Multiband Excitation Vocoder

DANIEL W. GRIFFIN AND JAE S. LIM, FELLOW, IEEE

Abstract—**In this paper, we present a new speech model which we refer to as the Multiband Excitation Model. In this model, the short-time spectrum of speech is modeled as the product of an excitation spectrum and a spectral envelope. The spectral envelope is some smoothed version of the speech spectrum and the excitation spectrum is represented by a fundamental frequency, a voiced/unvoiced (V/UV) decision for each harmonic of the fundamental, and the phase of each harmonic declared voiced. In speech analysis, the model parameters are estimated by explicit comparison between the original speech spectrum and the synthetic speech spectrum. In speech synthesis, we synthesize the voiced portion of speech in the time domain and the unvoiced portion of speech in the frequency domain. To illustrate one potential application of this new model, we develop an 8 kbit/s Multiband Excitation Vocoder. Informal listening clearly indicates that this vocoder provides high quality speech reproduction for both clean and noisy speech without the "buzziness" and severe degradation in noise typically associated with vocoder speech. Diagnostic Rhyme Tests (DRT's) were performed as a measure of the intelligibility of this 8 kbit/s vocoder. For clean speech with an average DRT score of 97.8 when uncoded, the coded speech has an average DRT score of 96.2. For speech with wide-band random noise with an average DRT score of 63.1 when uncoded, the coded speech has an average DRT score of 58.0. When the V/UV decision for each harmonic of the fundamental is replaced by one V/UV decision for each frame with all other parameters identical to the 8 kbit/s Multiband Excitation Vocoder, the DRT scores obtained are 96.0 for clean speech and 46.0 for the noisy speech case.**

I. INTRODUCTION

THE problem of analyzing and synthesizing speech has a large number of applications, and as a result has received considerable attention in the literature. One class of speech analysis/synthesis systems (vocoders) which have been extensively studied and used in practice are based on an underlying model of speech. For this class of vocoders, speech is analyzed by first segmenting speech using a window such as a Hamming window. Then, for each segment of speech, the excitation parameters and system parameters are determined. The excitation parameters consist of the voiced/unvoiced decision and the pitch period. The system parameters consist of the spectral envelope or the impulse response of the system. In order to synthesize speech, the excitation parameters are used to synthesize an excitation signal consisting of a periodic impulse train in voiced regions or random noise in unvoiced regions. This excitation signal is then filtered using the estimated system parameters.

Even though vocoders based on this class of underlying speech models have been quite successful in synthesizing intelligible speech, they have not been successful in synthesizing high quality speech. The poor quality of the synthesized speech is, in part, due to fundamental limitations in the speech models and, in part, due to inaccurate estimation of the speech model parameters. As a consequence, vocoders have not been widely used in applications such as time-scale modification of speech, speech enhancement, or high quality bandwidth compression.

One of the major degradations present in vocoders employing a simple voiced/unvoiced model is a "buzzy" quality especially noticeable in regions of speech which contain mixed voicing or in voiced regions of noisy speech. Observations of the short-time spectra indicate that these speech regions tend to have regions of the spectrum dominated by harmonics of the fundamental frequency and other regions dominated by noise-like energy. Since speech synthesized entirely with a periodic source exhibits a "buzzy" quality, and speech synthesized entirely with a noise source exhibits a "hoarse" quality, it is postulated that the perceived "buzziness" of vocoder speech is due to replacing noise-like energy in the original spectrum with periodic "buzzy" energy in the synthetic spectrum. This occurs since the simple voiced/unvoiced excitation model produces excitation spectra consisting entirely of harmonics of the fundamental (voiced) or noise-like energy (unvoiced). Since this problem is a major cause of quality degradation in vocoders, any attempt to significantly improve vocoder quality must account for these effects.

The degradation in quality of vocoded noisy speech is accompanied by a decrease in intelligibility scores. For example, Gold and Tierney [7] report a DRT score of 71.4 for the Belgard 2400 kbit/s vocoder in F15 noise down 18.7 points from a score of 90.1 for the uncoded (5 kHz bandwidth, 12 bit PCM) noisy speech. In clean speech, a score of 86.5 was reported for the Belgard vocoder, down only 10.3 points from a score of 96.8 for the uncoded speech. They call the additional loss of 8.4 points in this noise condition the "aggravation factor" for vocoders. One potential cause of this "aggravation factor" is that vocoders which employ a single voiced/unvoiced decision for the entire frequency band eliminate potentially important acoustic cues for distinguishing between

Manuscript received April 7, 1987; revised February 18, 1988. This work was supported in part by Rome Air Development Center under Contract F19628-85-K-0028 and in part by the Advanced Research Projects Agency monitored by the ONR under Contract N00014-81-K-0742.

D. W. Griffin was with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with Sanders Associates, Nashua, NH 03061.

J. S. Lim is with the Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

IEEE Log Number 8821851.

frequency regions dominated by periodic energy due to voiced speech and those dominated by aperiodic energy due to random noise.

A number of mixed excitation models have been proposed as potential solutions to the problem of "buzziness" in vocoders. In these models, periodic and noise-like excitations are mixed which have either time-invariant or time-varying spectral shapes. In models with time-invariant spectral shapes, a mixture ratio controls the relative amplitudes of a periodic source and a noise source with fixed spectral envelopes [13], [14]. In models with time-varying spectral shapes, voiced/unvoiced decisions or ratios control large contiguous regions of the spectrum [5], [16], [14]. The boundaries of these regions are usually fixed and have been limited to relatively few (one to three) regions. Observations by Fujimara [5] of "devoiced" regions of frequency in vowel spectra in clean speech, together with our observations of spectra of voiced speech corrupted by random noise, argue for a more flexible excitation model than those previously developed. In addition, we hypothesize that humans can discriminate between frequency regions dominated by harmonics of the fundamental and those dominated by noise-like energy and employ this information in the process of separating voiced speech from random noise. Elimination of this acoustic cue in vocoders based on simple excitation models may help to explain the significant intelligibility decrease observed with these systems in noise [7]. To account for the observed phenomena and restore potentially useful acoustic information, a function giving the voiced/unvoiced mixture versus frequency is desirable.

One recent approach which has become quite popular is the Multipulse LPC Model [1]. In this model, Linear Predictive Coding (LPC) is used to model the spectral envelope. The excitation signal is modeled by multiple pulses per pitch period. One method for reducing the number of bits required to code the excitation signal is to allow only a small number of pulses per pitch period and then code the amplitudes and locations of these pulses. The amplitudes and locations of the pulses are estimated to minimize a weighted squared difference between the original Fourier transform and the synthetic Fourier transform. One drawback of this approach is that the pulses are placed to minimize the fine structure differences between the frequency bands of the original Fourier transform and the synthetic Fourier transform regardless of whether these bands contain periodic or aperiodic energy. It seems important to obtain a good match to the fine structure of the original spectrum in frequency bands containing periodic energy. However, in frequency bands dominated by noise-like energy, it seems important only to match the spectral envelope and not spend bits on the fine structure. Consequently, it appears that a more efficient coding scheme would result from matching only the periodic portions of the spectrum with pulses, and then coding the rest as frequency dependent noise which can then be synthesized at the receiver.

Inaccurate estimation of speech model parameters has also been a major contributor to the poor quality of vocoder synthesized speech. For example, inaccurate pitch estimates or voiced/unvoiced estimates often introduce very noticeable degradations in the synthesized speech. In noisy speech, the frequency of these degradations increases dramatically due to the increased difficulty of the speech model parameter estimation problem. Consequently, a high quality speech analysis/synthesis system must have both an improved speech model and robust methods for accurately estimating the speech model parameters.

In this paper, we present a new speech model, referred to as the Multiband Excitation Model, in which the band around each harmonic of the fundamental frequency is declared voiced or unvoiced. In addition, we develop accurate and robust estimation methods for the parameters of this new speech model and describe methods to synthesize speech from the model parameters. To illustrate a potential application of the new speech model, we develop an 8 kbit/s vocoder and evaluate its performance. Both informal listening and intelligibility tests show that the 8 kbit/s vocoder developed has very good performance both in speech quality and intelligibility, particularly for noise speech.

In Section II, our new Multiband Excitation (MBE) Model for modeling both clean and noisy speech is described. In Section III, methods for estimating the parameters of the MBE Model are developed. Section IV discusses methods for synthesizing speech from these model parameters. In Section V, the MBE analysis/synthesis system is applied to the development of a high quality 8 kbit/s vocoder. Results of informal listening as a measure of quality and Diagnostic Rhyme Tests as a measure of intelligibility are presented for this 8 kbit/s vocoder.

II. Multiband Excitation Speech Model

Due to the quasi-stationary nature of a speech signal $s(n)$, a window $w(n)$ is usually applied to the speech signal to focus attention on a short time interval of approximately 10–40 ms. The windowed speech segment $s_w(n)$ is defined by

$$s_w(n) = w(n)s(n). \tag{1}$$

The window $w(n)$ can be shifted in time to select any desired segment of the speech signal $s(n)$. Over a short time interval, the Fourier transform $S_w(\omega)$ of a windowed speech segment $s_w(n)$ can be modeled as the product of a spectral envelope $H_w(\omega)$ and an excitation spectrum $|E_w(\omega)|$,

$$\hat{S}_w(\omega) = H_w(\omega)\left|E_w(\omega)\right|. \tag{2}$$

As in many simple speech models, the spectral envelope $|H_w(\omega)|$ is a smoothed version of the original speech spectrum $|S_w(\omega)|$. The spectral envelope can be represented by linear prediction coefficients [17], cepstral coefficients [21], formant frequencies and bandwidths [24], or samples of the original speech spectrum [3]. The repre-

sentational form of the spectral envelope is not the dominant issue in our new model. However, the spectral envelope must be represented accurately enough to prevent degradations in the spectral envelope from dominating quality improvements achieved by the addition of a frequency dependent voiced/unvoiced mixture function. An example of a spectral envelope derived from the noisy speech spectrum of Fig. 1(a) is shown in Fig. 1(b).

The excitation spectrum in our new speech model differs from previous simple models in one major respect. In previous simple models, the excitation spectrum is totally specified by the fundamental frequency ω_0 and a voiced/unvoiced decision for the entire spectrum. In our new model, the excitation spectrum is specified by the fundamental frequency ω_0 and a frequency dependent voiced/unvoiced mixture function. In general, a continuously varying frequency dependent voiced/unvoiced mixture function would require a large number of parameters to represent it accurately. The addition of a large number of parameters would severely decrease the utility of this model in such applications as bit-rate reduction. To reduce this problem, the frequency dependent voiced/unvoiced mixture function has been restricted to a frequency dependent binary voiced/unvoiced decision. To further reduce the number of these binary parameters, the spectrum is divided into multiple frequency bands and a binary voiced/unvoiced parameter is allocated to each band. This new model differs from previous models in that the spectrum is divided into a large number of frequency bands (typically 20 or more), whereas previous models used three frequency bands at most [5]. Due to the division of the spectrum into multiple frequency bands with a binary voiced/unvoiced parameter for each band, we have termed this new model the Multiband Excitation Model.

The excitation spectrum $|E_w(\omega)|$ is obtained from the fundamental frequency ω_0 and the voiced/unvoiced parameters by combining segments of a periodic spectrum $|P_w(\omega)|$ in the frequency bands declared voiced with segments of a random noise spectrum $|U_w(\omega)|$ in the frequency bands declared unvoiced. The periodic spectrum $|P_w(\omega)|$ is completely determined by ω_0. One method for generating the periodic spectrum $|P_w(\omega)|$ is to take the Fourier transform magnitude of a windowed impulse train with pitch period P. In another method, the Fourier transform of the window is centered around each harmonic of the fundamental frequency and summed to produce the periodic spectrum. An example of $|P_w(\omega)|$ corresponding to $\omega_0 = 0.045\pi$ is shown in Fig. 1(c). The V/UV information allows us to mix the periodic spectrum with a random noise spectrum in the frequency domain in a frequency-dependent manner in representing the excitation spectrum.

The Multiband Excitation Model allows noisy regions of the excitation spectrum to be synthesized with 1 V/UV bit per frequency band. This is a distinct advantage over simple harmonic models in coding systems [19] where noisy regions are synthesized from the coded phase requiring around 4 or 5 bits per harmonic. In addition, when

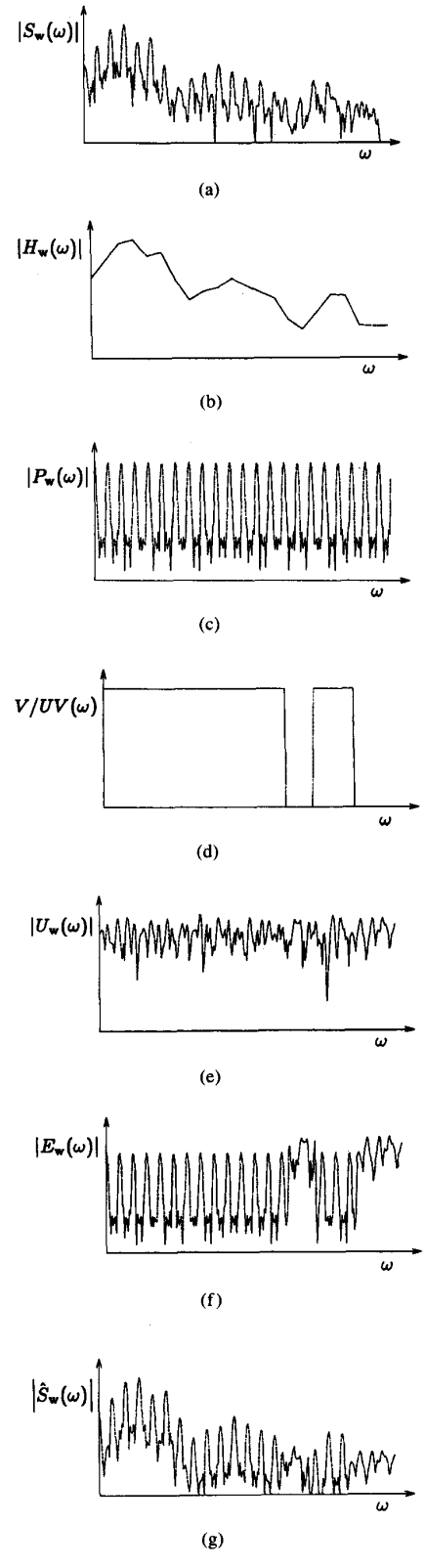

Fig. 1. Illustration of Multiband Excitation Model. (a) Original spectrum. (b) Spectral envelope. (c) Periodic spectrum. (d) V/UV information. (e) Noise spectrum. (f) Excitation spectrum. (g) Synthetic spectrum.

the pitch period becomes small with respect to the window length, noisy regions of the excitation spectrum can no longer be well approximated with a simple harmonic model.

An example of V/UV information is displayed in Fig. 1(d) with a high value corresponding to a voiced decision. An example of a typical random noise spectrum $|U_w(\omega)|$ used is shown in Fig. 1(e). The excitation spectrum $|E_w(\omega)|$ derived from $|S_w(\omega)|$ in Fig. 1(a) using the above procedure is shown in Fig. 1(f). The spectral envelope $|H_w(\omega)|$ is represented by one sample $|A_m|$ for each harmonic of the fundamental in both voiced and unvoiced regions to reduce the number of parameters. When a densely sampled version of the spectral envelope is required, it can be obtained by linearly interpolating between samples. The synthetic speech spectrum $|\hat{S}_w(\omega)|$ obtained by multiplying $|E_w(\omega)|$ in Fig. 1(f) by $|H_w(\omega)|$ in Fig. 1(b) is shown in Fig. 1(g).

It is possible [9] to synthesize high quality speech from the synthetic speech spectrum $|\hat{S}_w(\omega)|$. However, this algorithm introduces a significant delay and requires considerable computation. Consequently, we have included the phase of harmonics declared voiced as additional model parameters to avoid these problems.

The parameters that we use in our model, then, are the spectral envelope, the fundamental frequency, the V/UV information for each harmonic, and the phase of each harmonic declared voiced. The phases of harmonics in frequency bands declared unvoiced are not included since they are not required by the synthesis algorithm (Section IV).

III. Speech Analysis

In many approaches [17], [21], [2], [6], [25], the algorithms for estimation of excitation parameters and estimation of spectral envelope parameters operate independently. These parameters are usually estimated based on some reasonable but heuristic criterion without explicit consideration of how close the synthesized speech will be to the original speech. This can result in a synthetic spectrum quite different from the original spectrum.

In our approach, the excitation and spectral envelope parameters are estimated simultaneously so that the synthesized spectrum is closest in the least squares sense to the spectrum of the original speech. This approach can be viewed as an "analysis by synthesis" method [22].

Estimation of all of the speech model parameters simultaneously would be a computationally prohibitive problem. Consequently, the estimation process has been divided into two major steps. In the first step, the pitch period and spectral envelope parameters are estimated to minimize the error between the original spectrum $|S_w(\omega)|$ and the synthetic spectrum $|\hat{S}_w(\omega)|$. Then, the V/UV decisions are made based on the closeness of fit between the original and the synthetic spectrum at each harmonic of the estimated fundamental.

The parameters of our speech model can be estimated by minimizing the following error criterion:

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\,|S_w(\omega)| - |\hat{S}_w(\omega)|\,\right]^2 d\omega \qquad (3)$$

where

$$|\hat{S}_w(\omega)| = |H_w(\omega)||E_w(\omega)|. \qquad (4)$$

This error criterion was chosen since it performed well in our previous work [8]. In addition, this error criterion yields fairly simple expressions for the optimal estimates of the sample $|A_m|$ of the spectral envelope $|H_w(\omega)|$. Frequency dependent weighting functions can be applied to the original spectrum prior to minimization to emphasize high SNR regions. Other error criteria could also be used. For example, the error criterion given by

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|S_w(\omega) - \hat{S}_w(\omega)\right|^2 d\omega \qquad (5)$$

can be used to estimate both the magnitude and phase of the samples A_m of the spectral envelope.

A. Estimation of Pitch Period and Spectral Envelope

The objective is to choose the pitch period and spectral envelope parameters to minimize the error of (3). In general, minimizing this error over all parameters simultaneously is a difficult and computationally expensive problem. However, we note that for a given pitch period, the best spectral envelope parameters can be easily estimated. To show this, we divide the spectrum into frequency bands centered each harmonic of the fundamental frequency. For simplicity, we will model the spectral envelope as constant in this interval with a value of A_m. This allows the error criterion of (3) in the interval around the mth harmonic to be written as

$$\varepsilon_m = \frac{1}{2\pi}\int_{a_m}^{b_m}\left[\,|S_w(\omega)| - |A_m||E_w(\omega)|\,\right]^2 d\omega \qquad (6)$$

where the interval $[a_m, b_m]$ is an interval with a width of the fundamental frequency centered on the mth harmonic of the fundamental. The error ε_m is minimized at

$$|A_m| = \frac{\int_{a_m}^{b_m}|S_w(\omega)||E_w(\omega)|\,d\omega}{\int_{a_m}^{b_m}|E_w(\omega)|^2\,d\omega}. \qquad (7)$$

The corresponding estimate of A_m based on the error criterion of (5) is given by

$$A_m = \frac{\int_{a_m}^{b_m}S_w(\omega)E_w^*(\omega)\,d\omega}{\int_{a_m}^{b_m}|E_w(\omega)|^2\,d\omega}. \qquad (8)$$

For voiced frequency intervals, the envelope parameters are estimated by substituting the periodic transform $P_w(\omega)$ for the excitation transform $E_w(\omega)$ in (7) and (8).

Note that the A_m obtained has both magnitude and phase. An efficient method for obtaining a good approximation for the periodic transform $P_w(\omega)$ in this interval is to precompute samples of the Fourier transform of the window $w(n)$ and center it around the harmonic frequency associated with this interval.

For unvoiced frequency intervals, the envelope parameters are estimated by substituting idealized white noise (unity across the band) for $|E_w(\omega)|$ in (7) which reduces to averaging the original spectrum in each frequency interval. For unvoiced regions, only the magnitude of A_m is estimated since the phase of A_m is not required for speech synthesis.

For adjacent intervals, the minimum error for entirely periodic excitation $\tilde{\varepsilon}$ for the given pitch period is then computed as

$$\tilde{\varepsilon} \approx \sum_m \tilde{\varepsilon}_m \tag{9}$$

where $\tilde{\varepsilon}_m$ is ε_m in (6) evaluated with the $|A_m|$ of (7). In this manner, the spectral envelope parameters which minimize the error ε can be computed for a given pitch period P. This reduces the original multidimensional problem to the one-dimensional problem of finding the pitch period P that minimizes $\tilde{\varepsilon}$.

Experimentally, the error $\tilde{\varepsilon}$ tends to vary slowly with the pitch period P. This allows an initial estimate of the pitch period near the global minimum to be obtained by evaluating the error on a coarse grid. In practice, the initial estimate is obtained by evaluating the error $\tilde{\varepsilon}$ for integer pitch periods. In this initial coarse estimation of the pitch period, the high-frequency harmonics cannot be well matched so the frequency weighting function applied to the original spectrum is chosen to deemphasize high frequencies.

Since integer multiples of the correct pitch period have spectra with harmonics at the correct frequencies, the error $\tilde{\varepsilon}$ will be comparable for the correct pitch period and its integer multiples. Consequently, once the pitch period which minimizes $\tilde{\varepsilon}$ is found, the errors at submultiples of this pitch period are compared to the minimum error and the smallest pitch period with comparable error is chosen as the pitch period estimate. This feature can be used to reduce computation by limiting the initial range of P over which the error $\tilde{\varepsilon}$ is computed to long pitch periods.

To accurately estimate the voiced/unvoiced decisions in high-frequency bands, pitch period estimates more accurate than the closest integer value are required [10]. More accurate pitch period estimates can be obtained by using the best integer pitch period estimate chosen above as an initial coarse pitch period estimate. Then, the error is minimized locally to this estimate by using successively finer evaluation grids. The final pitch period estimate is chosen as the pitch period which produces the minimum error in this local minimization. The pitch period accuracies that can be obtained using this method are given in [10].

To illustrate our new approach, a specific example will be considered. In Fig. 2(a), 256 samples of female speech sampled at 10 kHz are displayed. This speech segment was windowed with a 256 point Hamming window, and an FFT was used to compute samples of the spectrum $|S_w(\omega)|$ shown in Fig. 2(b). Fig. 2(c) shows the error $\tilde{\varepsilon}$ as a function of pitch period P. The error E is smallest for $P = 85$, but since the error for the submultiple at $P = 42.5$ is comparable, the initial estimate of the pitch period is chosen as 42.5 samples. If an integer pitch period estimate is desired, the error is evaluated at pitch periods of 42 and 43 samples, and the integer pitch period estimate is chosen as the pitch period with the smaller error. If noninteger pitch periods are desired, the error $\tilde{\varepsilon}$ is minimized around this initial estimate using a finer evaluation grid. Fig. 2(d) shows the original spectrum overlayed with the synthetic spectrum for the final pitch period estimate of 42.48 samples. For comparison, Fig. 2(e) shows the original spectrum overlayed with the synthetic spectrum for the best integer pitch period estimate of 42 samples. This figure demonstrates the mismatch of the high harmonics obtained if only integer pitch periods are allowed.

To obtain the maximum sensitivity to regions of the spectrum containing pitch harmonics when large regions of the spectrum contain noise-like energy, the expected value of the error $\tilde{\varepsilon}$ should not vary with the pitch period for a spectrum consisting entirely of noise-like energy. However, since the spectral envelope is sampled more densely for longer pitch periods, the expected error is smaller for longer pitch periods. This bias toward longer pitch periods can be calculated [10], and an unbiased error criterion ε_{UB} is developed by multiplying the error $\tilde{\varepsilon}$ by a pitch period dependent correction factor to produce

$$\varepsilon_{UB} = \frac{\int_{-\pi}^{\pi} |S_w(\omega) - \hat{S}_w(\omega)|^2 \, d\omega}{\left(1 - P \sum_{n=-\infty}^{\infty} w^4(n)\right) \int_{-\pi}^{\pi} |S_w(\omega)|^2 \, d\omega}. \tag{10}$$

To obtain this result, the window $w(n)$ was normalized to have unit energy. The error criterion ε_{UB} has been normalized so that the minimum is near zero for a purely periodic signal and is near one for a noise signal. This unbiased error criterion significantly improves the performance for noisy speech.

In practice, these computations are performed by replacing integrals of continuous functions by summations of samples of these functions. However, evaluating the error criterion for all possible integer pitch periods in order to obtain an initial fundamental frequency estimate can be quite computationally expensive. Reasonable approximations [10] lead to a substantially more efficient method for computing ε_{UB}:

$$\varepsilon_{UB} \approx \frac{\sum_{n=-\infty}^{\infty} w^2(n) s^2(n) - \Psi(P)}{\left(1 - P \sum_{n=-\infty}^{\infty} w^4(n)\right) \int_{-\pi}^{\pi} |S_w(n)|^2 \, d\omega} \tag{11}$$

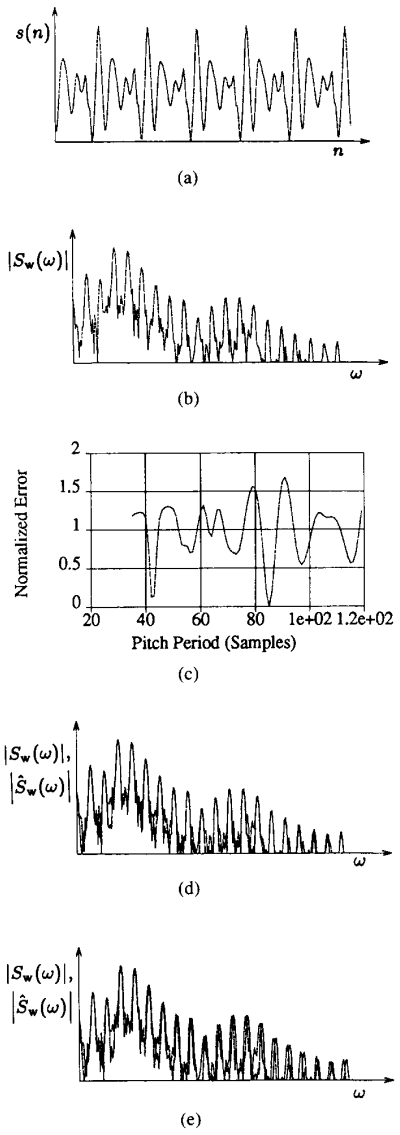

Fig. 2. Estimation of model parameters. (a) Speech segment. (b) Original spectrum. (c) Error versus pitch period. (d) Original and synthetic ($P = 42.48$). (e) Original and synthetic ($P = 42$).

where

$$\Psi(P) = P \sum_{k=-\infty}^{\infty} \phi(kP) \tag{12}$$

and $\phi(m)$ is the autocorrelation function of $w^2(n)s(n)$ given by

$$\phi(m) = \sum_{n=-\infty}^{\infty} w^2(n)s(n)w^2(n-m)s(n-m). \tag{13}$$

Minimizing (11) over P is equivalent to maximizing (12). This technique is similar to the autocorrelation method, but considers the peaks at multiples of the pitch period instead of only the peak at the pitch period. This suggests a computationally efficient method for maximizing $\Psi(P)$ over all integer pitch periods by computing the autocorrelation function using the fast Fourier transform (FFT) and then summing samples spaced by the pitch period. It should be noted that, in practice, the summations of (12) are finite due to the finite length of the window $w(n)$. For a rectangular window, the result given by (12) and (13) reduces to the result given in Wise *et al.* [27]. Since this autocorrelation domain method is somewhat less accurate than the frequency domain method discussed earlier [10], the frequency domain method is used to refine the initial coarse fundamental estimate provided by the autocorrelation domain method.

B. Pitch Tracking

Pitch tracking methods can easily be incorporated in this analysis system. Many pitch tracking methods employ a smoothing approach to reduce gross pitch errors. One problem with these techniques is that in the smoothing process, the accuracy of the pitch period estimate is degraded even for clean speech. One pitch tracking method which we have found particularly useful, in practice, for obtaining accurate estimates in clean speech and reducing gross pitch errors under very low signal-to-noise ratios, is based on a dynamic programming approach. There are three pitch track conditions to consider: 1) the pitch track starts in the current frame, 2) the pitch track terminates in the current frame, and 3) the pitch track continues through the current frame. We have found that the third condition is adequately modeled by one of the first two. We wish to find the best pitch track starting or terminating in the current frame. We will look forward and backward N frames where N is small enough that insignificant delay is encountered ($N = 3$ corresponding to 60 ms is typical). The allowable frame-to-frame pitch period deviation is set to D samples ($D = 2$ is typical). We then find the minimum error paths from N frames in the past to the current frame, and from N frames in the future to the current frame. We then determine which of these paths has the smallest error, and the initial pitch period estimate is chosen as the pitch period in the current frame in which this smallest error path terminates. The error along a path is determined by summing the errors at each pitch period through which the path passes. Dynamic programming techniques [20] are used to significantly reduce the computational requirements of this procedure.

C. Estimation of V/UV Information

The voiced/unvoiced decision for each harmonic is made by comparing the normalized error over each harmonic of the estimated fundamental to a threshold. When the normalized error over the mth harmonic

$$\xi_m = \frac{\tilde{\varepsilon}_m}{\frac{1}{2\pi}\int_{a_m}^{b_m} |S_w(\omega)|^2 \, d\omega} \tag{14}$$

is below the threshold, this region of the spectrum matches that of a periodic spectrum well and the mth harmonic is

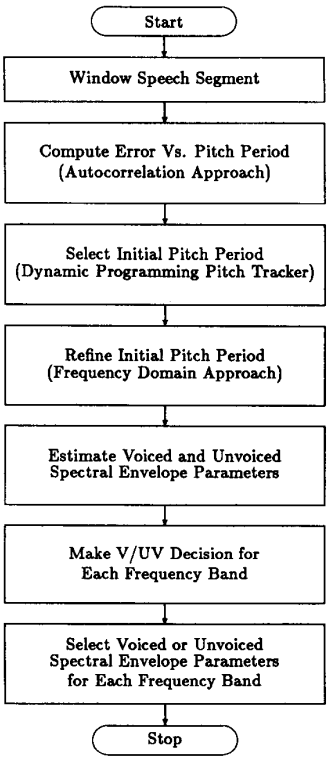

Fig. 3. Analysis algorithm flowchart.

marked voiced. When ξ_m is above the threshold, this region of the spectrum is assumed to contain noise-like energy. A threshold value of 0.2 works well in practice. After the voiced/unvoiced decision is made for each frequency band, the voiced or unvoiced spectral envelope parameter estimates are selected as appropriate.

D. Analysis Algorithm

The analysis algorithm that we use in practice consists of the following steps (see Fig. 3).

1) Window a speech segment with the analysis window.

2) Compute the unbiased error criterion of (10) versus pitch period using the efficient autocorrelation domain approach (11). This error is typically computed for all integer pitch periods from 20 to 120 samples for a 10 kHz sampling rate.

3) Use the dynamic programming approach described in Section III-B to select the initial pitch period estimate. This pitch tracking technique improves tracking through very low signal-to-noise ratio (SNR) segments while not decreasing the accuracy in high SNR segments.

4) Refine this initial pitch period estimate by minimizing (10) using the more accurate frequency domain pitch period estimation method described in Section III-A.

5) Estimate the voiced and unvoiced spectral envelope parameters using the techniques described in Section III-A.

6) Make a voiced/unvoiced decision for each frequency band in the spectrum. The number of frequency bands in the spectrum can be as large as the number of harmonics of the fundamental present in the spectrum.

7) The final spectral envelope parameter representation is composed by combining voiced spectral envelope parameters in those frequency bands declared voiced with unvoiced spectral envelope parameters in those frequency bands declared unvoiced.

IV. Speech Synthesis

In the previous two sections, the Multiband Excitation Model parameters were described, and methods to estimate these parameters were developed. In this section, an approach to synthesizing speech from the model parameters is presented. There exist a number of methods for synthesizing speech from the spectral envelope and excitation parameters. One approach is to generate a sequence of synthetic spectral magnitudes from the estimated model parameters. Algorithms [8] for estimating a signal from the synthetic short-time Fourier transform magnitude (STFTM) are expensive computationally and require a processing delay of approximately 1 s. This delay is unacceptable in most real-time speech bandwidth compression applications, and we have not considered this approach further.

In another approach, which we refer to as the frequency domain approach, an excitation transform is constructed by combining segments of a periodic transform in frequency bands declared voiced with segments of a noise transform in frequency bands declared unvoiced. The noise transform segments are normalized to have an average magnitude of unity. A spectral envelope is constructed by linearly interpolating between the spectral envelope samples $|A_m|$. The phase of the spectral envelope in voiced frequency bands is set to the phase of envelope samples A_m. A synthetic STFT is then constructed as the product of the excitation transform and the spectral envelope. The weighted overlap-add algorithm [8] can then be used to estimate a signal with STFT closest to this synthetic STFT in the least-squares sense. A problem can arise with this method when voiced speech is synthesized for large window shifts (large window shifts are required to reduce the bit-rate in speech coding applications). Since the voiced portion of the synthesized signal is modeled as a periodic signal with constant fundamental over the entire frame, when large window shifts are used, a large change in fundamental frequency from one frame to the next causes time discontinuities in the harmonics of the fundamental in the STFTM.

A third apprach to synthesizing speech, which we refer to as the time domain approach, involves synthesizing the voiced and unvoiced portions in the time domain and then adding them together. The voiced signal can be synthesized as the sum of sinusoidal oscillators with frequencies at the harmonics of the fundamental and amplitudes set by the spectral envelope parameters. This technique has the advantage of allowing a continuous variation in fun-

damental frequency from one frame to the next eliminating the problem of time discontinuities in the harmonics of the fundamental in the STFTM. The unvoiced signal can be synthesized as the sum of bandpass filtered white noise.

The time domain method was selected for synthesizing the voiced portion of the synthetic speech. This method was selected due to its advantage of allowing a continuous variation in fundamental frequency from frame to frame. The frequency domain method was selected for synthesizing the unvoiced portion of the synthetic speech. This method was selected due to the ease and efficiency of implementation of a filter bank in the frequency domain with the fast Fourier transform (FFT) algorithm.

A block diagram of our current speech synthesis system is shown in Figs. 4–7. First, the spectral envelope samples are separated into voiced or unvoiced spectral envelope samples depending on whether they are in frequency bands declared voiced or unvoiced (Fig. 4). Voiced envelope samples in frequency bands declared unvoiced are set to zero, as are unvoiced envelope samples in frequency bands declared voiced. Voiced envelope samples include both magnitude and phase, whereas unvoiced envelope samples include only the magnitude.

Voiced speech is synthesized from the voiced envelope samples by summing the outputs of a band of sinusoidal oscillators running at the harmonics of the fundamental frequency (Fig. 5):

$$\hat{s}_v(t) = \sum_m A_m(t) \cos\left(\theta_m(t)\right). \tag{15}$$

The amplitude function $A_m(t)$ is linearly interpolated between frames with the amplitudes of harmonics marked unvoiced set to zero. The phase function $\theta_m(t)$ is determined by an initial phase ϕ_0 and a frequency track $\omega_m(t)$ as follows:

$$\theta_m(t) = \int_0^t \omega_m(\xi)\, d\xi + \phi_0. \tag{16}$$

The frequency track $\omega_m(t)$ is linearly interpolated between the mth harmonic of the current frame and that of the next frame by

$$\omega_m(t) = m\omega_0(0)\frac{(S-t)}{S} + m\omega_0(S)\frac{t}{S} + \Delta\omega_m \tag{17}$$

where $\omega_0(0)$ and $\omega_0(S)$ are the fundamental frequencies at $t = 0$ and $t = S$, respectively, and S is the window shift. The initial phase ϕ_0 and frequency deviation $\Delta\omega_m$ parameters are chosen so that the principal values of $\theta_m(0)$ and $\theta_m(S)$ are equal to the measured harmonic phases in the current and next frame. When the mth harmonics of the current and next frames are both declared voiced, the initial phase ϕ_0 is set to the measured phase of the current frame, and $\Delta\omega_m$ is chosen to be the smallest frequency deviation required to match the phase of the next frame. When either of the harmonics is declared unvoiced, only the initial phase parameter ϕ_0 is required to match the phase function $\theta_m(t)$ with the phase of the voiced harmonic ($\Delta\omega_m$ is set to zero). When both harmonics are declared unvoiced, the amplitude function $A_m(t)$ is zero over the entire interval between frames so any phase function will suffice.

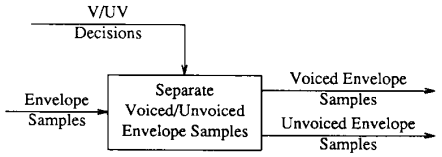

Fig. 4. Separation of envelope samples.

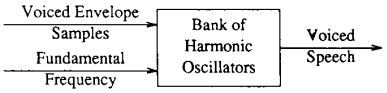

Fig. 5. Voiced speech synthesis.

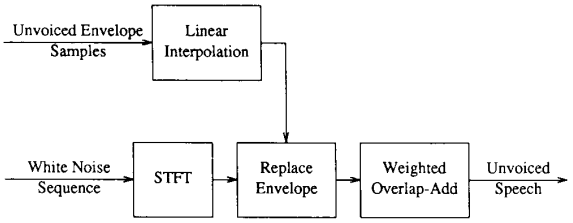

Fig. 6. Unvoiced speech synthesis.

Large differences in fundamental frequency can occur between adjacent frames due to word boundaries and other effects. In these cases, linear interpolation of the fundamental frequency between frames is a poor model of fundamental frequency variation and can lead to artifacts in the synthesized signal. Consequently, when fundamental frequency changes of more than 10 percent are encountered from frame to frame, the voiced harmonics of the current frame and the next frame are treated as if followed and preceded respectively by unvoiced harmonics.

Unvoiced speech is synthesized from the unvoiced envelope samples by first synthesizing a white noise sequence. For each frame, the white noise sequence is windowed and an FFT is applied to produce samples of the Fourier transform (Fig. 6). In each unvoiced frequency band, the noise transform samples are normalized to have unity magnitude. The unvoiced spectral envelope is constructed by linearly interpolating between the envelope samples $|A_m|$. The normalized noise transform is multiplied by the spectral envelope to produce the synthetic transform. The synthetic transforms are then used to synthesize unvoiced speech using the weighted overlap-add method.

The final synthesized speech is generated by summing the voiced and unvoiced synthesized speech signals (Fig. 7).

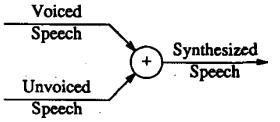

Fig. 7. Speech synthesis.

V. Development of 8 kbit/s Multiband Excitation Vocoder

Among many applications of our new model, we considered the problem of bit-rate reduction for speech transmission and storage. In a number of speech coding applications, it is important to reproduce the original clean or noisy speech as closely as possible. For example, in mobile telephone applications, users would like to be able to identify the person on the other end of the phone and are usually annoyed at any artificial sounding degradations. These degradations are particularly severe for most vocoders when operating in noisy environments such as a moving car. Consequently, for these applications, we are interested in both the quality and intelligibility of the reproduced speech. In other applications, such as a fighter cockpit, the message is of primary importance. For these applications, we are interested mainly in the intelligibility of the reproduced speech.

To demonstrate the performance of the Multiband Excitation Speeck Analysis/Synthesis System for this problem, an 8 kbit/s speech coding system was developed. Since our primary goal is to demonstrate the high performance of the Multiband Excitation Model and the corresponding speech analysis methods, conventional parameter coding methods have been used to facilitate comparison with other systems.

The major innovation in the Multiband Excitation Speech Model is the ability to declare a large number of frequency regions as containing periodic or aperiodic energy. To determine the advantage of this new model, the Multiband Excitation Vocoder operating at 8 kbit/s was compared to a system using a single V/UV bit per frame (Single Band Excitation Vocoder). The Single Band Excitation (SBE) Coder employs exactly the same parameters as the Multiband Excitation Speech Coder, except that one V/UV bit per frame is used instead of 12 and is a degenerate case of the MBE Coder (one frequency band). Although this results in a somewhat smaller bit rate for the SBE Coder (7.45 kbit/s), we wished to maintain the same coding rates for the other parameters in order to focus the comparison on the usefulness of the V/UV information rather than particular modeling or coding methods for the other parameters.

A. Coding of Speech Model Parameters

A 25.6 ms Hamming window was used to segment 4 kHz bandwidth speech sampled at 10 kHz. The estimated speech model parameters were coded at 8 kbit/s using a 50 Hz frame rate. This allows 160 bits per frame for coding the harmonic magnitudes and phases, fundamental frequency, and voiced/unvoiced information. The number of bits allocated to each of these parameters per frame is displayed in Table I. The fundamental frequency is coded using 9 bits with uniform quantization. As discussed in Section IV, phase is not required for harmonics declared unvoiced. Consequently, bits assigned to phases declared unvoiced are reassigned to the magnitude. When all harmonics are declared voiced, 45 bits are assigned for phase coding and 94 bits are assigned for magnitude coding. At the other extreme, when all harmonics are declared unvoiced, no bits are assigned to phase and 139 bits are assigned for magnitude coding.

Coding of Harmonic Magnitudes: The harmonic magnitudes are coded using the same techniques employed by channel vocoders [11] (Fig. 8). In this method, the logarithms of the harmonic magnitudes are encoded using adaptive differential PCM across frequency. The log-magnitude of the first harmonic is coded using 5 bits with a quantization step size of 2 dB. The number of bits assigned to coding the difference between the log-magnitude of the mth harmonic and the coded value of the previous harmonic (within the same frame) is determined by summing samples of the bit density curve of Fig. 9 over the frequency interval occupied by the mth harmonic. The available bits for coding the magnitude are then assigned to each harmonic in proportion to these sums. The quantization step size depends on the number of bits assigned and is listed in Table II.

Coding of Harmonic Phases: When generating the STFT phase, the primary consideration in high quality synthesis is to generate the STFT phase so that the phase difference from frame to frame is consistent with the fundamental frequency in voiced regions. Obtaining the correct relative phase between harmonics is of secondary importance for high quality synthesis. However, results of informal listening indicate that incorrect relative phase between harmonics can cause a variety of perceptual differences between the original and synthesized speech especially at low frequencies.

Fig. 10 shows the method used for phase coding. The phases of harmonics declared voiced are encoded by predicting the phase of the current frame from the phase of the previous frame using the average fundamental frequency for the two frames. Then, the difference between the predicted and estimated phase for the current frame is coded starting with the phases of the low-frequency harmonics. The difference between the predicted and estimated phase is set to zero for any uncoded voiced harmonics to maintain a frame-to-frame phase difference consistent with the fundamental frequency. The phases of harmonics in frequency regions declared unvoiced do not need to be coded since they are not required by the speech synthesizer.

The phase differences for voiced regions are expected to cluster around zero due to the influence of the fundamental frequency. Phase difference histograms were computed for several frequency bands. These histograms were used to develop 13 level Lloyd–Max quantizers [15], [18], by minimizing the average quantization error.

TABLE I
BIT ALLOCATION PER FRAME

Parameter	Bits
Fundamental Frequency	9
Harmonic Magnitudes	139-94
Harmonic Phases	0-45
Voiced/Unvoiced Bits	12
Total	160

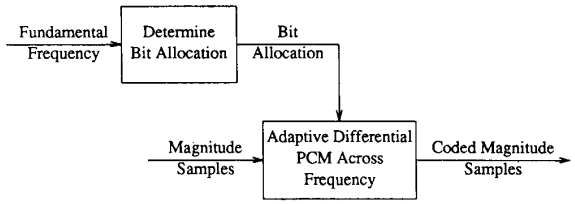

Fig. 8. Coding of magnitudes.

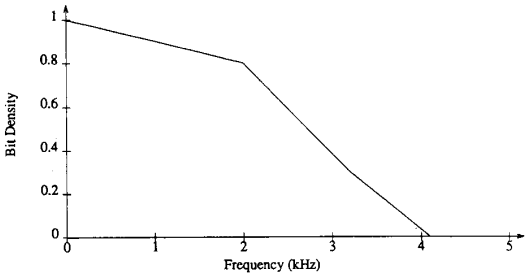

Fig. 9. Magnitude bit density curve.

TABLE II
QUANTIZATION STEP SIZES

Bits	Step Size (dB)	Min (dB)	Max (dB)
1	8	-4	4
2	6.5	-9.75	9.75
3	5	-17.5	17.5
4	3	-22.5	22.5
5	2	-31	31
6	1	-31.5	31.5
7	0.5	-31.75	31.75
8	0.25	-31.875	31.875

Coding of V/UV Information: The voiced/unvoiced information can be encoded using a variety of methods. We have observed that voiced/unvoiced decisions tend to cluster in both frequency and time due to the slowly varying nature of speech in the STFTM domain. Run-length coding can be used to take advantage of this expected

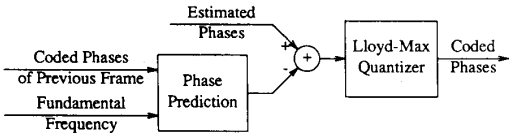

Fig. 10. Coding of phases.

clustering of voiced/unvoiced decisions. However, run-length coding requires a variable number of bits to exactly encode a fixed number of samples. This makes implementation of a fixed rate coder more difficult.

A simple approach to coding the voiced/unvoiced information with a fixed number of bits while providing good performance was developed (Fig. 11). In this approach, if N bits are available, the spectrum is divided into N equal frequency bands and a voiced/unvoiced bit is used for each band. The voiced/unvoiced bit is set by comparing a weighted sum of the normalized errors of all of the harmonics in a particular frequency band to a threshold. When the weighted sum is less than the threshold, the frequency band is set to voiced. When the weighted sum is greater than the threshold, the frequency band is set to unvoiced. The sum is weighted by the estimated harmonic magnitudes as follows:

$$E_k = \frac{\sum_m |A_m| \xi_m}{\sum_m |A_m|} \tag{18}$$

where m is summed over all of the harmonics in the kth frequency band.

Coding—Implementation: The 8 kbit/s MBE Coder was implemented on a MASSCOMP computer (68020 CPU) in the C programming language. The entire system (analysis, coding, synthesis) required approximately 1 min of processing time per second of input speech on this general purpose computer system. The increased throughput available from special purpose architectures and conversion from floating point to fixed point should make these algorithms implementable in real time with several Digital Signal Processing (DSP) chips.

B. Quality—Informal Listening

Informal listening was used to compare a number of speech sentences processed by the 8 kbit/s Multiband Excitation Vocoder and the 7.45 kbit/s Single Band Excitation Vocoder. For clean speech, the speech sentences coded by the MBE Vocoder did not have the slight "buzziness" present in some regions of speech processed by the SBE Vocoder. Fig. 12(a) shows a spectrogram of the sentence. "He has the bluest eyes" spoken by a male speaker.

In this spectrogram, darkness is proportional to the log of the energy versus time (0–2 s, horizontal axis) and frequency (0–5 kHz, vertical axis). Periodic energy is typified by the presence of parallel horizontal bars of darkness which occur at the harmonics of the fundamental

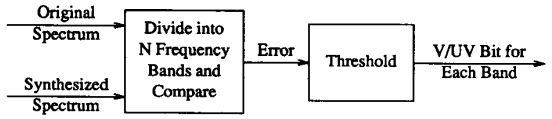

Fig. 11. Coding of V/UV information.

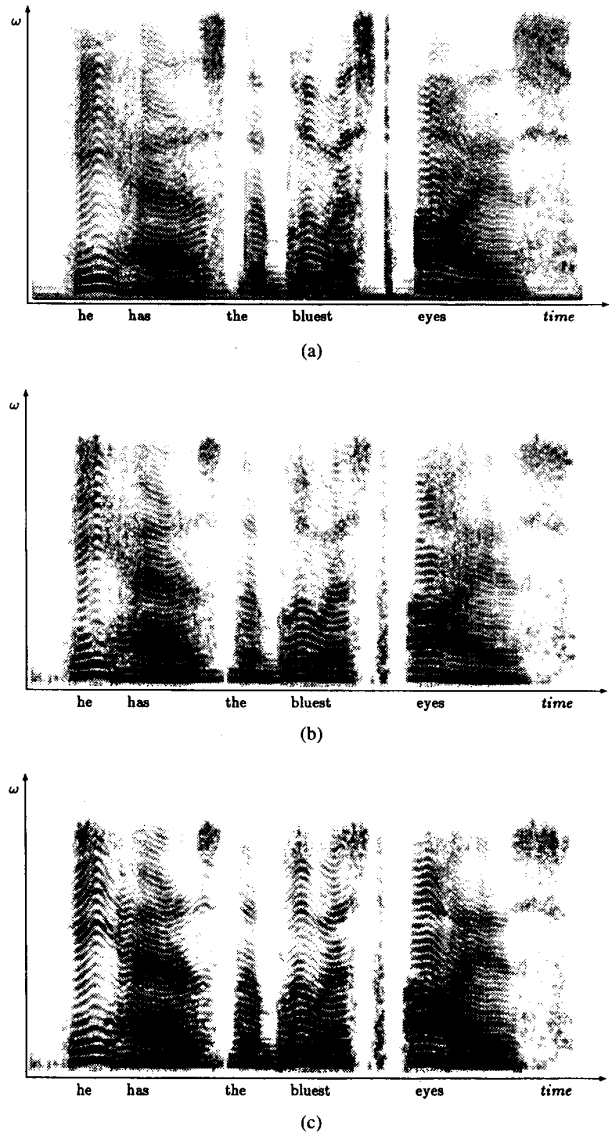

Fig. 12. Clean speech spectrograms. (a) Uncoded speech. (b) MBE vocoder. (c) SBE vocoder.

frequency. One region of particular interest is the /h/ phoneme in the word "has." In this region, several harmonics of the fundamental frequency appear in the low-frequency region, while the upper frequency region is dominated by aperiodic energy. The Multiband Excitation Vocoder operating at 8 kbit/s reproduces this region quite faithfully using 12 V/UV bits [Fig. 12(b)]. The SBE Vocoder declares the entire spectrum voiced and replaces the aperiodic energy apparent in the original spectrogram with harmonics of the fundamental frequency [Fig. 12(c)]. This causes a "buzzy" sound in the speech synthesized by the SBE Vocoder which is eliminated by the MBE Vocoder. The MBE Vocoder produces fairly high quality speech at 8 kbit/s. The major degradation in these two systems (other than the "buzziness" in the SBE Vocoder) is a slightly reverberant quality due to the large synthesis windows (40 ms triangular windows) and the lack of enough coded phase information.

For speech corrupted by additive random noise [Fig. 13(a)], the SBE Vocoder (Fig. 13(c)] had severe "buzziness" and a number of voiced/unvoiced errors. The severe "buzziness" is due to replacing the aperiodic energy evident in the original spectrogram by harmonics of the fundamental frequency. The V/UV errors occur due to dominance of the aperiodic energy in all but a few small regions of the spectrum. The voiced/unvoiced threshold could not be raised further without a large number of the totally unvoiced frames being declared voiced. The noisy speech sentences processed by the Multiband Excitation Vocoder [for example, see Fig. 13(b)] did not have the severe "buzziness" present in the Single Band Excitation Speech Coder and did not seem to have a problem with voiced/unvoiced errors since much smaller frequency regions are covered by each V/UV decision. In addition, the sentences processed by the MBE Vocoder sound very close to the original noisy speech.

C. *Intelligibility—Diagnostic Rhyme Tests*

The Diagnostic Rhyme Test (DRT) was developed to provide a measure of the intelligibility of speech signals. The DRT is a refinement of earlier intelligibility tests such as the Rhyme Test developed by Fairbanks [4] and the Modified Rhyme Test developed by House *et al.* [12]. The form of the DRT used here is described in detail in Voiers [26]. The DRT score is adjusted to remove the effects of guessing so that random guessing would achieve a score of zero on average. No errors in a DRT correspond to a score of 100.

The DRT was employed to compare uncoded speech with the 8 kbit/s Multiband Excitation Vocoder (12 V/UV bits per frame) and the Single Band Excitation Vocoder (1 V/UV bit per frame). Two conditions were tested: 1) clean speech, and 2) speech corrupted by additive white Gaussian noise. Based on the informal listening in the previous section, we expect the scores for the two vocoders to be very close for clean speech since only a slight quality improvement was noted for this case. For noisy speech, the MBE Vocoder provides a significant quality improvement over the SBE Vocoder which leads us to expect a measurable intelligibility improvement. The noise level was adjusted to produce approximately a 5 dB signal-to-noise ratio in the noisy speech. However, since amplitudes of the words on the DRT tapes differed significantly from each other, the SNR varied substantially from word to word. In these tests, we are interested in the

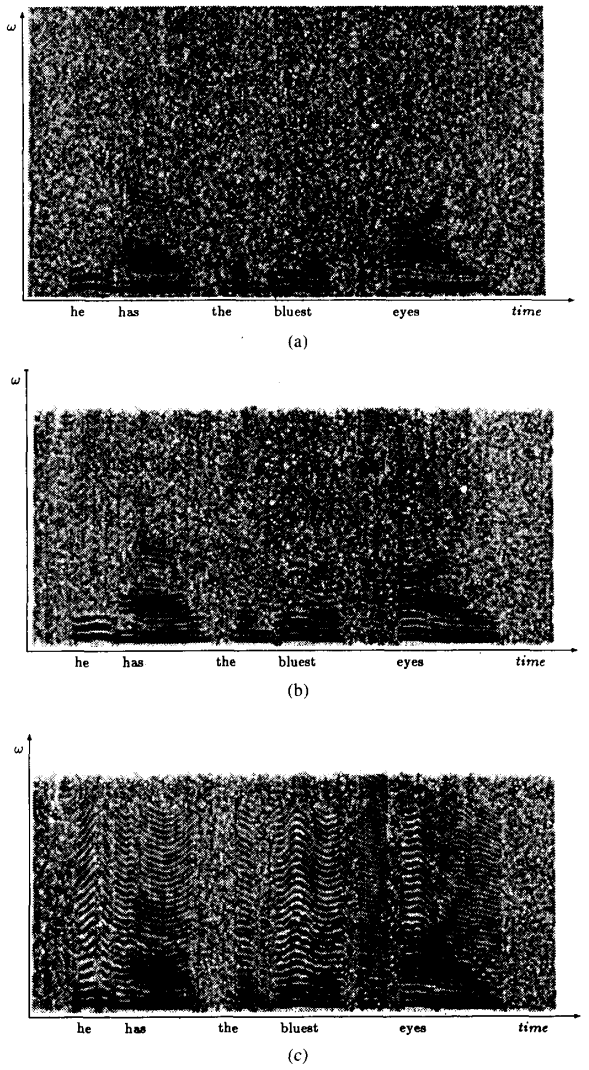

Fig. 13. Noisy speech spectrograms. (a) Uncoded speech. (b) MBE vocoder. (c) SBE vocoder.

relative performance of the vocoders in the same background noise which makes the noise level uncritical.

DRT test tapes for three speakers for each of the six conditions (2 SNR's × 3 coding conditions) were submitted to RADC for evaluation. The DRT's performed by RADC employed experienced listeners in a fairly controlled environment. The resulting DRT scores are presented for clean speech in Table III and for noisy speech in Table IV.

For clean speech, as expected, a couple of points are lost going from uncoded to coded due to low-pass filtering inherent in the vocoders and degradations introduced by coding. Also, the intelligibility scores are approximately the same for the MBE Vocoder and the SBE Vocoder.

For noisy speech, the MBE Vocoder performs an average of about 12 points better than the SBE Vocoder while

TABLE III
DRT SCORES–CLEAN SPEECH

System	Type	Speaker			Average
		CH	JE	RH	
Uncoded	Mean	98.2	96.6	98.7	97.8
	S. D.	.33	.55	.38	.30
8 kbps MBE	Mean	97.0	94.4	97.1	96.2
	S. D.	.54	.39	.33	.35
7.45 kbps SBE	Mean	96.9	94.1	96.9	96.0
	S. D.	.44	.55	.81	.44

TABLE IV
DRT SCORES–NOISY SPEECH

System	Type	Speaker			Average
		CH	JE	RH	
Uncoded	Mean	67.5	52.6	69.3	63.1
	S. D.	1.3	1.6	1.5	1.8
8 kbps MBE	Mean	60.8	48.7	64.5	58.0
	S. D.	1.4	1.4	1.8	1.6
7.45 kbps SBE	Mean	50.3	37.9	49.9	46.0
	S. D.	.94	2.3	1.8	1.6

performing only about 5 points worse than the uncoded noisy speech. This demonstrates the utility of the extra voiced/unvoiced bands in the Multiband Excitation Vocoder.

VI. CONCLUSION

In this paper, we presented a new speech model. We also presented methods for estimating the speech model parameters and methods for synthesizing speech from the estimated speech model parameters. The model was applied to the development of a high quality 8 kbit/s vocoder, and its performance was evaluated through both informal listening and DRT tests. The results indicate that the Multiband Excitation Model has a defnite advantage over a single band excitation model.

There are various ways to improve the performance of the 8 kbit/s Multiband Excitation Vocoder. For example, the method we employed in coding the estimated model parameters is somewhat crude, and we have not devoted much effort to optimizing the coding method. Some additional efforts have the potential to improve the system performance significantly.

In addition to speech coding, the Multiband Excitation Vocoder has potential usefulness in various other applications. Since the Multiband Excitation Model separately estimates spectral envelope and excitation parameters, it can be applied to problems requiring modifications of these parameters. For example, in the application of enhancement of speech spoken in a helium–oxygen mixture, a nonlinear frequency warping of the spectral envelope is

desired without modifying the excitation parameters [23]. Other applications include time-scale modification (modification of the apparent speaking rate without changing other characteristics) and pitch modification. Since the Multiband Excitation Model appears to provide an intelligibility improvement over a system employing a single voiced/unvoiced decision for the entire spectrum, this model may also prove useful for the front ends of speech recognition systems.

REFERENCES

[1] B. S. Atal and J. R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Apr. 1982, pp. 614-617.

[2] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*. New York: Dover, 1958.

[3] H. Dudley, "The vocoder," *Bell Labs Rec.*, vol. 17, pp. 122-126, 1939.

[4] G. Fairbanks, "Test of phonemic differentiation: The rhyme test," *J. Acoust. Soc. Amer.*, vol. 30, pp. 596-600, 1958.

[5] O. Fujimara, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, pp. 68-72, Mar. 1968.

[6] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, no. 2, pt. 2, pp. 442-448, Aug. 1969.

[7] B. Gold and J. Tierney, "Vocoder analysis based on properties of the human auditory system," *M. I. T. Lincoln Lab. Tech. Rep.* TR-670, Dec. 1983.

[8] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 236-243, Apr. 1984.

[9] ——, "A new model-based speech analsys/synthesis system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL, Mar. 26-29, 1985, pp. 513-516.

[10] D. W. Griffin, "Multiband excitation vocoder," Ph.D. dissertation, M.I.T., Cambridge, MA, 1987.

[11] J. N. Holmes, "The JSRU channel vocoder," *Proc. IEEE*, vol. 127, pt. F, no. 1, pp. 53-60, Feb. 1980.

[12] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Amer.*, vol. 37, pp. 158-166, 1965.

[13] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method," in *Rep. 6th Int. Congr. Acoust.*, Tokyo, Japan, 1968, pp. C17-20, Paper C-5-5.

[14] S. Y. Kwon and A. J. Goldberg, "An enhanced LPC vocoder with no voiced/unvoiced switch," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 851-858, Aug. 1984.

[15] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129-137, Mar. 1982.

[16] J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins, "A mixed-source excitation model for speech compression and synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1978, pp. 163-166.

[17] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[18] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, 2, pp. 7-12, Mar. 1960.

[19] R. J. McAulay and T. F. Quatieri, "Mid-rate coding based on a sinusoidal representation of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 236-243, Apr. 1984.

[20] C. S. Myers and L. R. Rabiner, "Connected digit recognition using a level-building DTW algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 351-363, June 1981.

[21] A. V. Oppenheim, "A speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Amer.*, vol. 45, pp. 458-465, Feb. 1969.

[22] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[23] M. A. Richards, "Helium speech enhancement using the short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 841-853, Dec. 1982.

[24] A. E. Rosenberg, R. W. Schafer, and L. R. Rabiner, "Effects of smoothing and quantizing the parameters of formant-coded voiced speech," *J. Acoust. Soc. Amer.*, vol. 50, no. 6, pp. 1532-1538, Dec. 1971.

[25] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 24-33, Feb. 1977.

[26] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technol.*, Jan./Feb. 1983.

[27] J. D. Wise, J. R. Capiro, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 418-423, Oct. 1976.

Daniel W. Griffin was born in Detroit, MI, on December 18, 1960. He received the B.S. degree in computer engineering from the University of Michigan in 1981 and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, both in electrical engineering, in 1983 and 1987, respectively.

He is currently with the Advanced Signal Processing Group at Sanders Associates, Nashua, NH. His research interests include digital signal processing and speech processing.

Jae S. Lim (S'76-M'78-SM'83-F'86) received the S.B., S.M., E.E., and Sc.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1974, 1975, 1978, and 1978, respectively.

He joined the M.I.T. Faculty in 1978 as an Assistant Professor, and is currently an Associate Professor in the Department of Electrical Engineering and Computer Science. While on leave from M.I.T., he was a Research Staff member at the M.I.T. Lincoln Laboratory during 1978-1979, and a Visting Researcher at the Woods Hole Oceanographic Institute in 1986. His research interests include digital signal processing and its applications to speech and image processing. He has contributed more than 90 articles to journals and conference proceedings. He is the Editor of a reprint book, *Speech Enchancement* (1982), a Co-editor (with A. Oppenheim) of *Advanced Topics in Signal Processing* (1987), and the author of *Two-Dimensional Signal and Image Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1988). He also contributed three chapters to the books edited by M. Ekstrom (1984), T. Kailath (1985), and T. Huang (1986).

Dr. Lim is the winner of three prize paper awards, one from the Boston Chapter of the Acoustical Society of America in 1976, and two from the IEEE ASSP Society in 1979 (ASSP Paper Award) and in 1985 (ASSP Senior Award). He is also a co-recipient of the 1984 Harold E. Edgerton Faculty Achievement Award, and the recipient of the 1984 M.I.T. Graduate Student Council's EECS Department Teaching Award. He is a member of Eta Kappa Nu and Sigma Xi.