certain that the first five components are signal components, this algorithm would be sufficient. However, owing to background noise and model inaccuracies it is possible that at least one of the components found will be "spurious". These components typically are relatively small in amplitude and are close in frequency to dominant signal components.

We can exploit the property of noise masking to aid in detecting these spurious components. One way to describe noise masking is that the amplitude needed to hear a noise component added to a signal is higher when its frequency is near that of a signal component. Using this as a guide, an audibility threshold function can be defined which we use as a criterion for classifying a component as either signal or spurious.

Assessing the masking threshold of a signal like speech, which has multiple components, is an empirical process about which it is difficult to make general statements. However, given \( k-1 \) components we may easily construct a masking threshold function \( T_{k-1}(\omega) \) which preserves the qualitative aspect of noise masking by these components as

\[
T_{k-1}(\omega) = \alpha \sum_{\omega_1}^{\omega_{k-1}} A_i e^{-\beta(\omega_1-\omega)^2}, \quad (62)
\]

which is a weighted sum of Gaussian functions with equal variances. When \( \alpha \) is increased, \( T_{k-1}(\omega) \) tends toward greater masking. A value of \( \alpha = .5 \) is the minimum value at which significant signal components are not masked, hence we use this value. The "masking bandwidth" of a single signal component is determined by \( \beta \), and by the aforementioned experiment it is set to correspond to a half-power bandwidth of 100 Hz.

Having defined this masking threshold function, given the \( k \)th component, if its amplitude is such that \( A_k^2 > T_{k-1}(\omega_k) \), then the component is classified as signal, used in the fundamental frequency estimation, and the threshold function is updated by

\[
T_k(\omega) = T_{k-1}(\omega) + \alpha A_k^2 e^{-\beta(\omega_1-\omega)^2}. \quad (63)
\]

Otherwise, the component is classified as spurious and disregarded.

3 CODING

The previous sections described an overlap-add, quasi-harmonic sinusoidal model for speech signals and detailed algorithms for determining the model parameters and fundamental frequency of a frame of speech. For the purposes of low bit rate speech coding it is necessary to quantize the time-varying gain \( \sigma(n) \) and for each frame \( i \) the fundamental frequency \( \omega_i \) and the model parameters \( \{A_i\} \), \( \{\omega_i\} \) and \( \{\phi_i\} \). This section discusses methods for accomplishing this which incorporate error spectrum weighting.
3.1 Frequency Parameters: \( \{ \omega_k^f \} \)

It is possible to efficiently code the frequency parameters \( \{ \omega_k^f \} \) by exploiting the quasi-harmonic structure of speech. Recall from Equation 41 in Section 2.2 that \( \omega_k^f \approx k \omega_0^f \). Given this it is possible to significantly reduce the number of bits required to code the frequency parameters by coding the differential frequencies \( \{ \Delta \omega_k^f \} \) rather than the frequencies themselves. In fact, experiments have indicated that very little subjective distortion results from quantizing the component frequencies to the harmonic frequencies with which they are associated, resulting in a harmonic sinusoidal model in which the only frequency information we need to code is the fundamental frequency. Referring to Equation 40 this quantization changes the expression for \( \hat{s}_i(n) \) to

\[
\hat{s}_i(n) = \sum_{k=1}^{K} A_i \cos( k \omega_0^f n + \phi_i )
\]  

(64)

3.2 Amplitude Parameters: \( \{ A_i \} \)

Denoting by \( \hat{s}_i(n) \) the approximation to \( s_i(n) \) which results when the amplitude and phase parameters of Equation 64 are quantized, one approach to effectively coding the sinusoidal model parameters is to define a norm \( \| \cdot \|_w \) then to code the parameters in such a way that \( \| \hat{s}_i(n) - s_i(n) \|_w \) is minimized. Atl and Schroeder [7] have proposed an error norm which is given in the frequency-domain by

\[
\| \hat{s}(e^{j\omega}) - \hat{s}(e^{j\omega}) \|_w^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} | W(e^{j\omega}) | | \hat{s}(e^{j\omega}) - \hat{s}(e^{j\omega}) |^2 d\omega.
\]

(65)

where \( W(e^{j\omega}) \) is a perceptual weighting function which is given in terms of the LPC spectrum \( H(e^{j\omega}) \) by

\[
W(e^{j\omega}) = \frac{H(e^{j\omega})/\gamma}{H(e^{j\omega})}
\]

(66)

where \( \gamma \) typically has a value of 0.8 at a sampling frequency of 8000 samples/sec. As shown in Figure 2, \( W(\omega) \) has dips in the formant regions of the spectrum and peaks in the inter-formant regions. As stated before, this weighting exploits noise masking by concentrating error energy in the vicinity of the formants, where signal energy is highest. By use of Parseval's relation this norm may be expressed in the time domain as

\[
\| \hat{s}_i(n) - s_i(n) \|_w^2 = \| s_i(n) \|_w^2
\]

(67)

\[
= \sum_{n=0}^{\infty} | s_i(n) |
\]

(68)

Since this is an infinite summation it cannot be evaluated. However, since both \( \hat{s}_i(n) \) and \( s_i(n) \) are periodic sequences with fundamental frequency \( \omega_0^f \), so is \( e_i(n) \). Thus

\[
P(e_i) = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} e_i(n) = \frac{1}{2} \sum_{k=1}^{K} | \beta_k |^2
\]

(72)

The expression for \( P(e_i) \) is simple to evaluate and provides a meaningful analog to the norm \( \| \cdot \|_w \) given above.

By representing the amplitude and phase parameters as vectors, it is possible to code these parameters using vector quantization, whereby the vectors are represented as members of fixed codebooks. The appropriate codebook entries are chosen using an ensemble search procedure to determine which entries result in minimum \( P(e_i) \). However, performing this search for both the amplitude and phase vectors together is very computationally intensive. For this reason we vector-quantize the amplitude and phase parameters separately, minimizing \( P(e_i) \) for each vector.

Figure 2. An example of the LPC spectrum \( H(\omega) \) of a frame of speech and its corresponding perceptual weighting function \( W(\omega) \) for a sampling rate of 8000 samples/sec.
It is easy to demonstrate that if
\[ \hat{s}_k(n) = \Re \{ \sum_{k=1}^{K} a_k e^{j\theta_k} e^{j\omega_k n} \}, \] (73)
where we have omitted superscripts \( i \), then
\[ w(n) \ast \hat{s}_k(n) = \Re \{ \sum_{k=1}^{K} W(e^{j\omega_k}) a_k e^{j\theta_k} e^{j\omega_k n} \}. \] (74)
If we denote an amplitude codebook entry as \( \{ A_k \} \) for \( 1 \leq k \leq K \) and the approximation to \( \hat{s}_k(n) \) with only amplitudes quantized as \( \hat{a}(\cdot) \), then clearly
\[ e_k(n) = w(n) \ast (\hat{s}_k(n) - \hat{a}_k(n)) \] (75)
In this case
\[ b_k = W(e^{j\omega_k})(A_k - \hat{A}_k) e^{j\theta_k}. \] (76)
Substituting this value into Equation 72 yields
\[ P(e_k^2) = \frac{1}{2} \sum_{k=1}^{K} |b_k|^2 = \frac{1}{2} \sum_{k=1}^{K} [W(e^{j\omega_k})]^2 (A_k - \hat{A}_k)^2. \] (77)
By using \( P(e_k^2) \) as the distance measure in vector quantizing the amplitude parameters as opposed to the unweighted mean-square error norm, the resulting signal errors are less perceptually significant, hence the subjective quality of the coded speech improves.

3.3 Phase Parameters: \( \{ \phi_k \} \)

In a similar fashion, denoting a phase codebook entry as \( \{ \hat{\phi}_k \} \), then
\[ e_n(n) = w(n) \ast (\hat{\phi}_k(n) - \hat{\phi}_k(n)) \] (79)
In this case
\[ \phi_k = W(e^{j\omega_k}) \hat{A}_k (e^{j\theta_k} - e^{j\hat{\theta}_k}). \] (80)
Substituting this value into Equation 72 yields
\[ P(e_n^2) = \frac{1}{2} \sum_{k=1}^{K} (|W(e^{j\omega_k})| \hat{A}_k)^2 (e^{j\theta_k} - e^{j\hat{\theta}_k})^2 \] (81)
\[ = \frac{1}{2} \sum_{k=1}^{K} (|W(e^{j\omega_k})| \hat{A}_k)^2 (e^{j\theta_k} - e^{j\hat{\theta}_k}) \]
\[ \times (e^{-j\theta_k} - e^{-j\hat{\theta}_k}) \]
\[ = \sum_{k=1}^{K} (|W(e^{j\omega_k})| \hat{A}_k)^2 (1 - \cos(\phi_k - \hat{\phi}_k)). \] (82)
Using \( P(e_n^2) \) as the distance measure in quantizing the phase vector has the same effect on subjective quality as does the perceptually weighted measure used in amplitude vector quantization.

3.4 Fundamental Frequency: \( \omega_k \)

In order to quantize the fundamental frequency \( \omega_k \) we take advantage of the fact that subjectively a similar error in the fundamental frequency made at a low pitch is more easily detected than one made at a high pitch. In other words, human perception is uniformly sensitive to frequency errors in the musical scale, which is logarithmic. For this reason, the logarithm of \( \omega_k \), rather than \( \omega_k \) itself, is uniformly quantised.

4 CONCLUSION

The sinusoidal model formulation presented in this paper is very well-suited to representing speech signals and provides knowledge of pertinent time-varying characteristics of speech such as pitch and short-time spectral information which is useful for speech coding. The model also has a particularly simple form which lends itself easily to analysis and includes an envelope signal which separately models syllabic volume changes, enhancing the performance of the model.

The analysis procedure presented provides greater accuracy than other techniques, resulting in higher quality synthetic and coded speech. In addition, as more components are added the synthetic speech signal is guaranteed to converge to the original speech signal. Using perceptual factors in coding the parameters of this model yields considerable improvements in the overall subjective performance of the coder. Results will be presented at the conference.

REFERENCES
