Multi-Band Excitation Vocoder

by

Daniel W. Griffin

B. S., University of Michigan (1981)
S. M., Massachusetts Institute of Technology (1983)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1987

Signature of Author. _____
Department of Electrical Engineering and Computer Science
February 23, 1987

Certified by _____
Jae S. Lim
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Multi-Band Excitation Vocoder

by

Daniel W. Griffin

Submitted to the Department of Electrical Engineering
and Computer Science on February 23, 1987 in partial fulfillment
of the requirements for the Degree of Doctor of Philosophy
in Electrical Engineering

Abstract

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 underlying speech model have been quite successful in synthesizing intelligible speech, they have not been successful in synthesizing high quality speech. For clean speech, the synthesized speech often exhibits a "buzzy" quality. For noisy speech, severe "buzziness" and other degradations often occur resulting in a large drop in intelligibility scores. The poor quality of the synthesized speech is, in part, due to the excitation models and the parameter estimation methods used

in existing vocoders.

This thesis presents the Multi-Band Excitation Vocoder which contains a speech model allowing the band around each harmonic of the fundamental frequency to be declared voiced or unvoiced. Accurate and robust estimation methods are developed for the parameters of this new speech model and methods for synthesizing speech from the model parameters are described. Methods for coding the speech model parameters are presented and an 8 kbps vocoder is developed.

This 8 kbps Multi-Band Excitation (MBE) Vocoder is compared with a more conventional Single Band Excitation (SBE) Vocoder (1 V/UV bit per frame) in terms of quality and intelligibility. Informal listening indicates that the "buzzy" quality of the SBE Vocoder is eliminated by the MBE Vocoder with the improvement being most dramatic in noisy speech. Intelligibility tests (Diagnostic Rhyme Tests) for speech corrupted by additive white noise (approximately 5 dB SNR) produced an average score of 58.0 points for the MBE Vocoder, 12 points better than the average score of 46.0 for the SBE Vocoder. In addition, the average score for the MBE Vocoder was only about 5 points below the average DRT score of 63.1 for the uncoded noisy speech. This represents a much smaller intelligibility decrease in noise than experienced by most vocoders.

Thesis Supervisor: Jae S. Lim
Title: Assoc. Professor of Elec. Eng.

Acknowledgements

I wish to thank Jae Lim for his guidance and many technical discussions. I wish to thank the members of the Digital Signal Processing Group at M.I.T. for providing an excellent technical environment. I wish to thank my parents for encouraging and supporting my academic interests. I wish to thank Webster Dove and Doug Mook for many needed mental and physical thesis diversions. I wish to thank my wife Janet for all of her help and encouragement. Finally, I gratefully acknowledge the financial support of Sanders Associates and M.I.T.

For My Wife

Janet

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Description

In a number of applications, introduction of speech models provides improved performance. For example, in applications such as bandwidth compression of speech, introduction of an appropriate speech model provides increased intelligibility at low bit rates when compared to typical direct coding of the waveform. The advantage of introducing a speech model is that the highly redundant speech waveform is transformed to model parameters with lower bandwidth. Examples of systems based on an underlying speech model (vocoders) include linear prediction vocoders, homomorphic

vocoders, and channel vocoders. In these systems, speech is modeled on a short-time basis as the response of a linear system excited by a periodic impulse train for voiced sounds or random noise for unvoiced sounds. 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. This class of speech models is chosen since the excitation and system parameters tend to vary slowly with time due to physical constraints on the vocal tract and its excitation sources. 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.

In addition to the lower bandwidth of the model parameters, speech models are often introduced to allow speech transformations through modification of the model parameters. For example, in the application of enhancement of speech spoken in a helium-oxygen mixture, a nonlinear fre-

quency warping of the spectral envelope is desired without modifying the excitation parameters [28]. Introduction of a speech model allows separation of spectral envelope and excitation parameters for separate processing which could not be directly applied to the speech waveform.

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 (Table 1.1) for the Belgard 2400 bps 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 frequency regions dominated by periodic energy due to voiced speech and those dominated by aperiodic energy due to random noise.

Vocoder	Clean Speech	F15 Noise
Uncoded	96.8	90.1
Belgard: 2400 bps	86.5	71.4
Belgard: Noise Excitation	86.4	66.3

Table 1.1: DRT Scores

Another important piece of information in Table 1.1 is that for clean speech, the DRT score remains about the same when an all-noise excitation is used in the Belgard Vocoder. However, for noisy speech, the DRT score drops about 5 points with the all-noise excitation. This indicates that the composition of the excitation signal can be important for intelligibility, especially in noisy speech.

As will be discussed in Section 1.2, in previous approaches to this problem the voiced/unvoiced decisions or ratios control large contiguous regions of the spectrum. These approaches are too restrictive to adequately model

many speech segments, especially voiced speech in noise.

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.

1.2 Background

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 excitation models having time-invariant spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with

fixed spectral envelopes. The mixture ratio controls the amplitudes of the periodic and noise sources. Examples of such models include Itakura and Saito [14], and Kwon and Goldberg [15]. In the excitation model proposed by Itakura and Saito, a white noise source is added to a white periodic source. The mixture ratio between these sources is estimated from the height of the peak of the autocorrelation of the LPC residual. Results from this model were not encouraging [17]. In one excitation model implemented by Kwon and Goldberg, a white periodic source and a white noise source with the mixture ratio estimated from the autocorrelation of the LPC residual are reported to produce "slightly muffled" and "hoarse" synthesized speech.

The primary assumption in these excitation models is that the spectral shapes of the periodic and noise sources is not time-varying. This assumption is often violated in clean speech. For example, inspection of the speech spectra in mixed voicing regions such as a typical /z/ (Figure 1.1) indicates that low frequencies exhibit primarily periodic excitation and the high frequencies exhibit primarily noise-like excitation. However, inspection of speech spectra in almost completely voiced regions such as a typical /a/ (Figure 1.2) indicate that a periodic source with a nearly flat spectral

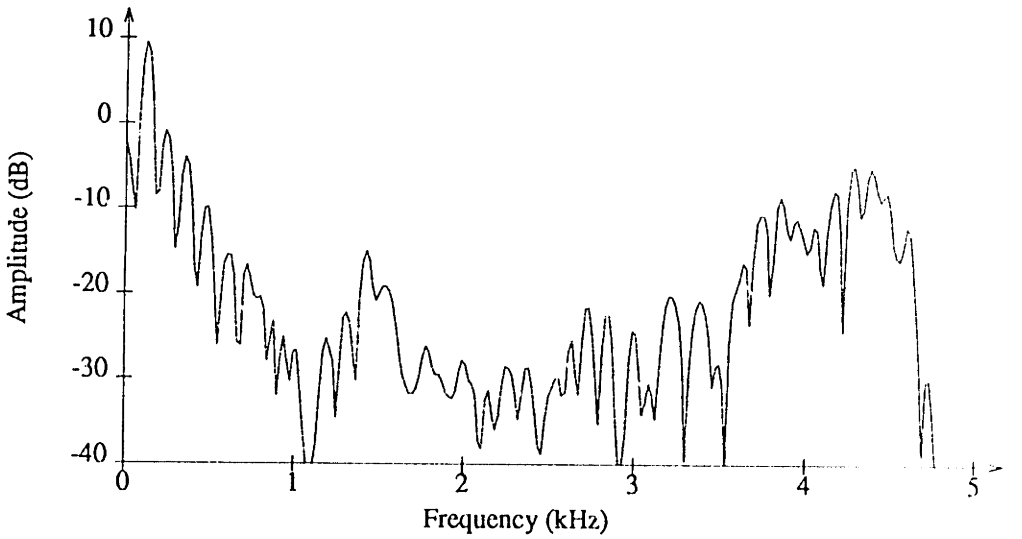

Figure 1.1: Spectrum of a /z/ Phoneme

envelope is required. Similarly, speech spectra in completely unvoiced regions such as a typical /t/ (Figure 1.3) indicate that a noise-like source with a flat spectral envelope is required. These observations indicate that periodic and noise sources with time-varying spectral shapes are required and help to explain the poor results obtained with the excitation models having time-invariant spectral shapes.

In excitation models having time-varying spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with time-varying spectral envelope shapes. Examples of such models include Fujimara [5], Makhoul et al. [17], and Kwon and Goldberg [15].

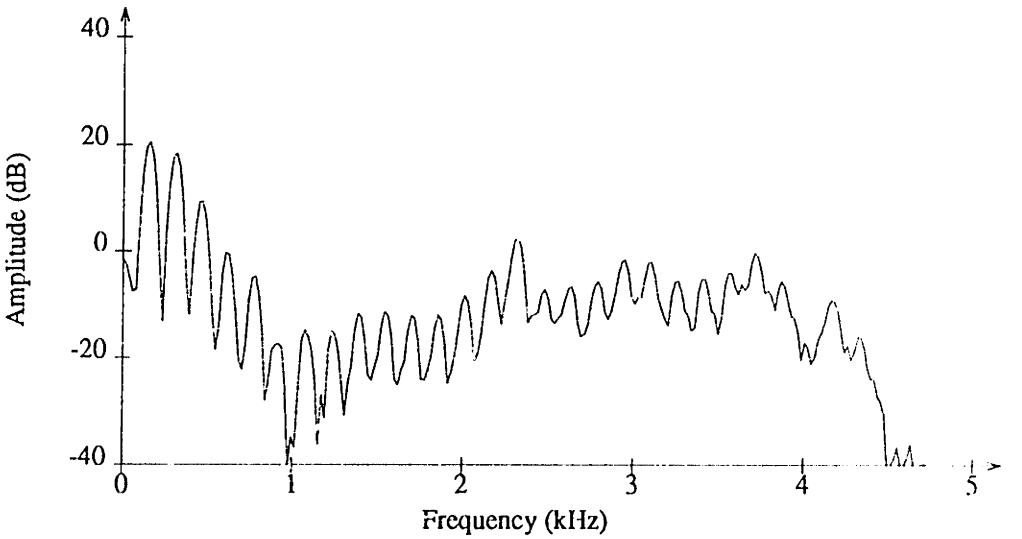

Figure 1.2: Spectrum of a /i/ Phoneme

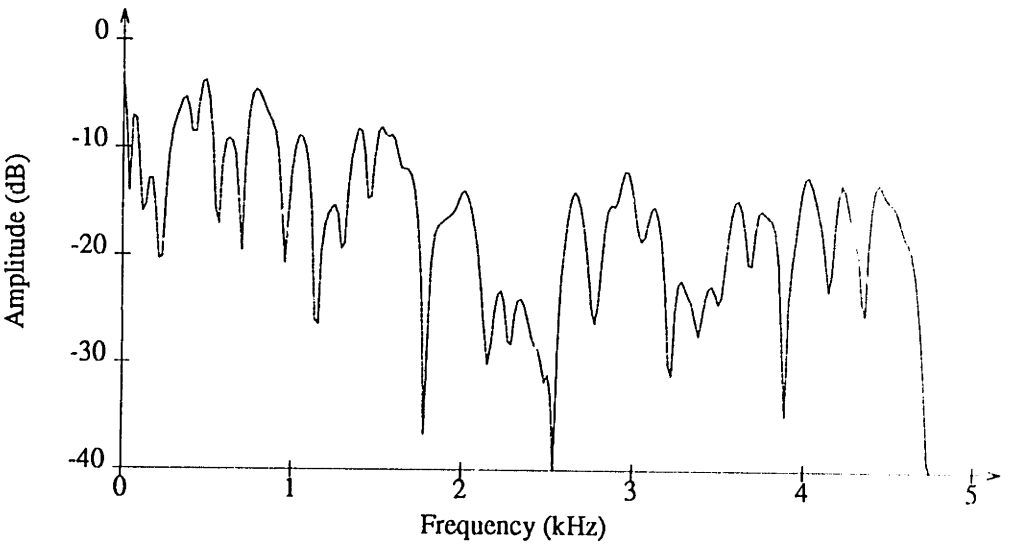

Figure 1.3: Spectrum of a /t/ Phoneme

In the excitation model propcsed by Fujimara, the excitation spectrum is divided into three fixed frequency bands. A separate cepstral analysis is performed for each frequency band and a voiced/unvoiced decision for each frequency band is made based on the height of the cepstrum peak as a measure of periodicity.

In the excitation model proposed by Makhoul et al., the excitation signal consists of the sum of a low-pass periodic source and a high-pass noise source. The low-pass periodic source was generated by filtering a white pulse source with a variable cut-off filter. Similarly, the high-pass noise source was generated by filtering a white noise source with a variable cut-off high-pass filter. The cut-off frequencies for the two filters are equal and are estimated by choosing the highest frequency at which the spectrum is periodic. Periodicity of the spectrum is determined by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level.

In a second excitation model implemented by Kwon and Goldberg, a pulse source is passed through a variable gain low-pass filter and added to itself, and a white noise source is passed through a variable gain high-pass filter and added to itself. The excitation signal is the sum of the resul-

tant pulse and noise sources with the relative amplitudes controlled by a voiced/unvoiced mixture ratio. The filter gains and voiced/unvoiced mixture ratio are estimated from the LPC residual signal with the constraint that the spectral envelope of the resultant excitation signal is flat.

In these excitation models, the voiced/unvoiced decisions or ratios control large contiguous regions of the spectrum. 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 argues 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 Multi-Pulse LPC model [1]. In this model, Linear Predictive Coding (LPC) is used to model the spectral envelope. The excitation signal consists of multiple pulses per pitch period instead of the standard LPC excitation consisting of one pulse per pitch period for voiced speech or a white noise sequence for unvoiced speech. With this model the original signal can be recovered by using one pulse per sample and setting the excitation signal to the LPC residual signal. However, coding the excitation signal for this case would require a prohibitively large number of bits. 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. This estimation procedure can be quite expensive computationally since the error criterion must be evaluated for all possible locations of each pulse introduced. 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.

1.3 Thesis Outline

In Chapter 2, our new Multi-Band Excitation Model for high quality modeling of clean and noisy speech is described. This model allows a large number of frequency bands to be declared voiced or unvoiced for improved modeling of mixed voicing and noisy speech. In Chapter 3, methods for estimating the parameters of this new model are developed. These methods estimate the excitation and spectral envelope parameters simultaneously so that the synthesized spectrum is closest in the least squares sense to the

original speech spectrum. This approach helps avoid the problem of the spectral envelope interfering with pitch period estimation and the pitch period interfering with the spectral envelope estimation. Chapter 4 discusses methods for synthesizing speech from these model parameters. In Chapter 5, we apply the MBE Model to the problem of bit-rate reduction for speech transmission and storage. Coding methods for the MBE Model parameters are presented which result in a high quality 8 kbps vocoder. High quality 8 kbps vocoders are of particular interest in applications such as mobile telephones. The 8 kbps MBE Vocoder is then evaluated using the results of informal listening as a measure of quality and Diagnostic Rhyme Tests (DRTs) as a measure of intelligibility. Finally, Chapter 6 discusses additional potential applications and presents some directions for future research for additional quality improvement and bit-rate reduction.

The objective of this thesis was to develop a better speech model for speech segments containing mixed voicing and for speech corrupted by noise. These speech segments tend to be degraded by systems using existing speech models. These degradations take the form of "buzziness" in the synthesized speech and a severe decrease in DRT scores for noisy speech. This objective was met through development of the Multi-Band Excita-

tion Model which allows the spectrum to be divided into many frequency bands, each of which may be declared voiced or unvoiced. When applied to the problem of bit-rate reduction, the MBE Model provided both quality and intelligibility improvements over a more conventional Single Band Excitation (SBE) Vocoder (1 V/UV bit per frame). In informal listening, the MBE Vocoder didn't have the "buzziness" present in the coded speech synthesized by the SBE Vocoder. An 8 kbps speech coding system was developed based on the MBE Model that provided a 12 point average DRT score improvement over the SBE Vocoder for speech corrupted by additive white noise. In addition, the average DRT score of the 8kbps MBE Vocoder was only about 5 points below the average DRT score of the uncoded noisy speech.

Chapter 2

Multi-Band Spectral Excitation Speech Model

2.1 Introduction

In Chapter 1, the need for a new speech model capable of overcoming the shortcomings of simple speech models for mixed voicing or in voiced regions of noisy speech was discussed. In the following section, our new Multi-Band Excitation Model is described for high quality modeling of clean and noisy speech.

2.2 New 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 10ms - 40ms. The windowed speech segment $s_w(n)$ is defined by

$$s_w(n) = w(n)s(n) \tag{2.1}$$

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 $|E_w(\omega)|$.

$$\hat{S}_w(\omega) = H_w(\omega)\,|E_w(\omega)| \tag{2.2}$$

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 [19], cepstral coefficients [25], formant frequencies and bandwidths [29], or samples of the original speech spectrum [3]. The representational 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 Figure 2.1(a) is shown in Figure 2.1(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

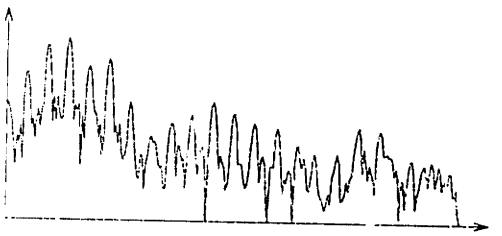

Fig. 2.1(a) - Original Spectrum

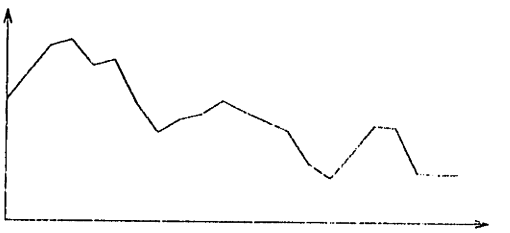

Fig. 2.1(b) - Spectral Envelope

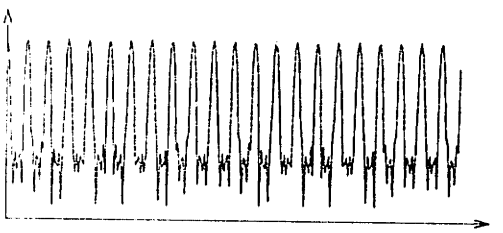

Fig. 2.1(c) - Periodic Spectrum

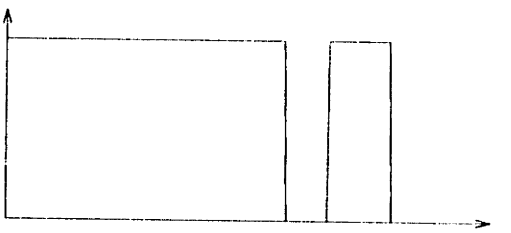

Fig. 2.1(d) - V/UV Information

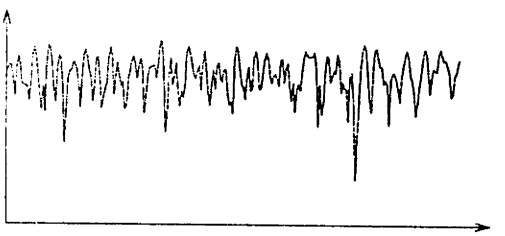

Fig. 2.1(e) - Noise Spectrum

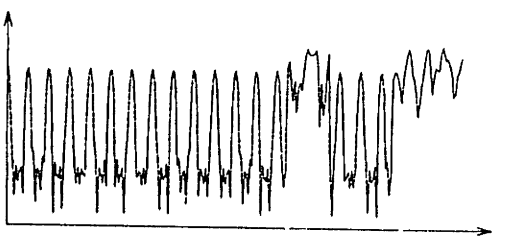

Fig. 2.1(f) - Excitation Spectrum

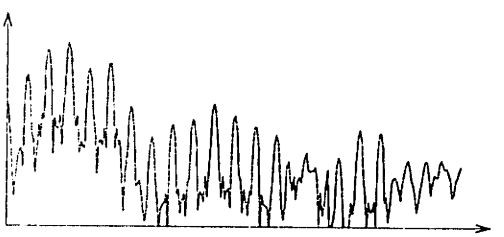

Fig. 2.1(g) - Synthetic Spectrum

Figure 2.1: Multi-Band Excitation Model - Noisy Speech

models in that the spectrum is divided into a large number of frequency bands (typically twenty 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 Multi-Band 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 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 $\omega_0 = .045\pi$ is shown in Figure 2.1(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 Multi-Band 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 [21] where noisy regions are synthesized from the coded phase requiring around 4 or 5 bits per harmonic. In addition, when 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 Figure 2.1(d) with a high value corresponding to a voiced decision. An example of a typical random noise spectrum used is shown in Figure 2.1(e). The excitation spectrum $|E_w(\omega)|$ derived from $|S_w(\omega)|$ in Figure 2.1(a) using the above procedure is shown in Figure 2.1(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 $\left|\hat{S}_w(\omega)\right|$ obtained by multiplying $|E_w(\omega)|$ in Figure 2.1(f) by $|H_w(\omega)|$ in Figure 2.1(b) is shown in Figure 2.1(g).

Additional examples of voiced, unvoiced, and mixed voicing segments of clean speech are shown in Figures 2.2 - 2.4. For voiced speech segments

(Figure 2.2), most of the spectrum is declared voiced. For unvoiced speech segments (Figure 2.3), most of the spectrum is declared unvoiced. For speech segments containing mixed voicing (Figure 2.4), regions containing periodic energy (harmonics of the fundamental frequency) are marked voiced and regions containing noise-like energy are marked unvoiced.

Based on the examples of Figures 2.1 - 2.4, it can be seen that some regions of the speech spectrum are dominated by harmonics of the fundamental frequency while others are dominated by noise-like energy depending on noise and speech production conditions. To account for this observed behavior, frequency bands with widths as small as the fundamental frequency should be individually declared voiced or unvoiced. This was the motivation for the Multi-Band Excitation Model.

It is possible [9] to synthesize high quality speech from the synthetic speech spectrum $\left|\hat{S}_w(\omega)\right|$. To use the above model for the purpose of developing a real time mid-rate speech coding system, however, it is desirable to introduce one additional set of parameters in our model. Specifically, the algorithm [8] that we have developed to synthesize speech from $\left|\hat{S}_w(\omega)\right|$ is an iterative procedure that estimates the phase of $\hat{S}_w(\omega)$ from $\left|\hat{S}_w(\omega)\right|$ and then synthesizes speech from $\left|\hat{S}_w(\omega)\right|$ and the estimated phase of $\hat{S}_w(\omega)$. This

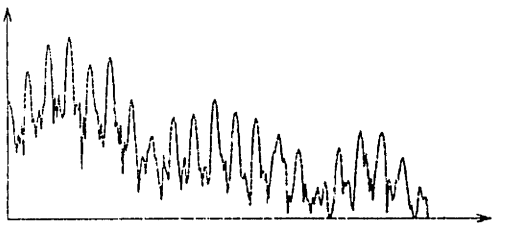

Fig. 2.2(a) - Original Spectrum

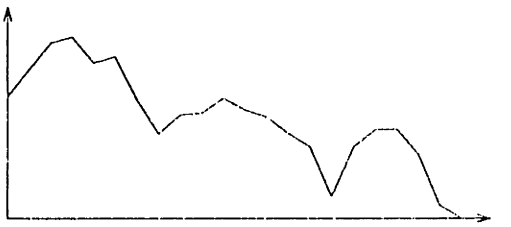

Fig. 2.2(b) - Spectral Envelope

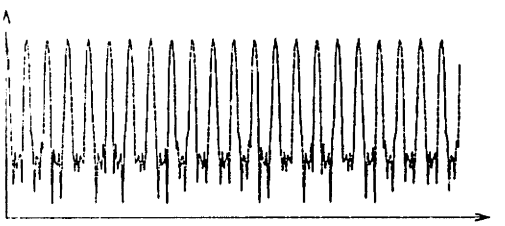

Fig. 2.2(c) - Periodic Spectrum

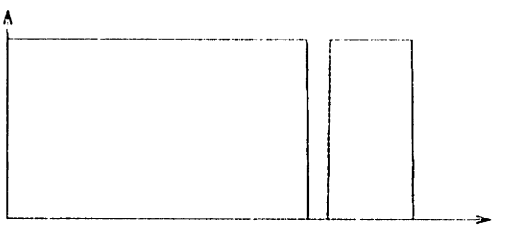

Fig. 2.2(d) - V/UV Information

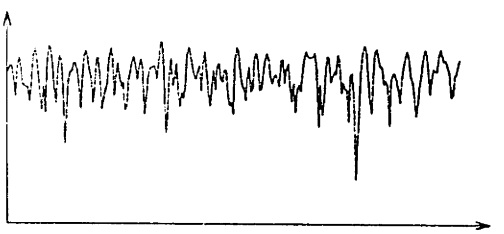

Fig. 2.2(e) - Noise Spectrum

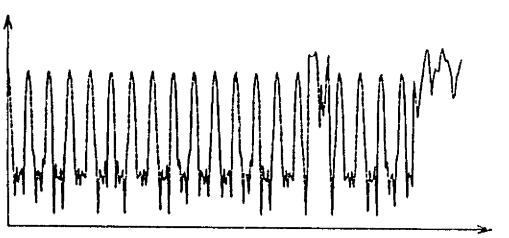

Fig. 2.2(f) - Excitation Spectrum

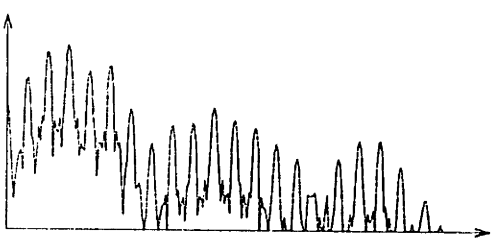

Fig. 2.2(g) - Synthetic Spectrum

Figure 2.2: Multi-Band Excitation Model - Voiced Speech

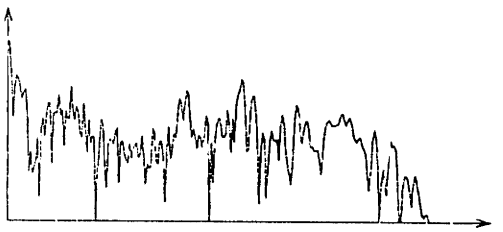

Fig. 2.3(a) - Original Spectrum

Fig. 2.3(b) - Spectral Envelope

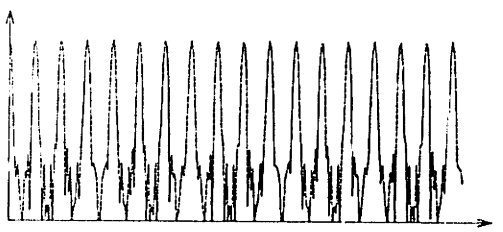

Fig. 2.3(c) - Periodic Spectrum

Fig. 2.3(d) - V/UV Information

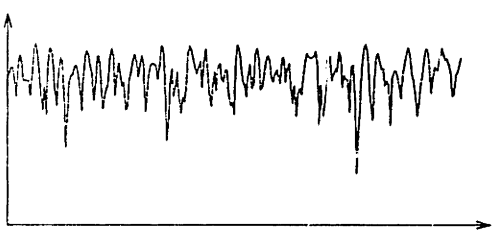

Fig. 2.3(e) - Noise Spectrum

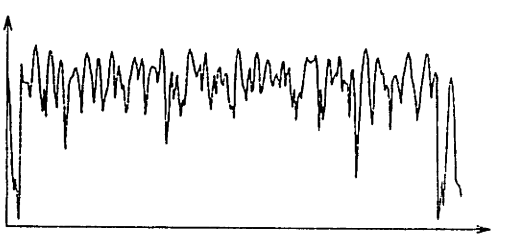

Fig. 2.3(f) - Excitation Spectrum

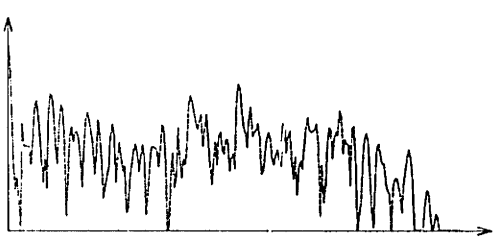

Fig. 2.3(g) - Synthetic Spectrum

Figure 2.3: Multi-Band Exc'tation Model - Unvoiced Speech

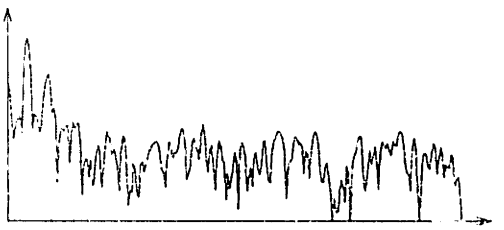

Fig. 2.4(a) - Original Spectrum

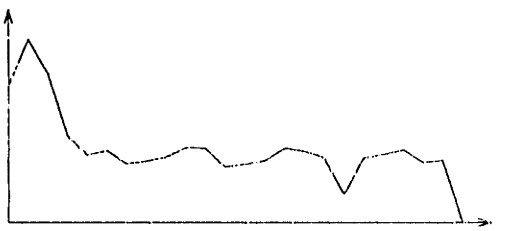

Fig. 2.4(b) - Spectral Envelope

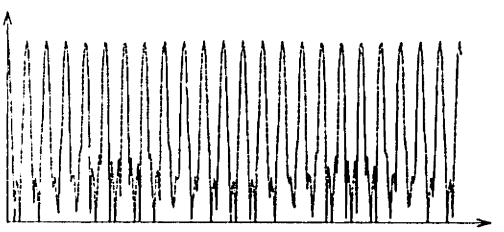

Fig. 2.4(c) - Periodic Spectrum

Fig. 2.4(d) - V/UV Information

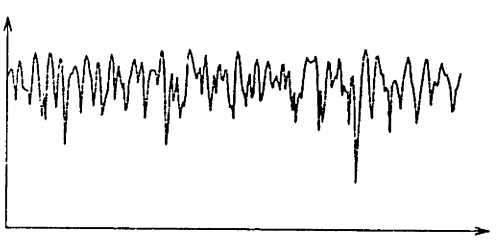

Fig. 2.4(e) - Noise Spectrum

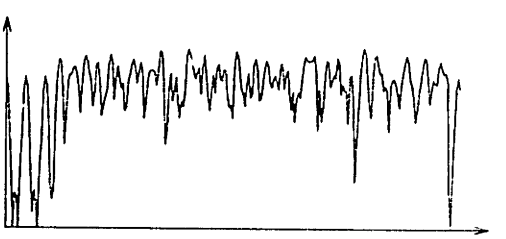

Fig. 2.4(f) - Excitation Spectrum

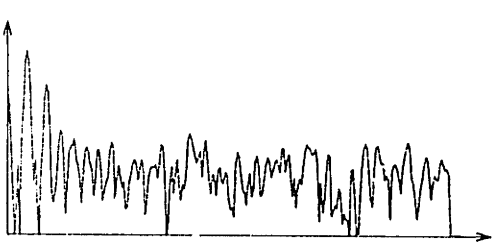

Fig. 2.4(g) - Synthetic Spectrum

Figure 2.4: Multi-Band Excitation Model - Mixed Voicing

algorithm requires a delay of more than one second and a fairly accurate representation of $\left|\hat{S}_w(\omega)\right|$. In applications such as time scale modification of speech where these limitations are not serious and determining the desired phase of $\hat{S}_w(\omega)$ is not easy, the algorithm that synthesizes speech from $\left|\hat{S}_w(\omega)\right|$ has been successfully applied. In applications such as real time speech coding, however, a delay of more than one second may not be acceptable and furthermore, the desired phase of $\hat{S}_w(\omega)$ can be determined straightforwardly. Due to the above considerations, we introduce an additional set of model parameters, namely, the phase of each harmonic declared voiced. We have chosen to include the phase in the samples of the spectral envelope A_m rather than the excitation spectrum $|E_w(\omega)|$ for later notational convenience.

The sets of 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. From these sets of parameters, speech can be synthesized with little delay and significant computational savings relative to synthesizing speech from $\left|\hat{S}_w(\omega)\right|$ alone. The synthesis

of speech from these model parameters is discussed in Chapter 4.

Chapter 3

Speech Analysis

3.1 Introduction

In Chapter 2, the Multi-Band Excitation Speech Model was introduced. The parameters of our model are the spectral envelope, the fundamental frequency, V/UV information for each harmonic, and the phase of each harmonic declared voiced. To obtain high quality reproduction of both clean and noisy speech, accurate and robust methods for estimating these parameters must be developed. In the next section, existing methods for estimating the spectral envelope and fundamental frequency are discussed. The inadequacies of these existing techniques led to the development of an

integrated method (Section 3.3) for estimating the model parameters so that the difference between the synthetic spectrum and the original spectrum is minimized. Obtaining an initial fundamental frequency using this method can be quite expensive computationally. An alternative formulation in Section 3.4 is used to substantially reduce the computation required to obtain the initial fundamental frequency estimate to the order of an autocorrelation pitch detection method.

In Section 3.5, we calculate the fundamental frequency bias associated with minimizing the least-squares error criterion for a periodic signal in noise. We then normalize the error criterion by the calculated bias to produce an unbiased error criterion. This unbiased error criterion significantly improves the system performance for noisy speech.

In Section 3.6, the required pitch period (or fundamental frequency) accuracy is determined for accurate estimation of the voiced/unvoiced information in the Multi-Band Excitation Model. An efficient procedure for obtaining this accuracy based on the earlier sections of this chapter is then described.

Finally, in Section 3.7, a flowchart of the complete analysis algorithm is presented and discussed.

3.2 Background

In previous approaches, 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.

Previous approaches to spectral envelope estimation include Linear Prediction [19] (All-Pole Modeling), windowing the cepstrum [25] (smoothing the log magnitude spectrum), and windowing the autocorrelation function [2] (smoothing the magnitude squared spectrum). In these approaches, the pitch period often interferes with the spectral envelope estimation procedure. For example, for speech frames with short pitch periods, widely separated harmonics in the spectrum tend to cause pole locations and bandwidths to be poorly estimated in the Linear Prediction method. Methods that window the cepstrum or autocorrelation function obtain a poor envelope estimate for short pitch periods due to interference of the peak at the pitch period with the spectral envelope information present in the low time

portions of these signals.

Previous approaches to pitch period estimation include the Gold-Rabiner parallel processing method [6], choosing the minimum of the average magnitude difference function (AMDF) [30], choosing the peak of the autocorrelation of the Linear Prediction residual signal (SIFT) [18], choosing the peak of the cepstrum [24], and choosing the peak of the autocorrelation function [27]. In these approaches, the spectral envelope often interferes with the pitch period estimation procedure. For example, methods that choose the peak of the cepstrum or autocorrelation function often obtain a poor pitch period estimate for short pitch periods due to interference of the spectral envelope information present in the low-time portions of these signals with the pitch period peak. Ross et al. [30] remark in their description of the AMDF pitch detector that the limiting factor on accuracy is the inability to completely separate the fine structure from the effects of the spectral envelope.

In one technique for compensating for the spectral envelope before pitch detection (SIFT), a spectral envelope estimate (produced by Linear Prediction) is divided out of the spectrum (inverse filtering). In this approach, the spectrum is "whitened" in an attempt to reduce the effects of the spec-

tral envelope on pitch period estimation. However, this technique boosts low energy regions of the spectrum which tend to be dominated by noise-like energy which reduces the periodic signal to noise ratio. Consequently, although performance is improved by reducing the effects of the spectral envelope, performance is degraded by the reduction in the periodic signal to noise ratio.

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 [27].

3.3 Estimation of Speech Model Parameters

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 $\left|\hat{S}_w(\omega)\right|$. 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:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) [|S_w(\omega)| - \left|\hat{S}_w(\omega)\right|]^2 d\omega \tag{3.1}$$

where

$$\left|\hat{S}_w(\omega)\right| = |H_w(\omega)|\,|E_w(\omega)| \tag{3.2}$$

and $G(\omega)$ is a frequency dependent weighting function. 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 samples $|A_m|$ of the spectral envelope $|H_w(\omega)|$. Other error criteria could also be used. For example, the error criterion:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) \left|S_w(\omega) - \hat{S}_w(\omega)\right|^2 d\omega \tag{3.3}$$

can be used to estimate both the magnitude and phase of the samples A_m of the spectral envelope. These envelope samples are the magnitudes (Equation (3.1)) or magnitudes and phases (Equation (3.3)) of the harmonics for frequency bands declared voiced. These samples of the envelope are

sufficient for synthesizing speech in the voiced frequency bands using the algorithm described in Chapter 4. For frequency bands declared unvoiced, one sample of the spectral envelope per harmonic of the estimated fundamental is also used. This sample is obtained by sampling a smoothed version of the original spectrum $|S_w(\omega)|$. During synthesis, additional samples of the spectral envelope in unvoiced regions are required. These are obtained by linearly interpolating between the estimated samples in the magnitude domain.

3.3.1 Estimation of Pitch Period and Spectral Envelope

The objective is to choose the pitch period and spectral envelope parameters to minimize the error of Equation (3.1). 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 around each harmonic of the fundamental frequency. For simplicity, we will model the spectral enve-

lope as constant in this interval with a value of A_m. This allows the error criterion of Equation (3.1) in the interval around the m^{th} harmonic to be written as:

$$\mathcal{E}_m = \frac{1}{2\pi} \int_{a_m}^{b_m} G(\omega)[|S_w(\omega)| - |A_m|\,|E_w(\omega)|]^2 d\omega \tag{3.4}$$

where the interval $[a_m, b_m]$ is an interval with a width of the fundamental frequency centered on the m^{th} harmonic of the fundamental. The error $\mathcal{E}_m$ is minimized at:

$$|A_m| = \frac{\int_{a_m}^{b_m} G(\omega)\,|S_w(\omega)|\,|E_w(\omega)|\,d\omega}{\int_{a_m}^{b_m} G(\omega)\,|E_w(\omega)|^2\,d\omega} \tag{3.5}$$

The corresponding estimate of A_m based on the error criterion of Equation (3.3) is:

$$A_m = \frac{\int_{a_m}^{b_m} G(\omega) S_w(\omega) E_w^*(\omega) d\omega}{\int_{a_m}^{b_m} G(\omega)\,|E_w(\omega)|^2\,d\omega} \tag{3.6}$$

At this point, we could obtain estimates of the envelope parameters A_m from Equation (3.5) or Equation (3.6) if we knew whether this frequency band was voiced/unvoiced. If the frequency band contains primarily periodic energy, there will be energy centered at the harmonic of the fundamental with the characteristic window frequency response shape. Consequently, if the periodic spectrum $|P_w(\omega)|$ is used as the excitation spectrum $|E_w(\omega)|$

in this band a good match will be obtained. If the frequency band contains primarily aperiodic energy, there will be no characteristic shape. Aperiodic energy in the frequency band is perhaps best characterized by a lack of a good match when the periodic spectrum $|P_w(\omega)|$ is used as the excitation spectrum. Thus, by using $|P_w(\omega)|$ as the excitation spectrum at this point, the voiced/unvoiced (periodic/aperiodic) decision can be made based on the modeling error in this frequency band. After making the voiced/unvoiced decision the appropriate spectral envelope parameter estimate can be selected. For a voiced frequency band, the following estimates are obtained by substituting $|P_w(\omega)|$ for $|E_w(\omega)|$ in Equation (3.5) and Equation (3.6)

$$|A_m| = \frac{\int_{a_m}^{b_m} G(\omega)\,|S_w(\omega)|\,|P_w(\omega)|\,d\omega}{\int_{a_m}^{b_m} G(\omega)\,|P_w(\omega)|^2\,d\omega} \tag{3.7}$$

$$A_m = \frac{\int_{a_m}^{b_m} G(\omega) S_w(\omega) P_w^*(\omega)\,d\omega}{\int_{a_m}^{b_m} G(\omega)\,|P_w(\omega)|^2\,d\omega} \tag{3.8}$$

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 an unvoiced frequency band, we model the excitation spectrum as idealized white noise (unity across the band) which yields the following

estimate:

$$|A_m| = \frac{\int_{a_m}^{b_m} G(\omega)\,|S_w(\omega)|\,d\omega}{\int_{a_m}^{b_m} G(\omega)d\omega} \tag{3.9}$$

This estimate reduces to the average of the original spectrum in the frequency band when the weighting function $G(\omega)$ is constant across the band. Since the unvoiced spectral envelope parameters are not used in pitch period estimation, they only need to be computed after the final pitch period estimate is determined.

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

$$\tilde{\mathcal{E}} \approx \sum_m \tilde{\mathcal{E}}_m \tag{3.10}$$

where $\tilde{\mathcal{E}}_m$ is $\mathcal{E}_m$ in Equation (3.4) evaluated with the $|A_m|$ of Equation (3.7). In this manner, the spectral envelope parameters which minimize the error $\mathcal{E}$ can be computed for a given pitch period P. This reduces the original multi-dimensional problem to the one-dimensional problem of finding the pitch period P that minimizes $\tilde{\mathcal{E}}$.

Experimentally, the error $\tilde{\mathcal{E}}$ 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 prac-

tice, the initial estimate is obtained by evaluating the error $\tilde{\mathcal{E}}$ 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 $G(\omega)$ is chosen to de-emphasize high frequencies.

If the pitch period of the original speech segment is 40 samples, the associated normalized fundamental frequency is .025. We define normalized frequency as the actual analog frequency divided by the sampling frequency so that the normalized fundamental frequency is just the reciprocal of the pitch period in samples. Integer multiples of the correct pitch period (80, 120, ...) will have fundamental frequencies at integer submultiples of the correct fundamental frequency (.0125, .00833, ...). Every n^{th} (second, third, ...) harmonic of the n^{th} submultiple (.0125, .00833, ...) of the correct pitch period will lie at the frequency of one of the harmonics of the correct fundamental frequency. For example, Figure 3.1 shows the periodic spectrum $|P_w(\omega)|$ for pitch periods of 40 and 80 samples. Since every second harmonic of a fundamental frequency of .0125 are at the harmonics of a fundamental frequency of .025, the error $\tilde{\mathcal{E}}$ will be comparable for the correct pitch period and its integer multiples. Consequently, once the pitch period which minimizes $\tilde{\mathcal{E}}$ 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 $\tilde{\mathcal{E}}$ 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 (See Section 3.6). 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 and a frequency weighting function $G(\omega)$ which includes high frequencies. 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 Section 3.6.

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 $\tilde{\mathcal{E}}$ 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

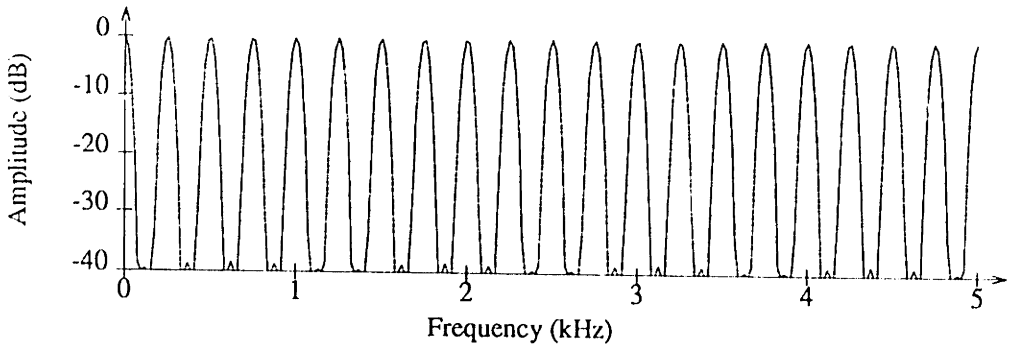

Figure 3.1(a) - Periodic Spectrum (Period=40)

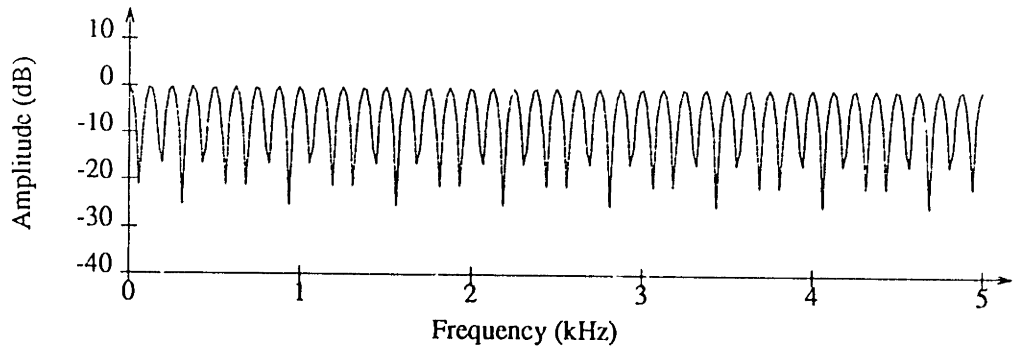

Figure 3.1(b) - Periodic Spectrum (Period=80)

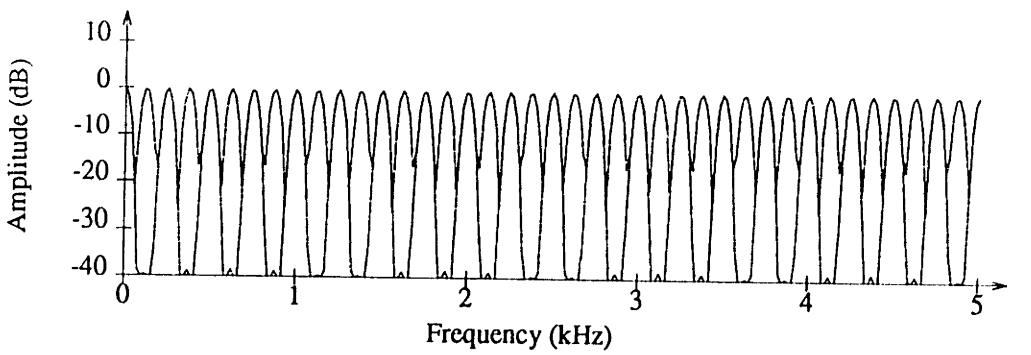

Figure 3.1(c) - Overlayed Periodic Spectra (Periods=40 and 80)

Figure 3.1: Pitch Period Doubling

expected error is smaller for longer pitch periods. This bias towards longer pitch periods is calculated in Section 3.5 and an unbiased error criterion is developed by multiplying the error $\tilde{\mathcal{E}}$ by a pitch period dependent correction factor. This correction factor is applied to the error $\tilde{\mathcal{E}}$ in Equation (3.10) prior to minimizing over the pitch period.

To illustrate our new approach, a specific example will be considered. In Figure 3.2(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 Figure 3.2(b). We use the property that the Fourier transform of a real sequence is conjugate symmetric [26] in order to compute these samples of the spectrum with a 256 point complex FFT. From the FFT, 255 complex points (samples of the Fourier Transform between normalized frequencies of 0 and .5) and 2 real points (at normalized frequencies of 0 and .5) are obtained. After the magnitude operation, there are 257 real samples of the spectrum between and including normalized frequencies of 0 and .5. Figure 3.2(c) shows the error $\tilde{\mathcal{E}}$ as a function of P with $G(\omega) = 1$ for frequencies less than 2 kHz and $G(\omega) = 0$ for frequencies greater than 2 kHz. The error E is smallest for $P = 85$, but since the error for the sub-

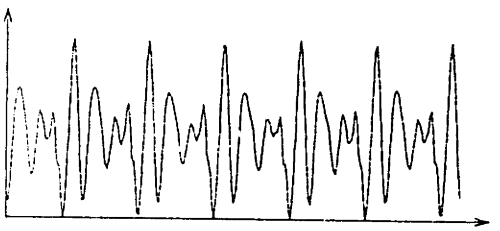

Fig. 3.2(a) - Speech Segment

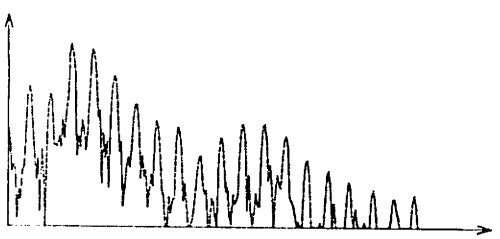

Fig. 3.2(b) - Original Spectrum

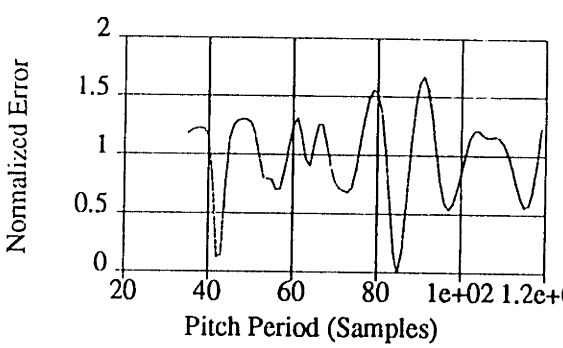

Fig. 3.2(c) - Error vs. Pitch Period

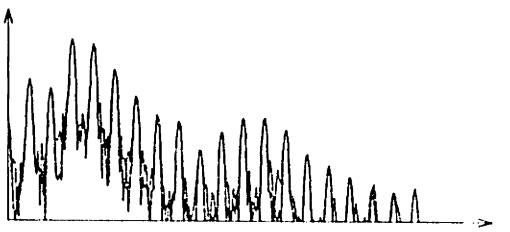

Fig. 3.2(d) - Original and Synthetic (non-integer P)

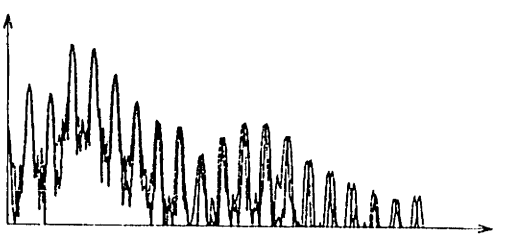

Fig. 3.2(e) - Original and Synthetic (Integer P)

Figure 3.2: Estimation of Model Parameters

multiple 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 non-integer pitch periods are desired, the error $\tilde{\mathcal{E}}$ is minimized around this initial estimate with $G(\omega)$ chosen to include the high frequencies. A typical weighting function $G(\omega)$ which we have used in practice is unity from 0 to 5 kHz. Figure 3.2(d) shows the original spectrum overlayed with the synthetic spectrum for the final pitch period estimate of 42.48 samples. For comparison, Figure 3.2(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.

Pitch track models can also be incorporated in this analysis system. For example, if the pitch period is not expected to change very much from one frame to the next, the error criterion can be biased to prefer pitch period estimates around the estimate for the previous frame. A pitch track model can also be used to reduce computation by constraining the possible pitch periods to a smaller region. In regions of speech where the normalized error

obtained by the best pitch period estimate is small, the periodic synthetic spectrum matches the original spectrum well and we can be relatively certain that the pitch period estimate in these regions is correct. The pitch track can then be extrapolated from such regions with our analysis method with the pitch track model incorporated.

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 periodic 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 60ms is typical). The al-

lowable 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 [22] are used to significantly reduce the computational requirements of this procedure.

3.3.2 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 m^{th} harmonic

$$\xi_m = \frac{\tilde{\mathcal{E}}_m}{\frac{1}{2\pi}\int_{a_m}^{b_m} G(\omega)\left|S_w(\omega)\right|^2 d\omega} \tag{3.11}$$

is below the threshold, this region of the spectrum matches that of a periodic spectrum well and the m^{th} harmonic is marked voiced. When ξ_m is above the threshold, this region of the spectrum is assumed to contain

noise-like energy. After the voiced/unvoiced decision is made for each frequency band the voiced or unvoiced spectral envelope parameter estimates are selected as appropriate.

In practice, these computations are performed by replacing integrals of continuous functions by summations of samples of these functions.

3.4 Alternative Formulation

By using a weighting function $G(\omega)$ which is one for all frequencies or by filtering the original signal, the error criterion of Equation (3.3) can be rewritten as:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| S_w(\omega) - \hat{S}_w(\omega) \right|^2 d\omega \tag{3.12}$$

In Section 3.3, the synthetic transform $\hat{S}_w(\omega)$ is the product of a spectral envelope and a periodic spectrum. Equivalently, the synthetic transform can be written as the transform of a periodic signal:

$$\hat{S}_w(\omega) = \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \tag{3.13}$$

where M is the largest integer such that $M\omega_0$ is in the frequency band $[-\pi, \pi]$ and $W(\omega)$ is the Fourier transform of the window function:

$$W(\omega) = \sum_{n=-\infty}^{\infty} w(n) e^{-j\omega n} \tag{3.14}$$

Equation (3.13) can be written in vector notation as

$$\hat{S}_{\mathrm{w}}(\omega) = \mathbf{w}^T \mathbf{a} \tag{3.15}$$

where

$$\mathbf{w} = \begin{bmatrix} W(\omega + M\omega_0) \\ W(\omega + (M-1)\omega_0) \\ \cdot \\ \cdot \\ \cdot \\ W(\omega - M\omega_0) \end{bmatrix} \tag{3.16}$$

and

$$\mathbf{a} = \begin{bmatrix} A_{-M} \\ A_{-M+1} \\ \cdot \\ \cdot \\ \cdot \\ A_M \end{bmatrix} \tag{3.17}$$

In this notation, the error criterion of Equation (3.12) can be expressed as:

$$\mathcal{E} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(\omega)|^2 \, d\omega - \mathbf{b}^H\mathbf{a} - \mathbf{a}^H\mathbf{b} + \mathbf{a}^H R\mathbf{a} \tag{3.18}$$

where

$$R = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{w}^*\mathbf{w}^T d\omega \tag{3.19}$$

and

$$\mathbf{b} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{w}^* S_w(\omega) d\omega \tag{3.20}$$

With this formulation, for a given fundamental frequency ω_0, minimizing the error criterion of Equation (3.12) results in the harmonic amplitude estimates A_m being the solution to the following linear equation:

$$R\mathbf{a} = \mathbf{b} \tag{3.21}$$

Using these amplitude estimates reduces the error of Equation (3.18) to:

$$\mathcal{E} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(\omega)|^2 \, d\omega - \mathbf{a}^H R\mathbf{a} \tag{3.22}$$

which is equivalent to:

$$\mathcal{E} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(\omega)|^2 \, d\omega - \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\hat{S}_w(\omega)\right|^2 d\omega \tag{3.23}$$

It should be noted that the synthetic transform $\hat{S}_w(\omega)$ of Equation (3.23) has been optimized over the harmonic amplitudes A_m and is therefore con-

strained to be evaluated at the optimal harmonic amplitudes for any particular fundamental frequency. We wish to minimize this error over all possible fundamental frequencies. This is equivalent to maximizing the second term over the fundamental frequency, since the first term is independent of fundamental frequency. This second term can be expressed independent of the harmonic amplitude estimates by applying Equation (3.21):

$$\Psi = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\hat{S}_{\mathrm{w}}(\omega)\right|^2 d\omega = \mathbf{a}^H R\mathbf{a} = \mathbf{b}^H R^{-1}\mathbf{b} \tag{3.24}$$

The window frequency responses are orthonormal if

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{w}^*\mathbf{w}^T d\omega = R = I \tag{3.25}$$

where I is the identity matrix. In order for orthonormality to hold, the window must be normalized so that

$$\sum_{n=-\infty}^{\infty} |w(n)|^2 = 1 \tag{3.26}$$

The window frequency responses are approximately orthonormal when the sidelobes of the window are small and the fundamental frequency is larger than the width of the main lobe so that the main lobes of window frequency responses at adjacent harmonics don't interact. For approximately

orthonormal window frequency responses, we have $R^{-1} \approx I$ which yields:

$$\Psi \approx \mathbf{b}^H \mathbf{b} \tag{3.27}$$

This approximation allows Ψ to be expressed in the time domain as

$$\Psi \approx \sum_{k=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} w^2(n)s(n)w^2(k)s(k) \sum_{m=-M}^{M} e^{-j\omega_0 m(n-k)} \tag{3.28}$$

For $\omega_0 M = \pi$, this simplifies to

$$\Psi \approx P \sum_{k=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} w^2(n)s(n)w^2(n-kP)s(n-kP) = P \sum_{k=-\infty}^{\infty} \phi(kP) \tag{3.29}$$

where $\phi(m)$ is the autocorrelation function of $w^2(n)s(n)$:

$$\phi(m) = \sum_{n=-\infty}^{\infty} w^2(n)s(n)w^2(n-m)s(n-m) \tag{3.30}$$

Thus, maximizing Ψ is approximately equivalent to maximizing a function of the autocorrelation function of the signal multiplied by the square of the analysis window. 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 Ψ 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 Equation (3.29) are finite due to the finite length of the window $w(n)$. Although this is a pseudo maximum likelihood pitch estimation method as in Wise et al. [33], it differs in that it is a frequency domain formulation rather than a time domain formulation. One advantage of this formulation is that a non-rectangular analysis window is allowed. For a rectangular window, the result given by Equations (3.29) and (3.30) reduces to the result given in Wise et al. [33].

More accurate pitch period estimates can be efficiently obtained by maximizing

$$\Psi \approx P \sum_{k=-\infty}^{\infty} \phi(\lfloor kP \rfloor) \tag{3.31}$$

over non-integer pitch periods where $\lfloor x \rfloor$ is defined as the largest integer not greater than x. Higher accuracy is obtained in this method due to the contributions of the peaks at multiples of the pitch period in the autocorrelation function.

Figure 3.3 shows a comparison of error versus pitch period for two different computation methods for a segment of speech with a pitch period of approximately 85 samples (The pitch period was determined by hand). The first method computes the error using the frequency domain approach given by Equation (3.10). The second method computes the error using

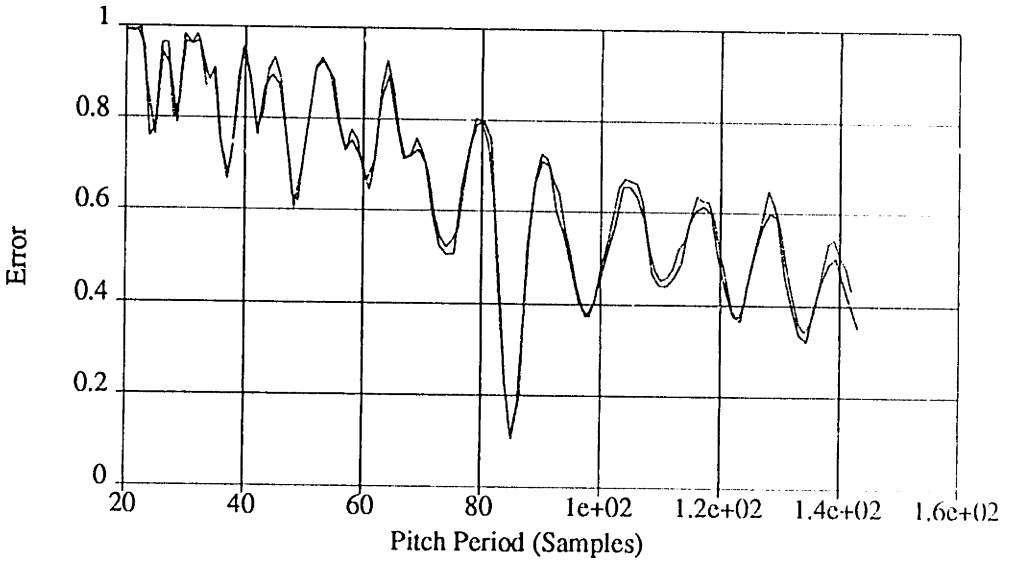

Figure 3.3: Comparison of Error Computation Methods

the autocorrelation approach described by the following equation:

$$\mathcal{E} \approx \sum_{n=-\infty}^{\infty} w^2(n)s^2(n) - P \sum_{k=-\infty}^{\infty} \phi(kP) \tag{3.32}$$

As can be seen from the figure, these two methods achieve approximately the same error curves. After estimating the pitch period using the autocorrelation domain approach, the spectral envelope parameters and the voiced/unvoiced parameters can be estimated as described in Section 3.3.1 and Section 3.3.2 for this specific pitch period.

3.5 Bias Correction

As discussed in Section 3.3 the expected value of the error of Equation (3.1) or Equation (3.3) is smaller for longer pitch periods since more free parameters are available for matching the original spectrum. This effect can be seen in Figure 3.3 as a general decrease in the error for larger pitch periods. To demonstrate this bias, we will calculate the expected value of the error $\mathcal{E}$ of Equation (3.12) for a periodic signal $p(n)$ in white noise $d(n)$:

$$s(n) = p(n) + d(n) \tag{3.33}$$

where

$$E[d(n)] = 0 \tag{3.34}$$

and

$$E[d(n)d(m)] = \sigma^2\delta(n-m) \tag{3.35}$$

The only constraints on the periodic signal $p(n)$ are that it has pitch period P so that

$$p(n+kP) = p(n) \tag{3.36}$$

where k is an integer.

Using Equation (3.23), the expected value of the error $\mathcal{E}$ of Equation

(3.12) evaluated at the optimal amplitude estimates for a given pitch period P is then:

$$E[\mathcal{E}] = E\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(\omega)|^2\, d\omega\right] - E\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\hat{S}_w(\omega)\right|^2 d\omega\right] \tag{3.37}$$

The first term in Equation (3.37) can be expressed in the time domain as:

$$E\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} |S_w(\omega)|^2\, d\omega\right] = E\left[\sum_{n=-\infty}^{\infty} s_w^2(n)\right] \tag{3.38}$$

For a window $w(n)$ normalized according to Equation (3.26) this reduces to:

$$E\left[\sum_{n=-\infty}^{\infty} s_w^2(n)\right] = \sigma^2 + \sum_{n=-\infty}^{\infty} w^2(n)p^2(n) \tag{3.39}$$

The second term in Equation (3.37) is the expected value of Ψ of Equation (3.24) which can be written as:

$$E[\Psi] \approx P \sum_{k=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} w^2(n)w^2(n-kP)E[s(n)s(n-kP)] \tag{3.40}$$

For $s(n)$ consisting of the sum of a periodic signal $p(n)$ of period P and white noise, Equation (3.40) reduces to:

$$E[\Psi] \approx \sigma^2 P \sum_{n=-\infty}^{\infty} w^4(n) + P \sum_{n=-\infty}^{\infty} w^2(n)p^2(n) \sum_{k=-\infty}^{\infty} w^2(n-kP) \tag{3.41}$$

For slowly changing window functions, the following approximation can be made:

$$P \sum_{k=-\infty}^{\infty} w^2(n-kP) \approx \sum_{n=-\infty}^{\infty} w^2(n) = 1 \tag{3.42}$$

This approximation reduces Equation (3.41) to:

$$E[\Psi] \approx \sigma^2 P \sum_{n=-\infty}^{\infty} w^4(n) + \sum_{n=-\infty}^{\infty} w^2(n)p^2(n) \tag{3.43}$$

By combining Equation (3.39) and Equation (3.43) a good approximation to the expected value of the error $\mathcal{E}$ of Equation (3.12) is obtained:

$$E[\mathcal{E}] \approx \sigma^2 \left(1 \quad P \sum_{n=-\infty}^{\infty} w^4(n)\right) \tag{3.44}$$

To determine the accuracy of the bias approximation given by Equation (3.44), error versus pitch period curves were computed for 100 different white noise segments and averaged together. This average error curve is shown in Figure 3.4 together with the bias approximation of Equation (3.44). As can be seen from the figure, the bias approximation is very close to the average error curve.

An unbiased error criterion is desired to prevent longer pitch periods from being consistently chosen over shorter pitch periods for noisy periodic signals. In addition, a normalized error criterion that is near zero for a purely periodic signal and is near one for a noise signal is desirable. The following error criterion is unbiased with respect to pitch period and is

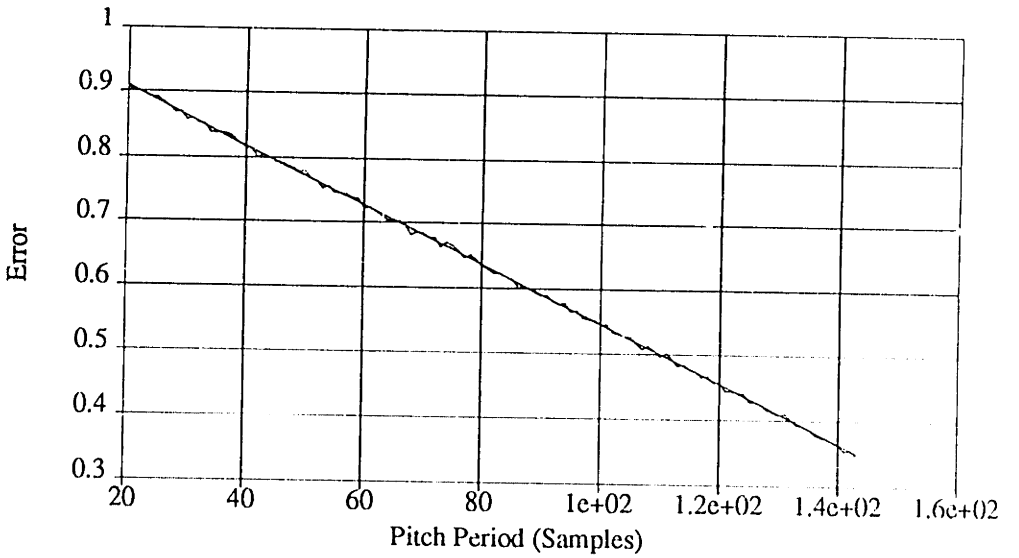

Figure 3.4: Average Error Versus Pitch Period

normalized appropriately:

$$\mathcal{E}_{UB} = \frac{\int_{-\pi}^{\pi} \left| S_w(\omega) - \hat{S}_w(\omega) \right|^2 d\omega}{\left(1 - P \sum_{n=-\infty}^{\infty} w^4(n)\right) \int_{-\pi}^{\pi} |S_w(\omega)|^2 \, d\omega} \tag{3.45}$$

It is important to note that the error criterion of Equation (3.45) is independent of the noise variance σ^2 so that estimation of the noise variance is not required. In addition, similar results can be seen to apply for colored noise by first applying a whitening filter to the original transform $S_w(\omega)$ and then removing it from the final result.

3.6 Required Pitch Period Accuracy

In Section 3.3.2 we described a method for estimating the voiced/unvoiced decisions for each harmonic by comparing the normalized error over each harmonic of the estimated fundamental to a threshold. The normalized error for each harmonic contains contributions due to the difference between the estimated harmonic frequency and the actual harmonic frequency as well as the contribution due to noise in the original signal. In this section, the required pitch period accuracy to prevent differences in the estimated and actual harmonic frequencies from dominating the normalized error is determined.

The normalized error between a harmonic of a perfectly periodic signal at normalized frequency f and a synthetic harmonic at estimated normalized frequency $\hat{f}$ depends on the difference Δf between the two frequencies. When the frequency difference Δf is near zero, the normalized error of Equation (3.11) is near zero. When the frequency difference Δf is large, the normalized error approaches one. Normalized error versus frequency difference is shown in Figure 3.5 for a 256 point square root triangular window. Figure 3.6 shows an expanded version of Figure 3.5 for small fre-

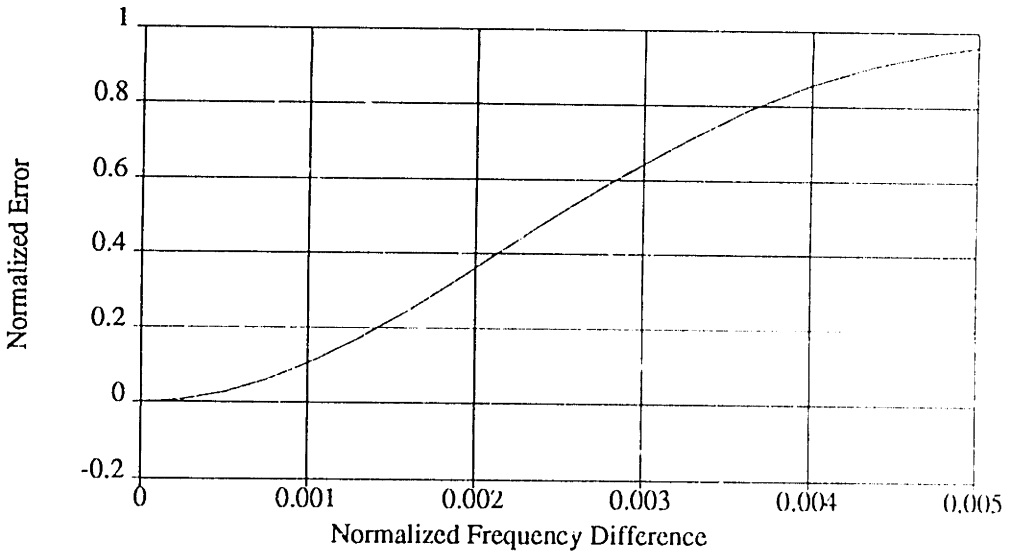

Figure 3.5: Normalized Error Versus Normalized Frequency Difference

quency differences. By listening to the synthesized speech, a good threshold for the Voiced/Unvoiced decision was determined to be approximately .2. Consequently, to prevent the normalized error from being dominated by an inaccurate pitch period estimate, by referring to Figure 3.6 we find that the maximum harmonic frequency difference should be smaller than about .001. The pitch period accuracy required to achieve a maximum harmonic frequency difference of .001 is shown in Figure 3.7.

The number of harmonics M of a normalized fundamental frequency f_0

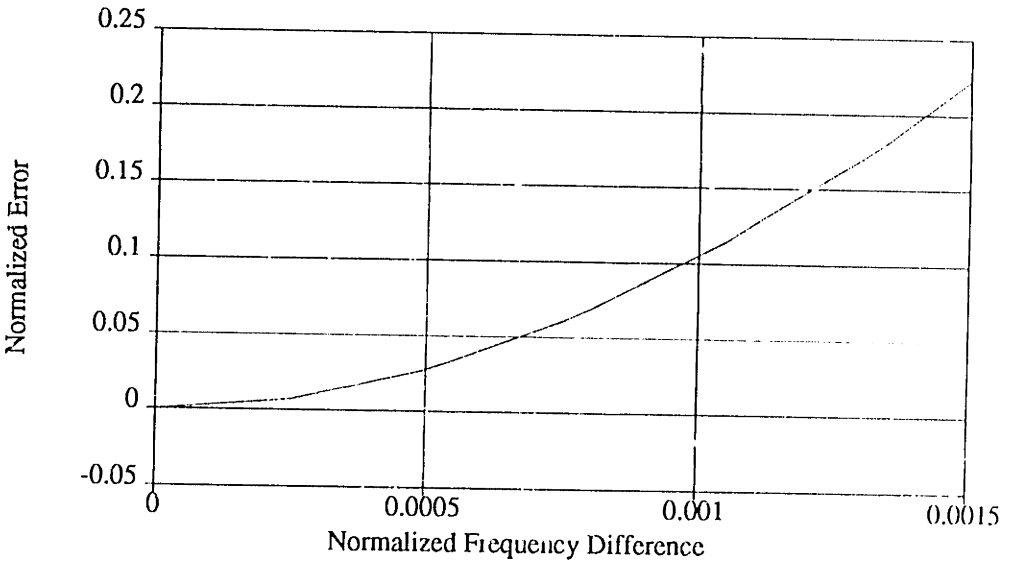

Figure 3.6: Normalized Error Versus Normalized Frequency Difference

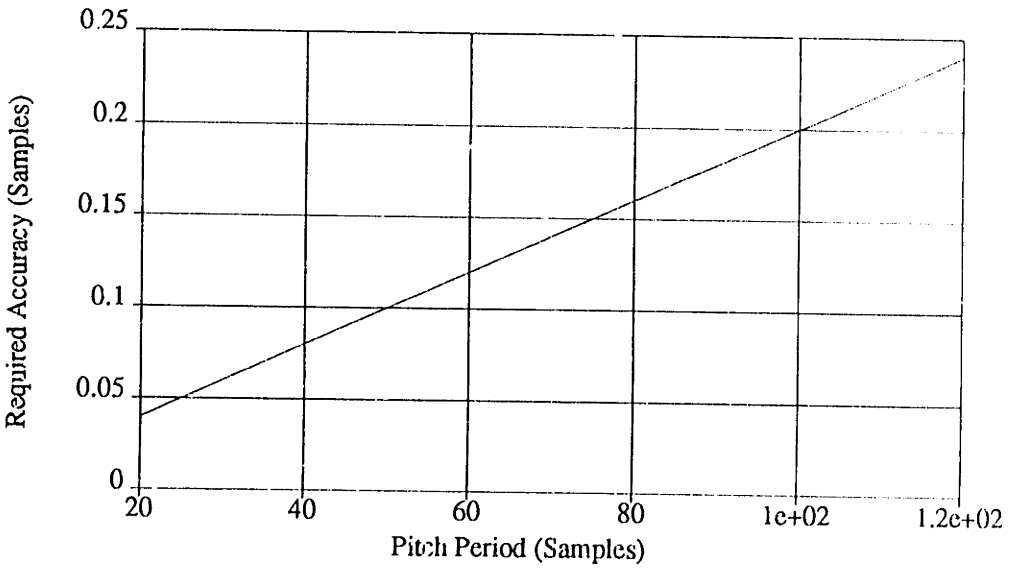

Figure 3.7: Required Pitch Period Accuracy

between normalized frequencies of zero and .5 is:

$$M = \left\lfloor \frac{1}{2f_0} \right\rfloor \tag{3.46}$$

So, the frequency deviation of the highest harmonic for an estimated fundamental of $\hat{f}_0$ and an actual fundamental of f_0 is:

$$\Delta f = \left\lfloor \frac{1}{2f_0} \right\rfloor (f_0 - \hat{f}_0) \tag{3.47}$$

In terms of pitch periods, Equation (3.47) becomes:

$$\Delta f \approx \frac{\Delta P}{2\hat{P}} \tag{3.48}$$

where ΔP is the difference between the actual and estimated pitch periods and the approximation comes from ignoring the floor function in Equation (3.47).

Figure 3.8 shows the smallest maximum harmonic frequency deviation attainable ($\Delta P = .5$) for a pitch detector which produces integer pitch period estimates. This figure clearly shows that the maximum harmonic frequency deviation significantly exceeds our desired value of .001 if only integer pitch periods are used. In addition, shorter pitch periods have significantly larger maximum harmonic frequency deviations than longer pitch periods.

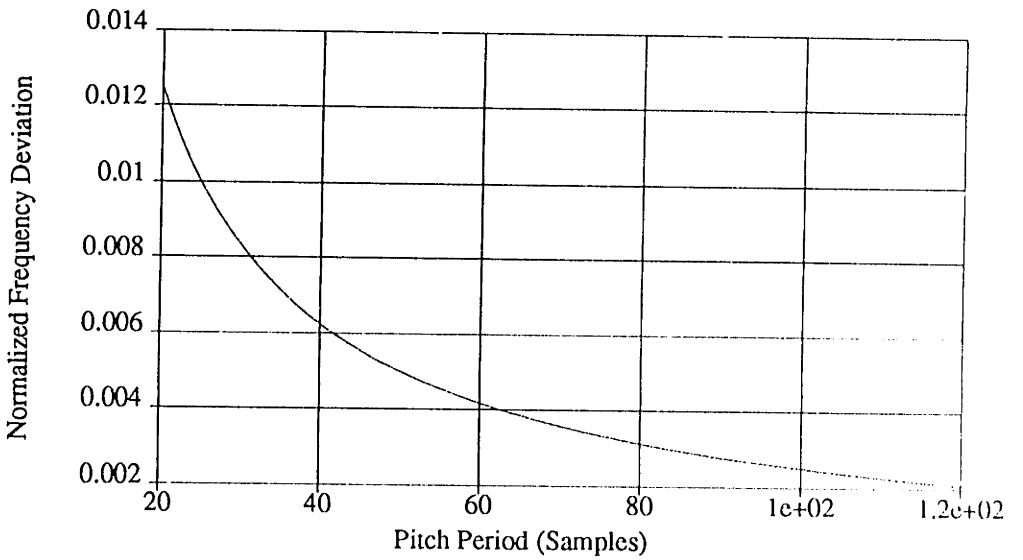

Figure 3.8: Smallest Maximum Harmonic Frequency Deviation for Integer Pitch Periods

In order to determine the accuracy of the autocorrelation domain method described in Section 3.4 and the frequency domain method described in Section 3.3.1, an experiment was conducted in which these techniques were used to estimate the pitch period of 6000 different synthesized periodic segments. The experiment consisted of generating 100 periodic segments for each of 60 different 2 sample intervals with center periods of 20 to 120 samples. The pitch periods of the segments were uniformly distributed in the 2 sample interval. The phases of the harmonics were random with a uniform distribution between $-\pi$ and π. The magnitudes of the harmonics

decreased linearly to zero at a frequency of half the sampling rate.

The maximum deviation and standard deviation of the pitch period estimates are shown in Figure 3.9 and Figure 3.10 for the autocorrelation domain and frequency domain methods. The corresponding maximum

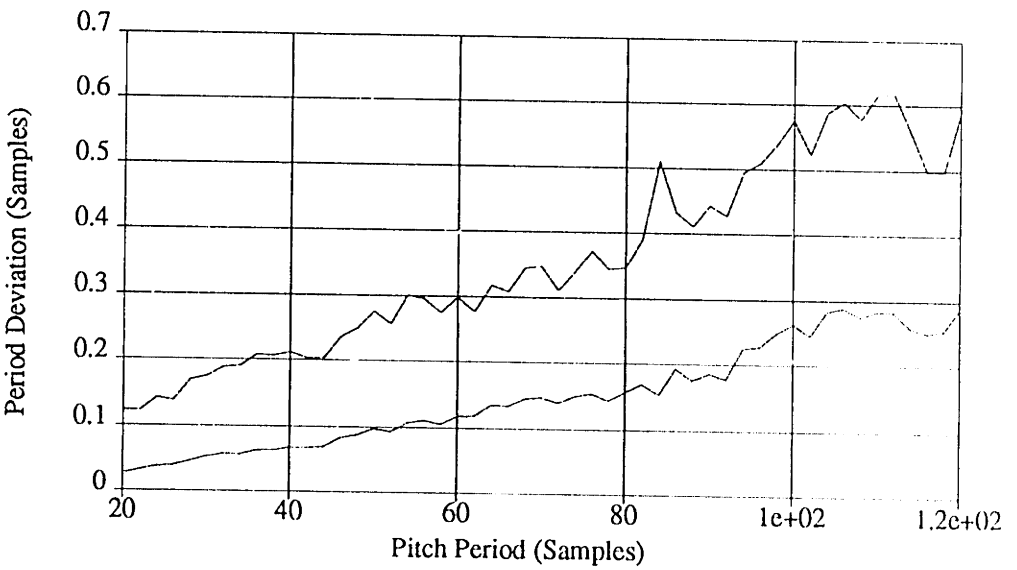

Figure 3.9: Pitch Period Deviation for Autocorrelation Domain Method

deviation and standard deviation of the frequency of the highest harmonic (in the normalized frequency range of 0 to .5) of the estimated fundamental are shown in Figure 3.11 and Figure 3.12 for the autocorrelation domain and frequency domain methods. These figures show that for this test, the frequency domain method provides pitch period estimates that are approximately 10 times more accurate than the autocorrelation method.

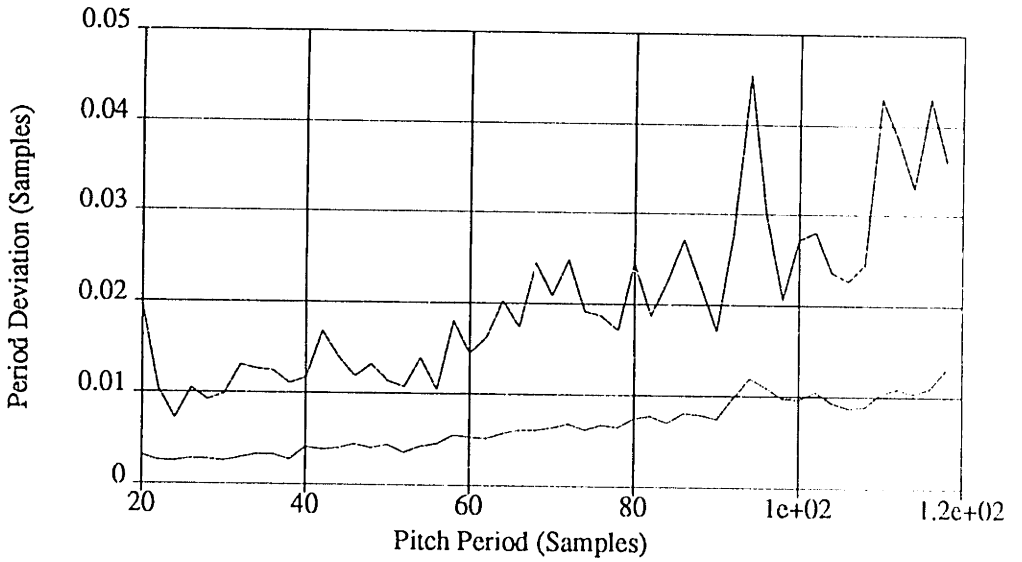

Figure 3.10: Pitch Period Deviation for Frequency Domain Method

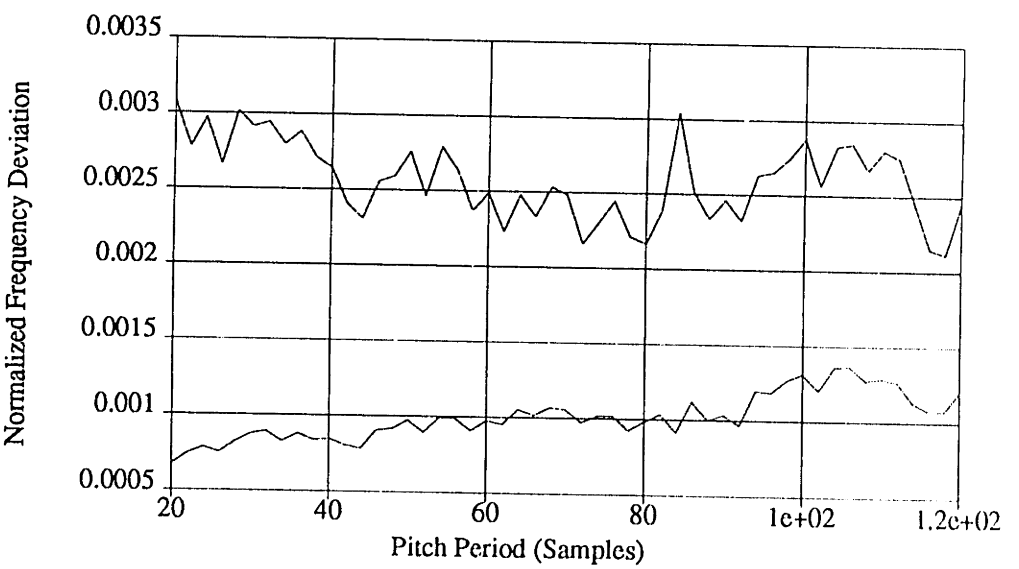

Figure 3.11: Frequency Deviation of Highest Harmonic for Autocorrelation Domain Method

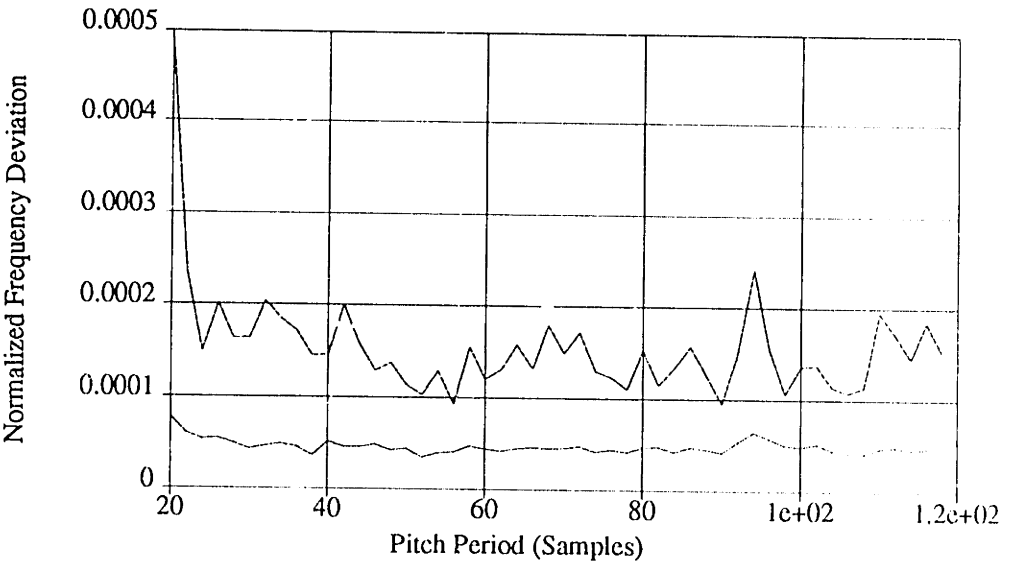

Figure 3.12: Frequency Deviation of Highest Harmonic for Frequency Domain Method

From Figure 3.11, it can be seen that the maximum harmonic frequency deviation for the autocorrelation method of approximately .003 is larger than our desired value of .001. The frequency domain method is capable of more than sufficient accuracy with a maximum harmonic frequency deviation near .0002. However, the autocorrelation method is significantly more efficient computationally due to the possibility of FFT implementation. Consequently, we use the computationally efficient autocorrelation domain method to obtain an initial pitch period estimate followed by the more accurate frequency domain method to refine the initial estimate.

3.7 Analysis Algorithm

The analysis algorithm that we use in practice consists of the following steps (See Figure 3.13):

1. Window a speech segment with the analysis window.

2. Compute the unbiased error criterion of Equation (3.45) vs. pitch period using the efficient autocorrelation domain approach described in Section 3.4. This error is typically computed for all integer pitch periods from 20 to 120 samples for a 10kHz sampling rate.

3. Use the dynamic programming approach described in Section 3.3.1 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 using the more accurate frequency domain pitch period estimation method described in Section 3.3.1.

5. Estimate the voiced and unvoiced spectral envelope parameters using the techniques described in Section 3.3.1.

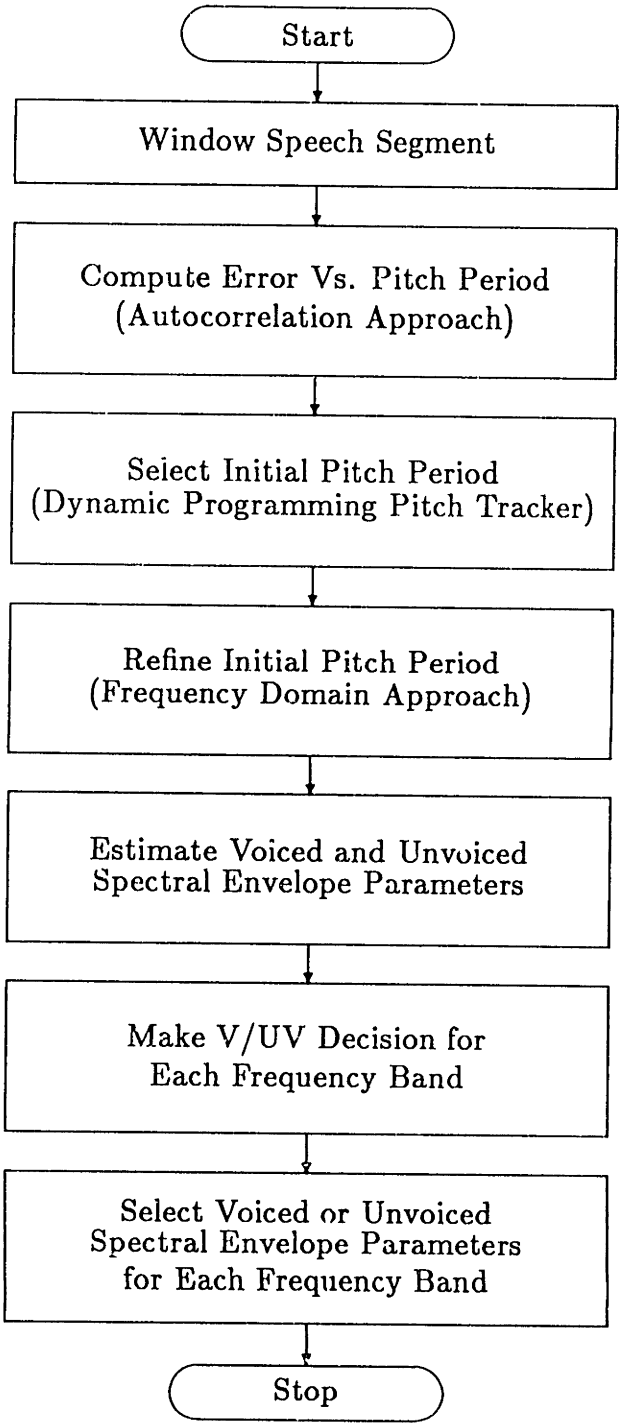

Figure 3.13: Analysis Algorithm Flowchart

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.

Chapter 4

Speech Synthesis

4.1 Introduction

In the previous two chapters, the Multi-Band Excitation Model parameters were described and methods to estimate these parameters were developed. In this chapter, 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. The following section discusses several applicable methods and selects one for generating the voiced portion of the synthesized speech and a second for generating the unvoiced portion of the synthesized speech. The details of our speech

synthesis algorithm are then presented in Section 4.3.

4.2 Background

Speech can be synthesized from the estimated model parameters using several different approaches. One approach is to generate a sequence of synthetic spectral magnitudes from the estimated model parameters. Then, algorithms for estimating a signal from this synthetic Short-Time Fourier Transform Magnitude (STFTM) are applied. In a second approach, a synthetic Short-Time Fourier Transform (STFT) is generated. Then, algorithms for estimating a signal from this synthetic STFT are applied. In a third approach, the synthetic speech signal is generated in the time domain from the speech model parameters.

A synthetic STFTM can be constructed from the Multi-Band Excitation model parameters by combining segments of a periodic spectrum in regions declared voiced with segments of a noise spectrum in regions declard unvoiced to generate the excitation spectrum. The noise spectrum segments are normalized to have an average magnitude per sample of unity. A densely sampled spectral envelope can be obtained by inter-

polating between the samples ($|A_m|$) of the spectral envelope. We have used a constant value set to $|A_m|$ in voiced regions and linear interpolation between adjacent samples ($|A_m|$) in unvoiced regions. The excitation spectrum is then multiplied by the densely sampled spectral envelope to generate the synthetic STFTM. Nawab has shown [23] that a signal can be exactly reconstructed from its STFTM under certain conditions. However, this algorithm requires the STFTM to be a valid STFTM (the STFTM of some signal). Due to the modeling and synthesis process, the synthetic STFTM is not guaranteed to be a valid STFTM. Consequently this algorithm cannot be successfully applied to this problem. Another algorithm, developed by Griffin and Lim [8] for estimating a signal from a modified STFTM has been successfully applied to this problem for the applications of analysis/synthesis and time-scale modification for both clean and noisy speech [9]. However, this algorithm is quite expensive computationally and requires a processing delay of approximately one second. This processing delay is unacceptable in most real-time speech bandwidth compression applications.

A synthetic STFT can be constructed from the Multi-Band Excitation model parameters by combining segments of a periodic transform in re-

gions declared voiced with segments of a noise transform in regions declard unvoiced. The noise tranform segments are normalized as in the previous paragraph and a densely sampled spectral envelope is generated. The phase of the samples in voiced regions is set to the phase of the spectral envelope samples A_m. 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. One problem with this approach is that the voiced portion of the synthesized signal is modeled as a periodic signal with constant fundamental over the entire frame. When small window shifts are used in the analysis/synthesis system, a fairly continuous fundamental frequency variation is allowed as observed in the STFTM of the original speech. However, when large window shifts are used (as is necessary to reduce the bit-rate for speech coding applications) the large potential change in fundamental frequency from one frame to the next causes time discontinuities in the harmonics of the fundamental in the STFTM.

A third approach to synthesizing speech invoives 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 fundamental 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.

4.3 Speech Synthesis Algorithm

A 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. A frequency domain (STFT) 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. Speech is then synthesized as the sum of the synthetic voiced signal and the synthetic unvoiced signal.

As discussed in the previous section, voiced speech can be synthesized

in the time domain as the sum of sinusoidal oscillators:

$$\hat{s}_v(t) = \sum_m A_m(t)\cos(\theta_m(t)) \tag{4.1}$$

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 $\omega_m(t)$.

$$\theta_m(t) = \int_0^t \omega_m(\xi)d\xi + \phi_0 \tag{4.2}$$

The frequency track $\omega_m(t)$ is linearly interpolated between the m^{th} harmonic of the current frame and that of the next frame as follows:

$$\omega_m(t) = m\omega_0(0)\frac{(S-t)}{S} + m\omega_0(S)\frac{t}{S} + \Delta\omega_m \tag{4.3}$$

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 m^{th} 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.

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.

The unvoiced speech has been generated by taking the STFT of a white noise sequence and zeroing out the frequency regions marked voiced. The samples in the unvoiced regions are then normalized to have the desired average magnitude specified by the spectral envelope parameters. The synthetic unvoiced speech can then be produced from this synthetic STFT using the weighted overlap-add method. It should be noted that this algorithm can synthesize the unvoiced portion of the synthetic speech signal on

a frame by frame basis for real-time synthesis.

4.4 Speech Synthesis System

A block diagram of our current speech synthesis system is shown in Figures 4.1 through 4.4. 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 (Figure 4.1). Voiced

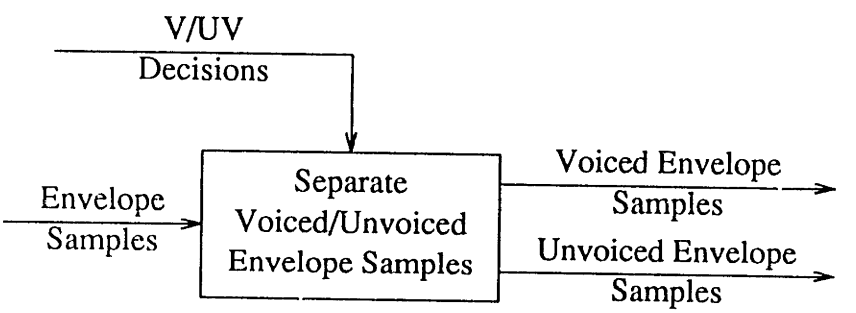

Figure 4.1: Separation of Envelope Samples

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 bank of sinusoidal oscillators running at the harmonics of the fundamental frequency (Figure 4.2). The amplitudes of the

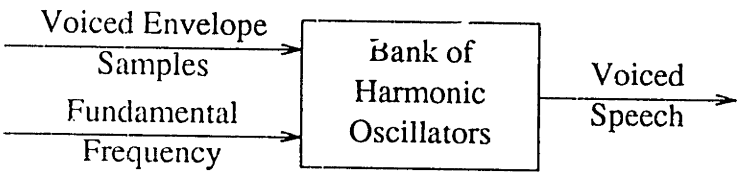

Figure 4.2: Voiced Speech Synthesis

oscillators are set to the magnitudes of the envelope samples with linear interpolation between frames. The phase tracks of the oscillators are adjusted to match the phases of the envelope samples.

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 (Figure 4.3). A sample of the spectral envelope is estimated in each frequency band by averaging together the magnitude of the FFT samples in that band. This spectal envelope is then replaced by the unvoiced spectral envelope generated from the unvoiced envelope samples. This unvoiced spectral envelope is obtained by linear interpolation between

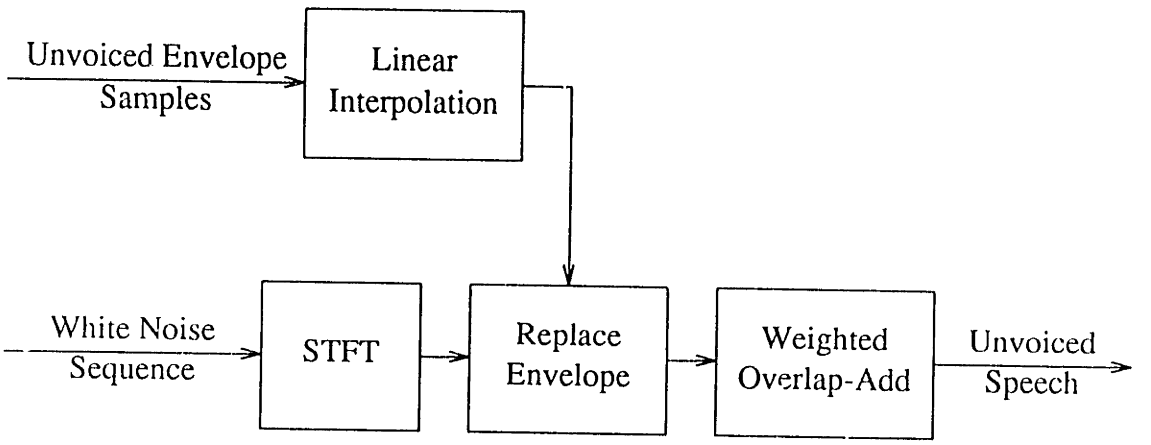

Figure 4.3: Unvoiced Speech Synthesis

the unvoiced envelope samples. These 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 (Figure 4.4).

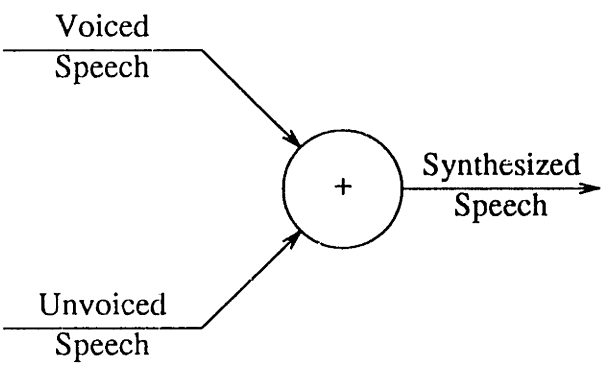

Figure 4.4: Speech Synthesis

Chapter 5

Application to the Development of a High Quality 8 kbps Speech Coding System

5.1 Introduction

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 Multi-Band Excitation Speech Analysis/Synthesis System for this problem, an 8 kbps speech coding system was developed. Since our primary goal is to demonstrate the high performance of the Multi-Band Excitation Model and the corresponding speech analysis methods, fairly conventional and simple parameter coding methods have been used to facilitate comparison with other systems. Even though simple coding methods have been used, the results are quite good.

The major innovation in the Multi-Band 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 Multi-Band Excitation Speech Coder operating at 8 kbps 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 Multi-Band Excitation Speech Coder except that one V/UV bit per frame is used instead of 12. Although this results in a somewhat smaller bit-rate for the more conventional coding system (7.45 kbps), 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. In addition, this avoids the problem of trying to optimally assign these 11 bits to coding the other parameters and the subsequent multitudes of DRT tests to evaluate all possible combinations.

5.2 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 kbps using a 50 Hz frame rate. This allows 160 bits per frame for coding of 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 5.1. As discussed in Chapter

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

Table 5.1: Bit Allocation per Frame

4, phase is not required for harmonics declared unvoiced. Consequently, bits assigned to phases declared unvoiced are reassigned to the magnitude. So, 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.

5.2.1 Coding of Harmonic Magnitudes

The harmonic magnitudes are coded using the same techniques employed by channel vocoders [11]. 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 m^{th} harmonic and the coded value of the previous harmonic (within the same frame) is determined by summing samples of the bit density curve of Figure 5.1 over the frequency interval occupied by the m^{th} harmonic. The available bits for coding the magnitude are then assigned to each harmonic in proportion to these sums. For example, Figure 5.2 shows the number of bits assigned to code each harmonic of a coded fundamental frequency of .01 (normalized frequency). The coded value of the fundamental is used so that the number of bits allocated to each harmonic can be determined at the receiver from the transmitted coded fundamental frequency. The number of bits assigned to each harmonic in Figure 5.2 is, in general, non-integer. For a non-integer number of bits, the integer part is taken and the fractional part is added to the bits

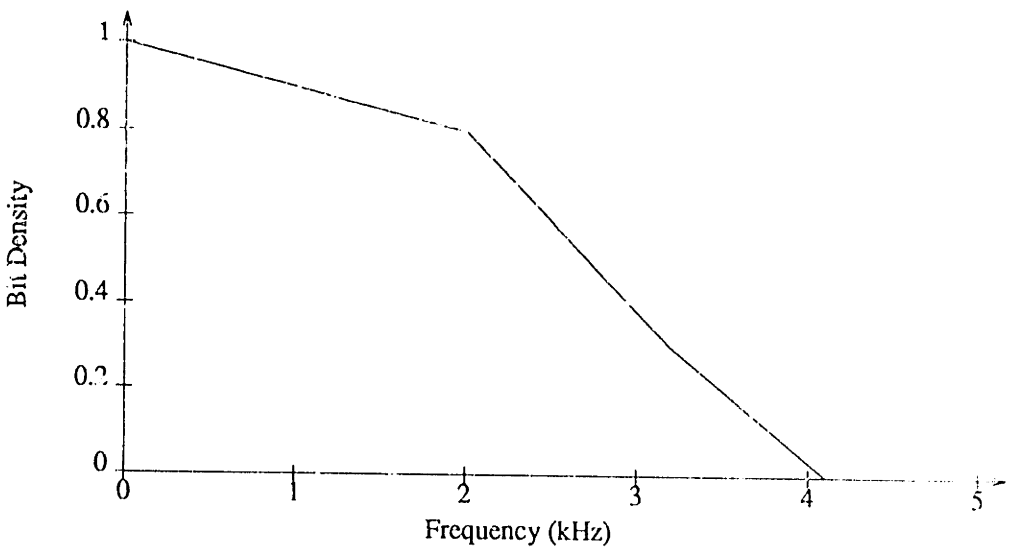

Figure 5.1: Magnitude Bit Density Curve

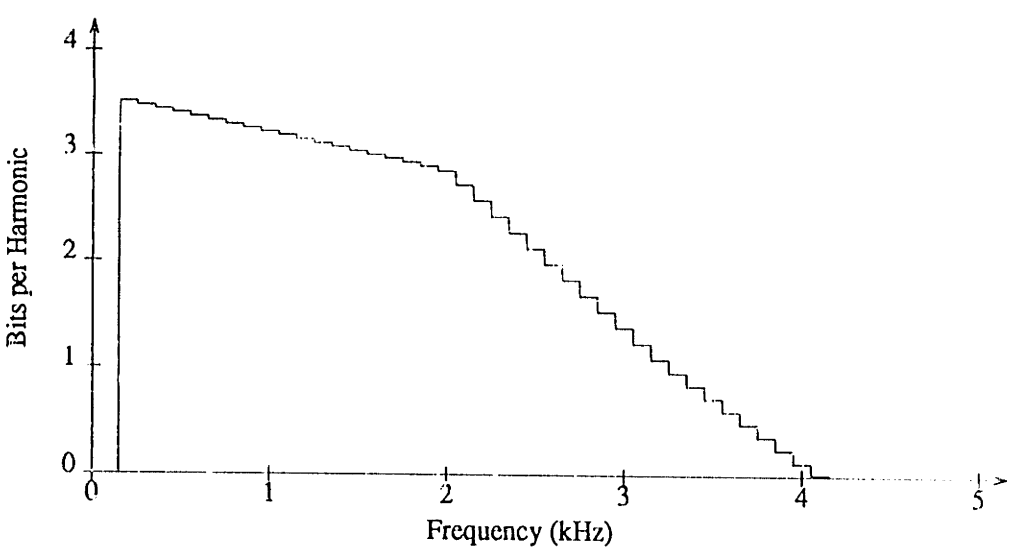

Figure 5.2: Magnitude Bits for Each Harmonic

assigned to the next harmonic. The quantization step size depends on the number of bits assigned and is listed in Table 5.2.

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

Table 5.2: Quantization Step Sizes

5.2.2 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. Consequently, 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. An example of phase coding is shown in Figures 5.3 through 5.6 for a frame of speech in which all frequency bands were declared voiced. 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 difference between the predicted and estimated phase can be coded using uniform quantization to code the first N harmonics between $-\pi$ and π. For the 8 kbps system, the phases of the first 12 harmonics (starting

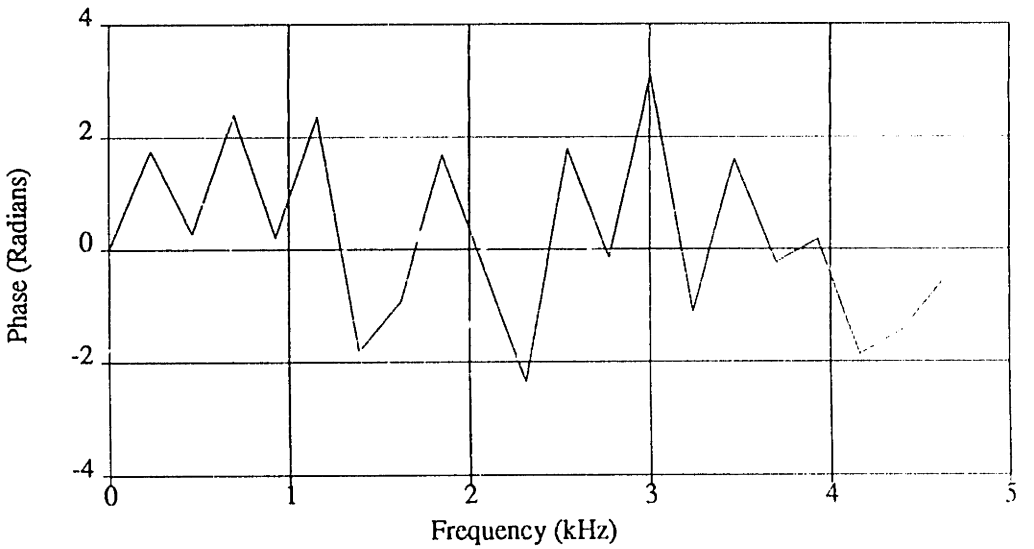

Figure 5.3: Estimated Harmonic Phases

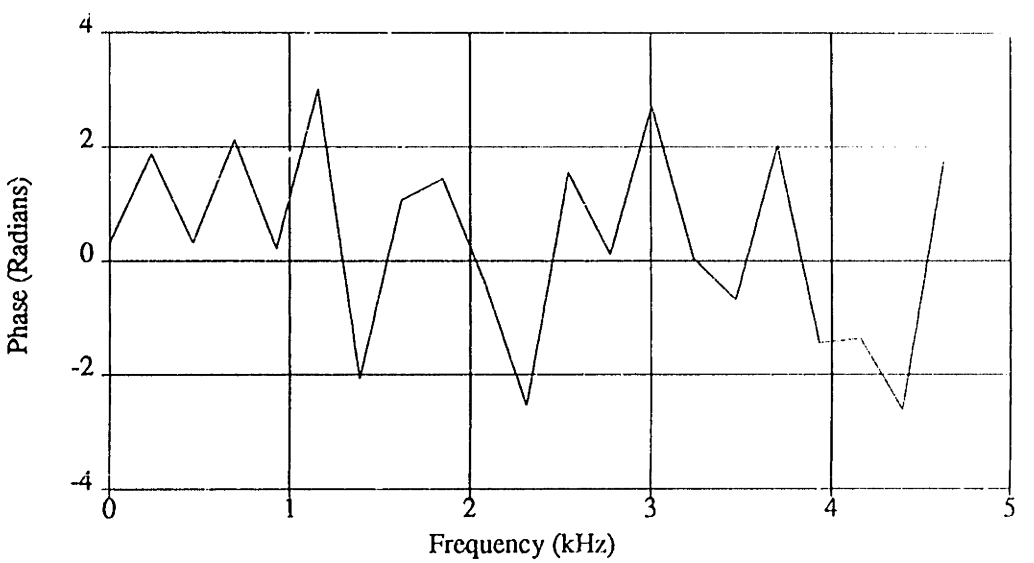

Figure 5.4: Predicted Harmonic Phases

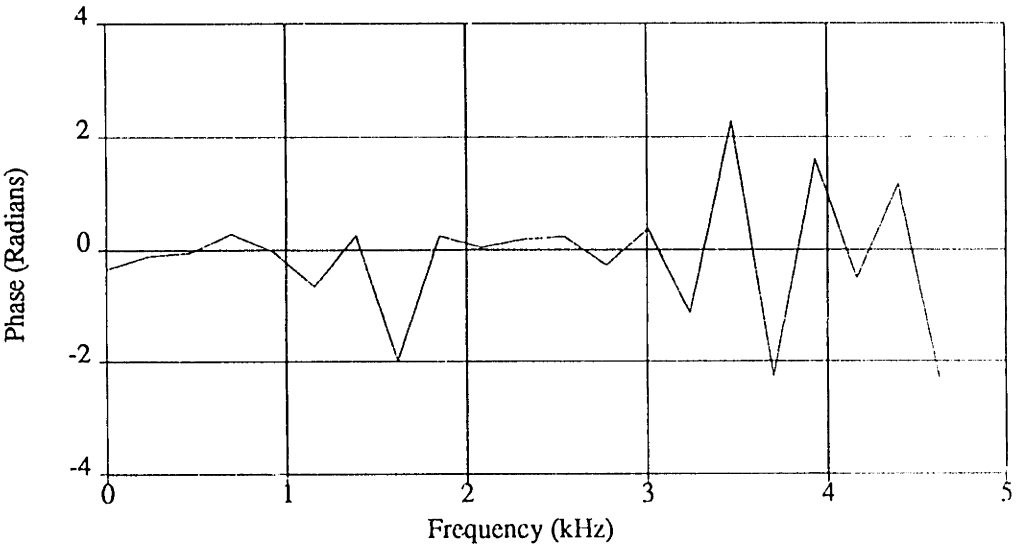

Figure 5.5: Difference Between Estimated and Predicted Phases

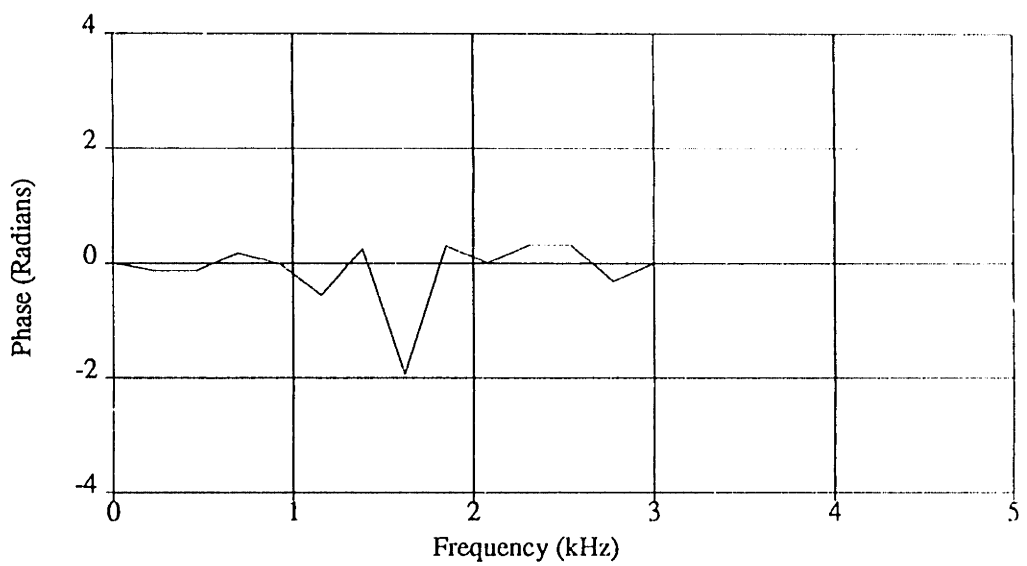

Figure 5.6: Coded Phase Differences

at low frequency) were coded using approximately 13 levels per harmonic. This coding method is simple and produces fairly good results. However, it fails to take advantage of the expected concentration of the phase differences around zero for consecutive voiced harmonics.

To show the distribution of phase differences for several frequency bands, six speech sentences were processed and the composite histograms generated. The phase differences accumulated were the difference between the predicted and estimated phase of the harmonics that were declared voiced in consecutive frames. As indicated in Figures 5.7 through 5.9, the phase differences tend to be concentrated around zero especially for low frequencies. For higher frequencies, the distribution tends to become more uniform as the phases of the higher frequency harmonics become less predictable.

Several methods are available for reducing the average number of bits required to code a parameter at a given average quantization error. In entropy coding [31], the parameter is uniformly quantized with L quantization levels and a symbol y_i is assigned to the i^{th} quantization level. The minimum average achievable rate to code these symbols is given by the

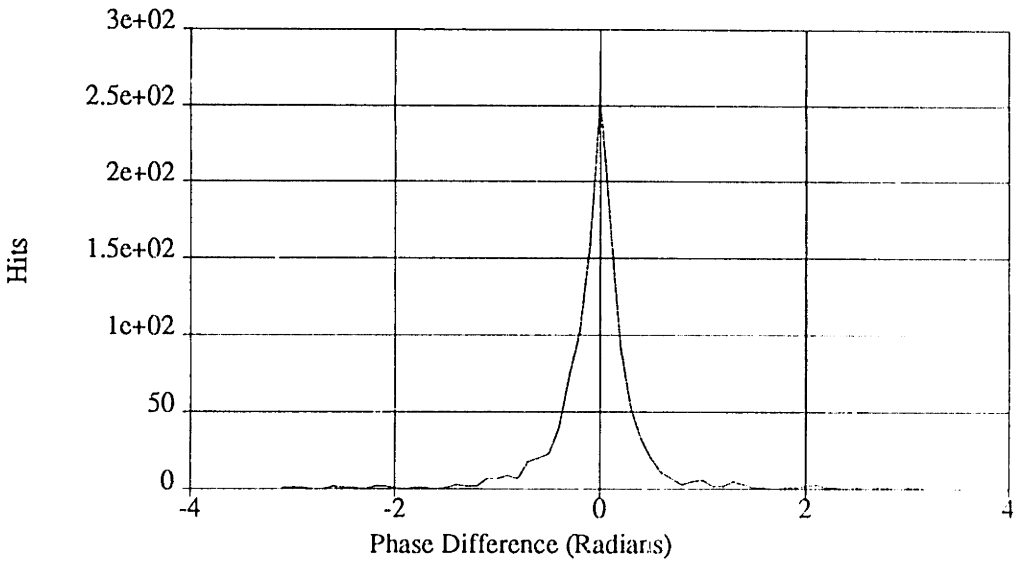

Figure 5.7: Phase Difference Histogram (60 - 500Hz)

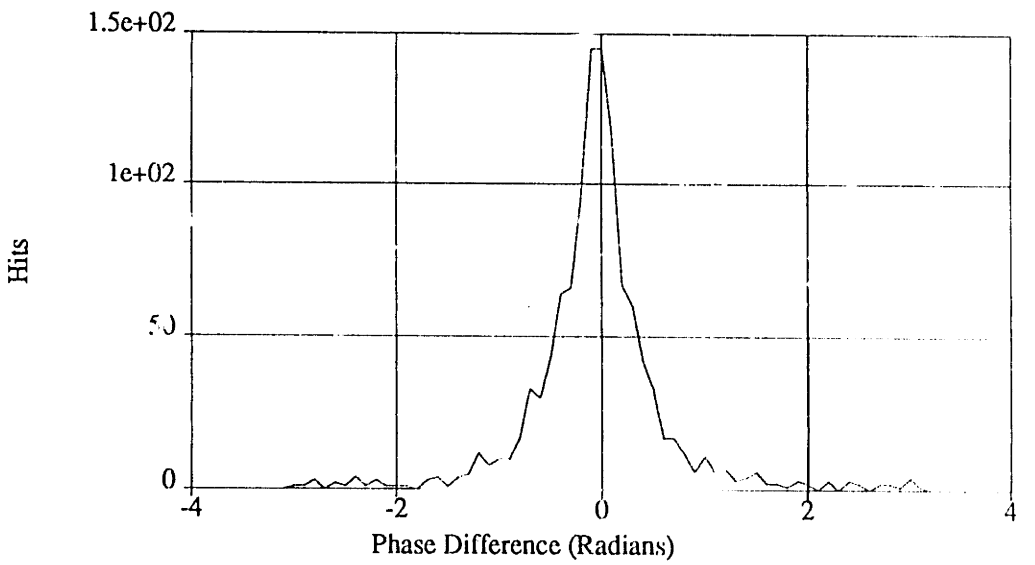

Figure 5.8: Phase Difference Histogram (.5 - 1.0kHz)

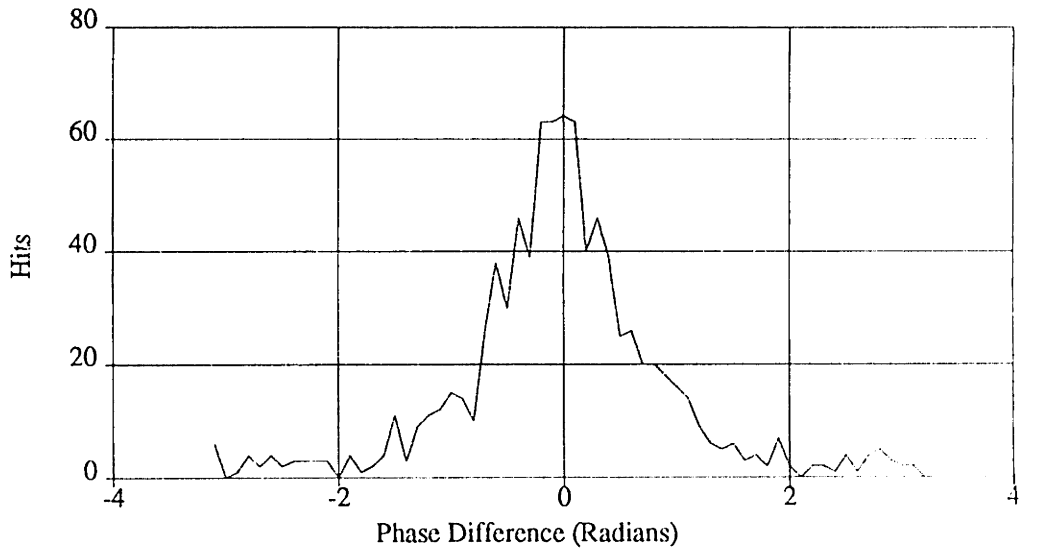

Figure 5.9: Phase Difference Histogram (1.0 - 1.5kHz)

entropy:

$$H(\mathbf{y}) = -\sum_{i=1}^{L} P(y_i) \log_2 P(y_i) \tag{5.1}$$

In entropy coding, the number of bits assigned to the symbol y_i is:

$$B_i \approx -\log_2 P(y_i) \tag{5.2}$$

so that shorter code words are used for more probable symbols. The approximation occurs in Equation (5.2) since $-\log_2 P(y_i)$ may not be an integer value. The resulting variable length code achieves an average rate close to the entropy. Constructive methods exist [13] for generating optimum variable length codes. The problem with entropy coding is that if a number of

improbable events occur closely spaced in time, a large delay is required to transmit the code words which can result in unacceptably long pauses in the synthesis end of a speech coding system in addition to requiring a large data buffer.

In Lloyd-Max quantization [16], [20], nonuniform quantization is used to minimize the average quantization error for a given number of quantization levels. An equal number of bits is then used to code each level. This coding method has the advantage of having fixed length code words. However, parameter values with low probability are often coded with a large quantization error.

An L level Lloyd-Max quantizer is specified by the end points x_i of each of the L input ranges and an output level y_i corresponding to each input range. We then define a distortion function

$$D = \sum_{i=1}^{L} \int_{x_i}^{x_{i+1}} f(x - y_i) p(x) dx \tag{5.3}$$

where $f(x)$ is some function (we used $f(x) = x^2$) and $p(x)$ is the input amplitude probability density. The objective is to choose the x_i's and the corresponding y_i's to minimize this distortion function. Several iterative methods exist [16], [20] for minimizing this distortion function.

Table 5.3 shows the reduction in quantization error in dB for a 13 level Lloyd-Max quantizer over a 13-level uniform quantizer. As expected, sig-

Freq (kHz)	Improvement (dB)
0.0 - 0.5	4.4
0.5 - 1.0	3.2
1.0 - 1.5	1.7
1.5 - 2.0	1.6
2.0 - 4.0	0.95

Table 5.3: Quantization Error Reduction

nificantly more improvement is obtained for the more predictable lower frequencies.

Due to the improved performance of the Lloyd-Max quantizer over a uniform quantizer and the advantage of fixed length code words over entropy coding, the Lloyd-Max quantizer was employed in the 8 kbps MBE Coder.

5.2.3 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 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. 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:

$$E_k = \frac{\sum_m |A_m| \xi_m}{\sum_m |A_m|} \tag{5.4}$$

where m is summed over all of the harmonics in the k^{th} frequency band.

5.3 Coding - Summary

The methods used for coding the MBE model parameters are summarized in Figures 5.10 through 5.13. The fundamental frequency is coded using uniform quantization (Figure 5.10).

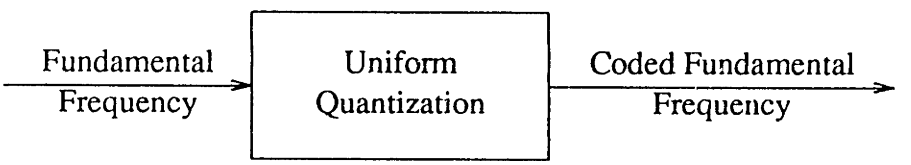

Figure 5.10: Fundamental Frequency Coding

The estimated phases are coded by predicting the phases of the current frame from the coded phases in the previous frame using the coded fundamental frequency (Figure 5.11). The difference between the predicted phases and the estimated phases are then coded using Lloyd-Max quantization. Only the phases of the M lowest frequency harmonics declared

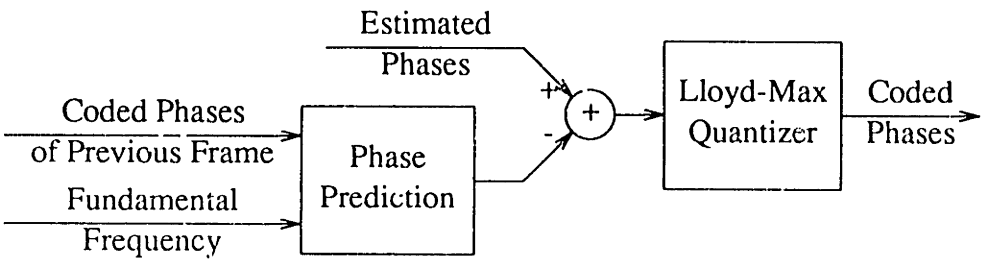

Figure 5.11: Coding of Phases

voiced are coded since these appear to be more important perceptually. The phases of harmonics declared unvoiced are not coded since they are not required by the synthesis algorithm and the bits allocated to them are used to code the magnitude samples.

The magnitude samples are coded by coding the lowest frequency magnitude sample using uniform quantization. The remaining magnitudes for the current frame are coded using adaptive differential PCM across frequency (Figure 5.12). The number of bits assigned to coding each magnitude sample is determined from the coded fundamental frequency by summing a bit distribution curve as described in Section 5.2.1.

The V/UV information is coded by dividing the original spectrum into N frequency bands ($N = 12$ for the 8 kbps system). The error (closeness of

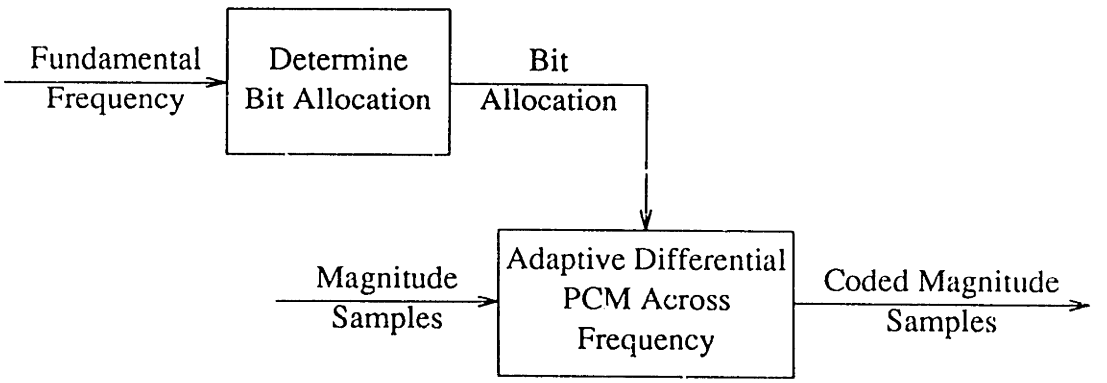

Figure 5.12: Coding of Magnitudes

fit) is determined between each frequency band of the original spectrum and the corresponding frequency band of the synthesized all-voiced spectrum (Figure 5.13). A threshold is then used to set a V/UV bit for each frequency

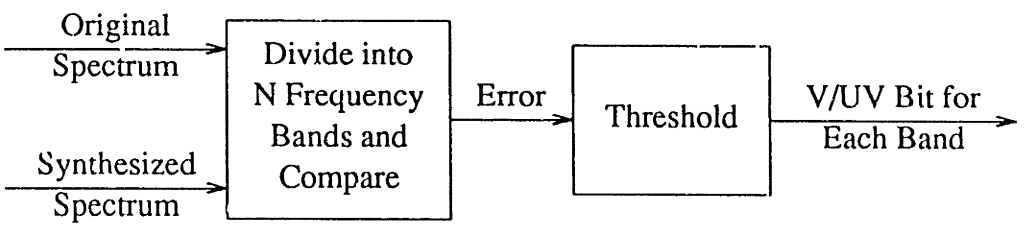

Figure 5.13: Coding of V/UV Information

band. When the error for a frequency band is below the threshold, the all-voiced synthetic spectrum is a good match for the original spectrum and this frequency band is declared voiced. When the error for a frequency band

is above the threshold, the all-voiced synthetic spectrum is a poor match for the original spectrum and this frequency band is declared unvoiced.

The 8kbps MBE Coder was implemented on a MASSCOMP computer (68020 CPU) in the C programming language. The entire system (analysis, coding, synthesis) requires approximately 1 minute 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.

5.4 Quality - Informal Listening

Informal listening was used to compare a number of speech sentences processed by the Multi-Band Excitation Speech Coder and the Single Band Excitation Speech Coder. For clean speech, the speech sentences coded by the MBE Speech Coder did not have the slight "buzziness" present in some regions of speech processed by the SBE Speech Coder. Figure 5.14 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

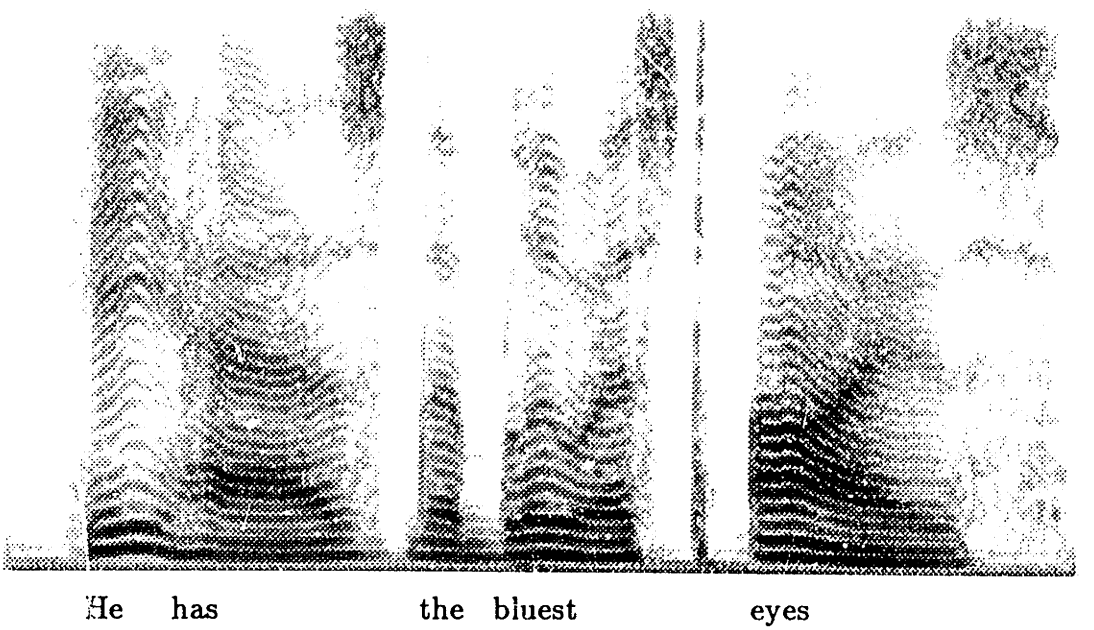

Figure 5.14: Uncoded Clean Speech Spectrogram

the energy versus time (0 - 2 seconds, 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 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 Multi-Band Excitation Vocoder op-

erating at 8kbps reproduces this region quite faithfully using 12 V/UV bits (Figure 5.15). The SBE Vocoder declares the entire spectrum voiced and

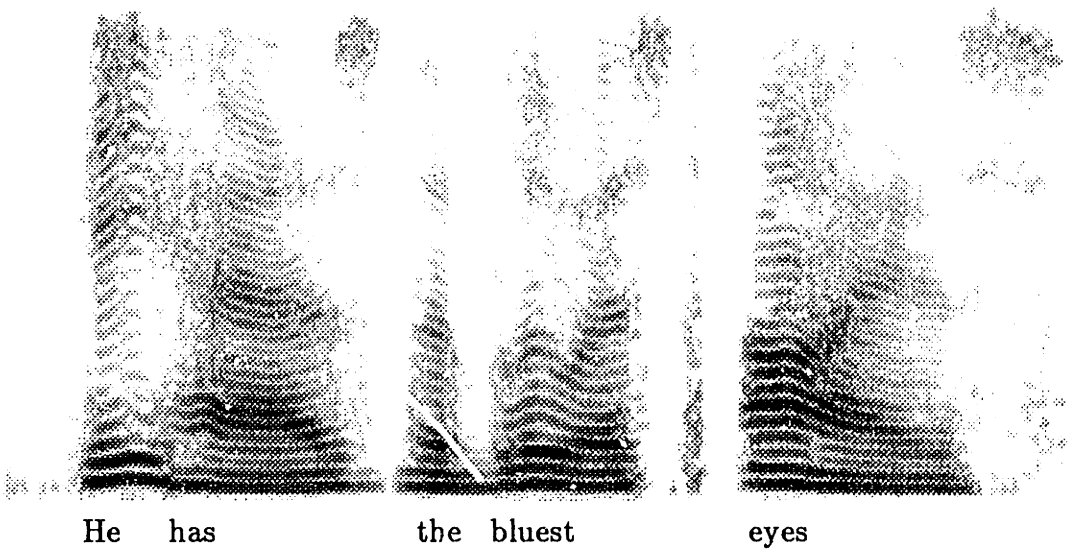

Figure 5.15: MBE Vocoder - Clean Speech Spectrogram

replaces the aperiodic energy apparent in the original spectogram with harmonics of the fundamental frequency (Figure 5.16). 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 kbps. The major degradation in these two systems (other than the "buzziness" in the SBE Vocoder) is a slightly reverberant quality due to

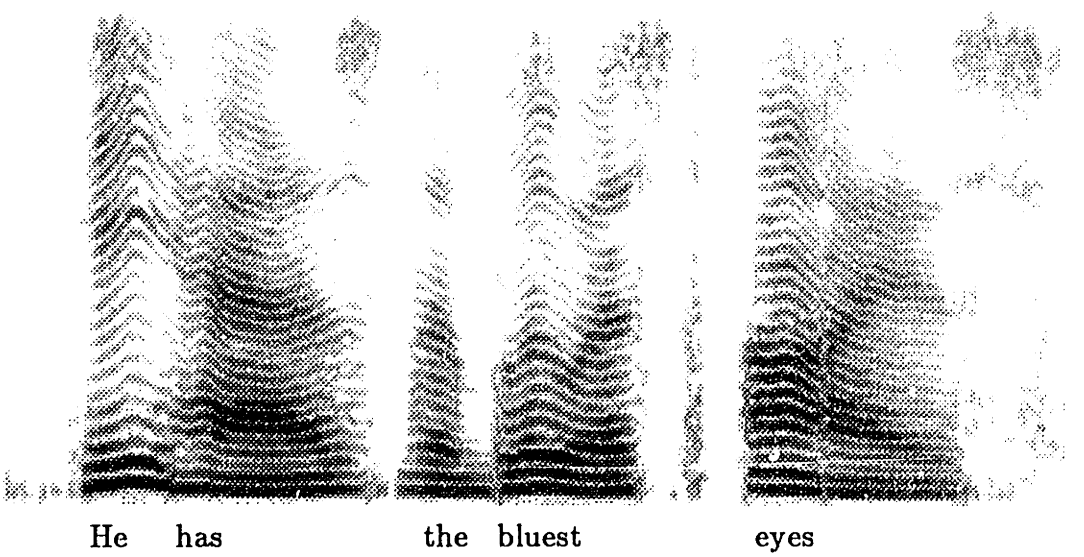

Figure 5.16: SBE Vocoder - Clean Speech Spectrogram

the large synthesis windows (40 ms triangular windows) and the lack of enough coded phase information.

For speech corrupted by additive random noise (Figure 5.17), the SBE Coding System (Figure 5.19) had severe "buzziness" and a number of voiced/unvoiced errors. The severe "buzziness" is due to replacing the

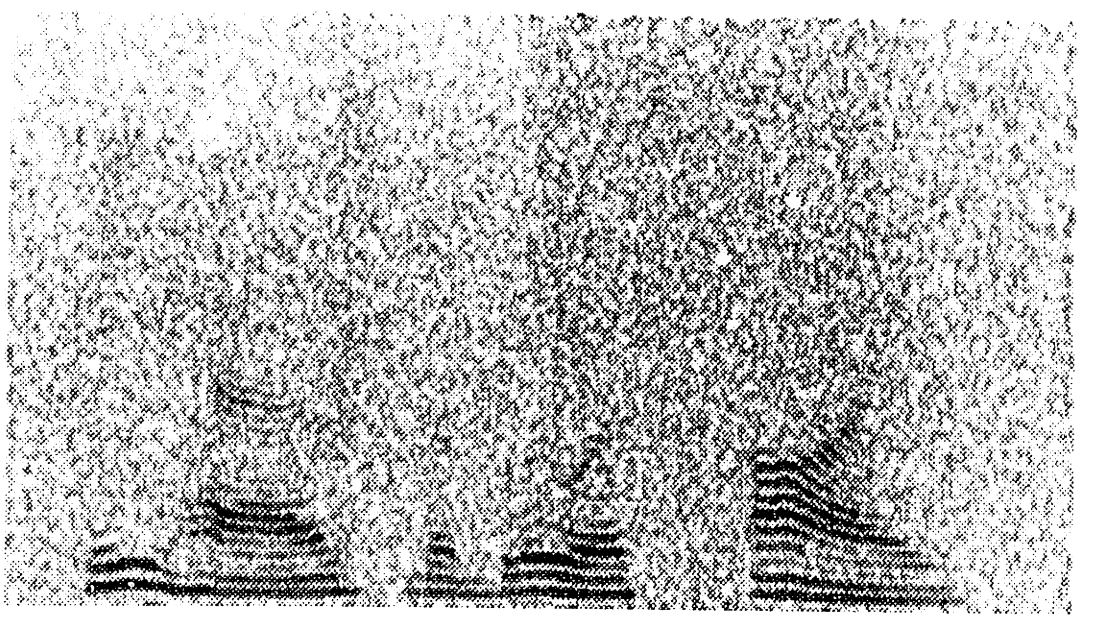

Figure 5.17: Uncoded Noisy Speech Spectrogram

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

the large synthesis windows (40 ms triangular windows) and the lack of enough coded phase information.

For speech corrupted by additive random noise (Figure 5.17), the SBE Coding System (Figure 5.19) had severe "buzziness" and a number of voiced/unvoiced errors. The severe "buzziness" is due to replacing the

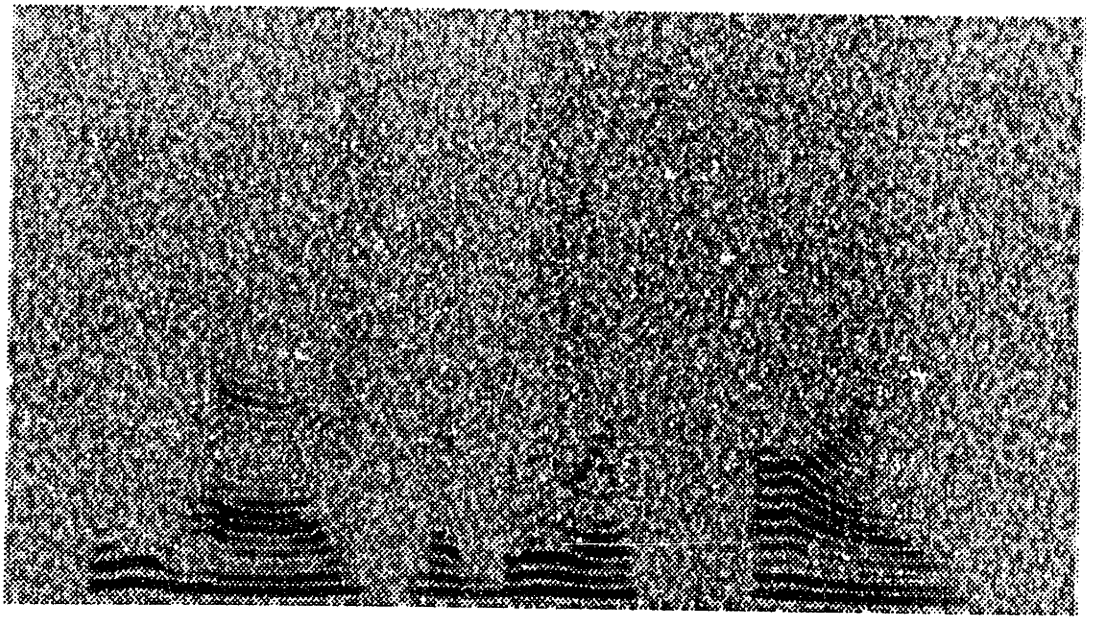

Figure 5.17: Uncoded Noisy Speech Spectrogram

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 Multi-Band Excitation Speech (for example, see Figure 5.18) Coder didn't have the severe "buzziness" present in the Single Band Excitation Speech Coder and didn't 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

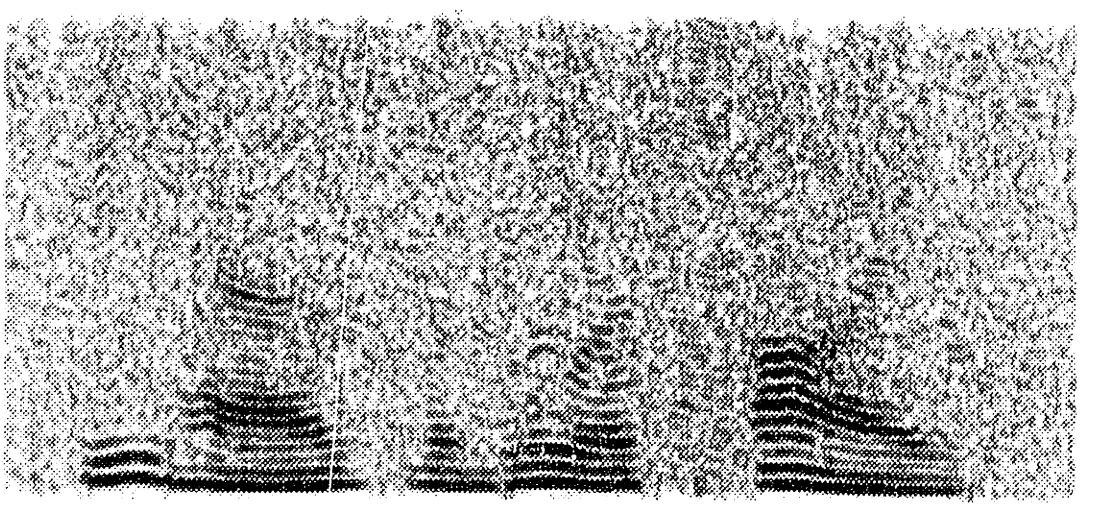

Figure 5.18: MBE Vocoder - Noisy Speech Spectrogram

Vocoder sound very close to the original noisy speech.

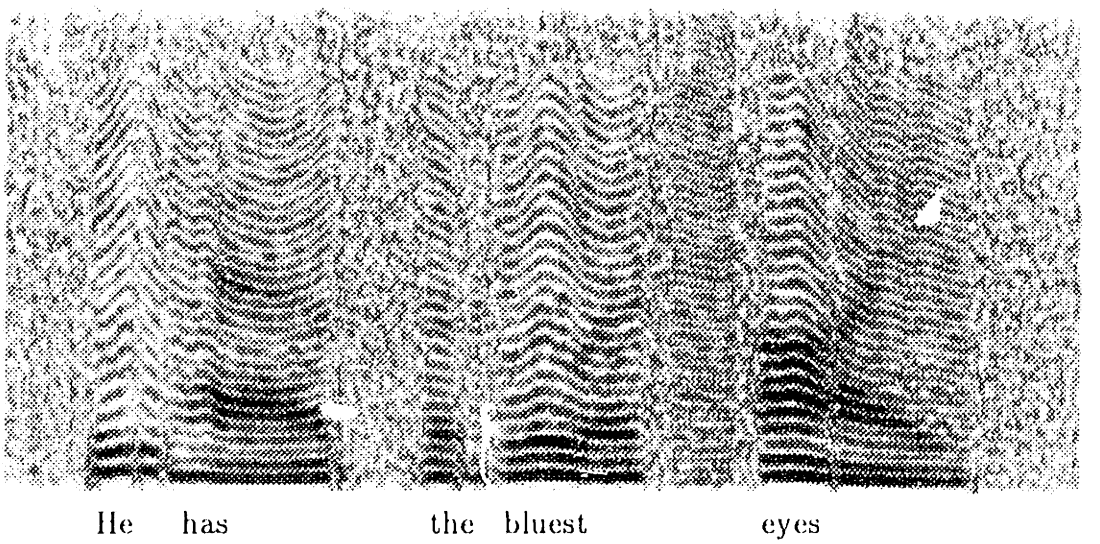

Figure 5.19: SBE Vocoder - Noisy Speech Spectrogram

5.5 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 [32].

The DRT consists of listening to a sequence of words spoken by the same speaker. Each of the words spoken is one of a set of two rhyming monosyllabic words. The listener must then choose which of the two words was spoken for each word in the sequence. The DRT word pairs were chosen so that only the initial consonant differs in order to minimize the effects of context. One DRT consists of 192 test words in addition to some filler words spoken by a single speaker and corresponds to approximately 7 minutes of speech. 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 corresponds to a score of 100.

The DRT was employed to compare uncoded speech with the 8 kbps Multi-Band 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 peak 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 relative performance of the vocoders in the same background noise which makes the noise level uncritical.

The DRT scores presented for clean speech (Table 5.4 and Figure 5.20) and noisy speech (Table 5.5 and Figure 5.21) were generated from three male speakers and 10 listeners. Figures 5.20 and 5.21 are bar graphs that show the average DRT scores and one standard deviation above and below them. Each of the 18 DRT tests taken by each listener was generated from an original set of 3 DRT tests (one for each speaker) by randomly rearrang-

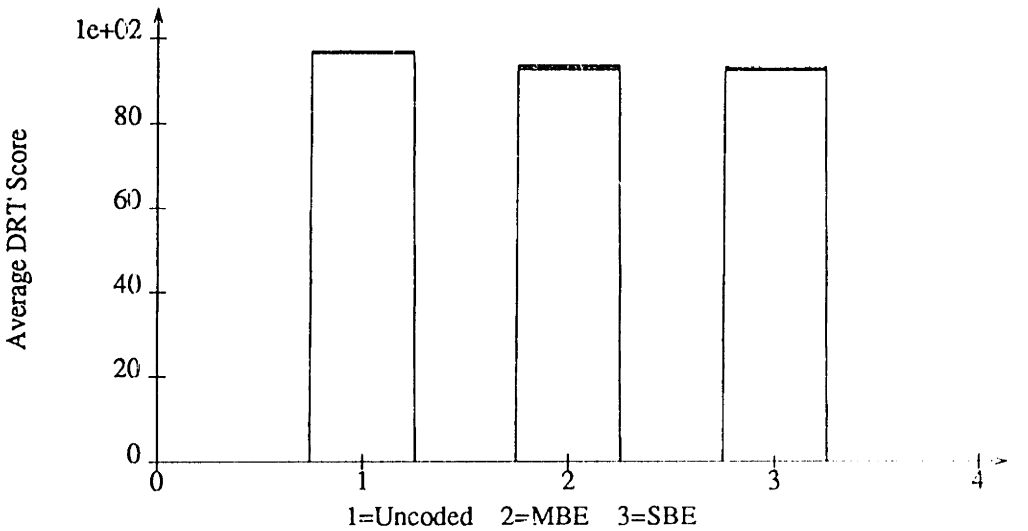

Figure 5.20: Average DRT Scores - Clean Speech

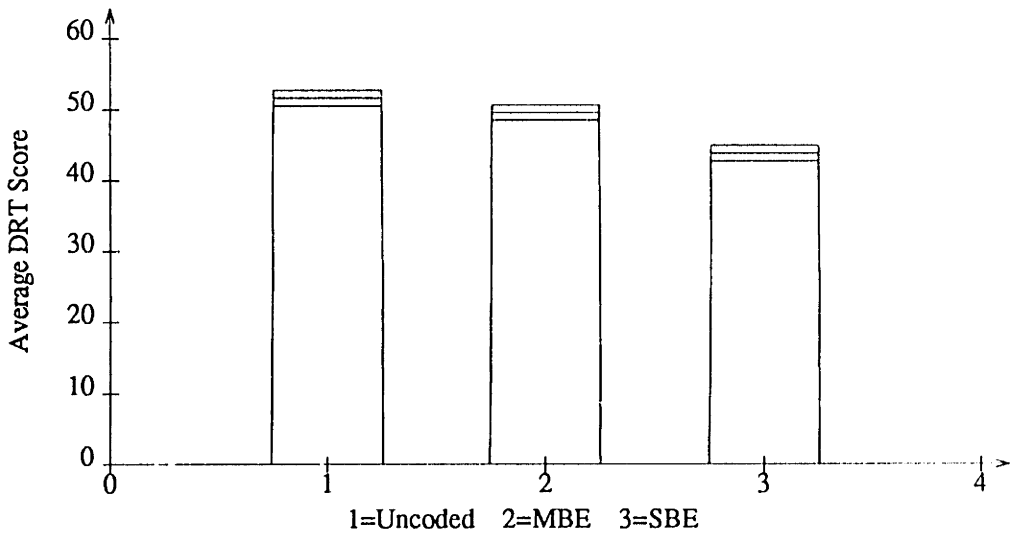

Figure 5.21: Average DRT Scores - Noisy Speech

System	Type	Speaker			Average
		CH	JE	RH	
Uncoded	Mean	97.6	95.7	97.3	96.9
	S. D.	.36	.50	.55	.28
8 kbps MBE	Mean	93.5	91.4	95.8	93.6
	S. D.	.90	1.1	.69	.53
Conventional	Mean	93.4	91.7	95.1	93.4
	S. D.	.84	1.1	.51	.49

Table 5.4: DRT Scores - Clean Speech

ing the word pair order for each test to prevent memorization by listeners. The listeners were inexperienced initially and were given 4-6 practice DRT tests until they became comfortable with the tests and produced reliable scores. The scores presented in the tables were computed by eliminating outliers in the original listeners' scores and then computing the mean and an estimate of the standard deviation of this mean assuming a Gaussian density for the listener scores. Outliers were eliminated by computing the average of the scores and removing the two scores furthest from the average.

System	Type	Speaker			Average
		CH	JE	RH	
Uncoded	Mean	56.5	43.6	54.9	51.7
	S. D.	1.8	1.8	2.1	1.1
8 kbps MBE	Mean	56.7	40.6	51.4	49.6
	S. D.	1.8	2.0	1.7	1.1
Conventional	Mean	49.6	35.0	47.1	43.9
	S. D.	1.5	2.2	1.8	1.1

Table 5.5: DRT Scores - Noisy Speech

The remaining 8 scores were then used to estimate the mean and standard deviation. Since the relative DRT scores are of primary interest, Tables 5.6 and 5.7 show the mean and standard deviation of the difference between the listeners' DRT scores for uncoded speech and speech processed by the two Vocoders.

For clean speech, as expected, several points are lost going from uncoded to coded due to lowpass filtering inherent in the vocoders and degradations introduced by coding. Also, the intelligibility scores are approximately the

Systems	Type	Speaker			Average
		CH	JE	RH	
Uncoded	Mean	3.3	4.3	1.4	3.0
- 8 kbps MBE	S. D.	.95	1.0	.58	.50
Uncoded	Mean	3.4	4.0	2.2	3.2
- SBE	S. D.	1.0	.72	.65	.46
8 kbps MBE	Mean	.1	-.26	.78	.2
- SBE	S. D.	.64	.44	.51	.31

Table 5.6: DRT Score Differences - Clean Speech

same for the MBE Vocoder and the SBE Vocoder.

For noisy speech, the MBE Vocoder performs an average of about 6 points better than the SBE Vocoder while performing only about 2.6 points worse than the uncoded noisy speech. This demonstrates the utility of the extra voiced/unvoiced bands in the Multi-Band Excitation Vocoder.

Systems	Type	Speaker			Average
		CH	JE	RH	
Uncoded	Mean	.4	3.8	3.5	2.6
- 8 kbps MBE	S. D.	.4	2.2	1.5	.90
Uncoded	Mean	9.6	8.4	7.8	8.6
- SBE	S. D.	1.6	2.1	1.2	.97
8 kbps MBE	Mean	8.8	4.9	4.3	6.0
- SBE	S. D.	1.2	1.5	1.6	.83

Table 5.7: DRT Score Differences - Noisy Speech

5.6 DRT Scores - RADC

DRT test tapes for each of the conditions tested in the previous section were submitted to RADC for independent evaluation. The DRTs performed by RADC employed experienced listeners in a fairly controlled environment. The resulting DRT scores are presented for clean speech in Table 5.8 and Figure 5.22. The DRT scores are presented for noisy speech in Table 5.9 and Figure 5.23). Figures 5.22 and 5.23 are bar graphs that show the average DRT scores and one standard deviation above and below them.

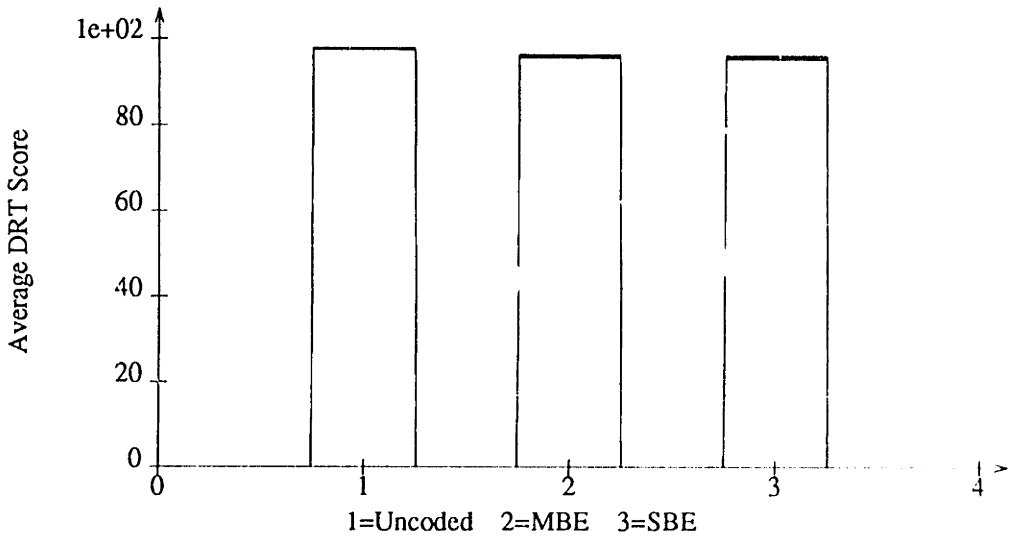

Figure 5.22: Average RADC DRT Scores - Clean Speech

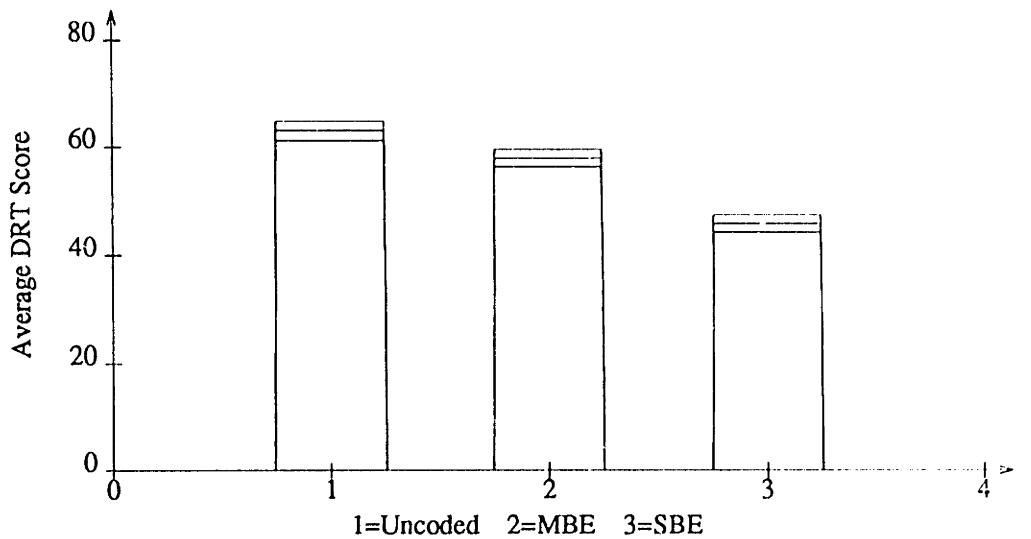

Figure 5.23: Average RADC DRT Scores - Noisy Speech

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
SBE	Mean	96.9	94.1	96.9	96.0
	S. D.	.44	.55	.81	.44

Table 5.8: RADC DRT Scores - Clean Speech

The RADC DRT scores confirm the trends noted in the previous section. For clean speech, the RADC DRT scores are slightly higher than those presented in the previous section due presumably to experienced listeners. Somewhat fewer DRT points are lost going from uncoded speech to coded speech than in the previous section. As in the previous section, the intelligibility scores for clean speech are approximately the same for the MBE Vocoder and the SBE Vocoder.

For noisy speech, the RADC DRT scores are significantly higher than

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
SBE	Mean	50.3	37.9	49.9	46.0
	S. D.	.94	2.3	1.8	1 6

Table 5.9: RADC DRT Scores - Noisy Speech

those presented in the previous section, probably due to experienced listeners, although the same trends are preserved. The MBE Vocoder performs an average of about 12 points better than the SBE Vocoder while performing only about 5 points worse than the uncoded noisy speech. This confirms the utility of the extra voiced/unvoiced bands in the Multi-Band Excitation Vocoder.

Chapter 6

Directions for Future Research

6.1 Introduction

In this thesis, we have considered in detail only the application of the Multi-Band Excitation Model to high quality speech coding. Some additional potential applications are discussed in Section 6.2. Improvements to the Multi-Band Excitation Speech Coding System can be made in a number of areas. Two areas of major importance are further improvement in quality and additional bit-rate reduction. Section 6.3 proposes some techniques for achieving these goals.

6.2 Potential Applications

Since the Multi-Band 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 non-linear frequency warping of the spectral envelope is desired without modifying the excitation parameters [28].

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

6.3 Improvement of the Speech Coding System

The quality of the Multi-Band Excitation Vocoder could be improved by elimination of the slightly reverberant quality of the 8 kbps vocoded speech. This degradation is due to the long synthesis windows (40 ms) used to accomplish the 50 Hz frame rate and the lack of enough coded phase information.

One approach to improving the quality and/or lowering the bit-rate would be to predict much of the phase information from the magnitude information. Since speech is often close to a minimum phase system excited by a periodic signal, a certain amount of phase information should be predictable from samples of the magnitude at the harmonics of the fundamental frequency. Since noise energy often dominates the signal in some frequency regions, this problem needs to be formulated as a best fit problem. For example, find the minimum phase signal which provides the best fit to the coded magnitude and several of the coded phases. A solution to this problem would allow the remaining phases at the receiver to be predicted from the coded phases and the coded magnitudes. If necessary, the

difference between the predicted phases and the actual phases could also be coded at the transmitter.

A second approach to improving the quality and/or lowering the bit-rate would be to take advantage of frame to frame correlation of the magnitude information. Speech usually consists of regions of slowly time-varying spectral magnitude bounded by short regions which change much more rapidly. One method for taking advantage of this would group frames into blocks and allocate more bits to rapidly varying sections of the block and fewer bits to more slowly varying sections. The blocks could be made fairly short (100-200ms) to avoid excessive coding and decoding delay.

Bibliography

[1] B. S. Atal and J. R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," *IEEE Int. Conf. on Acoust. Sp. & Sig. Proc.*, April 1982, pp. 614-617.

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

[3] H. Dudley, "The Vocoder," *Bell Labs Record*, Vol. 17, pp. 122-126, 1939.

[4] G. Fairbanks, "Test of Phonemic Differentiation: The Rhyme Test," *J. Acoust. Soc. Am.*, Vol. 30, pp. 596-600, 1958.

[5] O. Fujimara, "An Approximation to Voice Aperiodicity," *IEEE Trans. Audio and Electroacoust.*, pp. 68-72, March 1968.

[6] B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," *J. Acoust. Soc. Am.*, Vol. 46, No. 2, Pt. 2, pp. 442-448, August 1969.

[7] B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," *M. I. T. Lincoln Laboratory Technical Report*, TR-670, December 1983.

[8] D. W. Griffin and J. S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-32, no. 2, pp. 236-243, April 1984.

[9] D. W. Griffin and J. S. Lim, "A New Model-Based Speech Analsys/Synthesis System," *IEEE Int. Conf. on Acoust., Sp. & Sig. Proc.*, Tampa, Florida, March 26-29, 1985, pp. 513-516.

[10] D. W. Griffin and J. S. Lim, "A High Quality 9.6 kbps Speech Coding System," *IEEE Int. Conf. on Acoust., Sp. & Sig. Proc.*, Tokyo, Japan, 1986.

[11] J. N. Holmes, "The JSRU Channel Vocoder," *IEE Proc.*, Vol. 127, Pt. F, No. 1, February 1980, pp. 53-60.

[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. Am.*, Vol. 37, pp. 158-166, 1965.

[13] D. A. Huffman, "A Method for Construction of Minimum-Redundancy Codes," *Proc. IRE*, Vol. 40, no. 9, pp. 1098-1101, Sept. 1952.

[14] F. Itakura and S. Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," *Reports of 6th Int. Cong. Acoust.*, Tokyo, Japan, Paper C-5-5, pp. C17-20, 1968.

[15] S. Y. Kwon and A. J. Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," *IEEE Trans. on Acoutics, Speech, and Signal Processing*, vol. ASSP-32, no. 4, pp. 851-858, August 1984.

[16] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, Vol. IT-28, no. 2, pp. 129-137, March 1982.

[17] J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins, "A Mixed-Source Excitation Model for Speech Compression and Synthesis," *IEEE Int. Conf. on Acoust. Sp. & Sig. Proc.*, April 1978, pp. 163-166.

[18] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," *IEEE Trans. on Audio and Electoacoustics*, Vol. AU-20, No. 5, pp. 367-377, December 1972.

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

[20] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, Vol. IT-6, no. 2, pp. 7-12, March 1960.

[21] R. J. McAulay and T. F. Quatieri, "Mid-Rate Coding Based on a Sinusoidal Representation of Speech," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-32, no. 2, pp. 236-243, April 1984.

[22] C. S. Myers and L. R. Rabiner, "Connected Digit Recognition Using a Level-Building DTW Algorithm," *IEEE Trans. on Acoust., Speech and Signal Proc.*, Vol. ASSP-29, No. 3, pp. 351-363, June 1981.

[23] S. H. Nawab, "Signal Reconstruction from Short-Time Fourier Transform Magnitude," *Ph.D. Thesis*, M.I.T., 1983.

[24] A. M. Noll, "Cepstrum Pitch Determination," *J. Acoust. Soc. Am.*, Vol. 41, pp. 293-309, February 1967.

[25] A. V. Oppenheim, "A Speech Analysis-Synthesis System Based on Homomorphic Filtering," *J. Acoust. Soc. Am.*, Vol. 45, pp. 458-465, February 1969.

[26] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.

[27] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.

[28] M. A. Richards, "Helium Speech Enhancement Using the Short-Time Fourier Transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-30, pp. 841-853, December 1982.

[29] 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. Am.*, Vol. 50, No. 6, pp. 1532-1538, December 1971.

[30] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," *IEEE Trans. Acoust., Speech and Signal Proc.*, Vol. ASSP-25, No. 1, pp.

24-33, February 1977.

[31] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, Vol. 27, pp. 379-423 (Part I), pp. 623-656 (Part II).

[32] W. D. Voiers, "Evaluating Processed Speech using the Diagnostic Rhyme Test," *Speech Technology*, Jan./Feb. 1983.

[33] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum Likelihood Pitch Estimation," *IEEE Trans. on Acoust., Speech and Signal Proc.*, Vol. ASSP-24, No. 5, pp. 418-423, October 1976.