A 1.5 Kbps Multi-Band Excitation Speech Coder

by

Michael Shapiro Brandstein

Sc.B., Brown University, Providence, RI

(1988)

Submitted in Partial Fulfillment
of the Requirements for the
Degree of

Master of Science
in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

May 1990

Signature of Author _____.

Department of Electrical Engineering and Computer Science
May 9, 1990

Certified by _____ _____.

Professor Jae S. Lim
Thesis Supervisor

Accepted by _____.

Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

A 1.5 kbps Multi-Band Excitation Speech Coder

by

Michael Shapiro Brandstein

Submitted to the
Department of Electrical Engineering and Computer Science
on May 9, 1990 in partial fulfillment of the requirements
for the Degree of Master of Science

Abstract

In this thesis a 1.5 kbps speech coder based on the Multi-Band Excitation (MBE) speech model is presented. The system is comprised of several isolated elements. The first of these is the MBE analysis algorithm which estimates the MBE parameters. The parameters are then quantized, converted into a bit stream, and transmitted across a channel. At the receiver end a decoder regenerates these parameters and delivers them to the MBE synthesis routine where the synthesized speech is produced. The focus of this thesis is the coding of the model parameters with a bit rate constraint of 1.5 kbps. Given this restriction, quantization schemes employed in earlier, higher rate MBE coders were found to be unsatisfactory. A new coding method based upon Linear Predictive Coding (LPC) of the harmonic magnitudes and a Line Spectrum Pair (LSP) representation of the LPC coefficients is developed. A coder simulation based on this design is shown to obtain speech intelligibility on par with state-of-the-art 2.4 kbps speech coders. An informal listening comparison between the 1.5 kbps MBE coder and the government standard 2.4 kbps LPC-10e vocoder reveals comparable performance in high SNR conditions and a preference for the MBE coder in noisy environments.

The performance of this system further demonstrates the attractiveness of the MBE model to speech coding applications. The ability of the MBE model to accurately reproduce speech in a wide range of background environments provides a significant advantage over conventional speech modeling methods. The compactness of the MBE parameters and their potential to be efficiently quantized makes this model ideal for low rate systems. Finally, the computational and production costs of a real-time implementation of an MBE speech coder are small compared to systems producing similar quality. Real-time versions of the 2.4, 4.8, and 8.0 kbps MBE coders have been constructed with only a single DSP chip.

Thesis Supervisor: Jae S. Lim

Title: Professor of Electrical Engineering

Dedication

To Nanny.

My best friend and teacher.

Acknowledgments

A number of individuals have helped make this thesis possible. In particular, I would like to thank John Hardwick for the many helpful suggestions and for allowing me to benefit from his past research. I would also like to thank my thesis advisor, Professor Jae Lim, for his support and guidance throughout this endeavor. My appreciation also extends to the entire membership of the Digital Signal Processing Group. I thank you for making the laboratory such a nice place to learn.

This research project has been supported in part by the Rome Air Development Center. I have also received direct financial support from a National Science Foundation Graduate Fellowship. I would like to thank each of these organizations for assisting me in continuing with my education.

Finally, I acknowledge the contributions of my parents, Mom and Dad, who have been available in an extended consultational capacity for several years now. They are entirely responsible for the contents herein.

Contents

1 Introduction **8**

2 Multi-Band Excitation Speech Model **13**

3 Speech Analysis **18**

3.1 Pitch Estimation 19

3.1.1 Frequency Domain Pitch Detection 20

3.1.2 Autocorrelation Pitch Detection 21

3.2 Voiced/Unvoiced Decisions 22

3.3 Spectral Envelope Determination 23

4 Speech Synthesis **24**

4.1 Voiced Synthesis 24

4.2 Unvoiced Synthesis 25

5 Parameter Coding **27**

5.1 Spectral Magnitudes 27

5.1.1 LPC Spectral Modeling 28

5.1.2 LPC Quantization 38

5.2 Fundamental Frequency 43

5.3 Voicing Decisions 44

5.4 Summary 45

6 Performance Results **47**

6.1 Speech Intelligibility 47

6.2 Speech Quality 51

7 Conclusions **55**

7.1 Summary 55

7.2 Suggestions for Further Research 56

List of Figures

1.1 A Generic Vocoder System 10

2.1 Conventional Speech Model 14
2.2 MBE Speech Model 15
2.3 MBE Speech Spectra 17

4.1 Outline of the MBE Speech Coder 26

5.1 A 16-pole spectrum and its 16-pole LPC model generated using frequency points sampled at $.019\pi$ radian intervals 32
5.2 A 16-pole spectrum and its LPC estimation generated with a $.075\pi$ radian sampling interval 33
5.3 A 16-pole spectrum sampled at $.075\pi$ radian intervals and its 16-pole LPC estimate produced via linear interpolation in the log spectral domain 35
5.4 Quantizer Performance as a function of bit rate and LPC order . . . 42
5.5 Outline of the procedure for quantizing the harmonic magnitudes . . 43

6.1 Outline of the procedure used to generate the testing material for the LPC-10e vs. MBE side-by-side comparison 53

List of Tables

5.1 Average spectral distortion vs. LPC analysis order 36

5.2 Effects of enhancement methods on the spectral distortion scores of 3000 frames of speech 38

5.3 Parameter coding bit allocations 46

6.1 DRT Scores for 1.5 kbps system in clean speech 49

6.2 DRT Scores for several speech coders 50

Chapter 1

Introduction

A major application of speech processing research concerns the digital coding of a speech signal for efficient, secure storage and transmission. Speech is coded into a bit stream representation, transmitted over a channel, and then converted back into an audible signal. Distortions in the transmission channel may cause errors in the received bits which may necessitate the use of bit protection strategies during coding. The decoder is an approximate inverse of the encoder except that some of the information has been lost during coding due to the conversion of the analog signal into a digital bit stream. The discarded information is selected so as to minimize the total perceivable distortion and is a function of bit rate and coding methods. The speech is often coded in the form of parameters that represent the signal economically and with limited quality degradation.

Several wide categories of speech reproduction capability have been established. In order of decreasing quality these categories are broadcast, toll, communications, and synthetic. Broadcast quality refers to wide-bandwidth (usually 0-7000 Hz) high-quality speech with no perceptible noise. Toll quality describes the signal heard over a telephone network (200-3200 Hz range with SNR ratio greater than 30 dB). Communications quality is highly intelligible but has noticeable distortions. Synthetic speech is better than 80%-90% intelligible and suffers from substantial degrada-

tions. These typically include a "machinelike" sound, a "buzzy" background, and a lack of speaker identifiability. Currently, a minimum of 64000 bits per second (64 kbps) are required to obtain broadcast quality. Toll quality is generally available from systems operating in a range of 64 kbps to 10 kbps. Communications quality is produced by coders with bit rates as low as 4.8 kbps and synthetic quality is the case below 4.8 kbps.

As a general rule, coder complexity varies inversely with bit rate. The simplest are waveform coders which analyze, code, and reconstruct the speech on a sample by sample basis. These are further divided into time-domain waveform coders, which exploit waveform redundancies such as periodicity and slowly varying intensity, and spectral-domain waveform coders that take advantage of the non-uniform frequency distribution of the speech signal. At lower bit rates, more complex methods are required. These systems, known as voice coders or "vocoders", assume a speech production model. In particular, the speech signal is separated into information estimating the vocal tract shape and that involving the vocal tract excitation. The excitation and vocal tract parameters may then be coded separately with a large decrease in bit rate. Figure 1.1 presents an outline of the generic vocoder. While these techniques produce speech of limited quality, they outperform waveform coders at bit rates below 10 kbps. As a result, speech coding research for low rate applications has focussed primarily on model based approaches.

Most of the existing vocoders are based on the conventional speech model. For this class, speech is synthesized as the response of a time varying linear filter to some excitation, or equivalently, the speech spectrum is represented as the product of excitation and system spectra. The vocoder analyzes windowed portions of the speech sequence and estimates parameters which characterize the system filter and the excitation sequence for the selected segment. The excitation sequence is limited to two cases. For voiced speech the excitation is modeled as a periodic impulse train and for unvoiced speech it is specified as a white noise sequence. With this

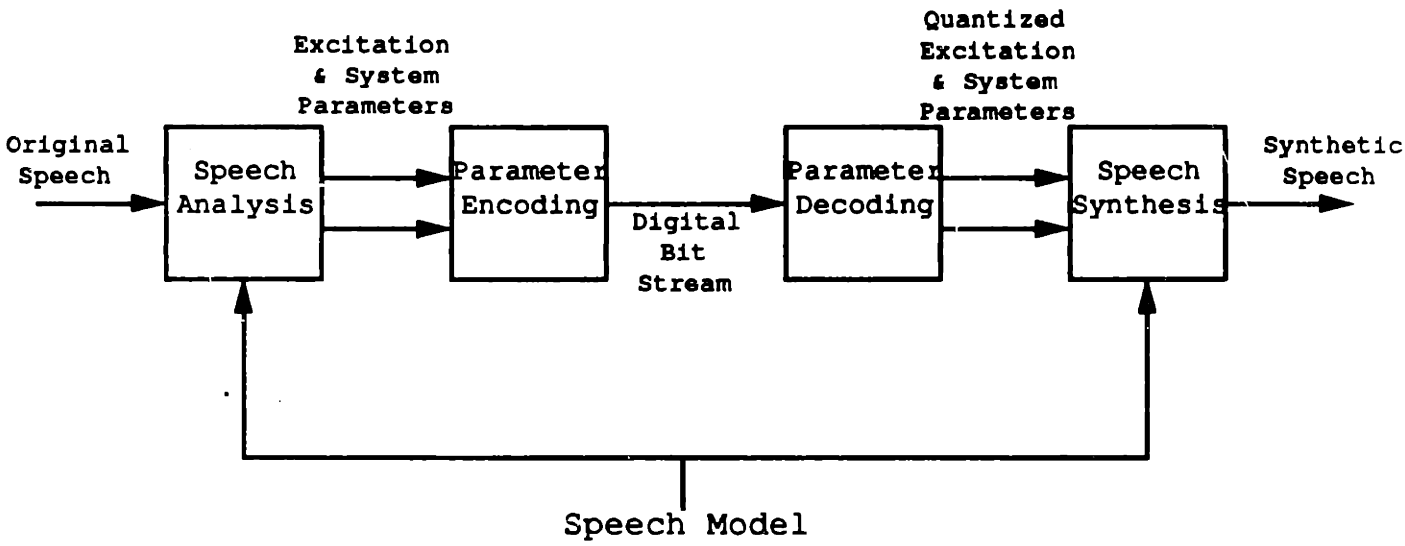

Figure 1.1: A Generic Vocoder System

classification the excitation parameters for each segment consist of a pitch period and a voiced/unvoiced (V/UV) decision. The system parameters are typically some representation of the spectral envelope or the impulse response of the vocal tract. The decoder uses the excitation parameters to generate either white noise for unvoiced segments or an impulse train with the desired pitch for voiced segments. This sequence is then passed through the filter specified by the system parameters and the output is the synthesized speech.

These systems are characterized by the methods used to estimate the transmission parameters. Some examples include the homomorphic vocoder which applies cepstral techniques to represent the system function and evaluate the pitch period, channel vocoders that use a series of bandpass filters to extract features of the spectral envelope, and LPC vocoders which model the system filter with a linear prediction polynomial. While some of these approaches outperform others due to their ability to estimate the necessary parameters, they are all limited by the appropriateness of the conventional speech model upon which they are based. Vocoders of this type are capable of producing intelligible speech, but they have not been

successful in synthesizing speech beyond synthetic quality. In addition, the performance of these systems is known to degrade rapidly in the presence of background noise. Considerable attention has been devoted to improving these systems. These improvements have focused primarily on the specification of the excitation signal after removal of the pitch structure, not at improving the accuracy of the underlying model. Examples of these approaches include code excited linear prediction (CELP) [7] and residual excited linear prediction (RELP) [8]. While these techniques have improved the quality, they have significantly increased algorithm complexity and raised the bit rate requirement.

In [1] a new speech model is presented which does not perform a binary voicing classification of the excitation sequence for a given analysis frame. Instead, the excitation is characterized by a number of V/UV decisions specified over a series of harmonic intervals. For this reason, the new approach has been termed the Multi-Band Excitation (MBE) speech model. This added degree of freedom allows each speech segment to be partially voiced and partially unvoiced. The result is increased flexibility in the selection of the excitation sequence, and consequently, the ability to more accurately model the original signal. When combined with newly developed means of estimating the pitch period and spectral envelope as described in [1], the MBE model has the potential to produce a robust, high-quality vocoder.

The applicability of the MBE speech model to high quality mid-rate and low-rate speech coding has been demonstrated in several systems. In [1] an 8.0 kbps MBE coder was developed and in [4] this figure was reduced to 4.8 kbps. Both of these analysis/synthesis systems have been shown to produce communications quality speech in a wide range of environments without a marked increase in computational complexity. The advantage of the MBE vocoder was most apparent from the lack of "buzziness" in noisy speech. Further attempts at lowering the coder bit rate appear to be successful. By applying quantization techniques similar to those used for the 4.8 kbps system, MBE coding at 2.4 kbps is possible with only a modest reduction

in quality from the higher rate systems [6].

The goal of this thesis was to investigate a means of applying the MBE speech model to coding applications below 2.4 kbps. Early tests demonstrated that the coding strategies used for earlier MBE coders were unsatisfactory at these reduced rates. In an attempt to maximize speech quality a number of alternative coding schemes were investigated. The focus of this research is to find a coding technique which is efficient enough to provide the desired speech quality at a given rate. The desired bit rate has been targeted at 1.5 kbps. This figure represents a significant reduction from earlier MBE coder bit rates while still having the potential to produce quality speech.

The remainder of this thesis will discuss the design and performance of a 1.5 kbps speech coder based on the MBE model. In chapter 2 the MBE speech model is described in greater detail. Chapter 3 discusses the analysis techniques used to estimate the MBE parameters while chapter 4 is dedicated to the synthesis of speech from these same parameters. Chapter 5 presents the new parameter coding schemes used in the development of the 1.5 kbps MBE coder. The quality of the speech produced by the vocoder is examined in chapter 6. Finally, chapter 7 concludes the thesis with several directions for further research.

Chapter 2

Multi-Band Excitation Speech Model

The quasi-stationary nature of speech requires that a speech signal $s(n)$ be analyzed over a short time duration, approximately 10 ms to 40 ms. A window $w(n)$ is applied to the sequence to focus attention on the desired interval. The windowed speech segment $s_w(n)$ is defined by

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

The sequence $w(n)$ is typically a Hamming or Kaiser window and may be shifted in time to select any desired segment of $s(n)$. The Fourier transform $S_w(\omega)$ of $s_w(n)$ can be modeled as the product of a spectral envelope $H_w(\omega)$ and an excitation spectrum $E_w(\omega)$:

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

The MBE speech model is similar to many simple speech models in the specification of the spectral envelope $H_w(\omega)$ as a smoothed version of the original speech spectrum $S(\omega)$. The primary difference between the MBE model and previous models lies in the form of the excitation spectrum. In previous models, the excitation spectrum is completely specified by the fundamental frequency ω_0 and a

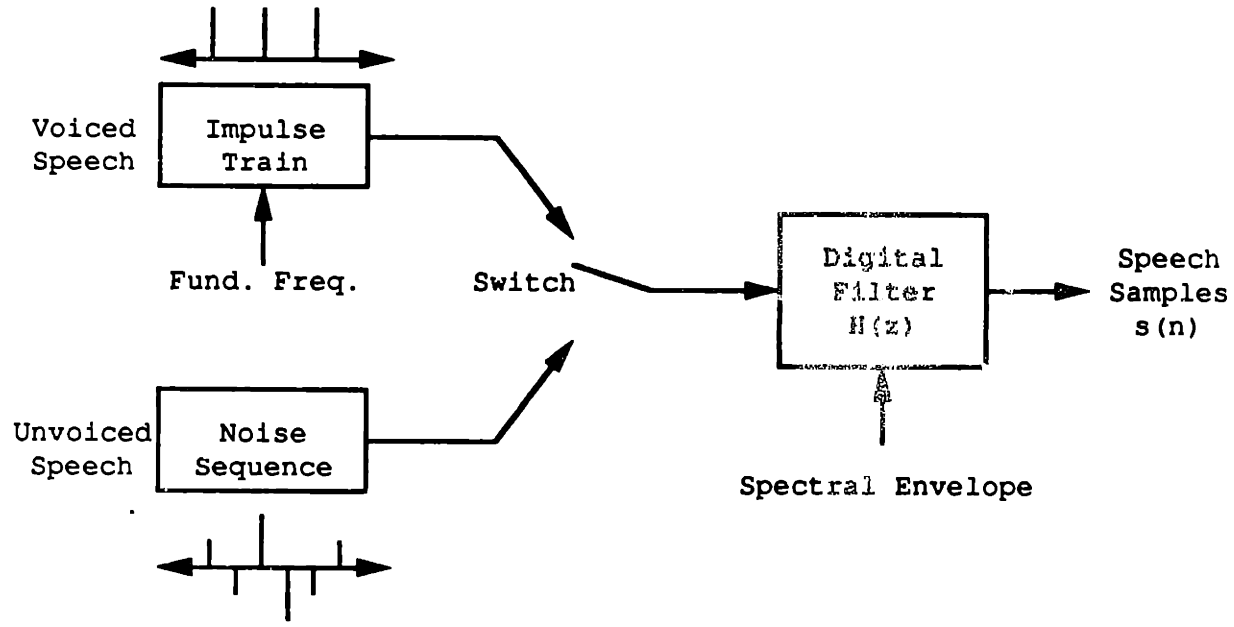

Figure 2.1: Conventional Speech Model

voiced/unvoiced decision for the entire analysis frame. For voiced segments $E_w(\omega)$ is set equal to $P_w(\omega)$, the Fourier transform of a windowed impulse train with a periodicity of $2\pi/\omega_0$ samples. Ignoring aliasing effects, $P_w(\omega)$ may be thought of as the sum of the Fourier transform of $w(n)$ centered at each harmonic of ω_0. Speech segments which do not possess a periodic structure are declared unvoiced and $E_w(\omega)$ is modeled as the spectrum of windowed white noise. As figure 2.1 demonstrates, the excitation signal in this conventional model is limited to only one of these two possibilities.

This approach is extremely limited in its ability to represent the full range of speech signals. Many speech segments have some frequency regions which are dominated by noise energy while others are filled with periodic voiced energy. This is especially true in mixed voicing segments of clean speech and in voiced segments of noisy speech. It has been shown that humans have the ability to discriminate between spectral regions dominated by harmonics of the fundamental frequency and those dominated by noise-like energy [2]. The elimination of these acoustic cues in vocoders employing this simple excitation model partially explains the synthetic

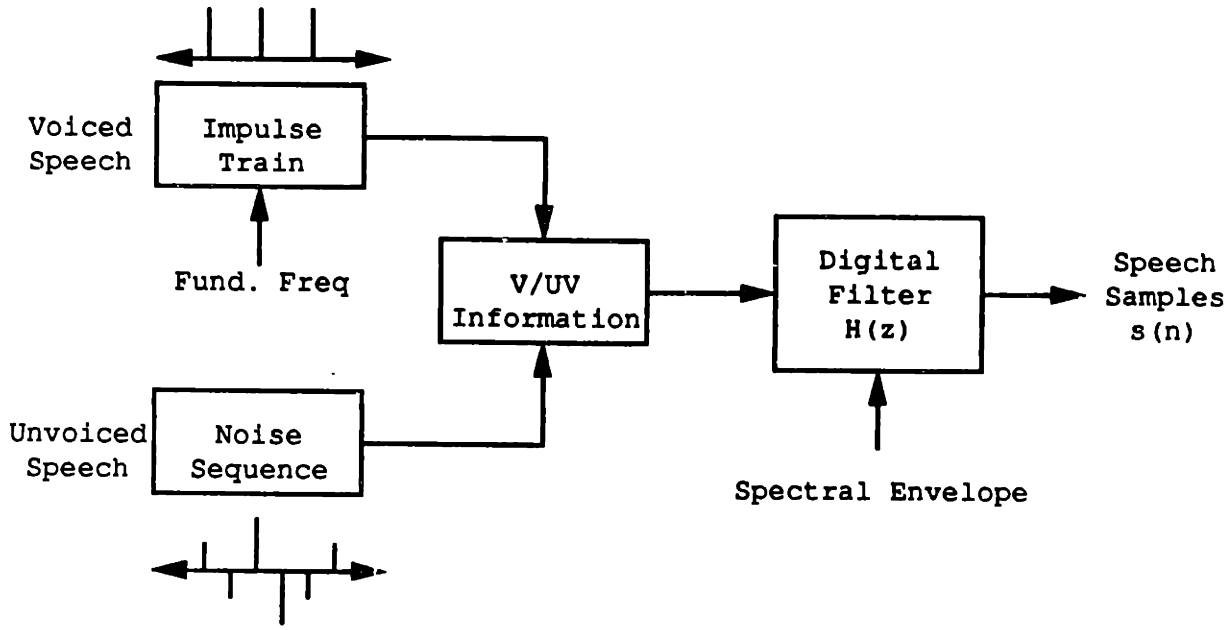

Figure 2.2: MBE Speech Model

quality of the generated speech and the significant intelligibility decrease observed in low SNR situations.

In the MBE model, the excitation spectrum is specified by the fundamental frequency ω_0 and a number of frequency dependent binary voiced/unvoiced decisions. The spectrum is divided into multiple frequency regions (typically 20 or more) and a voicing parameter is allocated to each band. In practice these regions are centered about each harmonic or group of harmonics of ω_0. The excitation spectrum $E_w(\omega)$ is obtained by combining segments of $P_w(\omega)$ in voiced frequency bands with portions of a random noise spectrum in the frequency bands declared unvoiced. Figure 2.2 outlines this new approach. The ability to specify voicing decisions over frequency bands with widths as small as the fundamental frequency allows the MBE speech model to produce a more accurate representation of the original speech spectrum than is possible with earlier speech models [1]. The result is high quality speech synthesis in a wide variety of environments.

Figure 2.3 demonstrates the procedure. In figure 2.3a the spectrum of a typical speech segment is shown. This is the DFT of $s_w(n)$ where $w(n)$ is a 256 point Ham-

ming window. Figure 2.3b presents the spectral envelope that has been calculated for this segment. $H_w(\omega)$ is represented by one sample A_m for each harmonic of the fundamental frequency in both voiced and unvoiced regions. The spectral envelope provides the scaling between $E_w(\omega)$ and the actual spectrum. The function $H_w(\omega)$ may be viewed as the frequency response that will map $E_w(\omega)$ into $\hat{S}_w(\omega)$ to effectively model $S_w(\omega)$. In the figure, the discrete harmonic magnitudes have been linearly interpolated to create a smooth contour containing the general shape of the original spectrum. For this segment, the pitch period has been estimated to be 80 samples at an 8 kHz sampling rate. The function $P_w(\omega)$ corresponding to this pitch period is shown in figure 2.3c. The voiced/unvoiced information is displayed in figure 2.3d. The frequency bands employed for these voicing decisions are a single harmonic in width. A high value on this graph represents a voiced region of the spectrum where $P_w(\omega)$ would be used for the excitation spectrum. Low values correspond to unvoiced frequency regions. Noise energy as shown in figure 2.3e is used as the excitation in these portions of the frequency spectrum. Finally, in figure 2.3f the voiced and unvoiced excitation spectra are combined and multiplied by the spectral envelope $H_w(\omega)$ to generate the synthetic speech spectrum $\hat{S}_w(\omega)$.

The parameters of the MBE speech model are the fundamental frequency, harmonic magnitudes, and voicing information as estimated for each analysis frame. The compactness of this representation in conjunction with the ability to produce high quality synthesized speech makes the MBE model ideal for low to mid rate coding applications.

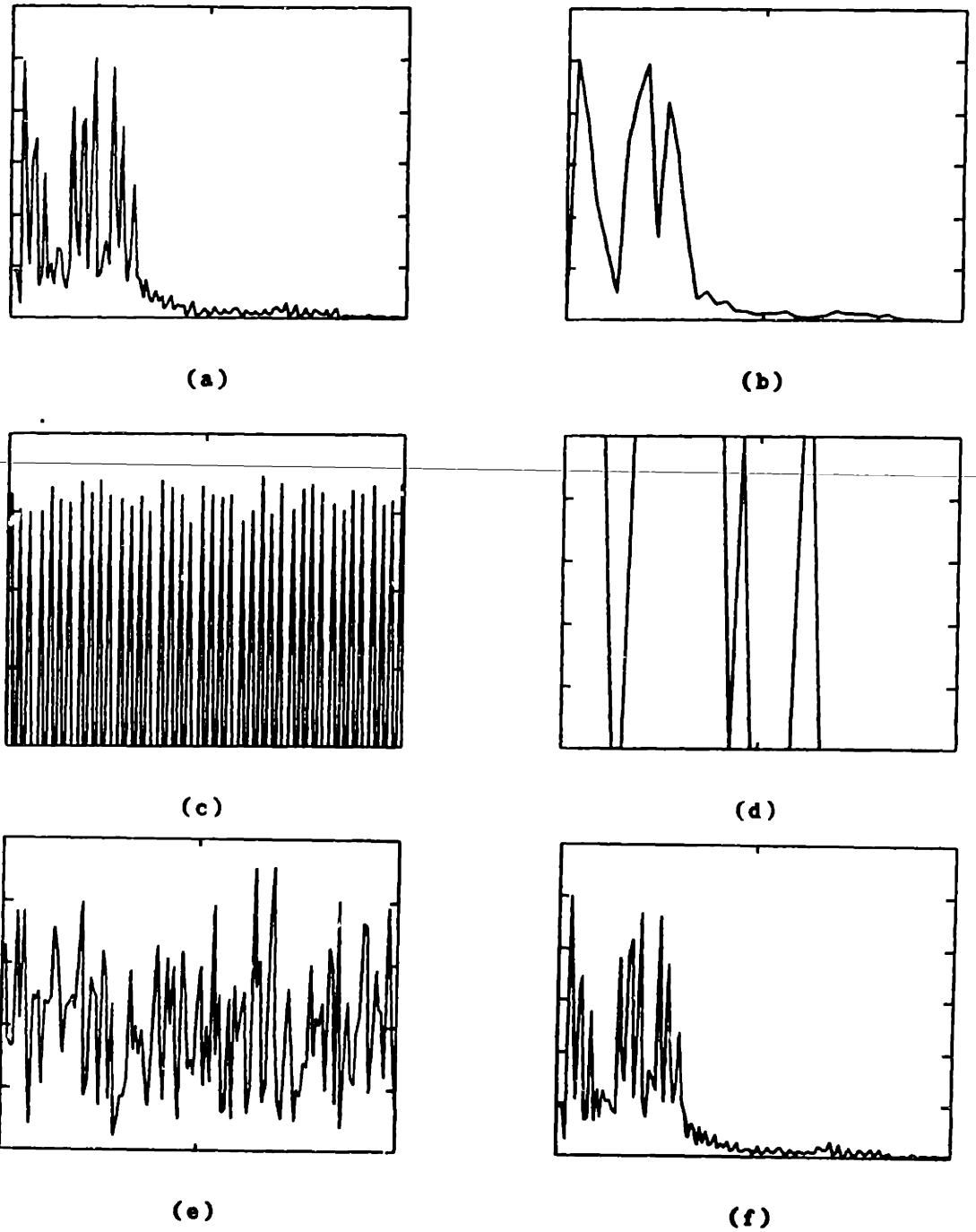

Figure 2.3: (a) Original speech spectrum, (b) Spectral envelope, (c) Periodic spectrum, (d) V/UV information, (e) Noise spectrum, (f) Synthetic speech spectrum.

Chapter 3

Speech Analysis

The parameters of the Multi-Band Excitation speech model are the spectral envelope, the fundamental frequency, and V/UV information for each frequency band. The methods used for estimating these parameters must be accurate and robust if high quality speech reproduction is to be achieved in both clean and noisy environments. The usual approach to extracting these parameters involves independent estimation of the excitation and system information. These algorithms are generally heuristic in nature and do not explicitly attempt to match the spectrum of the synthesized speech to that of the original. Often, the pitch structure present in spectrum interferes with the estimation of the spectral envelope and vice versa. In the MBE model an integrated approach to parameter estimation has been adopted. The excitation and spectral envelope parameters are evaluated simultaneously in an effort to model the spectrum of the original speech as accurately as possible in a least squares sense. This approach can be viewed as an "analysis by synthesis" method.

Estimation of all the model parameters simultaneously would require solving a highly non-linear optimization problem. In the interest of computational simplicity the estimation process has been divided into two steps. In the first step the speech segment is assumed to be entirely voiced. The pitch and spectral envelope are found

which minimize the difference between the original and synthetic spectra. V/UV decisions are then made based on the closeness of the fit between the spectra in each of the designated frequency bands. Areas below a difference threshold are classified as voiced, while those regions which vary exceedingly from the all-voiced synthetic spectrum are denoted as unvoiced. The harmonic magnitudes located within these unvoiced regions are set to the average magnitude of the original speech spectrum in their vicinity.

3.1 Pitch Estimation

In [1] it was shown that in order to obtain reliable voicing decisions, a high degree of accuracy is required in estimating the fundamental frequency of an analysis frame. Even a small disparity in this estimate, on order of 1 Hz, can produce large differences between the original and synthetic speech. This is especially true at the higher frequencies where small pitch errors are accentuated. To satisfy this requirement Griffin [1] developed a frequency domain approach to simultaneously estimate the pitch and spectral envelope. The technique attempts to match the synthetic and original speech spectra in a least squares sense and is capable of generating pitch estimates that are considerably more accurate than conventional algorithms [5]. While possessing sufficient resolution for this application, the frequency domain based pitch estimator requires that an error measure be calculated for each of the possible pitch periods. The computational requirements of this procedure make it unfeasible for a large range search.

Computational and accuracy considerations have necessitated the use of a two tiered approach to pitch period estimation. In the first stage an initial estimate of this figure is generated via an autocorrelation based algorithm. This technique is not sufficiently accurate to produce reliable voicing decisions, but does efficiently compute a rough estimate. The frequency domain method may then be employed

to search in fine increments over a small band centered at the initial pitch estimate. The following subsections detail each of these pitch detection algorithms.

3.1.1 Frequency Domain Pitch Detection

The parameters of the MBE model are estimated by minimizing the following error criterion:

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

where $G(\omega)$ is a frequency dependent weighting function and $|\hat{S}_w(\omega)|$ is the product of the excitation spectrum and the spectral envelope:

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

By assuming that the segment is entirely voiced with a fundamental frequency ω_0 the excitation is set equal to $P_w(\omega)$, the Fourier transform of a windowed impulse train with a periodicity of $2\pi/\omega_0$ samples. The spectrum is divided into frequency bands centered about each harmonic of ω_0 and the spectral envelope is modeled as being a constant of value A_m in the m^{th} interval. The error criterion in (3.1) may now be expressed for each harmonic interval as:

$$E_m = \frac{1}{2\pi}\int_{a_m}^{b_m} G(\omega)||S_w(\omega)| - A_m|P_w(\omega)||^2 d\omega \tag{3.3}$$

where the interval $[a_m, b_m]$ is an interval centered about the m^{th} harmonic of ω_0 and has a width equivalent to ω_0. Differentiating this expression with respect to A_m, the harmonic magnitude which minimizes the error criterion in this spectral region is evaluated to be:

$$A_m = \frac{\int_{a_m}^{b_m} G(\omega)|S_w(\omega)||P_w(\omega)|d\omega}{\int_{a_m}^{b_m} G(\omega)|P_w(\omega)|^2 d\omega} \tag{3.4}$$

The minimum error over the entire spectrum for a given fundamental frequency ω_0 and an entirely voiced excitation is then computed as:

$$E_{min}(\omega_0) = \sum_{m=0}^{M-1} E_{m_{min}} \tag{3.5}$$

where $E_{m_{min}}$ is calculated by evaluating the E_m of (3.3) with the optimal value of A_m given by (3.4). Equation (3.5) can then be used to select the best fundamental frequency out of a set of candidates.

Through this technique, the original multi-dimensional problem has been reduced to the one-dimensional problem of finding the value of ω_0 that minimizes $E_{min}(\omega_0)$. This method requires that a minimum error figure be calculated for each of the potential pitch periods and is therefore computationally burdensome. The advantage of the algorithm lies in its accuracy potential.

3.1.2 Autocorrelation Pitch Detection

In [1] an alternate means of minimizing equation (3.1) has been formulated in the time domain. This approach is approximately equal to the frequency domain algorithm, but constrained to integer pitch periods. The advantage of this method lies in the computational savings available through an efficient implementation.

The algorithm attempts to maximize the function $\Psi(P)$ given by

$$\Psi(P) = P \sum_{k=-\infty}^{\infty} \Phi(kP) \tag{3.6}$$

where P is an integer pitch period related to the fundamental frequency by $P = 2\pi/\omega_0$ and $\Phi(m)$ is the autocorrelation function of the signal multiplied by the square of the analysis window:

$$\Phi(m) = \sum_{n=-\infty}^{\infty} w^2(n)s(n)w^2(n-m)s(n-m) \tag{3.7}$$

This technique is similar to the autocorrelation method but considers the peaks at multiples of the pitch period, not just the peak at the pitch period. In practice, $\Phi(m)$ can be efficiently computed with an FFT and Ψ maximized over all integer pitch periods by summing samples of $\Phi(m)$ spaced by the pitch period. It should be noted that the summations in (3.6) and (3.7) are finite due to the finite length of the window $w(n)$.

While the integer value P that maximizes $\Psi(P)$ may be used as the initial pitch estimation, several processing steps are available for improving the accuracy and robustness of this estimate. The first of these involves the correction of a pitch related bias inherent in the error estimate. In [1] it was shown that $E_{min}(\omega_0)$ is biased such that longer pitch periods are favored over shorter ones. The expected value of this bias and a means for removing it are addressed in [1]. The continuity of the initial pitch estimate is then improved via a pitch tracker based on dynamic programming techniques. Several past and future analysis frames are incorporated in an effort to find the pitch track with the minimum total error. As a final step, the pitch estimate is checked against harmonic sub-multiples to ensure that this initial estimate is not an harmonic of the true pitch period. With these adjustments completed, the initial estimate is passed to the frequency domain pitch detector where the estimate is refined to a 1 Hz resolution necessary for the assessment of accurate voicing decisions.

3.2 Voiced/Unvoiced Decisions

One of the assumptions made while formulating the frequency domain pitch detection algorithm of section 3.1.1 was that the excitation is entirely voiced. If this assumption is correct for a given harmonic of ω_0, the difference between the original speech and the estimated spectrum over the harmonic interval will be relatively small. The voiced/unvoiced determination is therefore made by comparing the error figure, E_m of equation (3.3), to a predetermined threshold. Regions falling below this threshold are declared voiced while those exceeding it are unvoiced. In practice, the spectral regions of interest are not limited to a single harmonic. A single V/UV decision may be assigned to a group of harmonics with a similar procedure.

The value of this threshold is critical in order to obtain the correct mixing of voiced and unvoiced energy. A high threshold will produce speech that is reverber-

ant and mechanical due to the dominance of voiced harmonics. If the threshold is too low, the speech will sound hoarse and aspirated because of excessive unvoiced energy. In general, the appropriate value is determined heuristically through listening tests. It has been found that a frequency dependent threshold function produces superior speech than a fixed figure across the entire spectrum. This is motivated by the desire to limit voiced energy at high frequencies, particularly in noisy speech, while still permitting voiced harmonics at the lower end of the spectrum. In [4] a threshold function of this nature is presented. It is linear across frequency and downward sloping. The net effect is to make a voiced determination easier to achieve at the low frequencies.

3.3 Spectral Envelope Determination

The spectral envelope is assumed to be of a constant value across each harmonic interval. These harmonic magnitudes, A_m. may be thought of as the optimal scaling factor for mapping the estimated excitation spectrum into the spectrum of the original speech. For spectral regions declared voiced, the optimal value of A_m has already been determined in the process of estimating the fundamental frequency and is given by equation (3.4). For unvoiced frequency bands, the excitation spectrum is modeled as white noise of unity magnitude. For this case, equation (3.4) reduces to:

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

If the weighting function $G(\omega)$ is held constant across the spectral interval, this expression is equivalent to the average of the original spectrum in the desired frequency band.

Chapter 4

Speech Synthesis

The MBE synthesis algorithm employs separate techniques for generating the voiced and unvoiced portions of the synthetic speech. Voiced speech is synthesized in the time domain using a bank of tuned oscillators. A frequency domain approach is applied to produce the unvoiced frequency bands. These two contributions are then summed and the result is the synthetic speech. This chapter is intended to provide an overview of each algorithm. The details and the motivation behind each approach is discussed in [1].

4.1 Voiced Synthesis

For a particular speech segment, an oscillator is assigned to each harmonic which has been declared voiced. The output of the oscillator for the m^{th} harmonic may be expressed as:

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

The amplitude function, $A_m(t)$, is linearly interpolated between frames with the amplitudes of unvoiced harmonics set to zero. If the current frame is assumed to begin at $t = 0$ and the time between analysis frames is T then $A_m(t)$ is given by:

$$A_m(t) = A_m(0) + [A_m(T) - A_m(0)]\frac{t}{T} \tag{4.2}$$

where $A_m(0)$ is the m^{th} harmonic magnitude for the current segment and $A_m(T)$ is the value of the corresponding harmonic magnitude one frame into the future. This interpolation process assures a smooth transition across frame boundaries.

The phase function, $\Theta_m(t)$, is modeled with a second or third order polynomial. The coefficients of this polynomial are chosen such that $\Theta_m(t)$ is continuous across the frame boundaries at $t = 0$ and $t = T$. In addition the coefficients of $\Theta_m(t)$ are adjusted such that the frequency and phase of the m^{th} harmonic are matched at the frame boundaries. The details of this procedure are documented in [1].

For the low-rate coding problem addressed in this thesis, there are insufficient bits to code the harmonic phases. Therefore the phase matching constraint mentioned above is eliminated.

Once the oscillator parameters $A_m(t)$ and $\Theta_m(t)$ have been calculated for all the harmonics, the voiced synthetic speech for $t = 0$ to $t = T$ is generated by summing the contribution of each oscillator. The voiced speech may be expressed as follows:

$$\hat{s}_v(t) = \sum_{m=0}^{M-1} A_m(t) \cos(\Theta_m(t)). \tag{4.3}$$

4.2 Unvoiced Synthesis

In order to complete the synthesis procedure the unvoiced speech must be reconstructed. This is accomplished in the frequency domain by first calculating the spectrum of a windowed Gaussian noise sequence. Frequency bands corresponding to unvoiced harmonic intervals are then scaled by the appropriate harmonic magnitude. The remaining spectral regions corresponding to voiced harmonics are zeroed out and do not contribute any energy to the final signal. The inverse Fourier transform of this modified spectrum is calculated and the result becomes the unvoiced speech for the frame. A weighted overlap-add procedure similar to the one discussed in [9] is applied to combine the unvoiced speech signal with the unvoiced portions of neighboring analysis frames.

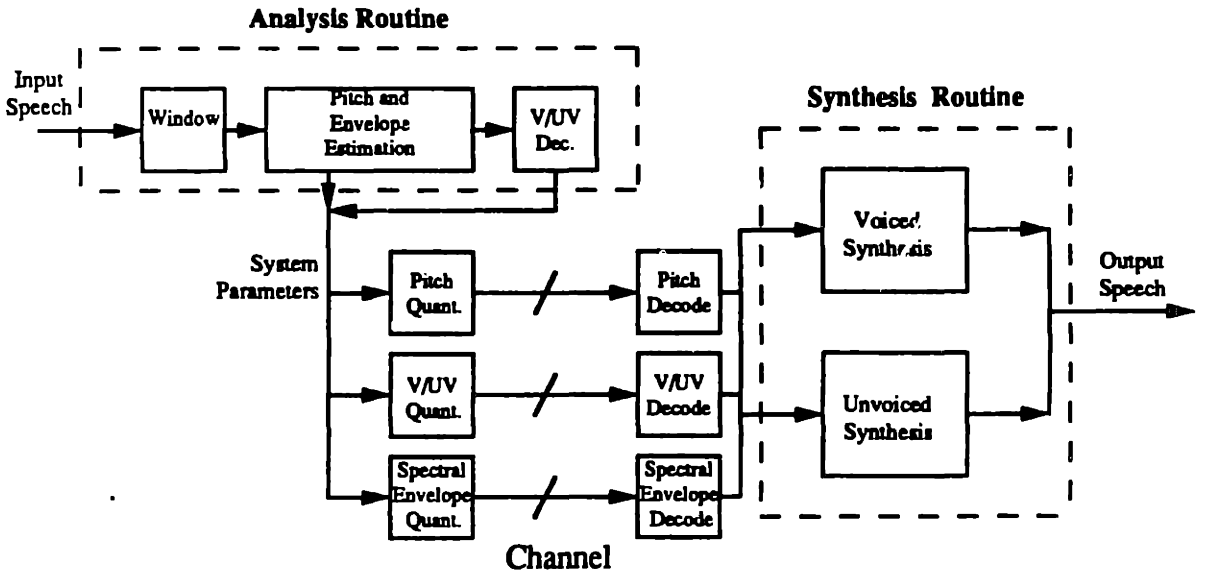

Figure 4.1: Outline of the MBE Speech Coder

The product of this procedure is added to the voiced speech component to complete the speech synthesis. Figure 4.1 presents an outline of the entire MBE speech coder. The analysis and synthesis routines have been discussed in this chapter and the last. The following chapter will concentrate on the quantization of the system parameters for the purpose of transmission over a binary communications channel.

Chapter 5

Parameter Coding

For a 30 msec frame interval there are 45 bits available for the quantization of the parameters of an entire speech frame. These parameters include the excitation information (the pitch period and voicing decisions) and the spectral information (the harmonic magnitudes of the spectral envelope). The excitation information requires a straightforward quantization and is not amenable to bit reducing techniques. The spectral information presents an entirely different situation. There are a number of methods available for dramatically improving the quantization efficiency. This chapter discusses the coding procedures employed in the 1.5 kbps system for quantizing these frame parameters.

5.1 Spectral Magnitudes

The number of harmonic magnitudes that must be quantized and transmitted for a given speech frame is a function of the estimated pitch period. This figure can vary from 12 harmonics in the case of a high-pitched voice to as much as 60 for an extremely low-pitched speaker. Roughly 32 of the 45 bits available are allocated to encoding these quantities.

In the 8.0 kbps MBE speech coder developed by Griffin[1], well over 100 bits are

available for representing the spectral envelope. At this rate ADPCM is a feasible quantization technique. The 4.8 kbps coder created by Hardwick[3] employs a variety of transform coding tactics designed to exploit the inter- and intraframe redundancies present in the harmonic magnitudes. Transform coding has proven successful at bit rates as low as 2.4 kbps, but its performance deteriorates soon thereafter. The quantization scheme required for this 1.5 kbps application must be significantly more efficient than those applied in these higher rate systems. It must also satisfy the restrictions imposed by a real-time implementation. These include computational complexity and coding delay limitations as well as channel error considerations.

Vector Quantization (VQ) was considered as a possible basis for the spectral envelope quantization. VQ achieves excellent quantization efficiency by utilizing both the linear and non-linear dependencies within a block of data. However, it demands prohibitive computational and storage requirements to achieve the desired results. These disadvantages prevented its use in this system. A more advantageous solution involves developing a mathematical model of the spectral envelope. It is then possible to represent the harmonic magnitudes with a limited set of parameters. LPC modeling appears to be a logical method to employ in this situation. Its benefits and limitations are discussed below.

5.1.1 LPC Spectral Modeling

Linear Predictive Coding (LPC) is one of the more popular forms of spectral estimation. It provides a compact yet precise representation of the spectral envelope without being computationally intensive. Its advantage in a speech coding application stems from the fact that the LPC coefficients may be effectively quantized at a low bit rate.

The LPC all-pole model has been found to accurately fit the spectral envelope of most speech segments, particularly non-nasal voiced sounds. For fricatives and

nasal sounds, the acoustic tube theory calls for both poles and zeros in the transfer function. However, if the polynomial order is adequate, all-pole modeling provides an adequate representation of these sounds as well.

Given the speech spectrum $S(\omega)$ or its power spectrum $P(\omega) = |S(\omega)|^2$, the goal of LPC spectral modeling is to fit $S(\omega)$ in some optimal manner by an all-pole spectrum $\hat{S}(\omega)$ or equivalently $\hat{P}(\omega) = |\hat{S}(\omega)|^2$. The all-pole model may be written as:

$$\hat{S}(\omega) = \frac{G}{A(\omega)} = \frac{G}{1 + \sum_{k=1}^{p} \alpha_k e^{-j\omega k}} \tag{5.1}$$

where G is a constant gain factor, p is the number of poles in the spectrum, and $A(\omega)$ is known as the inverse filter. We define an error measure E between $P(\omega)$ and $\hat{P}(\omega)$ as follows:

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) |A(\omega)|^2 d\omega \tag{5.2}$$

Note that in the time domain E is equivalent to the energy of the difference signal of $s(n)$ and $\hat{s}(n)$. An alternate derivation of the LPC equations is possible by considering the problem to be a minimization of the mean squared error between $s(n)$ and $\hat{s}(n)$.

The parameters $\{\alpha_k\}$ are determined by minimizing E with respect to α_k, i.e.

$$\frac{\delta E}{\delta \alpha_i} = 0, \qquad 1 \leq i \leq p \tag{5.3}$$

It can be shown [10] that these conditions reduce to

$$\sum_{k=1}^{p} \alpha_k R_{|i-k|} = -R_i, \qquad 1 \leq i \leq p \tag{5.4}$$

where

$$R_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) cos(k\omega) d\omega \tag{5.5}$$

This is a set of p linear equations in p unknowns which may be solved for $\{\alpha_k\}$. By exploiting the Toeplitz nature of the autocorrelation matrix, several efficient

recursive algorithms have been developed for the solution of this system. Similarly, it has been shown [10]that G may be calculated from

$$G^2 = R_{(1)} + \sum_{k=1}^{p} \alpha_k R_k \tag{5.6}$$

Equations (5.4) and (5.6) completely specify the parameters of the model spectrum $\hat{S}(\omega)$. For a spectrum $S(\omega)$ and desired number of poles, we first calculate the autocorrelation coefficients $\{R_i\}$ as specified in (5.5) and then determine $\{\alpha_k\}$ and G.

A few observations concerning the spectral matching properties of this procedure may be made. First, minimizing E is equivalent to minimization of the integrated ratio of $P(\omega)$ to $\hat{P}(\omega)$. As a result LPC modeling provides a better fit to spectral peaks than valleys. Though this property may have some advantages with spectral envelope estimation in the presence of pitch information, it creates a serious difficulty in modeling envelopes possessing a wide dynamic range. A second observation is that LPC spectral approximation is equally accurate at all frequencies. Human auditory perception has finer frequency resolution at the lower and middle regions of the audible spectrum. High resolution in the envelope approximation at the higher frequencies can result in preserving irrelevant high frequency details at the expense of the envelope approximation for the more important lower and middle range frequencies. Another consideration is the number of poles desired to achieve an accurate representation of the spectral envelope. The accuracy of the fit of $\hat{P}(\omega)$ to $P(\omega)$ increases as the order p increases. It can be shown that $\hat{P}(\omega) \rightarrow P(\omega)$ as $p \rightarrow \infty$. However, coding restrictions prevent the use of an arbitrarily large value for p. A compromise in the choice of p must be made that minimizes the total speech degradation due to the combined effects of the LPC modeling and the parameter quantization.

To this point we have only considered $S(\omega)$ to be a continuous function of frequency. However, in the majority of cases $S(\omega)$ is only available at a finite number of spectral samples. For these discrete cases we must redefine the error

measure E_d as a summation:

$$E_d = \frac{G^2}{N} \sum_{n=0}^{N-1} \frac{P(\omega_n)}{\hat{P}(\omega_n)} \tag{5.7}$$

where N is the total number of spectral points. Note that the $\{\omega_n\}$ need not be equally spaced. Following the same minimization procedure as in the continuous case, we again arrive at (5.4) for obtaining $\{\alpha_k\}$ [11]. However, the calculation of the autocorrelation coefficients must be redefined for the discrete case as

$$R_k = \frac{1}{N} \sum_{n=0}^{N-1} P(\omega_n) cos(k\omega_n). \tag{5.8}$$

When the $\{\omega_n\}$ are harmonics of a fundamental frequency ω_0 this expression reduces to:

$$R_k = \frac{1}{N} \sum_{n=0}^{N-1} P(\omega_n) cos(kn\omega_0). \tag{5.9}$$

This procedure suffers from a number of limitations. The most obvious is that the modified distortion measure E_d is a function of only the discrete frequencies $\{\omega_k\}$. The $\hat{P}(\omega)$ that is obtained is a minimization of the error between $\hat{P}(\omega)$ and $P(\omega)$ at only these frequencies. At other spectral locations, the performance of $\hat{P}(\omega)$ is not predictable.

If $P(\omega)$ is sparsely sampled, $\hat{P}(\omega)$ generated by this procedure is generally a poor estimate of the original spectral envelope. For an adequate spectral fit the number of frequency points must be large compared to the number of desired poles. This can be a significant problem with high-pitched voices. Figures 5.1 and 5.2 demonstrate this point. In figure 5.1 the solid curve is a 16-pole spectrum that has been sampled at $.019\pi$ radian intervals. At 8 kHz sampling this corresponds to a pitch of 78 Hz, which is not unusual for male speakers. The dashed curve is the 16-pole LPC model generated from these discrete spectral points. In figure 5.2, the same 16-pole spectrum is sampled at $.075\pi$ radians, equivalent to a 300 Hz pitch, and again the 16-pole spectrum is computed from the discrete samples. In the first case the estimated spectrum is a good fit to the original envelope. In the second

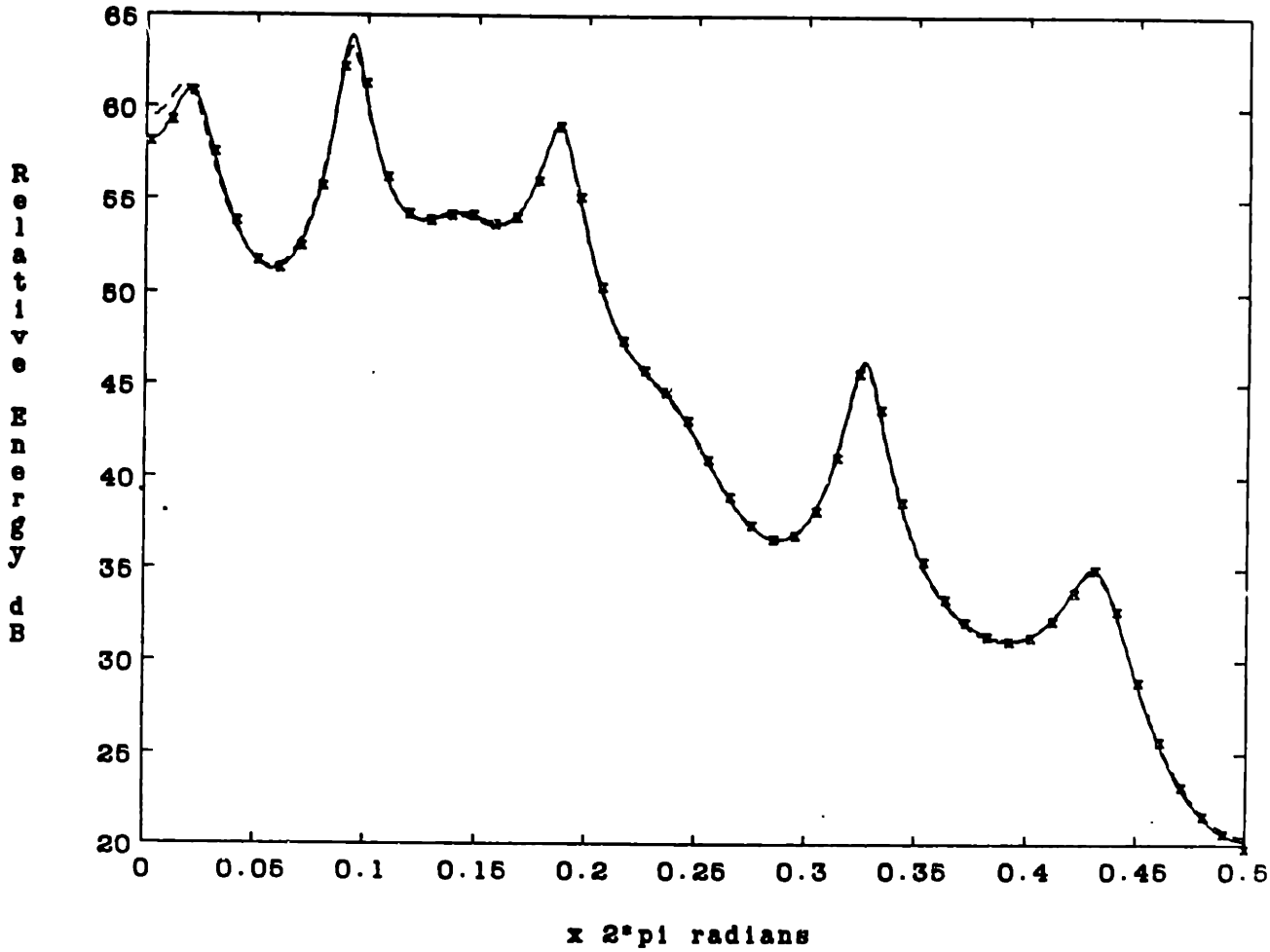

Figure 5.1: A 16-pole spectrum (solid line) and its 16-pole LPC model generated using frequency points sampled at $.019\pi$ radian intervals(dashed line).

example the spectrum is undersampled and a poor match is achieved. The types of discrepancies that can occur between the model and original spectrum in this situation include merging or splitting of pole peaks, and increasing or decreasing of pole frequencies and bandwidths [11]. Not only has the envelope estimation been corrupted, but the values of the estimated and original spectra at the sampled frequencies vary significantly.

This is a rather unfortunate result that stems from the correlation matching condition imposed by the LPC error criterion. LPC modeling equates the first $p+1$ autocorrelation coefficients of the original spectrum and the model all-pole spectrum. The $\{R_k\}$ calculated with (5.9) are an aliased version of the original

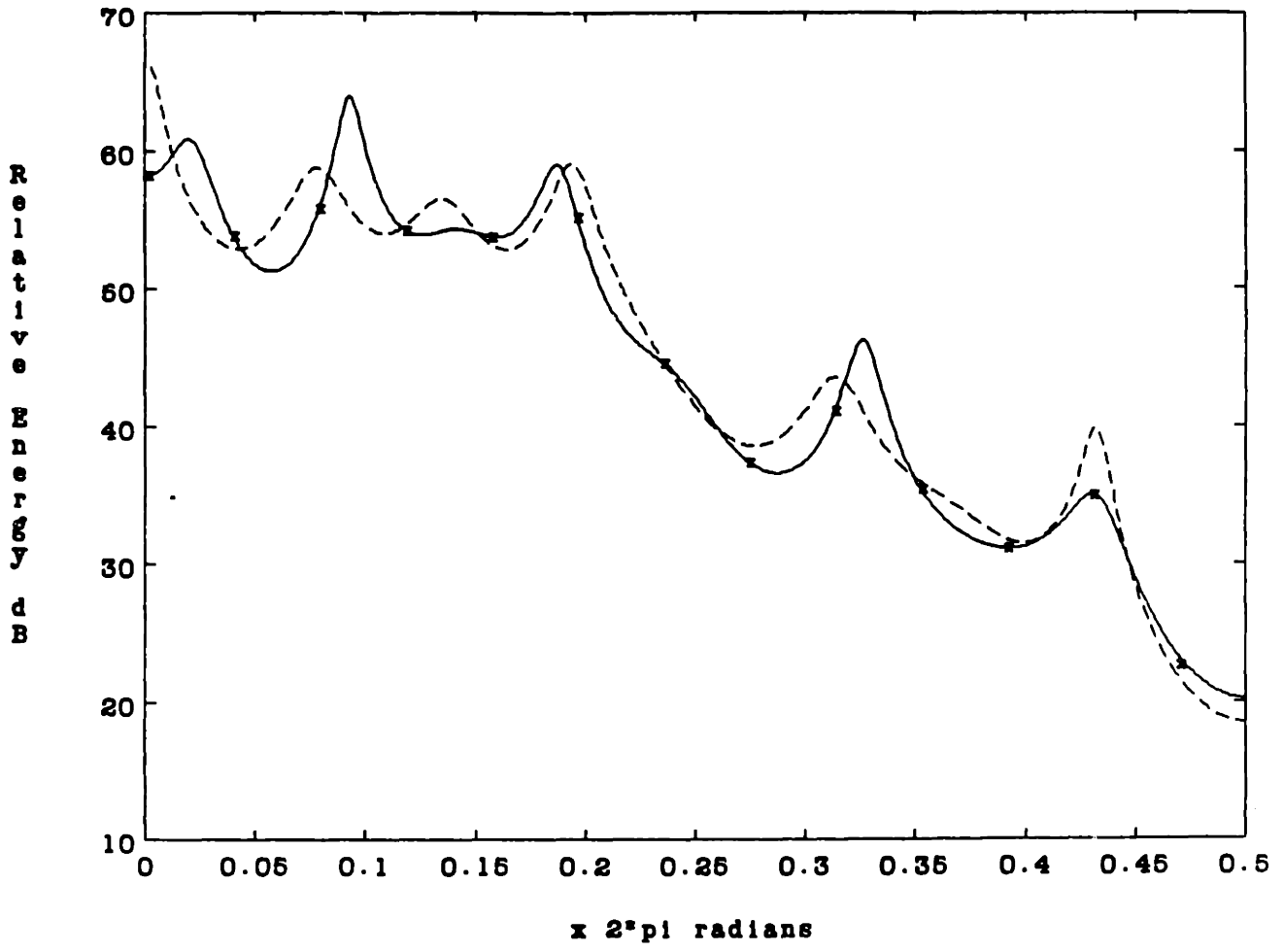

Figure 5.2: The 16-pole spectrum from figure 5.1 and its LPC estimation generated with a $.075\pi$ radian sampling interval.

autocorrelation coefficients and as a result the model will never match the original all-pole envelope. For low-pitched speech the aliasing of $\{R_k\}$ is small and the modeled spectrum is reasonable. With high-pitched speech this aliasing is severe and the resulting envelopes provide a poor fit to the original.

Other error criteria such as the Takura-Saito distortion measure have been shown to obtain better estimates than this procedure [12]. However, these algorithms require solving a set of non-linear equations that are not guaranteed to globally converge. More importantly, these methods do not perform as well as the interpolation technique to be discussed shortly.

We will now explore LPC spectral estimation in the context of mathematically

modeling the harmonic magnitudes generated by the MBE analysis. Computation of the autocorrelation coefficients directly from the discrete harmonic magnitudes has proven ineffective in all but the lowest pitched speech frames. Some sort of interpolation of the discrete spectral envelope is required to limit the aliasing effects described above. One solution is to generate the $\{R_k\}$ directly from the FFT of the original windowed speech segment. This assures that the spectral envelope has been sampled finely and the aliasing problem is corrected, but suffers from the pitch structure still present in the signal. The LPC must model the product of the spectral envelope with the excitation spectrum. In doing so the estimate of the envelope is degraded. This procedure fails to take advantage of the deconvolution properties of the MBE analysis. It would seem more logical to synthetically interpolate the spectrum directly from the estimated harmonic magnitudes themselves. In this way the pitch structure is removed and the LPC model is a direct estimate of the harmonic magnitudes.

Several interpolation methods have been pursued. These range from simple linear interpolation to averaged parabolic interpolation in the log spectral domain. The desired technique should generate a spectrum with LPC like qualities and still be computationally reasonable. Linear interpolation in the log spectral domain appears to be a suitable compromise [13]. The interpolated spectrum $Q(\omega)$ is calculated from

$$\log|Q(\omega)| = \log|S(\omega_k)| + [\log|S(\omega_{k+1})| - \log|S(\omega_k)|] \times \frac{\omega - \omega_k}{\omega_{k+1} - \omega_k} \qquad (5.10)$$

for $\omega_k \leq \omega < \omega_{k+1}$ where ω_k is the frequency of the k^{th} harmonic.

Figure 5.3 demonstrates the results obtained with this algorithm. The solid line is the 16-pole spectrum from figure 5.2 which has again been sampled at $.075\pi$ radian intervals. The harmonic samples are then interpolated via (5.10) and the result is used to calculate $\{R_k\}$ required for LPC analysis. The dashed line is the new 16-pole LPC estimate. This new estimate does an even worse job of tracking the envelope as a whole than its discretely generated counterpart in figure 5.2. Fortunately,

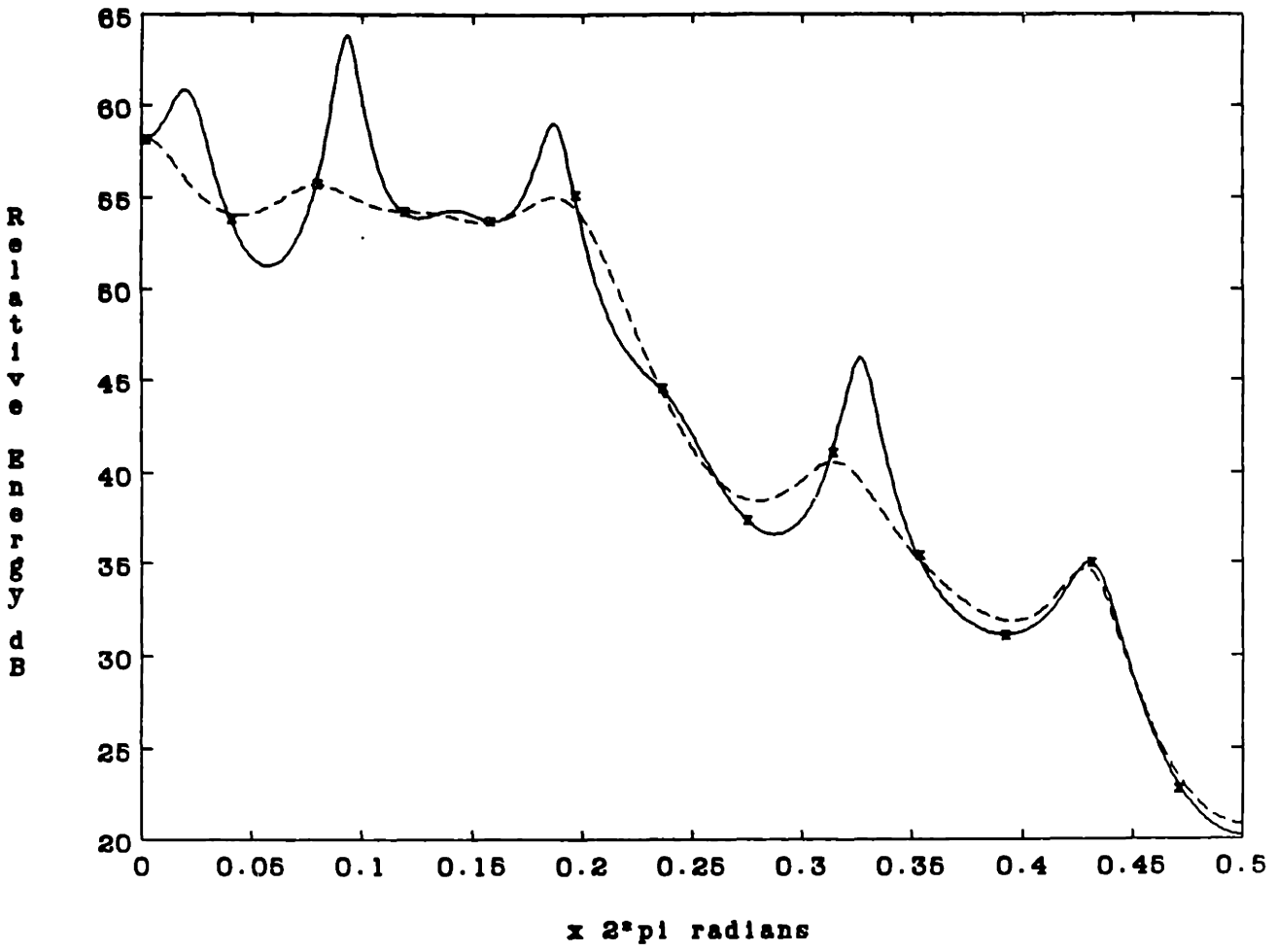

Figure 5.3: A 16-pole spectrum (solid line) sampled at $.075\pi$ radian intervals and its 16-pole LPC estimate (dashed) produced via linear interpolation in the log spectral domain.

we are not concerned with preserving the shape of the original spectrum, only with calculating an LPC model from which the harmonic magnitudes can be extracted. In this respect, the new estimate clearly outperforms the discretely generated model. The new LPC model accurately follows the harmonic magnitudes while missing everything else of interest in the spectrum. Note that there has been no attempt here to reconstruct the original 16-pole spectrum. Rather the objective has been to find the 16-pole LPC model that minimizes the difference between the original and estimate spectras over a series of discrete frequency points. This technique does not require the original spectrum to be all-pole in nature, and it is therefore extendable to modeling general speech spectra.

LPC Order	Spectral Distortion (dB^2)
6	3.69
8	2.37
10	2.11
12	0.48
14	0.38
16	0.14
18	0.12

Table 5.1: Average spectral distortion between the 16-pole spectrum in figure 5.3 and the LPC estimates of various analysis orders

The accuracy of the curve fit is a reflection of the number of poles used in the LPC analysis. Increasing p improves the model's representation of the harmonic magnitudes. An example of this is given in table 5.1. Here the same spectrum and harmonic samples as in figure 5.3 are used to generate LPC spectra of various orders. The degree of the spectral fit between the original and model harmonics is given by the average spectral distortion measure. Given two spectra, $S_1(\omega)$ and $S_2(\omega)$ known at N frequency points $\{\omega_k\}$, we define the average spectral distortion between the two as

$$S.D. = \frac{10^2}{N} \sum_{k=0}^{N-1} (\log |S_1(\omega_k)| - \log |S_2(\omega_k)|)^2 \quad dB^2. \tag{5.11}$$

This error measure was chosen for this application because of its ease of calculation and its good correspondence with subjective measures. In what follows, the average spectral distortion scores will be used as a means of comparing the relative quality of synthesized speech. Given a speech passage and two synthesized versions of the original speech, the segment with the lower spectral distortion score is, with few exceptions, perceived as the closer match of the two to the original. No attempt has been made here to relate these scores to any established absolute speech quality measures.

Two more steps have been taken to improve the accuracy of this procedure. Each stems from the observations made earlier concerning the spectral matching properties of LPC analysis. The first limitation dealt with the dynamic range of the original spectrum. If the range is too large, the LPC model will tend to favor the larger magnitude harmonics in its envelope estimation. To reduce these effects the original envelope is compressed prior to the LPC analysis. A typical compression method involves taking the cube root of each harmonic [13]. The second improvement attempts to account for the perceptual properties of the human auditory system. While the LPC spectral approximation is uniformly accurate across frequency, the human ear is more sensitive to low than to high frequencies. By warping the spectral axis during the interpolation process, it is possible to devote a larger portion of the total spectrum to the lower frequency regions while deemphasizing the less critical higher harmonics. The mel scale is a suitable warping function that has a basis in psychoacoustic theory [14].

Each of the steps described above has some effect on the overall fit of the LPC model to the original harmonic magnitudes. These results are summarized in table 5.2. Three thousand frames of speech representative of male and female voices under a variety of speaking conditions were processed with the MBE analysis algorithm. The harmonic magnitudes of each frame were then modeled with an 18^{th} order LPC polynomial. The autocorrelation coefficients were first calculated directly from the discrete harmonics. Each of the methods described above were then added in succession to the estimation procedure. The average spectral distortion statistics for each step in the experiment are displayed in table 5.2.

As the table indicates, the efforts at improving the LPC model estimate have been effective. Listening tests confirm this result. There is a marked audible improvement in the quality of the speech segments synthesized with the enhanced LPC estimate over those synthesized with the simple LPC estimate based solely on the unmodified discrete spectral harmonics. Listening tests also indicate that an

Estimation Method	Spectral Distortion (dB^2)	
	Mean	Standard Dev.
discrete harmonics	4.95	3.20
interpolation	3.95	2.71
interpolation and compression	3.46	2.15
interpolation, compression, and frequency warping	2.66	1.47

Table 5.2: Effects of enhancement methods on the spectral distortion scores of 3000 frames of speech

18-pole model is optimal. Speech quality improves with polynomial order up to 18 poles, beyond this value there is little to no audible benefit.

5.1.2 LPC Quantization

A number of methods exist for efficiently quantizing the LPC polynomial. The LPC coefficients $\{a_k\}$ from (5.1) are inappropriate for quantization because of their relatively large dynamic range and possible filter instability problems. Several alternative representations of the LPC coefficients exist and possess properties that make them quite amenable to coding applications. One such representation is the line spectrum pair (LSP). The LSP parameters have a limited dynamic range, do not suffer from instability problems, and can encode the LPC spectral information more efficiently than many other parameters. They have been shown to achieve a 30% bit rate savings over the popular log-area ratio LPC representation while producing equivalent speech quality [15].

For a given order p the LPC analysis produces the inverse filter

$$A(z) = 1 - \alpha_1 z^{-1} + \alpha_2 z^{-2} + \ldots + \alpha_p z^{-p}. \tag{5.12}$$

The LSP procedure involves defining two $(p+1)st$ order polynomials $P(z)$ and $Q(z)$ as

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{5.13}$$

$$Q(z) = A(z) + z^{-(p+1)} A(z^{-1}). \tag{5.14}$$

The function $A(z)$ may be reconstructed from $P(z)$ and $Q(z)$ by

$$A(z) = \frac{P(z) + Q(z)}{2}. \tag{5.15}$$

Equations (5.13) and (5.14) may be rewritten as

$$P(z) = (1 - z^{-1}) \prod_{i=2,4,\ldots p} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{5.16}$$

$$Q(z) = (1 - z^{-1}) \prod_{i=1,3,\ldots p-1} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \tag{5.17}$$

where $\omega_2 < \omega_4 < \ldots < \omega_p$ and $\omega_1 < \omega_3 < \ldots < \omega_{p-1}$. The roots of $P(z)$ and $Q(z)$ are of the form $e^{j\omega_i}$. The frequencies $\{\omega_k\}$ are known as the LSP parameters. The functions $P(z)$ and $Q(z)$ have two important properties [15]. All their roots lie on the unit circle and these roots alternate with each other along the unit circle. More specifically,

$$0 = \omega_0 < \omega_1 < \omega_2 < \ldots < \omega_p < \omega_{p+1} = \pi. \tag{5.18}$$

Note that $\omega_0 = 0$ and $\omega_{p+1} = \pi$ are fixed roots of $P(z)$ and $Q(z)$, respectively.

Calculating the LSP parameters involves solving two $\frac{p}{2}^{th}$ order polynomials. Computational savings can be achieved through the use of an iterative search algorithm or a DCT, but this method remains computationally intense in comparison with other LPC representations.

The ordering property and the limited range of the LSP parameters make them ideal for quantization schemes. It has also been observed that perturbing any given

LSP frequency produces highly localized effects in the resulting spectrum. It is therefore possible to regulate the quantization resolution with frequency. These properties dictate that a coding scheme which utilized the frequency differences between parameters, instead of the parameters themselves, would be effective. A number of these differential coding schemes exist, and they outperform non-differential coding schemes of similar complexity. These algorithms generally require between 30 and 40 bits to quantize a 10^{th} order LPC with 1 dB^2 average spectral distortion [16]. The drawback of differential quantizers lies in their sensitivity to channel errors. A single bit error will radically alter an entire speech frame.

The ordering property indicates that the LSP parameters within a frame are statistically dependent. Results show that there is a strong correlation between neighboring parameters within a frame as well as between the LSP parameters of adjacent frames. A coding technique which exploits these correlations would maintain enhanced coding efficiency. One such method is the discrete cosine transform (DCT). The DCT has been used successfully in image compression applications to produce nearly uncorrelated transform coefficients which may be encoded with greater efficiency. The DCT is not an optimal transform, but its computational requirements, data independence, and good decorrelation properties have contributed to its popularity in many applications. In [17] a 10^{th} order LSP quantizer is presented which is capable of achieving a $1dB^2$ average spectral distortion with only 21 quantization bits per frame. This coder employs a 2-dimensional DCT to fully exploit the intra- and inter-frame data redundancies. This algorithm requires a coding block ten frames in length and therefore possesses a coding delay which is excessive for a real-time application. A more realistic hybrid DCT-DPCM coding scheme is also discussed in [17]. This algorithm performs a 1-dimensional DCT to reduce intra-frame correlations followed by DPCM coding of the transform coefficients to utilize the interframe correlations. Twenty-five bits per frame are required to achieve a $1dB^2$ spectral distortion. While this scheme does not fully exploit the

temporal and spectral redundancies, it does achieve a minimal coding delay.

The DCT-DPCM quantization scheme was found to be the most effective coding scheme available given the real-time limitations and has been adopted for this application. We have performed experiments in which the DCT was replaced with an optimal Karhunen-Loeve transform (KLT). The performance increase was marginal and did not justify the extra computational expense or storage required for the KLT implementation. The DCT coefficients are reasonably insensitive to channel errors. When necessary, the more important coefficients may be protected against corruption. Under excessively poor channel conditions the DPCM stage may be removed entirely to limit the data degradation.

The spectral distortion scores provide a quantitative means of evaluating the error introduced in the LPC estimate during the coding process. The $1dB^2$ value is generally considered to be the cutoff beyond which the spectral errors become audible. The bit rates and average spectral distortion errors quoted above were all achieved with a 10^{th} order LPC. The optimal LPC order appears to be 18 poles. As one would expect, at a given bit rate quantizer performance degrades with increasing LPC order. Figure 5.4 illustrates this point. Here we have plotted spectral distortion scores versus bits per frame for a number of LPC polynomials. These figures were achieved using a hybrid DCT-DPCM LSP coder with Gaussian quantizers. Bit allocation is performed via an optimal allocation rule presented in [18] which is based on the variances of the individual transform coefficients. The analyzed speech consisted of 3000 frames of assorted male and female voices under a variety of speaking conditions. The 18^{th} order LPC crosses the $1dB^2$ mark at 35 bits per frame, well above the number of bits available. Therefore, if an 18-pole LPC is employed to model the harmonic magnitudes audible distortion will be introduced in the quantization process. On the other hand, a 10-pole LPC will not suffer any perceptible degradation during the quantization, but does not provide as accurate an estimate of the original spectral harmonics. Optimizing the synthesized speech

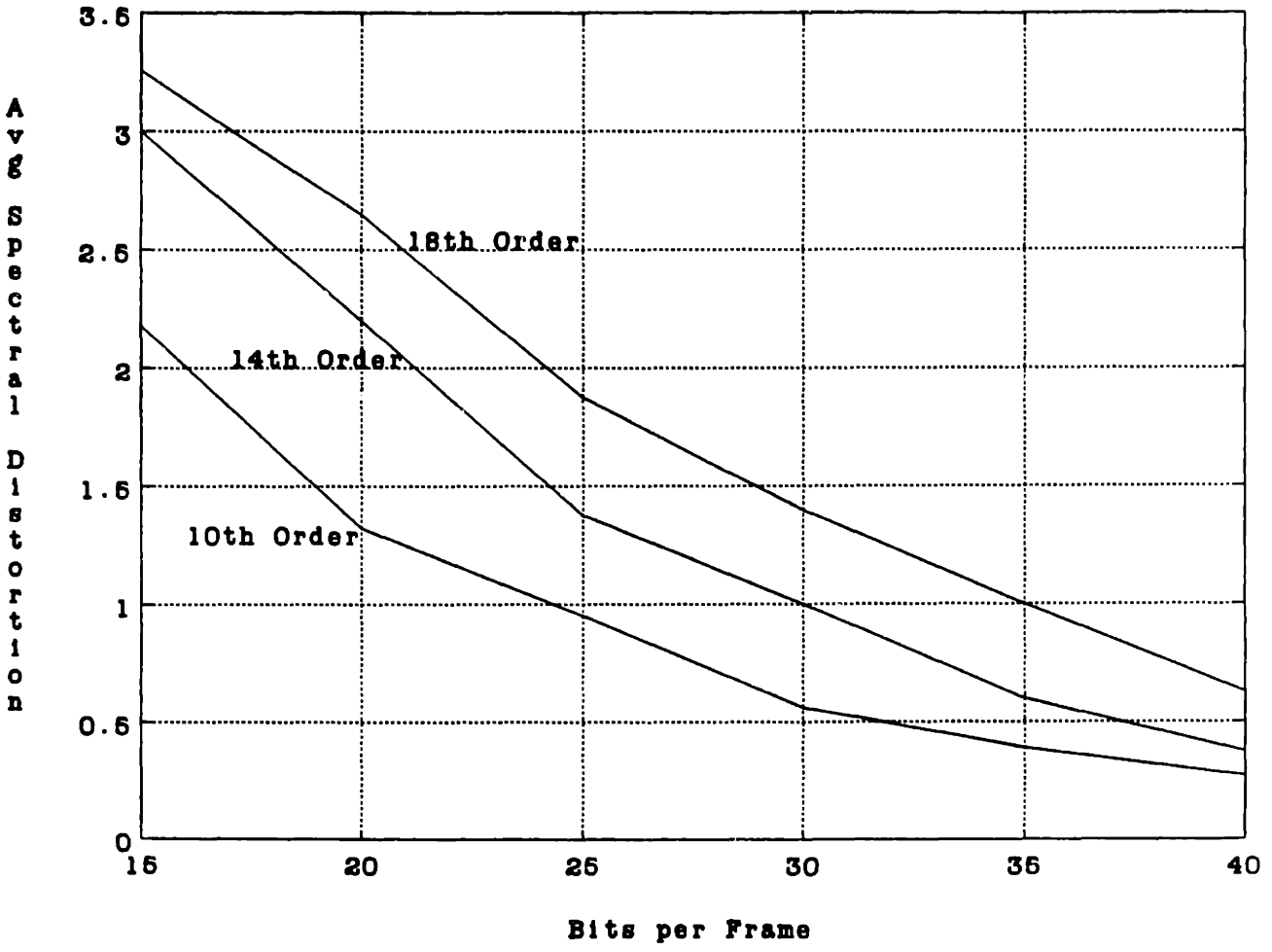

Figure 5.4: Quantizer Performance as a function of bit rate and LPC order

quality requires finding the LPC polynomial order that minimizes the combined degradation effects of the channel quantization and the modeling estimation. If the LPC gain parameter is differentially quantized in the log domain, 5 bits will be required, leaving approximately 27 bits for encoding the LSP parameters. At this rate, a 16^{th} order model was found to generate the highest quality synthesized speech and has been adopted for this application.

The entire procedure for encoding the harmonic magnitudes has been summarized in figure 5.4.

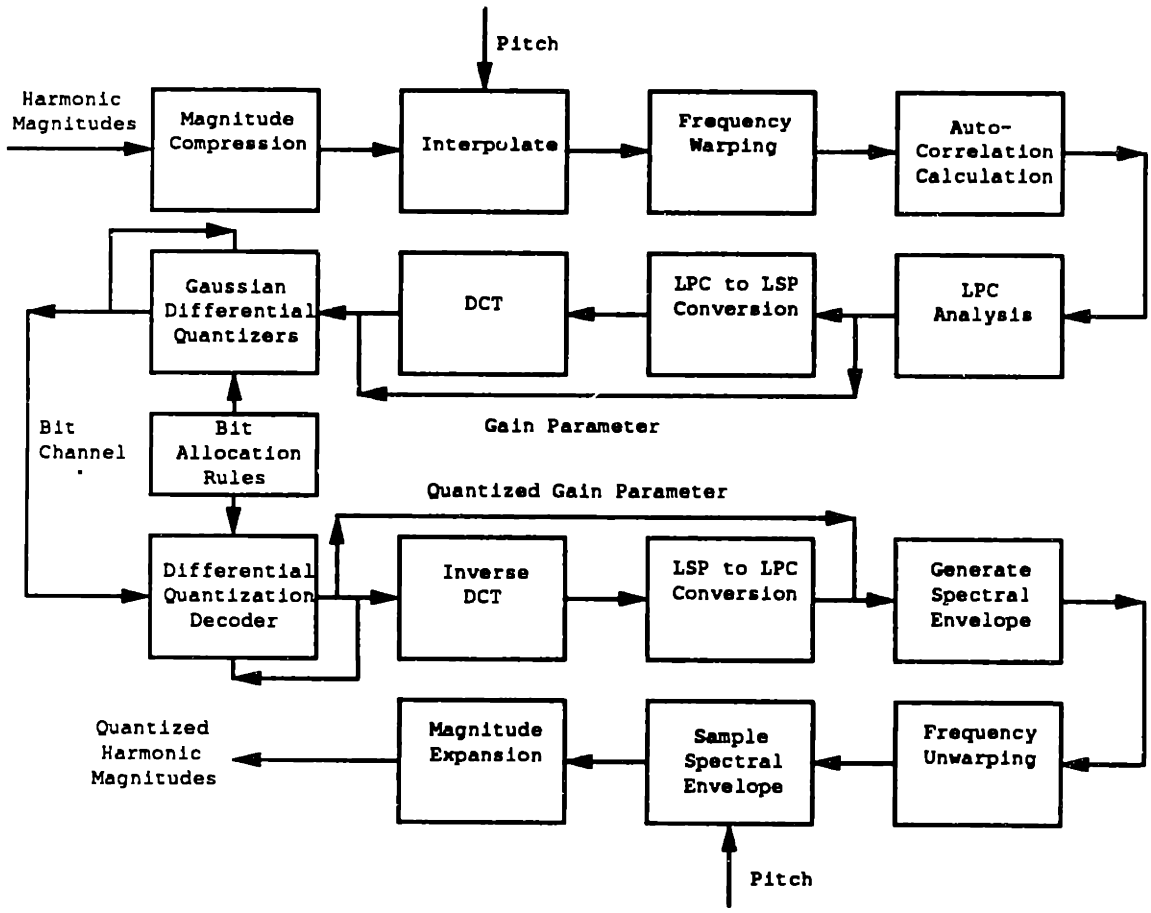

Figure 5.5: Outline of the procedure for quantizing the harmonic magnitudes

5.2 Fundamental Frequency

The MBE analysis algorithm estimates the fundamental frequency to a 1 Hz resolution in the frequency range 70 Hz to 370 Hz. This procedure itself may be viewed as a quantization process of the speech frame pitch. Further quantization of the parameter would produce audible effects in the synthesized speech [1]. Therefore, fixed length coding of the absolute fundamental frequency requires a quantizer with a minimum of 300 levels or 9 bits. An analysis of several speech segments reveals little variation between the fundamental frequencies of adjacent analysis frames. With a 30 msec frame interval approximately 88% of these interframe differences are less than 16 Hz. This result is due to slow time-varying nature of speech and the constraints imposed by the pitch tracking portion of the estimation algorithm. As

discussed in [4] a discrete Markov model has the potential to exploit this property to reduce the average number of bits required to encode this parameter. However, the number of Markov states necessary was found to increase the algorithm complexity significantly while saving just a few bits. A simpler but less efficient method involves having two encoding modes. If the fundamental frequency difference is less than 16 Hz, differential quantization is performed, requiring 5 bits. Otherwise, absolute quantization is done with 9 bits. This technique has a long term average of 5.48 bits/frame, but requires an additional mode bit, bringing this figure to 6.48 bits. This scheme achieves only marginal savings over fixed length coding, but has been adopted for this application. The accuracy of the fundamental frequency received at the MBE decoder/synthesizer is extremely critical for proper speech synthesis and for correctly decoding the other system parameters. Under poor channel conditions it is more economical to use an absolute encoding scheme with limited bit protection than to attempt to apply the differential quantizer described above.

5.3 Voicing Decisions

The voicing information consists of a set of binary classifications indicating the nature of different spectral regions. The size and structure of these regions has varied in past coders. In Griffen's 8 kbps system[1], the 4 kHz bandwidth is divided into 12 equal regions which are then individually classified as voiced or unvoiced. The 4.8 kps vocoder[4] makes these classifications in blocks of 3 consecutive harmonics. The number of voicing decision bits is limited to 12 and any harmonics beyond the 36^{th} are denoted unvoiced. These two schemes were found to produce nearly equivalent sounding speech. In the interest of bit savings, two voicing region partitioning methods have been investigated for this application. The first is similar to the 4.8 kbps method. Decisions are made in groups of 5 harmonics up to the 25^{th} harmonic and are declared unvoiced thereafter. The second is motivated by the observation

that the voiced harmonics tend to reside in the lower end of the spectrum while the unvoiced harmonics appear to be clustered in the upper spectral regions. The voicing decision consists of a single cutoff harmonic below which all the harmonics are voiced and above which they are unvoiced. The first scheme requires 3 to 5 bits/frame depending on the fundamental frequency and sounds equivalent to a 1.5 kbps coder with no voicing approximation. The second method was found to achieve comparable results with as little as 8 possible cutoff harmonics. The algorithm first makes voicing decisions in blocks of 3 harmonics and then choses the highest frequency voiced block as the cutoff. With only 3 bits required for implementation, this appears to be the method of choice.

5.4 Summary

Table 5.3 summarizes the bit allocations for the parameter coding techniques discussed in this chapter. A 30 msec frame interval has been chosen for the 1.5 kbps system. The choice of a frame rate is viewed as a compromise between the degradations produced by lowering the analysis rate and those introduced by the quantization process. Lowering the rate increases the number of bits available for encoding a given speech frame while reducing the sensitivity of the analysis algorithm to transient phenomena. At this rate, 45 bits are available for encoding the parameters of a single frame. The fundamental frequency requires either 5 or 9 bits depending on the state of the mode bit. The voicing decisions need 3 bits. The rest are allocated to the harmonic magnitudes. The LPC gain receives 5 and the LSP parameters use the remainder.

MBE Parameter	Number of Bits	
	Mode=0	Mode=1
Fundamental Frequency:		
mode bit	1	1
coding	5	9
Voicing Decisions	3	3
Harmonic Magnitudes:		
LPC gain	5	5
LSP parameters	31	27
Totals	45	45

Table 5.3: Parameter coding bit allocations

Chapter 6

Performance Results

The evaluation of model based speech coders has focussed on two principal issues. The first is intelligibility, or the degree to which basic speech sounds can be communicated through the system. The second is speech quality. This generally refers to the pleasantness or naturalness of the synthetic speech and the recognizability of the talker. At this point informal tests have been performed on the 1.5 kbps speech coder to evaluate each of the these attributes. This chapter will discuss the evaluation procedures and present the performance results.

6.1 Speech Intelligibility

Early attempts at measuring speech intelligibility were based on articulation tests in which a speaker was asked to recognize units of the transmitted speech. These speech units could be phonemes, monosyllables, words, or whole sentences. Because semantic and syntactic information enhance comprehension, these tests were limited in their ability to isolate particular speech features. The rhyme test was developed by Fairbanks [19] to eliminate the effects of context in evaluating intelligibility. By constraining the test sequence to monosyllabic words and limiting the listener responses to a set of rhyming words it is possible to test the intelligibility of a single

phoneme. The rhyme test was modified by Voiers [20] in the Diagnostic Rhyme Test (DRT). In the DRT the listener is limited to a choice of two possible words. Not only do these words rhyme, but they are further restricted to differ in just one distinctive feature of the leading phoneme. For example, the word pair *vast* - *fast* differ only by the presence of voicing in the initial phoneme of the first word. In this way the DRT measures not only the overall intelligibility of speech but also the types of errors that are prevalent in a given system. Six features are measured by the DRT: voicing, nasality, sustenation, sibilation, graveness, and compactness. Details of these feature classes are presented in [21].

One DRT test consists of 200 test words spoken by a single speaker at a rate of one every 1.33 seconds. The word pairs are presented visually either by use of a test booklet or a computer terminal. The features are tested in a specific sequence and the ordering of the words within the word pair is varied to eliminate any bias due to position. Scores may be obtained for the total test or for each feature. All scores are corrected for random guessing as follows:

$$S_j = \frac{R_j - W_j}{T_j} \times 100 \tag{6.1}$$

where S is the adjusted percent correct, R is the number of correct responses, W is the number of incorrect responses, T is the total number of tested items, and the subscript j refers to the distinctive feature class. With this correction, a listener who fills out a form at random will on the average obtain a score of zero.

An informal DRT test was conducted in our laboratory using DRT master tapes supplied by the Rome Air Development Center Speech Laboratory. The listening group consisted of 10 untrained native American speakers, 7 male and 3 female. Each was briefly familiarized with the testing equipment and procedure. The DRT tests were limited to the clean speech of a single male speaker for both the unprocessed recordings and the 1.5 kbps processed material. The results are presented in table 6.1. For these scores the high and low outliers have been eliminated and the remaining 8 scores were used to estimate the mean and standard deviation.

Feature	Mean Score	Std. Dev.
Voicing	98.6	3.2
Nasality	92.0	4.5
Sustenation	73.5	5.3
Sibilation	94.6	4.5
Graveness	88.0	6.0
Compactness	94.0	4.8
Total Coded	90.1	1.9
Total Uncoded	97.3	.55

Table 6.1: DRT Scores for 1.5 kbps system in clean speech

The overall DRT scores for the original clean speech and the 1.5 kbps coded material were calculated to be 97.3 and 90.1, respectively. There are two features with which the 1.5 kbps system achieves a score below 90. These are sustenation and graveness. Sustenation refers to the affrication feature of the phoneme. If the phoneme is articulated without a complete closure of the oral cavity then it is sustained, otherwise it is called interrupted. The initial consonant in the word *shoes* is sustained while the initial consonant of *choose* is not. Graveness has to do with the place of articulation. Grave phonemes are articulated primarily at the lips. The *p* in *pot* is grave and the *t* in *tot* is non-grave. The common characteristic of these two categories is the short duration in time of the acoustic cue necessary to distinguish each of these features. If this duration is small relative to the analysis window size, the proper cue will go undetected. This sort of limitation is common to many model based speech coders, particularly low rate systems where the bit constraint necessitates the use of a large analysis window.

Table 6.2 is offered as a means for comparing the 1.5 kbps system DRT results to those of several speech coders. The 8.0 kbps MBE system was developed by Griffin and has been presented in [1]. The 4.8 kbps MBE score is a result of a government

Speech Coder	DRT Score	Category
8.0 kbps MBE	96.2	Very Good
4.8 kbps MBE	91.3	Very Good
2.4 kbps MBE	90.1	Good
2.4 kbps STC	90.1	Good
2.4 kbps LPC10	90	Good
2.4 kbps Channel	85	Moderate
1.5 kbps MBE	90.1	Good

Table 6.2: DRT Scores for several speech coders

evaluation of 4800 bps voice coders [22]. The DRT scores for several 2.4 kbps are then listed. The first is a 2.4 kbps MBE coder recently produced by Meuse. The details of this vocoder and this DRT result are available in [24]. The second is a Sinusoidal Transform Coder (STC) based on the sinusoidal analysis/synthesis system developed by McAulay [25]. The third is the LPC-10 algorithm which has been adopted as the government 2.4 kbps standard [23]. Finally, a 2.4 kbps channel coder has been evaluated and presented in [20].

To aid in interpreting DRT results, categories of performance have been associated with these intelligibility scores [26]. Between 96 and 100 is regarded as "excellent", 91-96 is "very good", 87-91 is "good", 83-87 is "moderate", 79-83 is "fair", 75-79 is "poor", 70-75 is "very poor", and below 70 is considered unacceptable. Under this criteria both the 4.8 and 8.0 kbps systems achieve "very good" status. All the lower rate coders listed, with the exception of the channel coder, are considered "good". Several conclusions are apparent from this data. First, the 1.5 kbps speech coder described in this thesis has intelligibility equivalent to that of systems with nearly twice its bit rate. This is testament to the utility of the extra voiced/unvoiced bands in the MBE model and to the compactness of the model parameters. Second, the scores for the 1.5 kbps MBE coder and Meuse's 2.4 kbps

system are identical. Each vocoder is based on the MBE speech model. While there may be some disparity in the parameter estimation and speech synthesis procedures employed in the individual systems, this tends to indicate that a superior coding efficiency has been achieved by the lower bit rate application. This result demonstrates the merits of the coding techniques detailed in the previous chapter.

The DRT scores reported here have been generated under limited conditions and with unexperienced testing subjects. These current results are intended as a rough means of evaluation and comparison for the coder discussed in this thesis. More thorough third party DRT tests will be performed in the future. It should be noted that listener experience has a significant effect on DRT ratings. In the past, the scores produced by expert third party listeners have been superior to those evaluated informally in our laboratory for the same speech material.

6.2 Speech Quality

Several methods exist for rating speech quality. The most widely used direct method of subjective quality evaluation is the category judgement method which produces a mean opinion score (MOS). In this method, listeners rate the speech under test on a five-point scale ranging from excellent to unsatisfactory. Subjects are trained by a set of reference signals that exemplify each of the judgement categories, but are otherwise free to assign their own perceptual criteria to the evaluation. While this makes the test applicable to wide variety of speech distortions, it suffers greatly from personal bias. This limitation has motivated the development of several indirect judgement tets that rate speech over a range of specific perceptual qualities. The most widely accepted of these is the Diagnostic Acceptability Measure (DAM) which evaluates a speech signal on 16 separate scales encompassing signal, background, and total quality. Some of the class descriptors are "fluttering", "crackling", "muffling", "buzzing", and "hissing". Details of the DAM test are available in [27]. The

popularity of the DAM stems from its fine-grained parametric scoring, its reliability, and its consistency.

Both of the evaluation techniques discussed above require sufficient recording preparation and listener training to make their implementation in our laboratory unfeasible. While an official DAM test will be performed in the future, an interim means of evaluating the 1.5 kbps MBE coder speech quality has been devised. This involves a side-by-side comparison of the coder to an existing system. The reference system chosen for the study is the LPC-10e algorithm which has been adopted as the government's 2.4 kbps standard and has been implemented in the STU-III secure phone. While this study will not yield any absolute results, it will provide a direct perceptual comparison of this system to a well known and widely available vocoder.

The test tape generation procedure is outlined in figure 6.1. For the LPC processed speech, a source tape is played through a speaker into the handset of a STU-III, encoded with the LPC-10e algorithm and transmitted over phone lines to a second secure phone where it is decoded and recorded directly, bypassing the handset. For the 1.5 kbps material, the source tape is again played into the STU-III handset but transmitted without LPC processing. At the receiver end the unprocessed signal is recorded and used as input to software that simulates the 1.5 kbps MBE coder. The motivation behind this design is to recreate standard operating conditions as accurately as possible. The original audio material is of less than studio quality and the the original handset microphone and apparatus have been retained. The speech signal used as input to the MBE coder is subject to the same analog conditions available to the LPC algorithm.

Three sets of speaker conditions were evaluated:

- Male speaker with a quiet background (15 sentences).
- Female speaker with a quiet background (10 sentences).
- Male speaker with a noisy background (10 sentences).

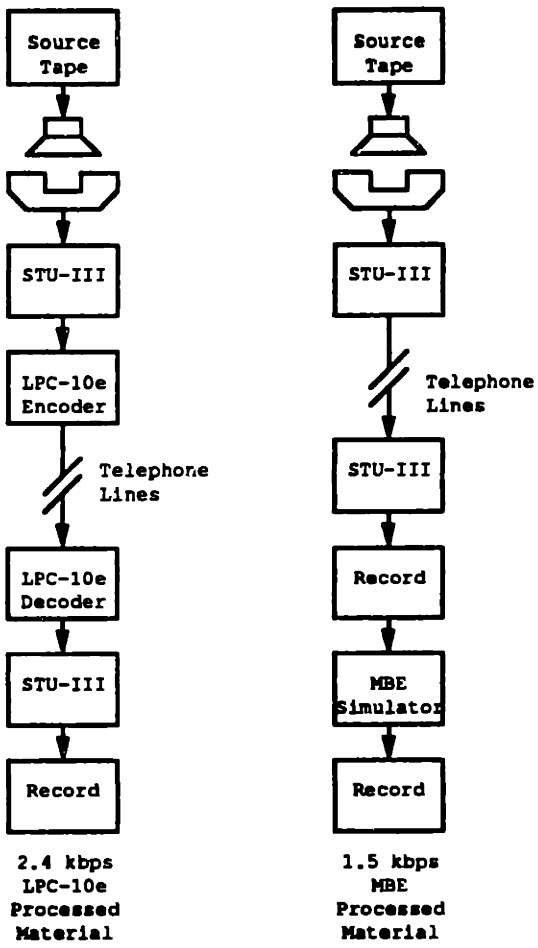

Figure 6.1: Outline of the procedure used to generate the testing material for the LPC-10e vs. MBE side-by-side comparison

The signals were segmented into sentences between 2 and 2.5 seconds in duration and paired with their counterpart from the opposite coder back-to-back with energy equalization and random ordering. Listeners were played each sentence pair twice and asked to pick the segment they preferred. Selection criteria was not specified, but listeners were asked to state the grounds for their decisions.

Preference results for the clean male and female speech varied greatly for individual test subjects and demonstrated a number of selection biases. In general, the 1.5 kbps system was found to sound clearer and more natural than the LPC algorithm, but possessed some noticeable artifacts. Listeners who preferred the LPC tended

to do so because of the artifacts present in the MBE speech. Those who selected the MBE complained that the LPC was muffled, mechanical, and unclear. While individual listeners usually demonstrated a bias towards a single system, the overall results show no average preference for either coder. The noisy sentences produced a more uniform response. Comments included those stated for the clean speech, but the predominant effect appears to be the reproduction of the background noise. The LPC noise was found to be extremely "buzzy" and considerably more noticeable than its MBE counterpart. Listeners stated that the buzziness interfered with the clarity and naturalness of the signal. In contrast, the MBE did a reasonable job of synthesizing the background conditions and maintaining natural sounding speech. Primarily for these reasons, the preference results largely favored the MBE coder for the noisy background condition.

The results of this experiment parallel those of a similar study performed by Griffin [1] in which DRT scores of the 8.0 kbps MBE coder and a 7.45 kbps single band excitation (SBE) coder were compared. While the coders produced nearly equivalent figures in clean speech, the MBE model clearly outperformed the SBE model when the speech was corrupted by additive noise. Informal listening tests of these two systems confirmed this disparity and provided quality comments similar to those presented for the experiment discussed here. The 1.5 kbps MBE and LPC-10e comparison indicates that these results are still valid at a much lower bit rate and further corroborates the utility of the extra voiced/unvoiced bands in the MBE model.

Chapter 7

Conclusions

7.1 Summary

In this thesis a 1.5 kbps speech coder based on the MBE speech model has been presented. The system is comprised of several isolated elements. The first of these is the MBE analysis algorithm which estimates the MBE parameters. The parameters are then quantized, converted into a bit stream, and transmitted across a channel. At the receiver end a decoder regenerates these parameters and delivers them to the MBE synthesis routine where the synthesized speech is produced. The focus of this thesis has been the coding of the model parameters with a bit rate constraint of 1.5 kbps. Given this restriction, quantization schemes employed in earlier, higher rate MBE coders were found to be unsatisfactory. A new coding method based upon LPC modeling of the harmonic magnitudes and an LSP representation of the LPC coefficients was developed. A coder simulation based on this design has been shown to obtain speech intelligibility on par with state-of-the-art 2.4 kbps speech coders. An informal listening comparison between the 1.5 kbps MBE coder and the government standard 2.4 kbps LPC-10e vocoder revealed comparable performance in high SNR conditions and a preference for the MBE coder in noisy environments.

The performance of this system further demonstrates the attractiveness of the

MBE model to speech coding applications. The ability of the MBE model to accurately reproduce speech in a wide range of background environments provides a significant advantage over conventional speech modeling methods. The compactness of the MBE parameters and their potential to be efficiently quantized makes this model ideal for low rate systems. Finally, the computational and production costs of a real-time implementation of an MBE speech coder are small compared to systems producing similar quality. Real-time versions of the 2.4, 4.8, and 8.0 kbps MBE coders have been constructed with only a single DSP chip [6].

7.2 Suggestions for Further Research

While significant work has gone into the development of an efficient encoding scheme, there is still room for improvement. The focus of the quantization issue has been the harmonic magnitudes. The spectral envelope is currently fit to a high order LPC polynomial for the purpose of applying extremely effective LPC quantization schemes. The selection of the appropriate LPC coefficients deserves a good deal of attention. Improving the ability of this polynomial to model the harmonic magnitudes at a given LPC order should be beneficial. It would allow for a reduction in the LPC order and a subsequent improvement in LSP quantization. The result would be higher quality synthesized speech. One consideration that has not been explored is matching the harmonic magnitudes to a pole-zero model. The LPC fit is all-pole in nature and, while it will model a general spectrum, it requires more parameters to do so than a system function consisting of both poles and zeros. While it is not clear if the quantization benefits obtained by reducing the number of poles would be offset by the inclusion of zeros, it does seem to be worthy of some attention. A second idea is motivated by the approximation made for encoding the voicing information. It was observed that the voicing decisions could be effectively modeled by a step function at a single frequency. Harmonics below this frequency

are classified voiced while those above are designated unvoiced. Employing distinct coding schemes for each of these frequency regions may be able to exploit the relative sensitivity of the human ear to each class of harmonics. One experiment along these lines involved modeling the voiced harmonics with a 10^{th} to 14^{th} order LPC polynomial and fitting the unvoiced harmonics with a 4^{th} to 6^{th} order LPC. A final suggestion concerns the quantization of the LSP parameters. Results indicate that non-linear quantization and VQ methods may be able to further exploit redundancies in these figures. At this point, the savings available have not justified the added computational expense. but further work may be worthy of pursuit.

Bibliography

[1] Daniel W. Griffen and Jae S. Lim, "Multi-Band Excitation Vocoder," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-36, pp. 1223-1235, Aug. 1988.

[2] 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.

[3] John C. Hardwick and Jae S. Lim, "A 4.8 KBPS Multi-Band Excitation Speech Coder," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 374-377, NY, NY, April 11-14, 1988.

[4] John C. Hardwick, "A 4.8 KBPS Multi-Band Excitation Speech Coder," *S.M. Thesis*, E.E.C.S. Department, M.I.T., 1988.

[5] Daniel W. Griffen and Jae S. Lim, "A New Pitch Detection Algorithm," /newblock *International Conference on Digital Signal Processing*, Florence, Italy, Sept. 5-8,1984.

[6] Michael S. Brandstein, Peter A. Monta, John C. Hardwick, and Jae S. Lim, "A Real-Time Implementation of the Improved MBE Speech Coder," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 5-8. Albuquerque, NM, April 3-6, 1990.

[7] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 937-940, Tampa, Florida, March 26-29, 1985.

[8] Bruce Fette, Wilburn Clark, and Cynthia Jaskie, "Experiments With A High Quality, Low Complexity 4800 bps Residual Excited LPC (RELP) Vocoder," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 263-266, NY, NY, April 11-14, 1988.

[9] Daniel W. Griffen and Jae S. Lim, "Signal Estimation From Modified Short-Time Fourier Transform," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-32. no.2, pp. 236-243, April 1984.

[10] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall: Englewood Cliffs, NJ, 1979.

[11] John Makhoul, "Spectral Linear Prediction: Properties and Applications," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-23, pp. 283-296, June 1975.

[12] Amro El-Jaroudi and John Makhoul, "Discrete All-Pole Modeling for Voiced Speech," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 320-323, Dallas, TX, April 6-9, 1987.

[13] Hynek Hermansky, Brian Hanson, Hisashi Wakita, and Hiroya Fujisaki, "Linear Predictive Modeling of Speech in Modified Spectral Domains," *Digital Proc. on Signals in Comm.*, IERE Conf., pp.55-62, 1985.

[14] Douglas O'Shaughnessy, *Speeech Communications: Human and Machine*, Addison-Wesley: Reading , MA, 1987.

[15] Frank Soong and Bing-Hwang Juang, "Line Spectral Pair (LSP) and Speech Data Compression," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 1.10.1-1.10.4, San Diego, CA, March 19-21, 1984.

[16] Noboru Sugamura and Nariman Farvardin, "Quantizer Design in LSP Speech Analysis and Synthesis," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 398-401, NY, NY, April 11-14, 1988.

[17] Nariman Farvardin and Rajiv Laroia, "Efficient Encoding of Speech LSP Parameters Using the Discrete Cosine Transformation," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 168-171, Glasgow, Scotland, May 23-26, 1989.

[18] N. S. Jayant and Peter Noll, *Digital Coding of Waveforms*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1984.

[19] G. Fairbanks, "Test of Phonemic Differentiation: The Rhyme Test," *Journal of the Acoustical Society of America*, vol. 30, no. 7, pp. 596-600, July, 1958.

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

[21] R. Jakobsen, C. G. M. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*, Cambridge, Mass., M.I.T.Press, 1967.

[22] David P. Kemp, Retha A. Sueda, and Thomas E. Tremain, "An Evaluation of 4800 BPS Voice Coders," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 200-203. Glasgow, Scotland, May 23-26. 1989.

[23] Thomas E. Tremain, "The Government Standard Linear Predictive Coding Algorithm: LPC10," *Speech Technology*, April 1982.

[24] Paul C. Meuse. "A 2400 bps Multi-Band Excitation Vocoder," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 9-12. Albuquerque, NM, Aril 3-6, 1990.

[25] Robert J. McAulay and Terrence Champion, "Improved Interoperable 2.4 kb/s LPC Using Sinusoidal Transform Coder Techniques," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 641-643, Albuquerque, NM, Aril 3-6, 1990.

[26] Caldwell Smith, "Relating the Performance of Speech Processors to Bit Error Rate," *Speech Technology*, Sept/Oct 1983.

[27] Schuyler R. Quackenbush, Thomas P. Barnwell III, and Mark A. Clements, *Objective Measures of Speech Quality*, Prentice Hall: Englewood Cliffs, NJ, 1988.