Frequency Domain Secondary Pulse Estimation

by

Jeffrey Crawford Dickerson

S.B., Massachusetts Institute of Technology (1994)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1995

Author .....
Department of Electrical Engineering and Computer Science
January 20, 1995

Certified by .....
Dr. Robert J. McAulay
Senior Staff Member, MIT Lincoln Laboratory
Thesis Supervisor

Accepted by .....
Prof. Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Frequency Domain Secondary Pulse Estimation

by

Jeffrey Crawford Dickerson

Submitted to the Department of Electrical Engineering and Computer Science
on January 20, 1995, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Science and Engineering

Abstract

The traditional model of speech production models voiced speech signals as the output of a minimum phase linear filter excited by a periodic impulse train. There exists a body of evidence which suggests that there are actually additional, secondary pulses in the excitation. This thesis investigates a frequency domain technique for estimating the locations and amplitudes of these secondary pulses. The estimator operates directly on the measured phase of the speech signal spectrum. While the estimator is shown to perform reliably on synthetic waveforms produced by exciting a minimum phase filter with a periodic excitation consisting of primary and secondary pulses, it fails when used on actual speech or even synthetic speech with non-impulsive excitation. The reasons for this failure are found to lie in extreme sensitivity of the signal spectrum to inaccuracies of the linear model. The results of exploratory experiments involving frequency domain speech coding using secondary pulse phase modelling are also presented.

Thesis Supervisor: Dr. Robert J. McAulay
Title: Senior Staff Member, MIT Lincoln Laboratory

Acknowledgments

I would first like to thank my thesis advisor Dr. Robert McAulay for allowing me the oppurtunity to conduct speech research at Lincoln Laboratory. I am also especially indebted to Elliot Singer for the countless hours he has spent assisting me in planning, conducting, and writing up my research. In addition I would particularly like to thank Charles Jankowski for putting up with my incessant speech and computer questions.

I would like to thank the members of the Speech Systems Technology group at MIT Lincoln Laboratories for their support.

I would also like to thank my friends and family who have supported me during the period of time I was working on this thesis. I would especially like to thank Elaine for putting up with my frantic moods as the deadline approached. My parents also supplied a great deal of support for me during this time.

Contents

List of Figures

Chapter 1

Background

1.1 Motivation for Secondary Pulse Estimation

An accurate model of speech production provides a useful basis for applications such as speech coding and speaker identification. The prevalent, traditional model of speech characterizes the physiological mechanisms which produce speech. Physiologically, speech is produced by the passage of an excitation through the vocal tract. The traditional model of speech views the vocal tract as a linear, time-varying minimum phase filter and segregates the glottal excitation into two narrowly defined categories [9]. For voiced speech, the excitation is modeled as a simple periodic pulse train. The frequency of this waveform is sometimes referred to as the pitch or fundamental frequency. Voiced speech occurs when the glottal excitation results from the vibration of the vocal cords. Vowel sounds are, in general, voiced. The other case, unvoiced speech, results from the passage of forced air through the vocal tract. As a result, unvoiced speech lacks the essential periodicity of voiced speech and the excitation is modeled very simply as white noise. This model is depicted in Figure 1-1.

The physical model of speech production has been used with considerable success in speech coding applications [10]. The speech parameters of this model also provide features which can be exploited for other applications such as speaker identification [8]. Despite this success, the model has significant shortcomings. In particular,

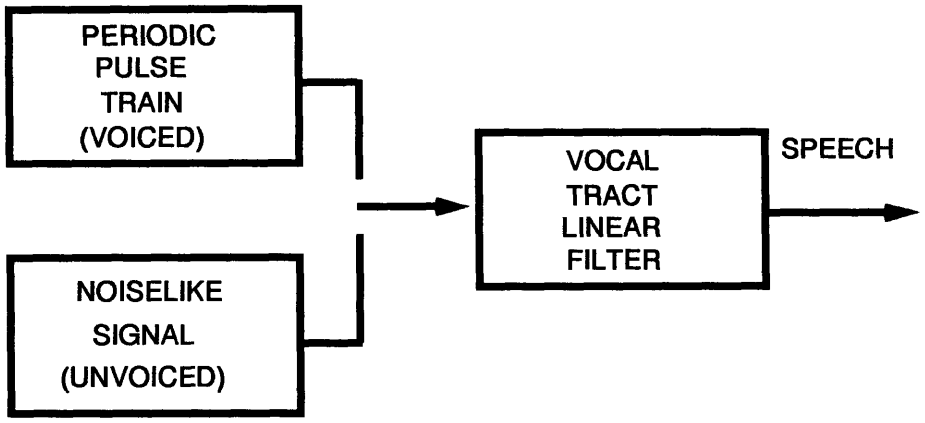

Figure 1-1: Traditional model of speech production

there are many utterances for which a strict voiced/unvoiced classification is inappropriate. In these situations, the speech synthesized by this model may sound harsh, buzzy, or unnatural.

1.1.1 Single secondary pulse estimation

There exists a body of empirical evidence which suggests the existence of additional pulses in the excitation for certain segments of voiced speech [2]. The prevalence of these pulses in speech is still a matter of dispute. While there are certain rare physiological phenomena which produce speech with obvious secondary pulses (i.e. diplophonia and vocal fry), in general, the existence of secondary pulses in common speech is the subject of ongoing debate among researchers. A sample of speech which contains obvious secondary pulses is shown in Figure 1-2. This segment of speech contains prominent secondary pulses occurring approximately two-thirds into each primary pitch period and their effects are visually apparent. One technique which does find prevalent secondary pulses relies on the high resolution Teager energy operator [8]. The source of most of these estimated secondary pulses is unknown; they may have a physiological basis or they may be an artifact of the signal processing. Finding an alternative technique for locating secondary pulses may prove useful for corroborating or refuting the physiological hypothesis.

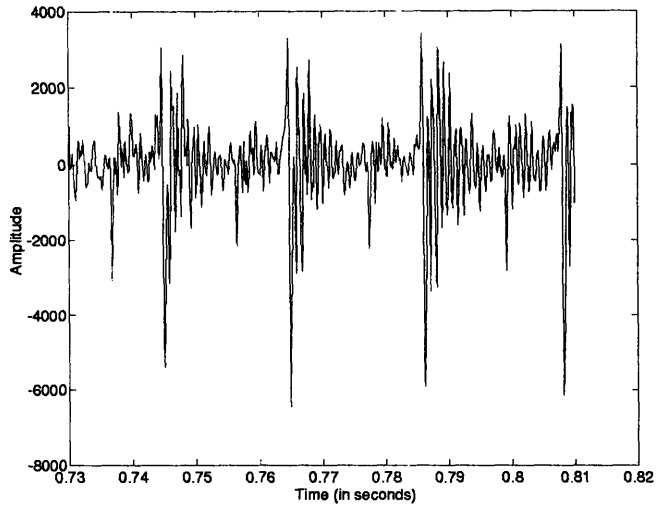

Figure 1-2: Speech with obvious secondary excitation pulses

Regardless of the physiological origins of the estimated secondary pulses, they remain useful as a tool for speaker identification (SID). Experiments have been performed which show that performance of a SID system on twenty females and twenty males drawn from the NTIMIT database [3] improved when the feature set was augmented with secondary pulse locations estimated by the Teager operator. Specifically, performance rose from 65% to 68% for the males and 62% to 65% for the females. Additional methods of estimating secondary pulse locations are thus well worth evaluating.

1.1.2 General case

Several speech coders have sought to improve on the binary excitation model in an attempt to produce more natural sounding speech. In multipulse coders [1], the glottal excitation is modeled not as a simple pulse train or white noise, but rather as a generic series of pulses with unconstrained positions and amplitudes. This model, which does not classify the speech as voiced or unvoiced, produces high quality output and specifically remedies the shortcomings of many previous systems when confronted with speech which is neither clearly voiced nor clearly unvoiced.

The primary focus of this thesis is the estimation of the location and amplitude of a single secondary pulse and their potential use as features for speaker identi-

fication; additional preliminary experiments in speech coding are also presented. In both cases, the sinusoidal transform system (STS) was used as a basic speech analysis/synthesis framework. The secondary pulse estimation technique proposed in this thesis relies explicitly on the fundamental model of speech as the output of a linear system. It seeks to estimate the location and amplitude of a secondary pulse by their effect on the measured phase of the short time spectrum of the speech itself, assuming accurate modeling of the vocal tract filter. This approach has been justified empirically since experiments have shown that measured phases provide information which enables virtually transparent speech coding [5]. Therefore, it would be highly desirable to develop a coder which accurately models these measured phases. Unfortunately, this has proven very difficult to do directly. A brief discussion of the sinusoidal transform system and the importance of phase in its representation of speech are presented in the next section.

1.2 Sinusoidal Transform System

1.2.1 Overview

Since the secondary pulse estimator was developed in the framework of STS, it is worthwhile discussing some of the salient characteristics of that system. STS models speech explicitly as a sum of sine waves with varying amplitudes, frequencies, and phases. This frequency domain representation of speech is attractive because it explicitly deals with the component parameters that make up the speech waveform. By imposing a linear speech production model on STS, the input to the linear system becomes a sum of sine waves. It is clear that in the case of voiced speech, the periodic pulse train used in the binary voiced-unvoiced excitation model can be represented by its Fourier series decomposition into a sum of harmonically related sine waves [6]. If the speech differs somewhat from the voiced model, the sine waves will, in general, be aharmonic. The validity of using the sum of sine waves approximation for unvoiced excitation is more difficult to establish since the sine

waves then seek to model a stochastic waveform as opposed to a deterministic one. A mathematical justification for this decomposition relies on the principles of the Karhunen-Loève expansion for arbitrary stochastic signals [7].

Since the input to the linear filter is a sum of sine waves, the output must also be a sum of sine waves with the same underlying frequencies but with different amplitudes and phases. This is the fundamental property of linear systems. The STS algorithm operates by extracting spectral information in order to construct a parametric representation of the original waveform. STS determines the spectral information through the application of the short time Fourier transform (STFT) (Figure 1-3 [5]). In the case of unvoiced speech, this pitch does not possess the

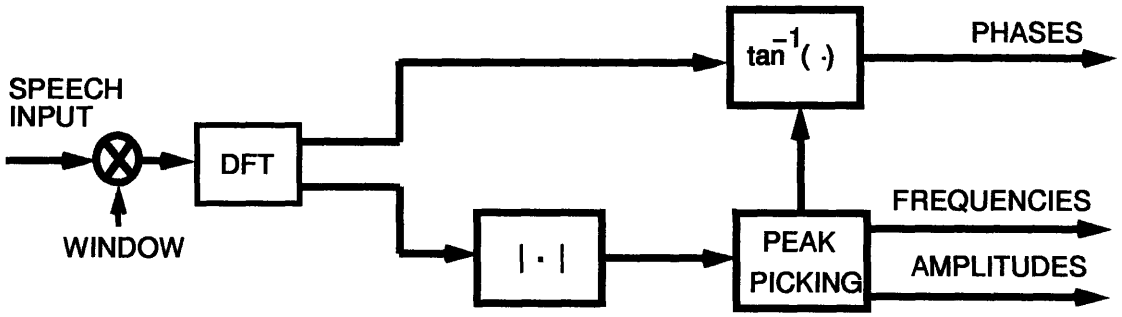

Figure 1-3: STS analysis system [5]

traditional physiological meaning. The spectrum produced by the STFT is analyzed and the most likely frequencies, amplitudes, and phases of the underlying sine waves are determined. The speech signal in the model can be written as

$$s(n) = \sum_{l=1}^{L} A_l e^{-j(n\omega_l + \theta_l)} \tag{1.1}$$

where $\{\omega_l\}$ is the set of underlying frequencies and $\{A_l\}$ and $\{\theta_l\}$ the corresponding sets of measured amplitudes and phases, respectively. It is appropriate to note that while a linear model for the vocal tract and the excitation waveform was used to

justify the sinusoidal decomposition of the waveform, this model does not enter explicitly into the Equation 1.1. In its most general form, STS needs make no assumptions about the excitation or vocal tract. In a real system, however, several assumptions are made in order to produce a practical coder.

1.2.2 Phase in STS

STS directly calculates the sets of frequencies, amplitudes, and phases and uses these to approximate the original speech waveform. By constraining these parameters in several ways, their calculation is greatly simplified. For instance, the frequencies ω_l are frequently taken to be strictly harmonic so that $\omega_l = l\omega_o$. Therefore, the entire set of frequencies can be completely specified by a single parameter ω_o, the pitch or fundamental frequency. As noted above, this harmonic model is a good representation of entirely voiced speech. It has also been empirically determined that the harmonic model leads to high quality synthesized speech provided measured phases are used [5]. Since direct coding of the measured phases requires a prohibitively large number of bits, strategies for representing the phase information more economically must be employed. One such approach relies on the linear model of speech production and views the speech as the output of a minimum-phase filter excited by a glottal waveform. The phases may then be decomposed into a sum of two components, one the phase of the excitation waveform at a particular frequency and the other the phase of the minimum-phase system function at the same frequency. This can be written

$$\theta_l = \Phi_s(\omega_l) + \theta_e(\omega_l) \tag{1.2}$$

where Φ_s is the minimum phase system phase and θ_e is the phase of the excitation.

Explicit transmission of the excitation phase is avoided in a coding system by using a model of the excitation waveform. The simplest model postulates the existence of a single primary pulse train as the excitation. This is the standard model for voiced speech. It can be easily shown that the spectrum of a periodic pulse train

has a strictly linear phase and that this phase is caused by the offset of the pulse train with respect to the origin of the analysis window. This model describes the excitation phase as

$$\hat{\theta}_e(\omega_l) = -n_o\omega_l \tag{1.3}$$

where n_o is the location of the primary pitch pulse within each frame. Use of this model does, however, lead to a noticeable degradation of the speech signal, hence the motivation for more sophisticated phase models.

If secondary pulses actually exist in the excitation, they will contribute in a predictable way to the failure of a linear phase model. The difference between the measured phase and the modeled phase can be viewed as a phase error or phase residual. This serves as a basis for estimating the positions and amplitudes of the secondary pulses. One can find which secondary pulse best models the phase residual in such a way as to provide a closer match to the waveform itself. This is the underlying idea of the secondary pulse estimator used in this thesis.

1.3 Derivation of Secondary Pulse Estimator

In this section, the analytic expressions used by the secondary pulse estimator are derived. Much of the development parallels that found in [5] for the estimation of n_o, the onset time of the primary pulse.

1.3.1 Derivation of signal error in terms of phase residual

If $\hat{s}(n)$ is an estimate of a signal s(n), then the mean square error of the signal estimate over $N+1$ time samples centered at $n = 0$ is

$$\epsilon = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} |s(n) - \hat{s}(n)|^2 \tag{1.4}$$

This expression can be manipulated into the following form:

$$\epsilon = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \left[|s(n)|^2 - 2\Re\left\{ s(n)\hat{s}^*(n) \right\} + |\hat{s}(n)|^2 \right] \tag{1.5}$$

Let the signal $s(n)$ being analyzed be written as

$$s(n) = \sum_{l=1}^{L} A_l e^{-\jmath(n\omega_l + \theta_l)} \tag{1.6}$$

and the signal estimate $\hat{s}(n)$ be written as

$$\hat{s}(n) = \sum_{l=-N/2}^{N/2} \hat{A}_l e^{-\jmath(n\hat{\omega}_l + \hat{\theta}_l)} \tag{1.7}$$

where $\{\hat{\omega}_l\}$, $\left\{\hat{A}_l\right\}$, and $\left\{\hat{\theta}_l\right\}$ are estimates of frequency, amplitude, and phase, respectively. If the amplitude and frequency estimates are exact, Equation 1.5 can be written

$$\epsilon = 2P_s - 2\sum_{l=1}^{L} A_l^2 \cos\left(\hat{\theta}_l - \theta_l\right) \tag{1.8}$$

where

$$P_s = \frac{1}{N+1} \sum_{l=1}^{L} A_l^2 \tag{1.9}$$

Now assume that $s(n)$ is the output of a linear system $H(e^{\jmath\omega})$ excited by an infinite impulse train $e_{1p}(n)$ with period $T_o = 2\pi/\omega_o$ and onset time n_o. Then $S(e^{\jmath\omega})$, the Fourier transform of $s(n)$, will be samples of $H(e^{\jmath\omega})$ at $\omega = l\omega_o$ for integer l, multiplied by a complex exponential $e^{-\jmath n_o \omega}$ corresponding to the onset time with respect to $n = 0$:

$$\begin{aligned} S(e^{\jmath\omega}) &= e^{-\jmath n_o \omega} H(e^{\jmath l\omega_o}) && (1.10) \\ &= e^{-\jmath n_o l\omega_o} H(e^{\jmath l\omega_o}) && (1.11) \\ &= H(e^{\jmath\omega}) E_{1p}(e^{\jmath\omega}) && (1.12) \end{aligned}$$

where E_{1p} is the Fourier transform of the excitation pulse train.

If $M-1$ secondary periodic pulse trains are added to the excitation, each with the same period as the primary pulse train but different amplitude and offset, the Fourier transform of the entire excitation will become

$$E_{Mp}(e^{\jmath\omega}) = E_{1p}(e^{\jmath\omega})\left[1+\sum_{k=2}^{M}\alpha_k e^{-\jmath n_k\omega}\right] \tag{1.13}$$

where n_k is the spacing between the k^{th} pulse train and the primary and α_k is the relative amplitude. The Fourier transform of the output $s_{Mp}(n)$ is

$$\begin{aligned} S_{Mp}(e^{\jmath\omega}) &= E_{Mp}(e^{\jmath\omega})H(e^{\jmath\omega}) && (1.14)\\ &= e^{-\jmath\omega n_o}\left[1+\sum_{k=2}^{M}\alpha_k e^{-\jmath n_k\omega}\right]\sum_{i=0}^{2\pi/\omega_o}\delta(\omega-i\omega_o)H(e^{\jmath\omega}) && (1.15)\end{aligned}$$

The phase of this expression is simply

$$\angle S_{Mp} = \left[-\omega n_o + \angle H(e^{\jmath\omega}) + \angle\left(1+\sum_{k=2}^{M}\alpha_k e^{-\jmath n_k\omega}\right)\right]\sum_{i=0}^{2\pi/\omega_o}\delta(\omega-i\omega_o) \tag{1.16}$$

The phase contribution of the secondary pulses is

$$\begin{aligned} \angle\left[1+\sum_{k=1}^{M}\alpha_k e^{-\jmath n_k\omega}\right] &= \arctan\frac{\Im\left[1+\sum_{k=2}^{M}\alpha_k e^{-\jmath n_k\omega}\right]}{\Re\left[1+\sum_{k=2}^{M}\alpha_k e^{-\jmath n_k\omega}\right]} && (1.17)\\ &= \arctan\frac{-\sum_{k=2}^{M}\alpha_k\sin(n_k\omega)}{1+\sum_{k=2}^{M}\alpha_k\cos(n_k\omega)} && (1.18)\end{aligned}$$

The estimate for the measured phase of the signal then becomes

$$\hat{\theta}_{total}(\omega_l) = \angle H(e^{\jmath\omega_l}) - \omega_l n_o + \arctan\frac{-\sum_{k=2}^{M}\alpha_k\sin(n_k\omega)}{1+\sum_{k=2}^{M}\alpha_k\cos(n_k\omega)} \tag{1.19}$$

Plugging Equation 1.19 into Equation 1.8 results in the following expression:

$$\epsilon\left(\{n_k\},\{\alpha_k\}\right) = 2P_s - 2\sum_{l=1}^{L}A_l^2\cos\left\{\hat{\theta}_{Mp}(\omega_l)+\xi_l\right\} \tag{1.20}$$

where

$$\hat{\theta}_{Mp}(\omega_l) = -\arctan\frac{-\sum_{k=2}^{M}\alpha_k \sin(n_k\omega_l)}{1+\sum_{k=2}^{M}\alpha_k\cos(n_k\omega_l)} \quad (1.21)$$

$$\xi_l = \theta_l + n_o\omega_l - \angle H\left(e^{j\omega_l}\right) \quad (1.22)$$

where ξ_l is the phase residual, or error, between the estimate of the total phase provided by the linear model and the actual measured phase θ_l, and $\hat{\theta}_{Mp}(\omega_l)$ is the phase contribution due to secondary pulses.

Since the secondary pulse parameters do not affect the first term of Equation 1.20, any optimization of them need only concern itself with the second term. Minimizing the error is then the same as maximizing the likelihood expression

$$\rho(\{n_k\},\{\alpha_k\}) = \sum_{l=1}^{L} A_l^2 \cos\left\{\hat{\theta}_{Mp}(\omega_l) + \xi_l\right\} \quad (1.23)$$

Using the trigonometric identity

$$\cos(\beta+\gamma) = \cos\beta\cos\gamma - \sin\beta\sin\gamma \quad (1.24)$$

Equation 1.23 becomes

$$\rho(\{n_k\},\{\alpha_k\}) = \sum_{l=1}^{L} A_l^2\left[\cos\hat{\theta}_{Mp}(\omega_l)\cos\xi_l - \sin\hat{\theta}_{Mp}(\omega_l)\sin\xi_l\right] \quad (1.25)$$

For a single secondary pulse train, this equation may be further simplified by constructing a right triangle in order to solve for the sine and cosine of $\hat{\theta}_{2p}(\omega_l)$ using the Pythagorean Theorem. This yields

$$\sin\hat{\theta}_{2p}(\omega_l) = \frac{\alpha_2\sin(n_2\omega_l)}{\sqrt{\alpha_2^2+1+2\alpha_2\cos(n_2\omega_l)}} \quad (1.26)$$

$$\cos\hat{\theta}_{2p}(\omega_l) = \frac{1+\alpha_2\cos(n_2\omega_l)}{\sqrt{\alpha_2^2+1+2\alpha_2\cos(n_2\omega_l)}} \quad (1.27)$$

Since this paper primarily concerns itself with this simple case, α_2 will be written as α and n_2 will be written as n_d. Substituting the equations above into the likelihood expression in Equation 1.23 yields

$$\rho(n_d, \alpha) = \sum_{l=1}^{L} \frac{A_l^2[(\alpha \cos n_d\omega_l + 1)\cos \xi_l - (\alpha \sin n_d\omega_l)\sin \xi_l]}{\sqrt{\alpha^2 + 1 + 2\alpha \cos(n_d\omega_l)}} \tag{1.28}$$

$$= \sum_{l=1}^{L} \frac{A_l^2 [\cos \xi_l + \alpha \cos (n_d\omega_l + \xi_l)]}{\sqrt{\alpha^2 + 1 + 2\alpha \cos(n_d\omega_l)}} \tag{1.29}$$

The secondary pulse estimator investigated in this paper operates by explicitly maximizing this equation over both relative amplitude and relative location. Note that the frequencies, while in general arbitrary, will for the purposes of this thesis be taken to be harmonic, meaning $\omega_l = l\omega_o$. This will create an artifact in the likelihood surface of Equation 1.29. This phenomenon is discussed in Section 3.1.

Chapter 2

Procedure

2.1 Synthetic vowels

In order to conduct controlled experiments, the estimator was first developed and tested on synthetic vowels. These vowels were formed by sending a pulse train through a simple allpole linear filter, corresponding directly to the linear model of voiced speech production. Many different variations on this simple scheme were tested, including a vowel with no primary pulse offset and no secondary pulses, a vowel with a primary pulse offset and no secondary pulses, and many with excitations augmented by secondary pulses in a variety of positions and amplitudes. The performance of the estimator on these test cases would determine its best possible performance since these cases fit exactly into its theoretical model, with no unknown or uncontrolled factors potentially found in actual speech. In addition to these idealized cases, the performance of the estimator was evaluated for a vowel constructed by exciting the same allpole linear filter with a more generalized glottal function. This glottal excitation was formed by generating an excitation waveform with an open quotient of .5, a reasonable value for male speakers [4]. The waveform is produced by smearing the excitation waveform by a small amount in order to model the phenomenon of glottal opening more realistically than does the simple impulse train. The estimator was implemented by maximizing Equation 1.29 explicitly using a simple grid search. The likelihood function $\rho(n_d, \alpha)$ was calculated

at every point on a two-dimensional grid, where one dimension was relative amplitude and one was relative location. The relative amplitude levels were evenly spaced over the interval from zero to the primary pulse amplitude and the relative location values were evenly spaced over a single primary pitch period. The grid had 100 divisions in relative amplitude and 800 divisions in relative location. The coordinates corresponding to the maximum value of the likelihood function were chosen as the maximum likelihood estimate of the secondary pulse parameters.

2.2 Secondary pulse estimation for speech

The secondary pulse estimator was then tested on an actual speaker. The speaker chosen exhibited strong secondary pulses, remarkable even upon visual examination of the speech waveform. The waveform is that shown in Figure 1-2. This speaker formed a good test case for the estimator since its performance could be readily determined from visual inspection. The results were examined at specific frames of the input speech to see what insights could be garnered from the system's behavior. The system used was essentially the same as that used on the synthetic vowels.

2.3 Speech coding and phase modeling

Finally, a series of exploratory experiments was performed using a sinusoidal transform coder which synthesized speech by combining the phase contributions of the system phase, onset time, and estimated secondary pulse. The synthesized speech was evaluated to determine whether this system improved the performance of a coder which used the simple phase model of Equation 1.3. The secondary pulse estimation technique was also extended so as to allow calculation of a phase residual assuming an arbitrary number of secondary pulses. The technique employed suboptimal estimation using an iterative analysis-synthesis loop (see Figure 2-1). Initially, the estimator determines the location and amplitude of a single secondary pulse. Its phase contribution is then removed from the total phase and the pulse

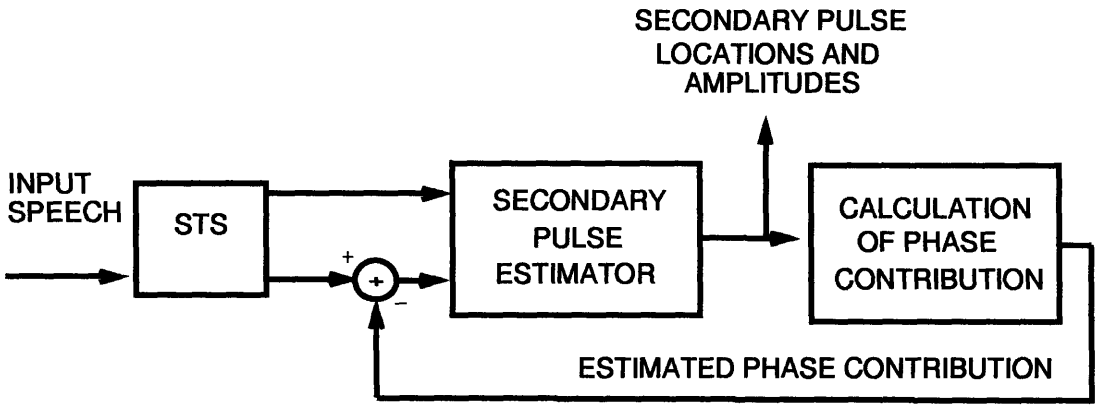

Figure 2-1: Iterative multiple pulse estimation loop

estimation process is repeated. The phase contribution from this new pulse is incorporated into an updated phase residual and the process begins again until the desired number of pulses have been estimated. One anticipated problem with the analysis- synthesis loop is the lack of spectral amplitude modeling. When the system removes the phase contribution of a particular pulse train, it does not likewise remove the spectral magnitude contribution of that pulse train. While the spectral magnitude occurs only as a weighting of the terms in Equation 1.29, this effect may be significant, especially when trying to estimate a large number of additional pulse trains.

Chapter 3

Synthetic Vowel Experiments

3.1 Impulsive Excitation

The secondary pulse estimator was first evaluated on an artificial vowel produced with no secondary pulse excitation and no primary pulse offset. This vowel had a fundamental frequency of 200 Hz (5 ms pitch period) and the synthesized speech was sampled at 8 kHz. The time waveform of this vowel is shown in Figure 3-1. The

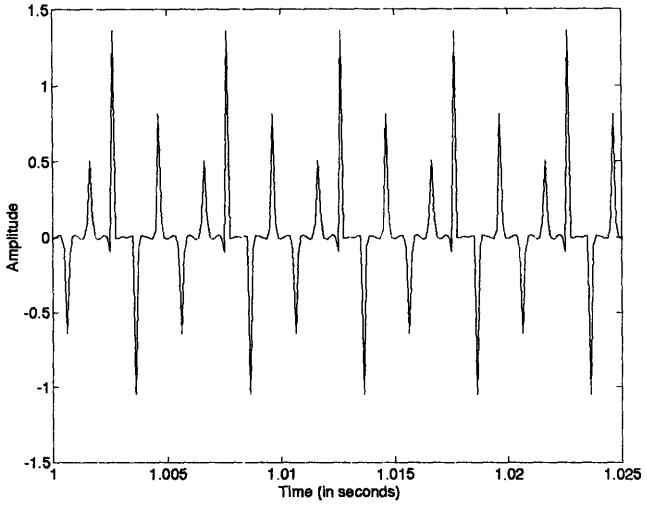

Figure 3-1: Time waveform of a synthetic vowel formed with an impulsive excitation and no secondary pulses

waveform was windowed around the location of one of its primary pulses and an STFT was computed. This STFT performed a 512 point discrete Fourier transform

of a 15 ms Hamming windowed segment of the waveform. The magnitude of this transform is shown in Figure 3-2 and the phase in 3-3. The phase of the minimum

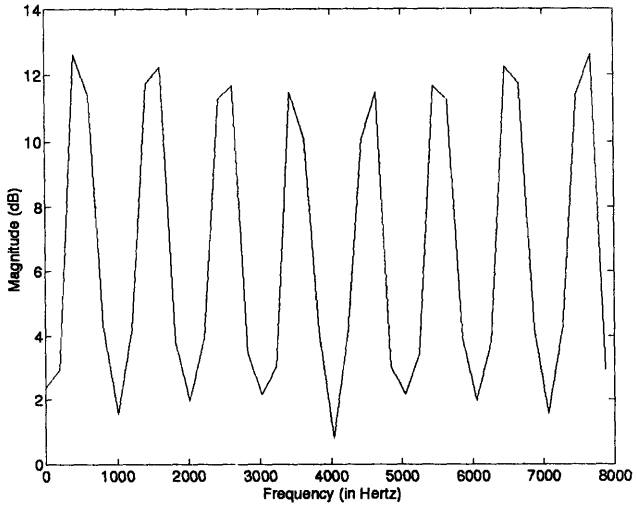

Figure 3-2: Magnitude of the STFT of synthetic waveform

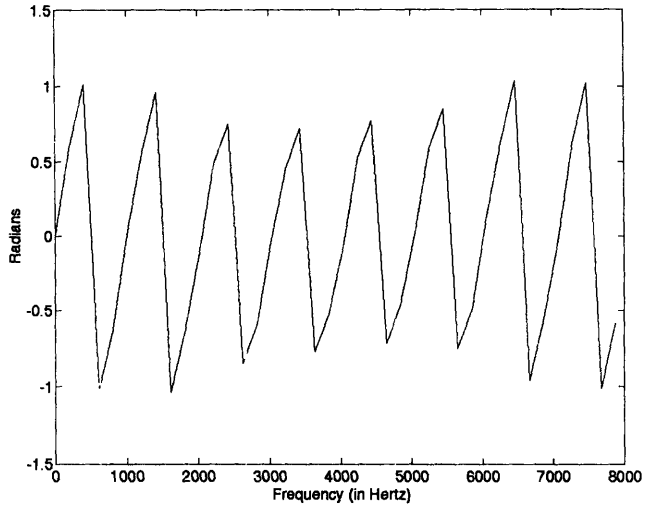

Figure 3-3: Phase of the STFT of the synthetic waveform

phase system function (Figure 3-4) was then subtracted from the measured phase to give a phase residual (Figure 3-5) corresponding to Equation 1.22. As can be readily seen, the phase residual is very nearly zero for all frequencies. Any deviation from zero can be attributed to computer round-off errors and computational limitations. This phase residual was then used by the estimator to locate a single secondary pulse using Equation 1.29. Several cross-sections of the likelihood function computed by the estimator are shown in Figures 3-6, 3-7, 3-8, and 3-9.

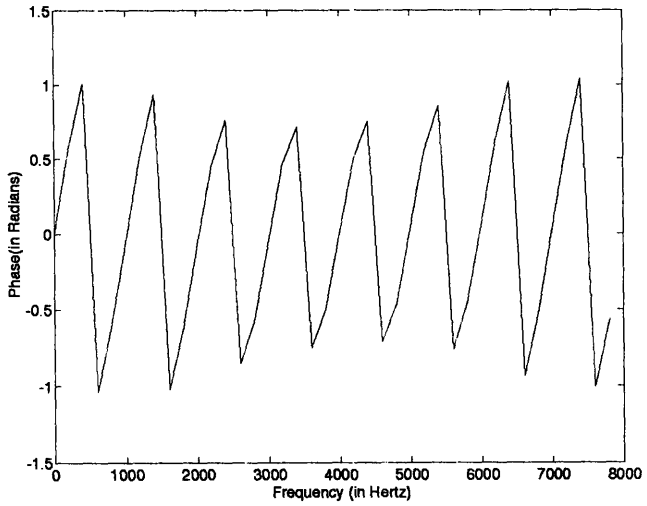

Figure 3-4: Phase of the system function

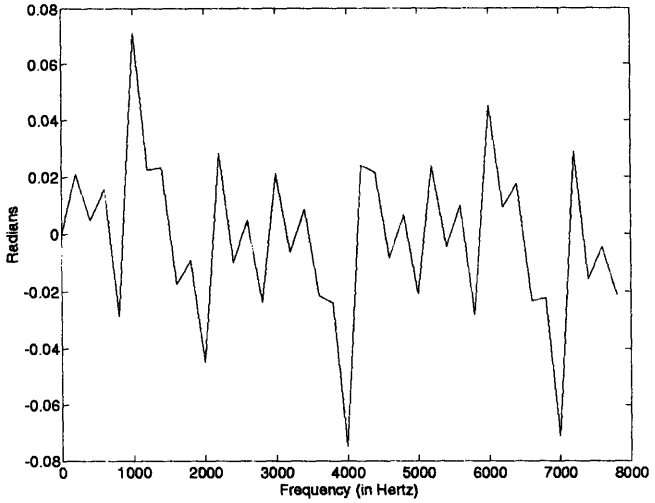

Figure 3-5: Phase residual

These figures show cross-sections of the three dimensional surface described by 1.29. The estimated location and relative amplitude are the coordinates of the global maximum of this entire surface. Since the figures present two dimensional slices of the surface, they do not characterize it fully and must be interpreted carefully. The peaks of the functions in the figures show that the estimator is correctly determining that there are no secondary pulses for this case. In Figures 3-6 and 3-7, the peaks occur at the beginning, middle, and end of the pitch period (5 ms). The peaks at the beginning and end both correspond to a secondary pulse train coincident with the primary one. In other words, the estimator does not identify a distinct

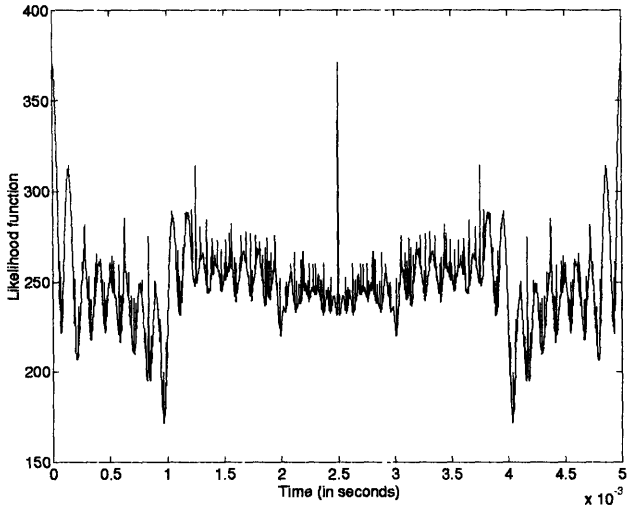

Figure 3-6: Likelihood function for speech vs. offset with no secondary pulses (slice at $\alpha = .98$)

secondary pulse train. The peak in the middle is an artifact of the processing and is discussed in some depth below. The cross-sections in Figures 3-8 and 3-9 have peaks at zero relative amplitude thus also demonstrating the success of the estimator at recognizing that no secondary pulses are present.

There is, however, a spurious peak occurring approximately halfway into the pitch period. This peak is an artifact of the use of phase at harmonic frequencies as the basis for locating a secondary pulse. A secondary pulse occurring exactly in the middle of the pitch period actually has no effect on the phase of the resulting signal and therefore produces no phase residual. This can be shown by examining Equation 1.18 for the single secondary pulse case with harmonic amplitudes. Under these conditions, the equation becomes

$$\theta_{2p} = \arctan \frac{-\alpha \sin(n_d l \omega_o)}{1 + \alpha \cos(n_d l \omega_o)} \tag{3.1}$$

where θ_{2p} is the phase contribution due to the single secondary pulse train. If n_d is taken to be at half the pitch period, $n_d = \frac{T_o}{2} = \frac{\pi}{\omega_o}$. Substituting these expressions into Equation 3.1 yields

$$\theta_{2p} = \arctan \frac{-\alpha \sin \pi l}{1 + \alpha \cos \pi l} \tag{3.2}$$

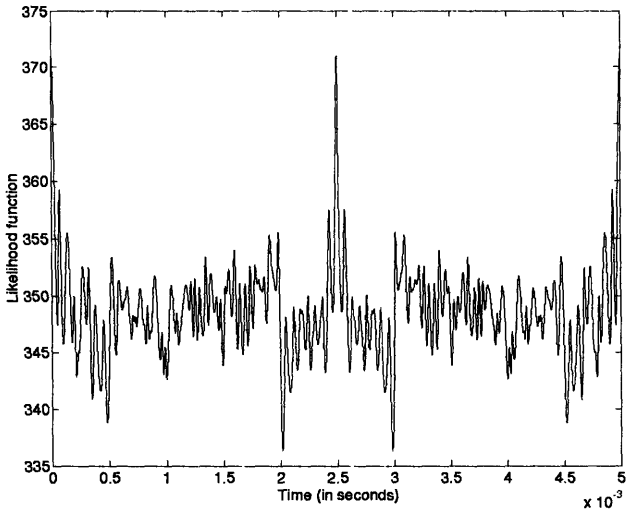

Figure 3-7: Likelihood function for speech vs. offset with no secondary pulses (slice at $\alpha = .5$)

$$= \arctan 0 \tag{3.3}$$

$$= 0 \tag{3.4}$$

for all l. Hence, for a signal with no phase residual, this estimator will be as likely to identify a pulse of arbitrary amplitude in the center of the pitch period as one of zero amplitude elsewhere, even though there was actually no such secondary pulse in the excitation. Similarly, in a signal with an actual secondary pulse at half the pitch period, this estimator would be as likely to conclude there were no secondary pulses whatsoever (Figure 3-10). As can be seen from the figure, the estimator would be as likely to find no secondary pulse as the correct one at one half the pitch period.

The secondary pulse estimator was also evaluated for synthetic vowels with a variety of secondary pulse locations and amplitudes and produced accurate results (ignoring the artifact mentioned above). A few illustrative examples of the likelihood functions are shown in Figures 3-11, 3-12, 3-13, and 3-14. In all these cases, the estimator correctly locates the relative offset of the true secondary pulse train, at 4 ms for Figure 3-12 and at 2 ms for Figure 3-14. It also provided good estimates for the relative amplitude, α, at .8 for Figure 3-11 and at .42 in Figure 3-13, although the correct relative amplitude in the second case was .5. Note that the

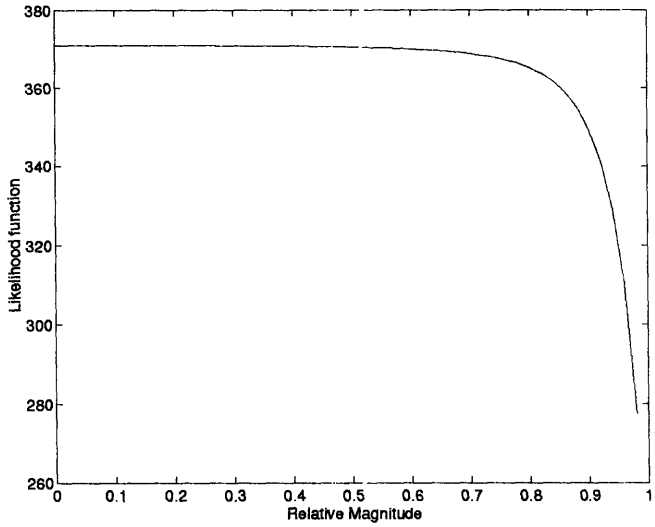

Figure 3-8: Likelihood function for speech vs. rel. amplitude with no secondary pulses (slice at $n_d = 2.5ms$)

likelihood functions tend to be quite smooth as the estimated amplitude is varied with n_d fixed, as shown in Figures 3-11 and 3-13. However, they can be extremely ragged when α is fixed and n_d is varied as in Figures 3-12 and 3-14. This suggests that much finer grid resolution be used for the location than for the relative amplitude.

3.2 Non-impulsive glottal excitation

The secondary pulse estimator was also used on synthetic speech created with a non-impulsive excitation. This non-impulsive excitation significantly affected the phase of the output speech. For the purposes of comparison, the measured phases for a waveform produced with a secondary pulse of relative amplitude .5 and location 2 ms, both with impulsive and non-impulsive excitation, are shown in Figures 3-15 and 3-16 respectively.

There is little resemblance between the two phase functions. When the estimator attempted to locate the secondary pulse in this system, it produced an incorrect result, estimating the secondary pulse to have a relative amplitude of 1 and an offset

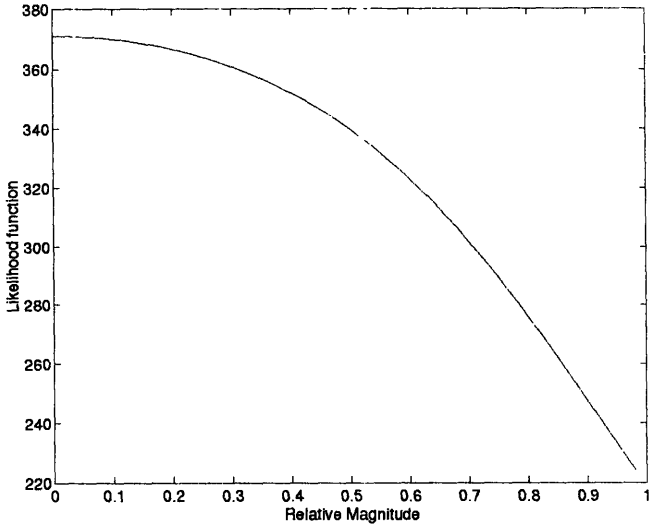

Figure 3-9: Likelihood function vs. rel. amplitude for speech with no secondary pulses (slice at $n_d = 4ms$)

of 4.2 ms. Cross-sections of the likelihood surface, at the true location and relative amplitude of the secondary pulse, are shown in Figures 3-17 and 3-18. Note that while Figure 3-18 does show a peak around the correct value of 2 ms, this was not a global peak for the entire surface.

In general, when the estimator was evaluated on synthetic speech created with a non-impulsive excitation, its performance was consistently poor. The phase perturbation caused by this more realistic glottal opening model effectively prevents the estimator from making a correct estimate. The smearing of impulsive excitation can be modeled as the addition of zeros into the previously allpole system function and the effect of these zeros on the total system phase may not be distinguishable from that of a true secondary pulse. Since the non-impulsive excitation model more closely represents true speech than the impulsive excitation model, the poor performance of the estimator on the non-impulsive synthetic speech anticipates poor performance for real speech, despite the estimator's success for impulsive synthetic speech.

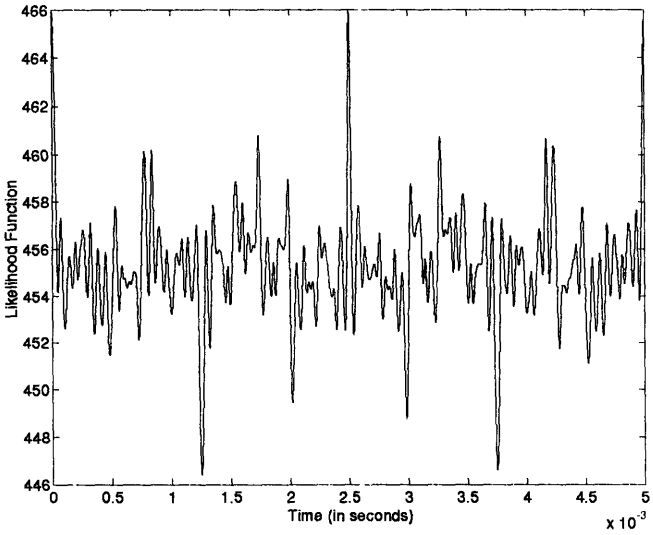

Figure 3-10: Likelihood function vs. offset for a secondary pulse occurring at half the pitch period $\alpha = .3$

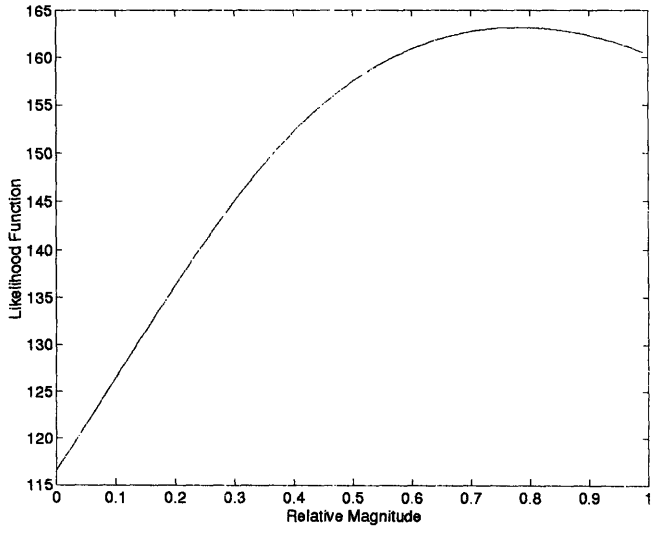

Figure 3-11: Likelihood function vs. relative amplitude, secondary pulse at $\alpha = .8, n_d = 4ms$ (slice at $n_d = 4ms$)

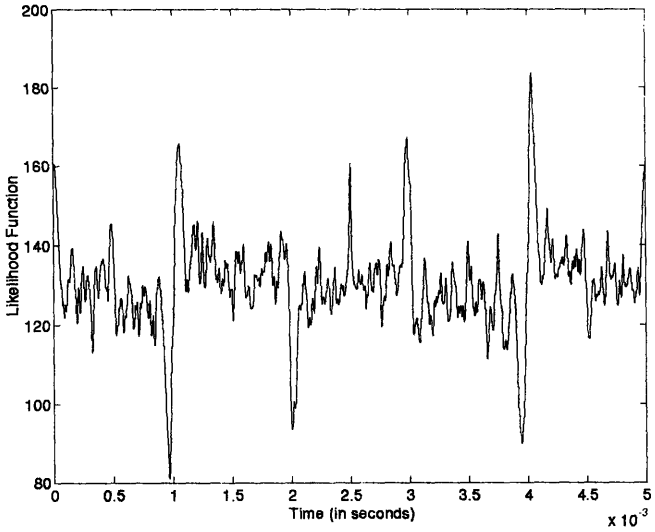

Figure 3-12: Likelihood function vs. offset, secondary pulse at $\alpha = .8, n_d = 4ms$ (slice at $\alpha = .8$)

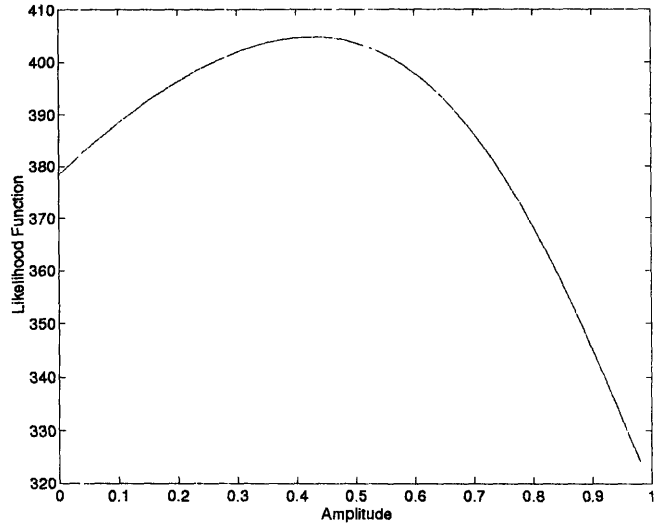

Figure 3-13: Likelihood function vs. relative amplitude, secondary pulse at $\alpha = .5, n_d = 2ms$ (slice at $n_d = 2ms$)

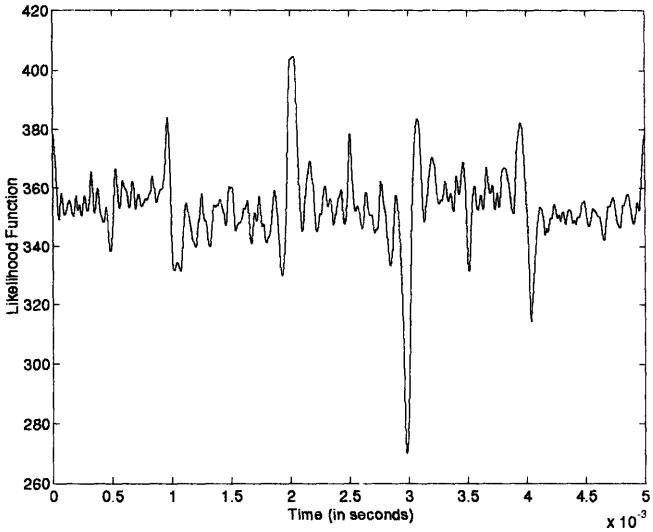

Figure 3-14: Likelihood function vs. offset, secondary pulse at $\alpha = .5, n_d = 2ms$ (slice at $\alpha = .5$)

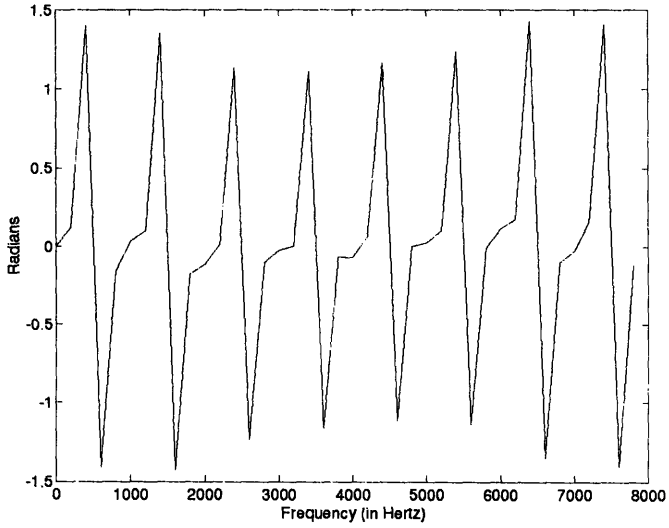

Figure 3-15: Measured phase for impulsive excitation and secondary pulse at $\alpha = .5, n_d = 2ms$

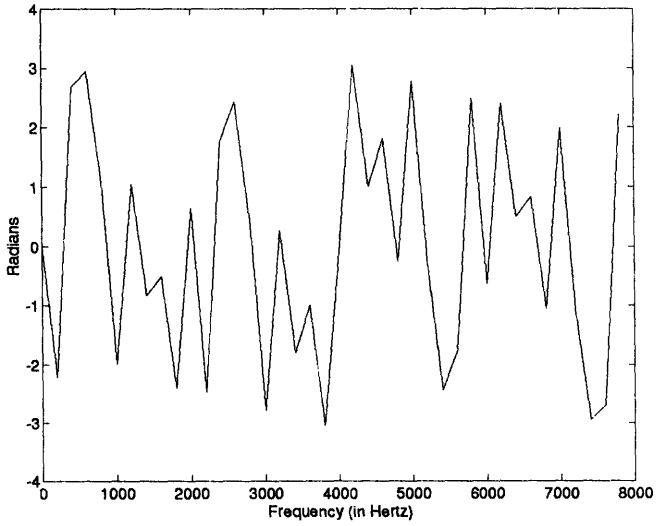

Figure 3-16: Measured phase for non-impulsive excitation and secondary pulse at $\alpha = .5, n_d = 2ms$

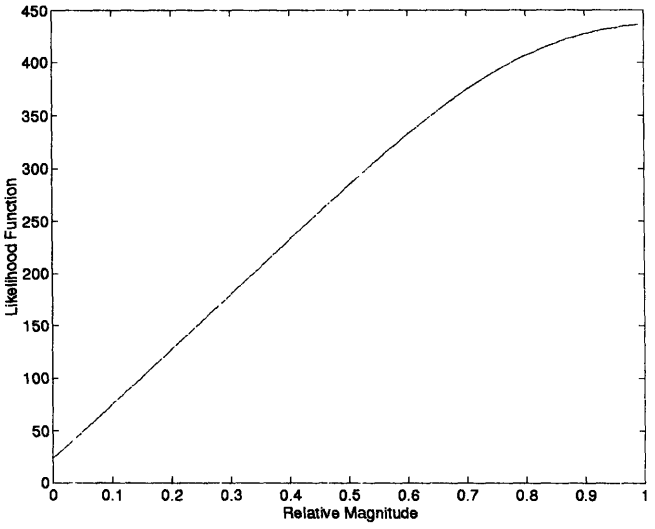

Figure 3-17: Likelihood function vs. relative amplitude, non-impulsive excitation, secondary pulse at $\alpha = .5, n_d = 2ms$ (slice at $n_d = 2ms$)

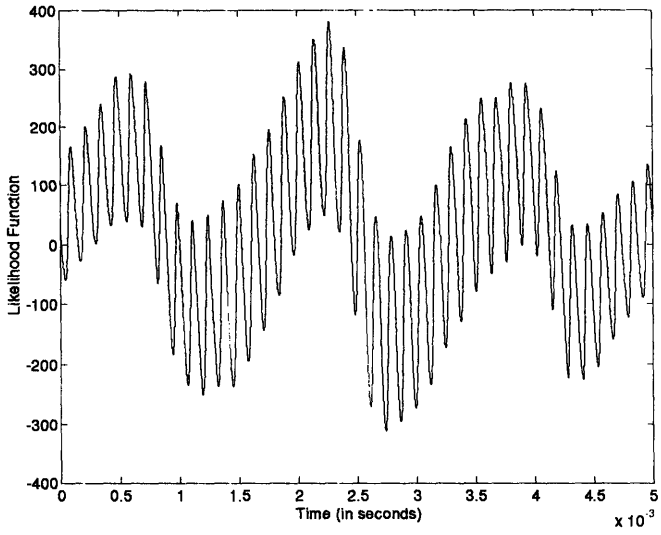

Figure 3-18: Likelihood function vs. offset, non-impulsive excitation, secondary pulse at $\alpha = .5, n_d = 2ms$ (slice at $\alpha = .5$)

Chapter 4

Speech Experiments

4.1 Single secondary pulse estimation

In the next series of experiments, the performance of the estimator was evaluated on actual human speech. A segment of the utterance used to test the system is shown in Figure 4-1. Also in this figure are the pitch estimates produced by a sinusoidal transform coder. This coder was also used to provide estimates of the system phase and linear phase components. The secondary pulses are obvious from visual inspection of the speech waveform. For the purposes of analysis, the speech waveform was broken up into a number of analysis frames. The size of these analysis frames varied depending on the pitch estimate. Each frame was windowed for subsequent spectral analysis. This analysis used a 512-pt discrete Fourier transform. A more detailed examination of one of the frames provides additional insight.

The frame starting at .8 seconds is strongly voiced and clearly contains secondary pulses. For this frame, the pitch was correctly estimated to be approximately 50 Hz and the frame length was chosen to be around 5 milliseconds. The measured phase of the windowed waveform for this frame is shown in Figure 4-2. The system phase and linear phase of the system are shown in Figures 4-3 and 4-4 respectively.

The corresponding phase residual is shown in Figure 4-5. The time waveform in this frame and the estimated secondary pulse locations are presented in Figure 4-6.

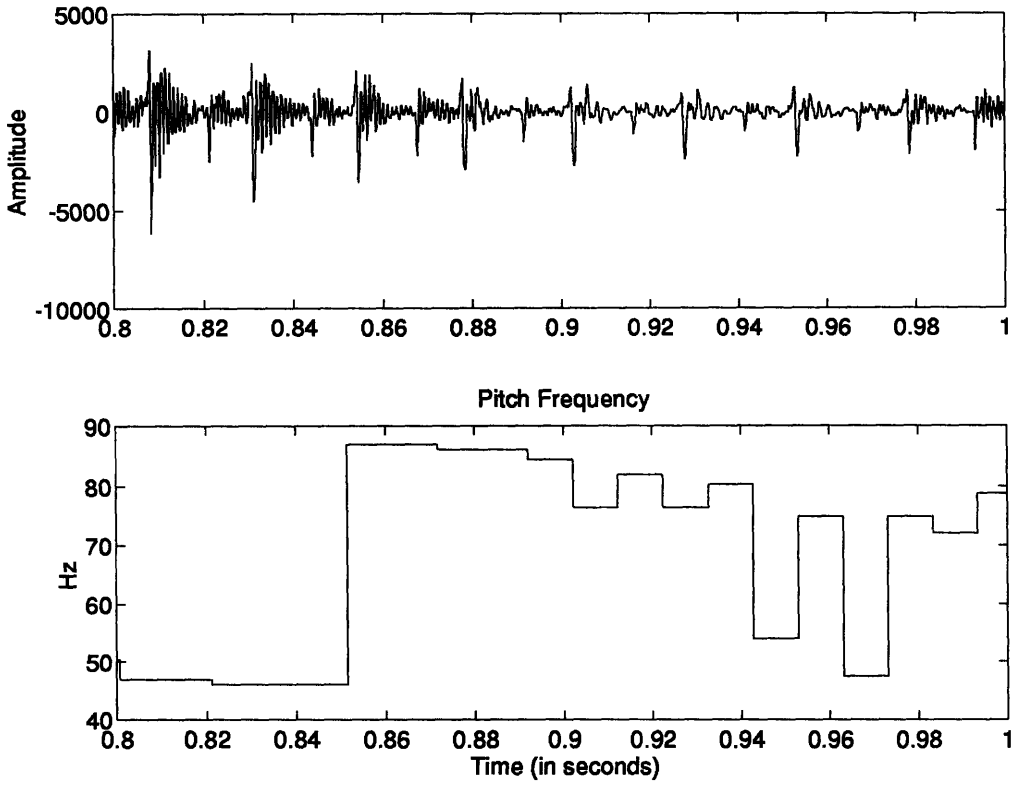

Figure 4-1: Time Waveform and pitch estimate for a portion of speech

(The secondary pulse amplitudes are not meaningful on this graph since the time waveform shown is the speech itself whereas the secondary pulses reflect the excitation.) Some representative slices of the likelihood function are shown in Figures 4-7 and 4-8. Note that the offset is shown normalized by the estimated pitch period. The slices for a fixed α (variable n_d) are more similar to the likelihood functions of the synthetic vowels formed with a non-impulsive glottal shape (Figures 3-6 and 3-7) than those formed by an impulsive glottal shape (Figure 3-18). The deviations from the ideal case have again caused the estimator to fail, even with a waveform

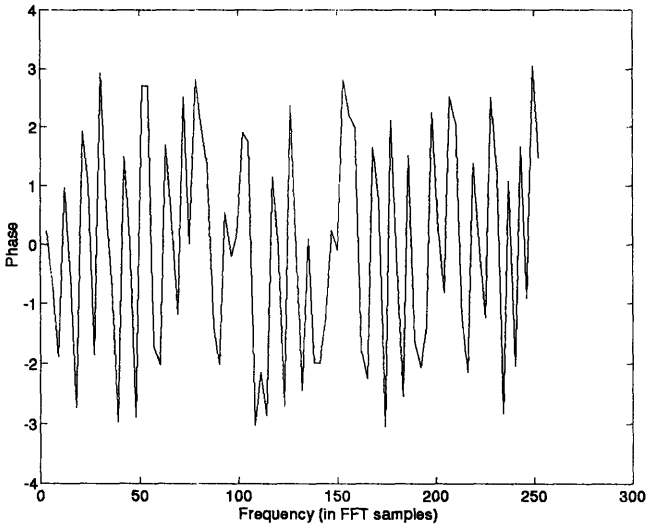

Figure 4-2: Measured phase of waveform at .8 seconds

with definite, distinguishable secondary pulses. The basic shape of the likelihood function is similar for most of the frames in this waveform.

There are several explanations for the failure of the estimator to identify the true location of the secondary pulse. First, the speech coder may have incorrectly chosen the onset time for the primary pulse. This would potentially introduce a strong linear phase component into the phase residual, leading to spurious secondary pulse estimates. Second, the combination of a true secondary pulse and zeros introduced by the glottal excitation may lead to a phase residual which the estimator cannot directly model with a single secondary pulse. The optimal location, in terms of waveform matching, need not be anywhere near the actual secondary pulse. The estimator is therefore unreliable at finding the true locations of secondary pulses. A final source of error for this system is faulty pitch estimation. In the vicinity of .85 seconds, the pitch estimator begins to identify the secondary pulses as primary pulses and doubles its estimate of the fundamental frequency. In these cases, the secondary pulse estimator necessarily fails.

The essential shortcoming of this secondary pulse estimator apparently lies in the difficulty in separating the effects of secondary pulses and zeros in the system

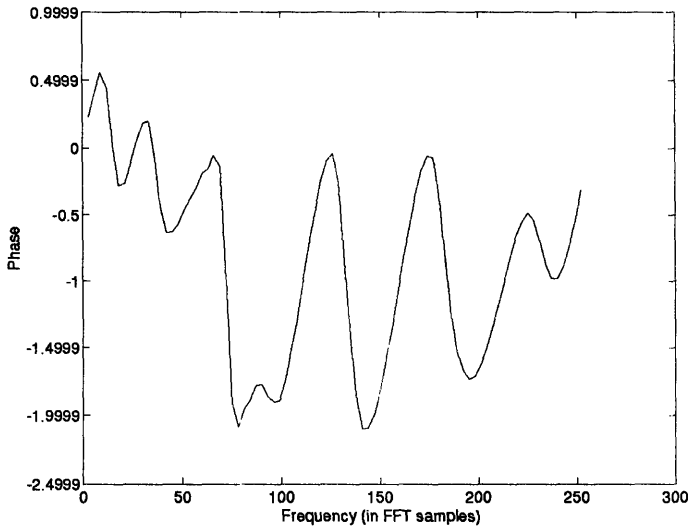

Figure 4-3: System phase component of waveform at .8 seconds

function or excitation on the phase. The total phase cannot be easily decomposed into components resulting from the two different sources. The phase function is a complicated non-linear function, so it is difficult to extract its components reliably. This estimator makes no attempt to model any effect other than that from a single impulsive secondary pulse and appears to be unable to handle even a small perturbation from its model.

4.2 Speech Coding and Phase Modeling

A limited set of experiments was conducted to determine the usefulness of the secondary pulse estimator in speech coding. The goal of a speech coder is to develop a compact representation of speech which produces output which is perceptually close to the original. For sinusoidal transform coders it would be desirable to formulate an accurate, compact model for the component sine wave phases since they have proven very difficult to code directly [5]. The phase of the minimum phase system function is generally easy to transmit, so it is natural to use the phase decomposition of Equation 1.2. The problem then becomes the familiar one of modeling

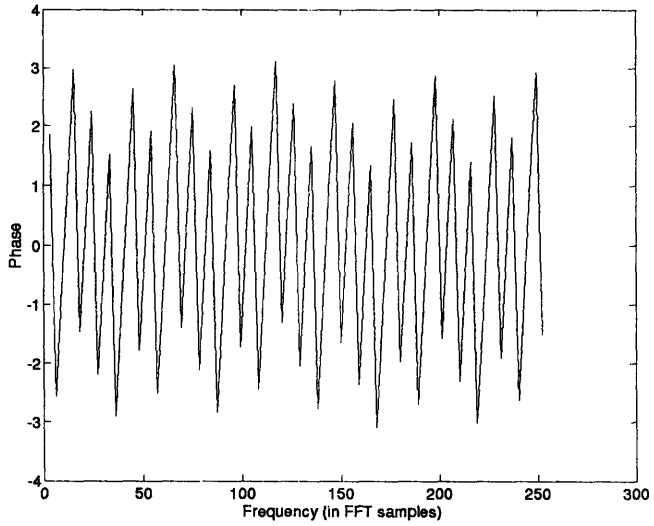

Figure 4-4: Linear phase component of waveform at .8 seconds

the excitation phase. The linear model of Equation 1.3 has been used effectively to represent the excitation phase, but it causes perceptually significant degradation of the synthesized speech. Unfortunately, listeners are very sensitive to small variations in the parameter n_o. Thus implementation of Equation 1.3 requires not only a good onset estimator but also very fine quantization of the onset time and a correspondingly large number of bits. By postulating the presence of secondary pulses, it is possible to augment the linear phase model with the phase contributions of several additional pulse trains. These phase contributions could then be represented compactly by the analyzer as a set of locations and relative amplitudes. Given these parameters, the synthesizer could then use Equation 3.1 to reconstruct the phase contribution from each pulse train to add to the linear phase model to form a total phase estimate. Since the success of this coding scheme depends not on the existence of actual secondary pulses in the excitation but rather the accuracy of the secondary pulse model in representing phase, this estimator, even without locating the position of an actual secondary pulse, may still produce a useful though physiologically meaningless representation of the phase residual. Since the derivation of the likelihood function explicitly minimized signal error rather than phase error, it should always produce an approximation to the original which is superior in the mean-squared-error sense to that of the simple linear phase model.

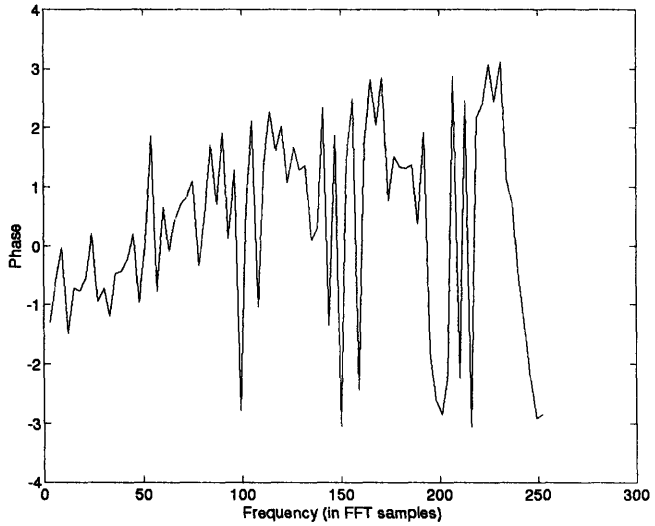

Figure 4-5: Phase residual of waveform at .8 seconds

A coder which added the phase contribution of a single secondary pulse to its phase estimate was implemented and tested on a small number of sentences for both male and female speakers. In all cases, the speech synthesized with this system was perceptually indistinguishable from that synthesized with the simple phase model. The system was then expanded to estimate eight secondary pulse trains iteratively for each analysis frame. This system was evaluated using a single test sentence and the resulting speech, while recognizable, was considerably degraded. The reasons for this failure are unclear. It is known that the perceptual quality of synthesized speech is very sensitive to small phase errors [5], so it is certainly possible that some level of phase quantization introduced by the system degraded the speech. Additional research is needed to support this assertion. In addition, by breaking up the search for additional pulses into an iterative process, the results may no longer be globally optimal. There is some justification for a suboptimal iterative procedure in that it is a variant of the successful system used in time-domain multipulse coders [1]. In addition, the sensitivity of phase across the whole spectrum to each excitation pulse would tend to increase the difficulty in segregating the effects of each pulse. The iterative analysis synthesis loop also only subtracts off the phase effects of

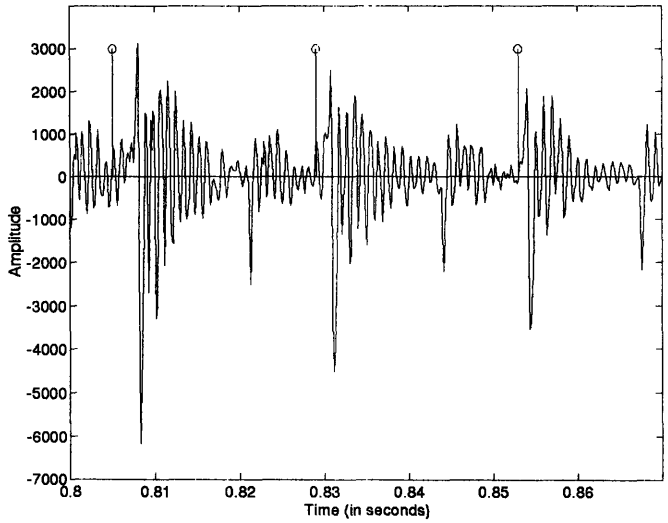

Figure 4-6: An analysis frame of speech with superimposed estimated secondary pulse locations

each estimated pulse. Any effects of each additional pulse on the magnitude of the frequency spectrum remain unmodeled.

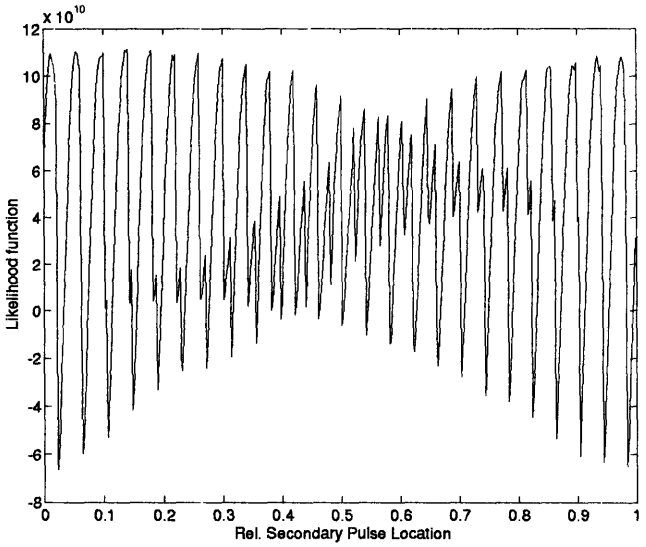

Figure 4-7: Likelihood function vs. normalized offset, speech waveform at .8 seconds (slice at $\alpha = .98$)

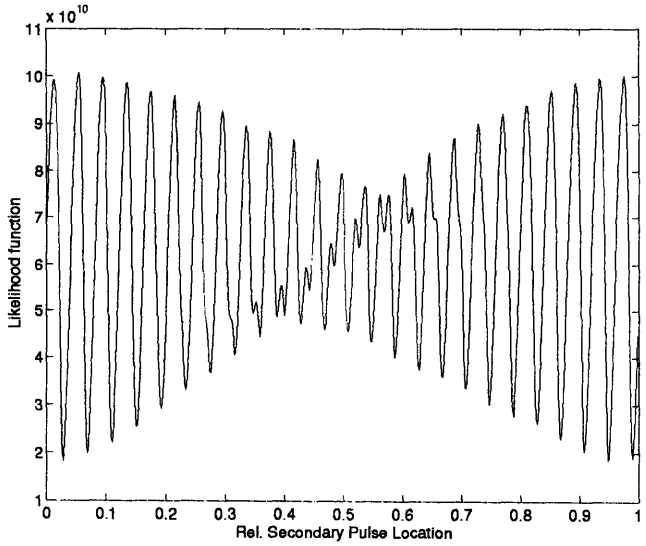

Figure 4-8: Likelihood function vs. normalized offset, speech waveform at .8 seconds (slice at $\alpha = .5$)

Chapter 5

Summary

5.1 Secondary pulse estimation

The experiments described in Chapters 3 and 4 highlight the deficiency of this estimator as an estimator of physiological secondary pulses. While the estimator performed correctly for the idealized case of a synthesized vowel, it was unable to pick out very pronounced secondary pulses from segments of human speech reliably. This failure may result from the fundamental nature of the phase function and its components. Unmodeled components of the phase seem to have a serious effect on the estimator's performance as shown in its failure for non-impulsive synthetic speech. For human speech, the linear model, with or without secondary pulses, is a reasonable perceptual approximation; however, it is far from exact. The reliance of this estimator on phase appears to make it especially sensitive to these errors. The failure of the estimator on synthetic vowels formed with a non-impulsive excitation corroborates this conclusion.

It is important to note that the estimator was designed specifically to find the location and amplitude of a secondary pulse such that its phase contribution would perceptually approximate the phase residual of a speech segment. In the presence of real speech, this does not necessarily have to correspond to an actual secondary pulse in the glottal excitation. Essentially, the derivation of the likelihood function relied on the estimator being part of a waveform and phase modeler, rather than

an estimator of physical excitation secondary pulses. For the idealized case of a synthetic vowel with impulsive excitation, the two problems are equivalent since the components of the speech signal are known exactly and forced to fit the linear model; however, for real speech, the two problems are no longer necessarily equivalent as additional unknown or uncharacterized effects are introduced.

5.2 Speech coding and phase modeling

The speech coding experiments described in Chapter 4 were part of an exploratory investigation into the viability of the multiple pulse model in coding phase residuals for a frequency domain sinusoidal coder. The success of multiple pulse models in the time domain, especially the multipulse method [1], would suggest that a successful equivalent system could be realized in the frequency domain. However, there are many issues and complexities involved in designing such a system. Clearly, the system used in this series of experiments is not sophisticated enough to produce natural sounding speech. The experiments examining the use of the estimator to pinpoint the locations of actual secondary pulses show that phase is not a reliable indicator of secondary pulses. No definite conclusions have been drawn to suggest the viability of using secondary pulses to model phase residuals. The results of these experiments show that the simplest multiple pulse model does not work. They also highlight the difficulty of finding parametric models to represent the phase function of speech signals. The relationship between the frequency domain phase of a waveform and the waveform itself is a nonlinear one and therefore simple yet precise models are difficult to identify.

Bibliography

[1] B.S. Atal and J.R. Remde. A new model of lpc excitation for producing natural-sounding speech at low bit rates. In *Proc. Int. Conf. on Acoustics, Speech and Signal Proc.*, volume 1, pages 614–617, May 1982.

[2] J.N. Holmes. Formant excitation before and after glottal closure. In *Conf. Rec. 1976 IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pages 39–42, April 1976.

[3] C.R. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, pages 109–112, 1990.

[4] D.H. Klatt and L.C. Klatt. Analysis, synthesis, and perception of voice quality in female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857, 1990.

[5] R.J. McAulay and T.F. Quatieri. Low-rate speech coding based on the sinusoidal model. In S. Furui and M.M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 165–208. Marcel-Dekker, 1992.

[6] A.V. Oppenheim and R.W. Schafer. *Digital Signal Processing.* Prentice-Hall, Englewood Cliffs, N.J., 1975.

[7] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw Hill, second edition, 1984.

[8] T.F. Quatieri, C.R. Jankowski, and D.A. Reynolds. Energy onset times for speaker identification. *IEEE Signal Processing Letters*, 1(11):160–162, 1994.

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

[10] T. Tremain. The government standard linear predictive coding algorithm: Lpc-10. *Speech Technology*, 1(2):40–49, April 1982.