Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification

by

Michael David Plumpe

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

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1997

Author .....
Department of Electrical Engineering and Computer Science
January 17, 1997

Certified by .....
Thomas F. Quatieri Jr.
Senior Staff, MIT Lincoln Lab
Thesis Supervisor

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

Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification

by

Michael David Plumpe

Submitted to the Department of Electrical Engineering and Computer Science
on January 17, 1997, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Speech production has long been viewed as a linear filtering process, as described by Fant in the late 1950's [10]. The vocal tract, which acts as the filter, is the primary focus of most speech work. This thesis develops a method for estimating the source of speech, the glottal flow derivative. Models are proposed for the coarse and fine structure of the glottal flow derivative, accounting for nonlinear source-filter interaction, and techniques are developed for estimating the parameters of these models. The importance of the source is demonstrated through speaker identification experiments.

The glottal flow derivative waveform is estimated from the speech signal by inverse filtering the speech with a vocal tract estimate obtained during the glottal closed phase. The closed phase is determined through a sliding covariance analysis with a very short time window and a one sample shift. This allows calculation of formant motion within each pitch period predicted by Ananthapadmanabha and Fant to be a result of nonlinear source-filter interaction during the glottal open phase [1]. By identifying the timing of formant modulation from the formant tracks, the timing of the closed phase can be determined. The glottal flow derivative is modeled using an LF model to capture the coarse structure, while the fine structure is modeled through energy measures and a parabolic fit to the frequency modulation of the first formant.

The model parameters are used in the Reynolds Gaussian Mixture Model Speaker Identification system with excellent results for non-degraded speech. Each category of source features is shown to contain speaker dependent information, while the combination of source and filter parameters increases the overall accuracy for the system. For a large dataset, the coarse structure parameters achieve 60% accuracy, the fine structure parameters give 40% accuracy, and their combination yields 70% correct identification. When combined with vocal tract features, the accuracy increases to 93%, slightly above the accuracy achieved with just vocal tract information. On smaller datasets of telephone-degraded speech, accuracy increases up to 20% when source features are added to traditional mel-cepstral measures.

Thesis Supervisor: Thomas F. Quatieri Jr.
Title: Senior Staff, MIT Lincoln Lab

Acknowledgments

I would like to thank my advisor, Tom Quatieri, for all of the excellent questions he asked and ideas he proposed during the past year and a half. He has been a great help from the beginning of my term project up through the completion of this document. Perhaps one of his best contributions is that he and I tend to think of things from different angles, hopefully I will carry his viewpoint along with my own after I leave MIT. I would also like to thank Doug Reynolds for helping me understand and use his speaker identification system. Thanks are also owed to all the members of the Speech System Technology Group who have helped me in so many ways. I would also like to thank all the people who thought it would be a shame for me to not get an advanced degree and all my friends in Seattle who signed their letters "quit school." These conflicting views enabled me to make up my own mind with a minimum of outside pressure.

And, of course, I wish to thank the sponsors of my research and my time here at MIT. I would like to thank the EECS department for awarding me a fellowship, which gave me the freedom to find an excellent group, advisor, and to choose a project which I have found very interesting. The Air Force sponsorship is also greatly appreciated, as it allowed me to focus my attention on research, which is my true interest.

Contents

1 Introduction **15**

1.1 Background 15

1.1.1 Linear Source/Filter Production Model 16

1.1.2 Linear Prediction 17

1.1.3 Inverse Filtering 19

1.1.4 Pre-emphasis 20

1.2 Motivation for the Use of the Glottal Flow in Speaker Identification . 21

1.3 Properties of the Glottal Flow Derivative 24

1.3.1 Coarse Structure 24

1.3.2 Fine Structure 27

1.4 Related Work 29

1.4.1 Previous Attempts at Estimating the Glottal Flow 29

1.4.2 Previous Uses of Source Information for SID 31

1.5 Thesis Contribution 32

1.6 Thesis Organization 33

2 The Glottal Flow Model **34**

2.1 Physical Model 34

2.1.1 Detailed Physiological Simulation 34

2.1.2 Simplified Production Model 35

2.2 Feature Model 41

2.2.1 Coarse Structure 41

2.2.2 Fine Structure 43

2.3 Summary 46

3 Calculation of the Glottal Flow Derivative Waveform Estimate 48

3.1 Determination of the Closed Phase 48

3.1.1 Initial Glottal Closure Estimate 49

3.1.2 Sliding Covariance Analysis 51

3.1.3 Measuring Formant Motion 56

3.1.4 High Pitch Speakers: Using Two Analysis Windows 61

3.2 From Closed Phase to Glottal Flow Derivative 62

3.2.1 Vocal Tract Response 62

3.2.2 Inverse Filtering 63

3.3 Examples 65

3.4 Summary 66

4 Estimating Coarse Structure 68

4.1 Formulation of the Estimation Problem 68

4.2 The NL2SOL Algorithm 70

4.3 Using the NL2SOL Algorithm 73

4.3.1 Difficulties with the NL2SOL Algorithm 73

4.3.2 Discarding Data 75

4.3.3 Initialization 76

4.4 Other Possible Approaches 76

4.5 Examples 77

4.6 Summary 77

5 Estimating Fine Structure 80

5.1 Modeling Ripple Through Formant Modulation 80

5.2 Time Domain Fine Structure 82

5.3 Examples 83

5.4 Summary 84

6 Speaker Identification Experiments **86**

6.1 Background 86

6.2 Difficulties with the Reynolds GMM SID System 87

6.3 Using Source Features for SID 88

6.4 SID for Degraded Speech 91

6.5 Summary 93

7 Conclusions **94**

7.1 Summary of Findings 94

7.2 Suggestions for Future Work 96

List of Figures

1-1 The Vocal Folds, as seen from above. (a): Configuration during quiet breathing. (b): Configuration for Vocal Fold Vibration [49] 22

1-2 Schematic source waveforms. (a): Glottal Area, (b): Corresponding Glottal Flow, (c): Glottal Flow Derivative, and (d): Log-spectrum of (c). 26

1-3 Ripple will be seen on the glottal flow derivative waveform due to source-filter interaction, as shown in this schematic representation. . . 27

1-4 Vowel /a/ showing truncation of the fourth cycle of the first formant 28

2-1 Equivalent circuit for a single formant load, from [29] 36

2-2 Norton equivalent circuit for single formant load showing time-varying elements, from [29] 37

2-3 Source-filter interaction causes modulation of formant frequencies and bandwidths when the glottis is open. Although these effects are non-linear, they can be approximated in a linear framework. This figure shows the function $g_o(t)$, which is proportional to the glottal area in (a), the formant frequency in Hz (b), the formant bandwidth in Hz (c), and speech waveforms generated with a time-invariant formant and with these formant modulations (d). Dashed : no formant modulation. Solid : formant modulation as shown in panels (b) and (c). . . 39

2-4 LF Model for the glottal flow derivative waveform 42

2-5 Often the glottal flow derivative exhibits a period containing a small amount of noise immediately after the return phase, followed by a period that shows ripple but no significant flow, followed by the standard glottal pulse and return phase. (a): Estimated Glottal Flow Derivative (solid) and overlaid LF model (dashed). (b): The error for the fitted LF model (see chapter 4), containing aspiration noise and ripple due to source-filter interaction 45

3-1 Inverse filtering speech using a pitch synchronous window and no pre-emphasis results in a waveform with sharp pulses occurring at the time of glottal closure. These pulses can be easily identified using a peak picking algorithm. (a): Speech waveform with a sequence of possible analysis windows. (b): Resulting inverse filtered waveform. 50

3-2 Flowchart of procedure to estimate glottal pulses. 52

3-3 The closed phase is identified by the region in which the first formant frequency is stable. Panel (a) shows the formant frequency tracks over several pitch periods of the speech in (b). Panel (b) also shows the first and last two analysis windows for one frame. Each formant value is calculated from a small window of speech samples. The closed phase is defined as every speech sample in the windows used to calculate the formant values in the stable region. x : F1, o : F2, + : F3. Formant values are shown at the end of the corresponding analysis window. Formant values of 0 are displayed for regions in which no value was calculated. 54

3-4 Glottal opening is identified by growing a small region in which the first formant frequency is stable until the next sample is greater than two standard deviations from the mean of the formants in the region. The procedure is illustrated in this flow chart. 57

3-5 Glottal Closure is identified similarly to glottal opening, except that the mean and standard deviation of the formant values within the region are not updated as the region is grown. The procedure is illustrated in this flow chart. 58

3-6 This plot shows the pole frequencies of the impulse response to a two pole system whose lower pole frequency changes at approximately sample 40. The frequencies are generated by a sliding covariance analysis. The higher pole has a larger bandwidth, causing the energy at that frequency to fall below the noise floor at approximately sample 20. When the filter is changed, the higher pole is excited due to the redistribution of energy that occurs when the characteristic modes of the response change. 65

3-7 Several examples of estimated glottal flow derivatives. The speech signals are above the corresponding glottal flow derivative waveforms. Each row represents a speaker. All the examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels. 67

4-1 LF Model for the glottal flow derivative waveform. (Repeat of figure 2-4 for convenience) 69

4-2 The parameters α and E_0 of the LF model for the glottal flow derivative waveform can be traded off for large values of α as demonstrated in this figure. A 10% increase in α and a 50% decrease in E_0 result in a squared error change of only 0.2% for this example. (a): Two superimposed LF waveforms, one with $\alpha = 7$ and $E_0 = 100$, the other with $\alpha = 7.7$ and $E_0 = 50$. (b): Difference between the two waveforms in (a). 75

4-3 Several examples of the LF model for the coarse structure in the estimated glottal flow derivative. The glottal flow derivatives are shown above the corresponding model waveforms. Each row represents a speaker. All the examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels. 78

5-1 Several examples of the LF model for the coarse structure in the estimated glottal flow derivative. The glottal flow derivatives are shown above the corresponding model waveforms. Each row represents a speaker. All the examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels. 85

7-1 Four examples of unusual glottal flow derivatives. In each case, the speech waveform is shown above the glottal flow derivative waveform. Superimposed on the glottal flow derivative waveform are small pulses indicating the timing of the glottal opening, closure, and pulse. Panels (a) and (b) show evidence of multiple pulses, likely due to the sudden onset of ripple. Panel (c) shows a case of a large amount of ripple, while panel (d) shows an error in identification of the closed phase and the resultant incorrect glottal flow derivative waveform. 101

List of Tables

3.1 Average Signal to Noise Ratios for several potential measures used in identifying the glottal opening. The closed phase was identified using the first formant frequency. 60

3.2 Average Signal to Noise Ratios for several potential measures used in identifying the glottal opening. The closed phase was identified using the second formant frequency. 60

4.1 Description of the seven parameters of the LF model for the glottal flow derivative waveform. 70

6.1 Speaker identification results for various combinations of the source parameters 90

6.2 Speaker identification results for mel-cepstral representations of the Glottal Flow Derivative (GFD) waveform and the modeled GFD waveform. 91

6.3 Speaker identification results for mel-cepstral representations of the speech signal, the Glottal Flow Derivative (GFD) waveform, the modeled GFD waveform, and combinations of the speech and source mel-cepstral data. All of the data was generated from the NTIMIT database. 92

Chapter 1

Introduction

As the source for voiced speech, the volume velocity airflow through the glottis, called the *glottal flow*, has a major impact on the characteristics of speech. The goal of this thesis is to estimate the glottal flow from speech waveforms, model the important features of the glottal flow, and use the model parameters for speaker identification. We start this chapter with a basic mathematical framework for the linear model of speech production. We then provide motivation for the importance of glottal flow characteristics and discuss, in general terms, the features of the source that we wish to model. This is followed by a description of previous analysis techniques and previous applications of source information to speaker identification. We then discuss the contributions of this thesis. Finally, we provide an outline for the remainder of the thesis.

1.1 Background

We begin with a mathematical framework for the classical linear speech production model.

1.1.1 Linear Source/Filter Production Model

Speech production is typically viewed as a linear filtering process which can be considered time invariant over short time intervals, such as 20ms. The vocal tract, with impulse response $h[n]$ is excited by a signal $e[n]$, and the speech signal $s[n]$ is the output of the vocal tract filter $h[n]$ filtered by the lip radiation $r[n]$:

$$s[n] = e[n] * h[n] * r[n] \tag{1.1}$$

$$S(z) = E(z)H(z)R(z), \tag{1.2}$$

where equation 1.1 is the discrete time representation of linear filtering, and equation 1.2 is its z-transform domain representation. Radiation can be approximated as a first difference operation for the frequencies of interest in speech, i.e., $r[n] = \delta[n] - \delta[n-1]$, and is typically included in the excitation function, as we shall do throughout this thesis, giving

$$\begin{aligned} s[n] &= (e[n] - e[n-1]) * h[n] \\ &= \tilde{e}[n] * h[n], \end{aligned} \tag{1.3}$$

where $\tilde{e}[n] = e[n] * r[n]$. In the z-domain we have

$$\begin{aligned} S(z) &= (1 - z^{-1})E(z)H(z) \\ &= \tilde{E}(z)H(z). \end{aligned} \tag{1.4}$$

For voiced speech, the excitation signal $e[n]$ is the volume velocity airflow through the vocal folds. By including radiation in the excitation function, the source becomes the derivative of the volume velocity. Henceforth, we drop the tilde notation and assume $e[n]$ to contain radiation.

It can be shown from acoustics that the vocal tract filter $H(z)$ is an all-pole filter

for vowels when the nasal passage is closed off from the vocal tract:

$$\begin{aligned} H(z) &= \frac{1}{\prod_{i=1}^{p}(1 - c_i z^{-1})} \\ &= \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \end{aligned}$$

where the vocal tract is represented as having p poles. In general, the poles will come in complex conjugate pairs since this is a real system. For an all-pole system, equation 1.3 becomes:

$$s[n] = e[n] + \sum_{i=1}^{p} a_i s[n-i]. \tag{1.5}$$

1.1.2 Linear Prediction

In order to estimate the filter $h[n]$ from the speech signal $s[n]$, we set up a least-squares minimization problem where we wish to minimize the error

$$e[n] = s[n] - \sum_{i=1}^{p} \alpha_i s[n-i], \tag{1.6}$$

where α_i are the calculated estimates of a_i. The total error is given by

$$E = \sum_{R} e^2[n], \tag{1.7}$$

where the error is to be minimized for the region R. Solutions of this minimization problem are called *linear prediction*, since the error $e[n]$ is the difference between the speech sample $s[n]$ and the value $\hat{s}[n]$ predicted by a linear combination of previous values of the signal $s[n]$. There are many different techniques of linear prediction, based on how $e[n]$ is calculated over the region R.

If we assume that the speech signal is zero outside of an interval $0 \leq n \leq N-1$, then the signal $e[n]$ will be non-zero only during the interval $0 \leq n \leq N+p-1$, which gives us the region R. This choice will give large errors at the start of the interval, since we are trying to predict non-zero speech samples from zero, as well as at the end, where we are trying to predict zero samples from non-zero data. These assumptions

result in the *autocorrelation method* of linear prediction, since the solution to this problem involves an autocorrelation matrix.

$$\boldsymbol{R}\vec{\alpha} = \vec{r}, \tag{1.8}$$

where the $(i,j)^{th}$ term of $\boldsymbol{R}$ is given by $r_{i,j}$, where

$$r_{i,j} = \sum_{n=0}^{N-1-|i-j|} s[n]s[n+|i-j|], \tag{1.9}$$

where $1 \leq i,j \leq p$. The two vectors are given by

$$\begin{aligned} \vec{\alpha} &= [\alpha_1, \alpha_2, \ldots, \alpha_p]^T, \text{ and} \\ \vec{r} &= [r_{0,1}, r_{0,2}, \ldots, r_{0,p}]^T. \end{aligned}$$

The primary benefit of the autocorrelation method is that it is guaranteed to produce a stable filter. The autocorrelation technique will calculate the correct filter only if the analysis window is of infinite length, due to the large errors at the beginning and end of the window. To help reduce the effects of using a finite data window, the data is typically windowed with a non-rectangular window.

If $e[n]$ is calculated over a finite region, with the appropriate speech samples before the window used in the calculation of $e[n]$, the solution to the minimization problem is called the *covariance method* of linear prediction:

$$\boldsymbol{\Phi}\vec{\alpha} = \vec{\psi}, \tag{1.10}$$

where the $(i,j)^{th}$ term of $\boldsymbol{\Phi}$ is given by $\phi_{i,j}$, where

$$\phi_{i,j} = \sum_{n=0}^{N-1} s[n-i]s[n-j] \ : \ 1 \leq i,j \leq p \tag{1.11}$$

and the two vectors are given by

$$\begin{aligned}\vec{\alpha} &= [\alpha_i, \alpha_2, \ldots, \alpha_p]^T, \text{ and} \\ \vec{\psi} &= [\phi_{0,1}, \phi_{0,2}, \ldots, \phi_{0,p}]^T.\end{aligned}$$

This matrix problem can be solved efficiently used Cholesky decomposition because the matrix $\mathbf{\Phi}$ has the properties of a covariance matrix.

The benefit of the covariance method is that with its finite error window, a correct solution will be achieved for any window length greater than p if no noise is present. Also, since the boundaries are handled correctly, a rectangular window can be used with no ill-effects. For a more detailed discussion of linear prediction, including derivations for the solutions given, see [45] or [37].

From a spectral standpoint, linear prediction attempts to match the power spectrum of the signal $s[n]$ to the predicted filter given by the α_i's. In particular, the error function $e[n]$ is given in the frequency domain by:

$$E(\omega) = \frac{P(\omega)}{\hat{P}(\omega)}, \tag{1.12}$$

where $P(\omega)$ is the power spectrum of the signal $s[n]$, and $\hat{P}(\omega)$ is the power spectrum of the estimated filter [37]. If the excitation function has a non-uniform spectrum, the α_i's calculated will be influenced to result in a spectrum $\hat{H}(z)$ that matches $H(z)E(z)$.

1.1.3 Inverse Filtering

Re-arranging equation 1.5, we can estimate the excitation signal $e[n]$ from the speech signal $s[n]$ and the estimated vocal tract response given by the α_i's:

$$\hat{e}[n] = s[n] - \sum_{i=1}^{p} \alpha_i s[n-i], \tag{1.13}$$

or in the frequency domain,

$$\begin{aligned} \hat{E}(z) &= S(z)\frac{1}{\hat{H}(z)} \\ &= E(z)H(z)\frac{1}{\hat{H}(z)}. \end{aligned} \tag{1.14}$$

These equations describe a process called *inverse filtering*, in which the estimated vocal tract response is removed from the speech to yield an estimate $\hat{e}[n]$ of the source function.

1.1.4 Pre-emphasis

Speech signals are commonly *pre-emphasized* before linear prediction analysis is performed. Pre-emphasis is the process of filtering the speech signal with a single zero high pass filter:

$$s_p[n] = s[n] - \beta_p s[n-1], \tag{1.15}$$

where β_p is the pre-emphasis coefficient. The value used for β_p is typically around 0.9 to 0.95.

While it is difficult to find reasoning for using pre-emphasis in the literature, we give two reasons here. As discussed above, the filter estimated by linear prediction will match the power spectrum of the combined excitation and vocal tract. The excitation has a spectral shape which has more energy at low frequencies than high frequencies, as will be seen below. In order to approximately remove the large-scale spectral contribution of the source, the speech signal is pre-emphasized. The resulting spectrum is a closer representation of the vocal tract response, and thus the filter calculated through linear prediction is a better match for the vocal tract response.

The other reasoning for pre-emphasis is an argument based on the spectral properties of the error function minimized. As can be seen in equation 1.12, the error is the ratio of the two power spectrum, which results in uniform spectral matching in a squared sense regardless of the energy at any particular frequency. Speech spectra are typically viewed on a log or dB plot, however, which will show better matching

for high energy regions of the spectrum than for low energy regions. Since speech tends to have a decrease in energy at high frequencies, the high-pass filter effect of pre-emphasis will help achieve more uniform spectral matching in a log sense across the entire spectrum.

1.2 Motivation for the Use of the Glottal Flow in Speaker Identification

We now give arguments for the importance of the glottal flow for speaker identification. During voicing, we consider the excitation function $e[n]$ to be the glottal flow derivative. In recent years, the importance of the glottal flow has been recognized and studied, especially for naturalness in speech synthesis systems [32], and for providing correlates to various vocal registers and speaking manners (loud, angry, etc.) [4, 5, 8, 26, 42]. In addition to these areas, there are many reasons that the glottal flow should be speaker dependent.

Figure 1-1 shows a diagram of the vocal folds and related structures as viewed from above. Videos of vocal fold vibration, such as [34], show large variations in the movement of the vocal folds from one individual to another. Perhaps the most basic of these is how completely the vocal folds close. For some individuals the vocal folds close completely, in what is referred to as *modal phonation.* In modal phonation, there is a period of time during which the vocal tract is completely separate from the lungs. For other individuals, there might be a small region at the arytenoid cartilages at the posterior end of the vocal folds that never closes completely, this is referred to as a *glottal* or *posterior chink.* Less common are openings at the anterior end of the vocal folds, at the center of the folds, or along their entire length [23]. Fixed glottal openings result in more aspiration noise, and tend to result in a higher amplitude of the fundamental as compared to the first harmonic, due to a slower glottal closure [32]. Incomplete glottal closure also results in other features which can be measured, such as the amount of source-filter interaction, which will be discussed below.

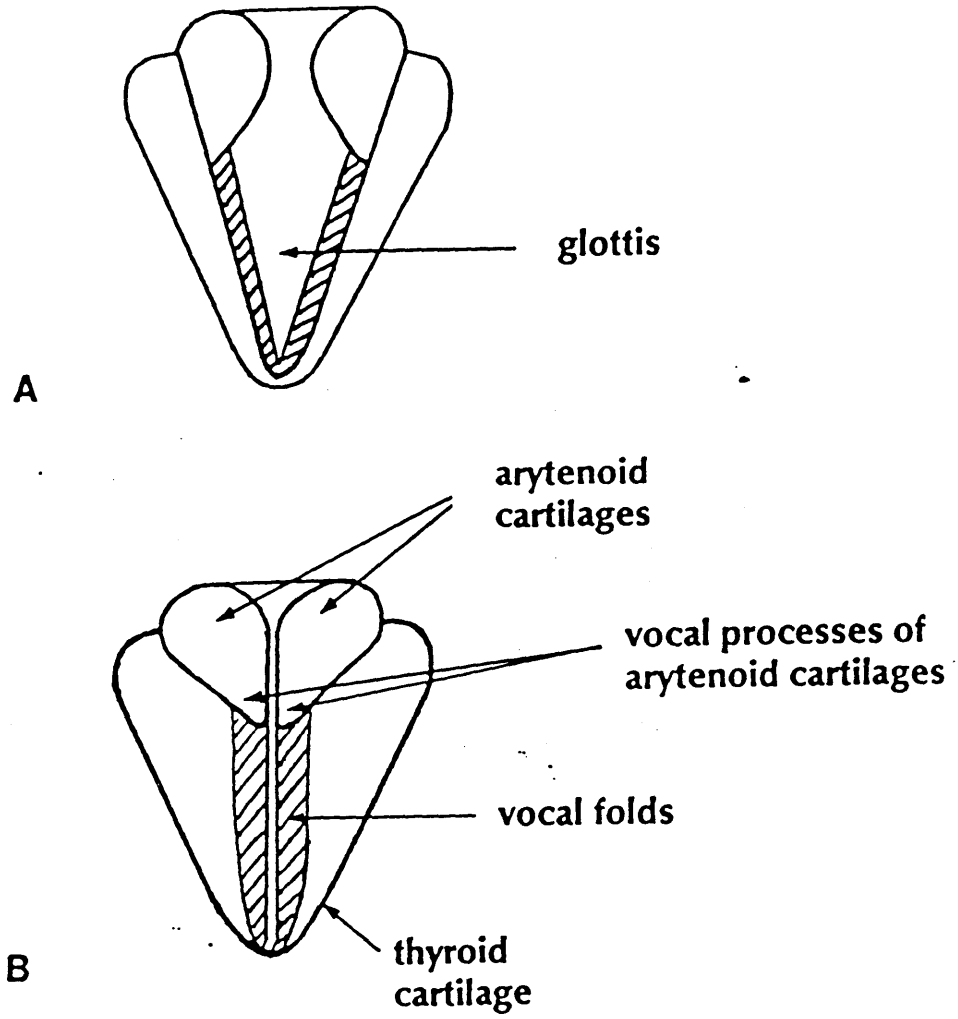

Figure 1-1: The Vocal Folds, as seen from above. (a): Configuration during quiet breathing. (b): Configuration for Vocal Fold Vibration [49]

The second most noticeable feature of vocal fold vibration is the manner in which they close. For some individuals, the cords close in a zipper-like fashion, while for others they close along the length of the vocal folds at approximately the same time. The speed of glottal closure determines the spectral content of the glottal source, since a rapidly closing glottis acts like an impulse and creates a source with a wide bandwidth. The more slowly the vocal folds close, the less energy is present at higher harmonics as compared to the fundamental, this is referred to as the *spectral tilt* of the source.

Another difference in the glottal opening is the configuration of the area of the opening. For some speakers, it may be approximately equal in width along the length of the glottis, such as in the case of *laryngealized* or *pressed* phonation, in which the arytenoid cartilages are pressed together. This results in both ends of the vocal folds being held close together, creating an opening shaped like a football. For speakers whose arytenoid cartilages are spread apart, a more triangle shaped opening will occur. Van den Berg has proposed an empirical formula for relating the steady-state pressure drop across glottis to the glottal flow:

$$\Delta P = (k\rho U^2)/(2A^2) + 12\mu DlU/A^3, \tag{1.16}$$

where ΔP is the pressure drop across the glottis, called the *trans-glottal pressure*, k is an experimentally determined constant, ρ is the density of air, U is the glottal flow, A is the glottal area, μ is the coefficient of viscosity, D is the glottal depth, and l is the length of the glottal opening. The first term, which dominates the equation, calculates the kinetic resistance of the flow, while the second term accounts for the viscous coupling. The kinetic resistance is the energy required to accelerate air from the large lung cavity through the narrow glottal opening and into the vocal tract. The viscous coupling term arises due to interaction of the flow with the walls of the glottis, and is dependent on not only the area of the glottal opening but also on the ratio l/A. Intuitively, a configuration that has a longer circumference for a given area will have a larger pressure drop due to viscous coupling with the walls of the glottis.

Thus the glottal flow will depend on the configuration of the glottal area.

While the influence of the viscous term is difficult to account for mathematically in a time-varying, nonuniform vocal tract, it seems evident that the configuration of the glottal opening influences the glottal flow. A narrow opening will have more viscous resistance, resulting in less flow with a maximum that is reached more quickly and maintained for a longer period of time. Source-filter interaction will be increased, since the viscous term is proportional to the trans-glottal pressure, while the kinetic term is proportional the the square root of the trans-glottal pressure. Any variation in pressure above the glottis will thus have a larger impact on the glottal flow.

Previous studies and video of vocal fold vibratory patterns give strong reason to believe that the motion of the vocal folds has speaker dependent characteristics. The glottal flow, which is closely related to the glottal opening, is modified in a predictable manner from these variations in vocal fold motion. As the source for voiced speech, an analysis system can be devised in which the glottal flow is estimated, and features identified which are useful for speaker identification, as well as other applications, such as more natural speech synthesis.

1.3 Properties of the Glottal Flow Derivative

This section describes the features of the glottal flow derivative waveform in general terms.

1.3.1 Coarse Structure

Vowel production can be viewed as a simple linear filtering problem, where the system is time invariant over short time periods. Under these assumptions, the glottal flow, acts as the source (figure 1-2), while the vocal tract acts as the filter. The glottis opens and closes pseudo-periodically at a rate between approximately 50 and 300 times per second. The period of time during which the glottis is open is referred to as the *open phase*, and the period of time in which it is closed is referred to as the *closed phase*. The *open quotient* is the ratio of the duration of the open phase to the

pitch period, and is generally between 30 and 70 percent. The closing of the glottis is particularly important, as this determines the amount of high frequency energy present in both the source and the speech, this period of time is called the *return phase* for reasons that will become evident later.

Under steady-state non-interactive conditions, the glottal flow would be proportional to the glottal area. The time-varying area of the glottis, and source-filter interaction modify the flow in several ways. The first change is the skewing of the glottal flow to the right with respect to the glottal area function. The air flowing through the glottis increases the pressure in the vocal tract, which causes loading of the glottal flow. This loading results in pulse skew to the right, as the loading slows down the acceleration of air through the glottis. Since closing the glottis eliminates loading, the glottal flow tends to end suddenly, as shown in figures 1-2a and 1-2b.

If we apply the radiation effect to the source rather than the output speech, the rapid closure caused by pulse skew results in a large negative impulse-like response at glottal closure, called the *glottal pulse*, as shown in figure 1-2. The glottal pulse is the primary excitation for speech, and has a wide bandwidth due to its impulse-like nature [7, 13]. From the glottal flow derivative, we can see the reasoning for the term return phase. After the peak of the glottal pulse, it takes some finite amount of time for the waveform to return to zero. Fant has shown that for one model of the return phase, the effect is to filter the source with a first order lowpass filter [15]. The more rapidly the glottis closes, the shorter the return phase. If a glottal chink or other DC glottal flow is present, the return phase will be lengthened.

We consider the glottal flow derivative as currently described to be the *coarse structure* of the source. The features of this source tend to have a smooth spectral content, and are of fixed positioning in relation to the glottal pulse. The extent of the features determines their timing in relation to the glottal pulse. For example, a glottis that closes slowly will result in a longer return phase, but it is not possible for the return phase to occur before the pulse. The *fine structure* of the source is now discussed. Elements of the source described as fine structure often have a more narrow spectral content, and their timing is not as clearly determined by the opening

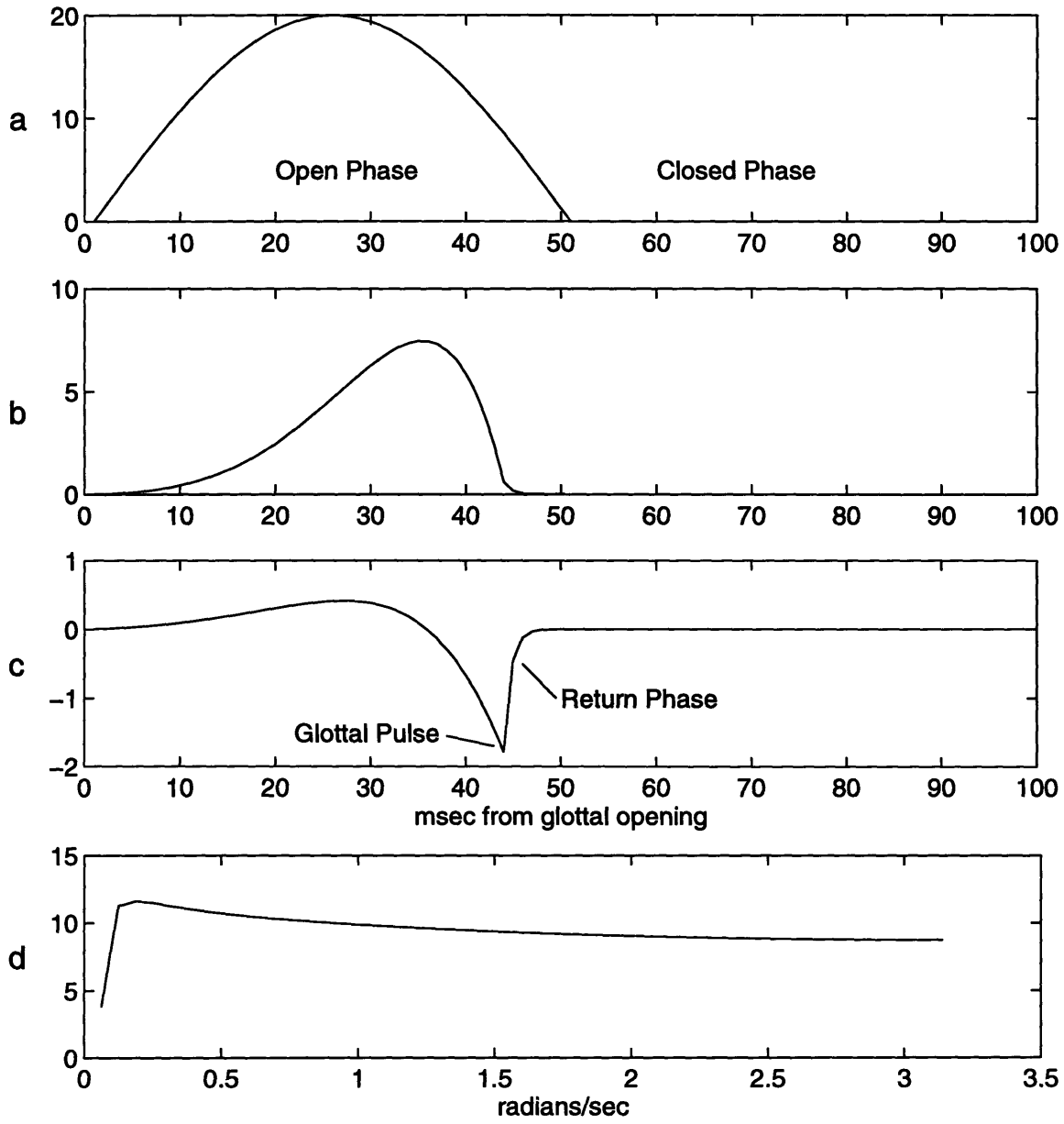

Figure 1-2: Schematic source waveforms. (a): Glottal Area, (b): Corresponding Glottal Flow, (c): Glottal Flow Derivative, and (d): Log-spectrum of (c).

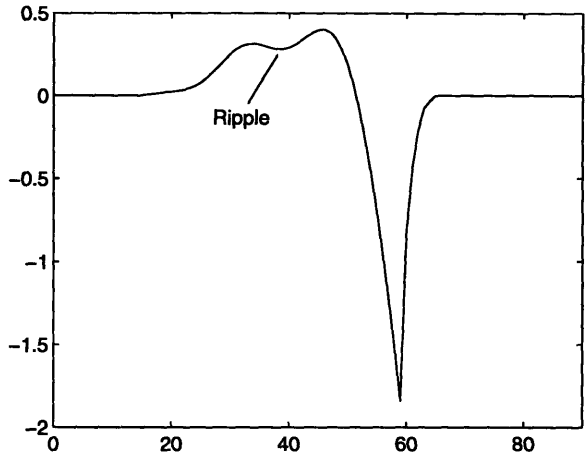

Figure 1-3: Ripple will be seen on the glottal flow derivative waveform due to source-filter interaction, as shown in this schematic representation.

of the glottis.

1.3.2 Fine Structure

Two primary sources of fine structure are considered in this thesis, ripple caused by source-filter interaction and aspiration noise. The pressure above the glottis, the *supra-glottal pressure*, is time varying, due to the formants of the vocal tract. The vocal tract will contain the decay of the previous glottal pulse when the glottis opens. This superposition of the decay of the vocal tract response from one pitch period into the next pitch period is the primary source for the energy that causes ripple. Figure 1-3 illustrates ripple on a glottal flow derivative waveform.

Since the airflow through the glottis is approximately proportional to the square root of the pressure drop across the glottis (ignoring the smaller viscous term in equation 1.16), the modulating supra-glottal pressure will interact with the flow in a nonlinear manner to create ripple in the glottal flow. The traditional assumption of a linearly separable source and filter must be discarded due to this interaction. The system will be linear only when the glottis is closed, eliminating source-filter interaction.

One clearly visible effect of the interaction that causes ripple is truncation of the response in the speech, as shown in figure 1-4. With the glottal closure as the primary

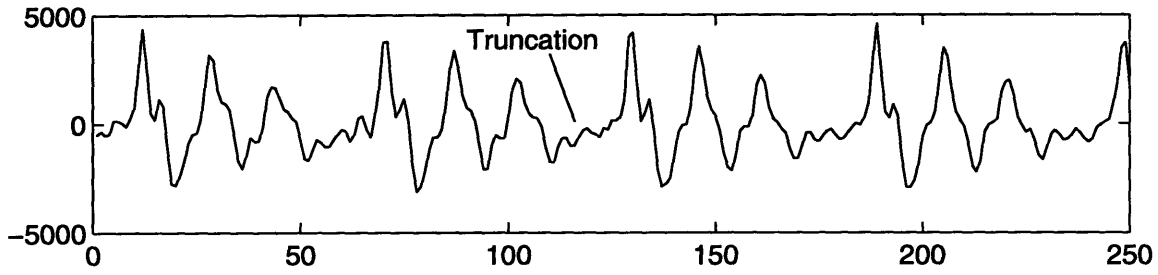

Figure 1-4: Vowel /a/ showing truncation of the fourth cycle of the first formant

excitation, the vocal tract decays for a period of time while the glottis is relatively closed, and any interaction is minimal. Once the glottis opens, the opening acts like a resistor to ground in an electrical circuit, allowing formant energy in the vocal tract to be absorbed by the lungs. This often causes a sudden drop in vocal tract energy, and a corresponding drop in speech output. If the analysis window used in linear prediction includes this truncated region, the vocal tract response estimated will be influenced by this sudden loss in energy. In particular, at a minimum we can expect the estimated formant bandwidths to be larger than they otherwise would be.

The timing and amount of ripple is dependent on the configuration of the glottal opening during both the open and closed phases. A narrow glottal opening will maximize ripple for a given area, since viscous coupling with the walls results in glottal flow that is more dependent on the trans-glottal pressure. If the glottis closes completely, there will be no ripple during the closed phase, otherwise ripple will always be present, though in varying magnitude. The manner of opening of the vocal folds will determine when ripple begins. If the folds open uniformly along their entire length, the viscous pressure drop will cause ripple before there is significant flow. Folds that open in a zipper-like fashion will start with a smaller amount of ripple, and will have significant flow almost immediately after the glottis begins to open.

Aspiration is similarly dependent on the glottal opening for its timing and magnitude. If the source of aspiration noise is airflow over the vocal folds, a long, narrow opening would tend to produce more aspiration noise than a triangle shaped opening, due to a larger surface area of the vocal folds. If the source of the aspiration noise is airflow hitting the epiglottis, the configuration of the glottis will not be as important,

but the aspiration will be filtered by a different system, since it occurs at a different point in the vocal tract.

1.4 Related Work

We will now discuss previous attempts at estimating the glottal flow from speech and previous uses of source information for speaker identification.

1.4.1 Previous Attempts at Estimating the Glottal Flow

By far the most common approach and least invasive approach to estimating the glottal flow is through inverse filtering. We would like to use linear prediction to estimate the vocal tract response, but the source and filter are not generally linearly separable, as we have seen. There can be no source-filter interaction when the glottis is closed, so estimation of the vocal tract must occur during the portion of each pitch period when the glottis is closed. Determining when the glottis is closed is the primary challenge in estimating the glottal flow using inverse filtering.

There are several common techniques to estimating the vocal tract response which will be discussed here. They include manually setting the vocal tract parameters rather than identifying the closed phase, using an electro-glottogram (EGG) to determine when the glottis is closed, and using automatic weighting procedures to discount or completely discard speech samples when some measure indicates the glottis might be open.

The system used at the Department of Speech Communication and Music Acoustics at the Royal Institute of Technology in Sweden requires the operator to specify the formant frequencies and bandwidths for each frame [22]. To do this, the operator looks at the resulting glottal flow waveform, adjusting the formants until the waveform has desired properties, such as minimal formant energy during the closed phase. In [22], Gobl indicates that an operator driven technique is required due to the sensitivity of the glottal flow derivative waveform to vocal tract parameter change. On the other hand, a system based on an operator choosing parameters to achieve

desired output characteristics will skew the output towards the operator's preconceived notions for the glottal flow. Also, techniques that require an operator are only practical in research settings.

Strik and Boves [50] use an automatic system that uses EGG data to determine the approximate timing of glottal closure. The do not attempt to find the glottal opening, and do not seem to consider their glottal closure estimate reliable. Rather than attempting to estimate the timing of these events, they take five fixed length windows (33, 34, 35, 36, and 37 samples) and five offsets relative to glottal closure (-2, -1, 0, 1, and 2), and perform linear prediction for each of the 25 combinations of window length and window offset. They average their results from each of these 25 analyses to achieve a final result. This routine is limited in that the analysis length is not adaptive, so it won't work well for high pitch speakers; and it is not rigorous, since it uses averaging to get decent results instead of accurately determining when the analysis should be performed.

Childers et al. have discussed two systems for estimating glottal flow [4, 6, 33]. One system uses EGG data to identify the glottal closed phase. Within this period, all possible analysis windows are used to estimate potential vocal tract filters, the vocal tract estimate with the minimum residue is considered the proper estimate. The other system weights previous speech samples based on the error in previous analysis windows. Regions with large error, such as the open phase, will be quickly de-weighted, while regions with small error, such as the open phase, will be used for a longer period of time. They indicate this works nearly as well as their EGG system.

Wong et al. developed a system that is similar to the one described in this thesis, but with a less rigorous theoretical background [54]. They argue that since the source is nonzero during the open phase, accurate vocal tract estimates can only be calculated during the closed phase. They do not appear to recognize the importance of source-filter interaction influencing the calculated vocal tract response during the open phase. To determine when the glottis opens and closes, they perform a sliding covariance analysis with a one sample shift. They use a function of the linear prediction error to identify the opening and closing of the glottis. Since the primary excitation occurs

at glottal closure, glottal closure is relatively easy to find. They mention that their technique has a much more difficult time identifying the glottal opening, likely due to the slowly increasing glottal flow. Cummings and Clements use a similar system in [8], but with the addition of some operator control over the particular analysis window chosen.

One of the more interesting approaches is that taken by Fujisaki and Ljungqvist in [19, 20]. Rather than estimating the vocal tract during the closed phase, they estimate the source and filter simultaneously. The glottal flow model they used did not take ripple into account, which will cause errors in estimation of the vocal tract. Also, since glottal flow models are not truly accurate representations of the flow, the vocal tract estimates will be further biased as the estimation routine attempts to make the glottal flow estimate match the model used.

1.4.2 Previous Uses of Source Information for SID

Source information has previously been used in only a few speaker identification systems [24, 41, 51]. One method is the use of the linear prediction residue. Any errors in modeling the vocal tract will show up in the residue, such as the error from attempting to model zeros in nasal sounds with an all-pole filter. The residue will also have some representation of the source, including phase and pitch information. The residue is thus useful for two primary reasons: errors in the linear prediction analysis and information inherently separate from the vocal tract.

Thevenaz and Hugli [51], on the other hand, argue for the use of the linear prediction residue for the simple reason that it is orthogonal to the predictor coefficients, and thus all the information contained in the residue is not contained in the coefficients. This mathematical argument has some non-intuitive results, such as that modeling the residue will decrease its usefulness, since the model parameters are not orthogonal to the predictor coefficients.

In a study designed to determine the importance of various features for speaker identification, Necioğlu, Clements, and Barnwell [41] use two source features, the spectral tilt and a "glottal pulse prototype approximation." The spectral tilt they

calculate is simply the tilt of a time-averaged normalized power spectrum for voiced frames. While this will contain the spectral tilt due to the source, it will also contain the spectral tilt due to the vocal tract, as the all-pole configuration of the vocal tract tends to have a spectral tilt regardless of the particular sound being spoken. The glottal pulse prototype approximation (GPP) is calculated by inverse filtering speech after the real poles and lowest frequency complex pole-pair are removed from a vocal tract estimate calculated without pre-emphasis. By performing linear prediction on non-pre-emphasized speech, they argue that the first complex pole-pair and any real poles model the spectral shape of the source, while the remaining pole-pairs model the vocal tract.

1.5 Thesis Contribution

This thesis contributes to speech science in three primary areas. First, it provides a reliable method for automatically estimating the glottal flow derivative waveform and estimating parameters for a model of the glottal flow derivative from a speech signal alone. Secondly, it illustrates the importance of the glottal flow for speaker identification. Finally, it helps to increase general understanding of the glottal flow, both through the particular variations observed in this study as well as easing future glottal flow studies through the techniques developed.

The techniques described in section 1.4.1 to estimate the glottal flow waveform from speech generally have a limitation that makes them unsuitable for practical applications. Some require an operator to set parameters on a frame by frame basis, others require the use of an electro-glottogram (a device that is attached to the side of the throat), while others don't work well enough or consistently enough for practical use. We use formant motion as predicted by theory [1] to accurately determine the glottal opening and the end of the return phase. This results in an automatically determined analysis window during which the speech signal is minimally influenced by the source and source-filter interaction. By estimating the vocal tract during this window and inverse filtering, we calculate an accurate estimate of the glottal flow

derivative.

Vocal tract information is further removed from the glottal flow through separating the fine and coarse structure of the source. Since ripple is due to interaction with formant energy, it will contain vocal tract information. By separating out the ripple, the coarse structure contains little vocal tract information, and the fine structure can be modeled in such a way that only glottal information is kept (by measuring the timing and magnitude of the fine structure, rather than its specific content). The speaker ID results which are presented show that the parameters calculated from the glottal flow derivative waveform are strongly speaker dependent.

A reliable, automatic approach to estimating and modeling the glottal flow derivative will help advance speech science by enabling the creation of large databases of glottal flow derivative data to be analyzed by both automatic and manual means. Simply observing some of the examples used in this thesis shows the wide variety of source waveforms that occur in practice.

1.6 Thesis Organization

In this chapter, we have provided a background for the discussion of speech production and analysis, a brief motivation for the study of this problem, a discussion of related works, and mentioned some of the more important contributions made by this thesis. In Chapter 2 we will discuss speech production in more detail, and develop a model to capture the important features of the glottal flow derivative. Chapter 3 discusses the techniques used to calculate an estimate of the glottal flow derivative waveform. Chapter 4 covers estimating the features of a model to capture the coarse structure of the glottal flow, while chapter 5 develops estimation of the fine structure of the glottal flow. Chapter 6 discusses the use of the model parameters for speaker identification. Finally, chapter 7 gives our conclusions and ideas for future direction in related research.

Chapter 2

The Glottal Flow Model

In section 1.3 we described the basic operation of the source for voiced speech. We now study the source in a more detailed, theoretical framework. A model is proposed for the purpose of extracting information from the glottal flow derivative waveform to use in speaker ID.

2.1 Physical Model

2.1.1 Detailed Physiological Simulation

The framework for this thesis comes primarily from the work by Ananthapadmanabha and Fant [1]. In order to better understand the glottal flow, a physiological model of speech production was developed, assuming a given glottal area function. The key feature of the physiological model is a nonlinear, time varying glottal impedance. Arguments are given for discarding the viscosity term from the glottal impedance equation 1.16, giving the steady-state equation

$$\Delta P = (1.1\rho/2A^2)U^2, \tag{2.1}$$

where ΔP is the pressure drop across the glottis, ρ is the density of air, A is the glottal area, U is the volume velocity, and 1.1 is an empirical constant combining the so-called entry drop and exit recovery coefficients, which are related to how much

energy is required to force the air into and out of the glottal opening.

An equivalent circuit with distributed parameters to simulate both the resonances of the lungs and vocal tract formants is developed for use in conjunction with the glottal impedance equation 2.1. The only time varying parameter in this system is the glottal area. This system yields a set of equations which must be solved simultaneously with a numerical iterative algorithm, yielding the true glottal flow, assuming a constant vocal tract.

Calculating the flow using these equations yields several interesting results. First, the glottal flow is primarily influenced by the first sub- and supra-glottal formants, and the higher supra-glottal formants to a lesser extent. The influence of higher sub-glottal formants is negligible. Secondly, pulse skew is seen as expected, with the addition that lower first formant frequencies cause more pulse skew. Finally, ripple is seen as a result of interaction with the varying supra-glottal pressure. This ripple tends to build up over multiple pitch periods, due to superposition. The exact nature of the ripple is dependent on the ratio of the pitch period to the period of the first formant, as this ratio will determine the phase of the first formant during the next glottal open phase.

2.1.2 Simplified Production Model

In light of the finding that only the first sub- and supra-glottal formants significantly influenced the glottal flow, a pseudo-equivalent circuit was developed with a single supra-glottal formant (figure 2-1). We will now analyze this circuit to develop a relation between the parameters and the glottal flow. The pressure drop across the glottis, described by equation 2.1, is rewritten here with more explicit labeling of terms.

$$P_{tg}(t) = \left(\frac{k\rho}{2A_g^2(t)}\right) U_g^2(t),$$

where $P_{tg}(t)$ is the trans-glottal pressure, k is 1.1, $A_g(t)$ is the glottal area, and $U_g(t)$ is the glottal flow. Using this equation with figure 2-1, the following expression is

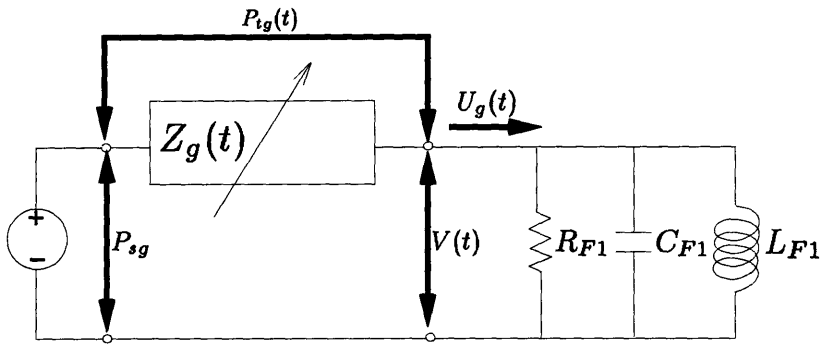

Figure 2-1: Equivalent circuit for a single formant load, from [29]

obtained:

$$C\frac{dV}{dt} + \frac{V}{R} + \frac{1}{L}\int V dt = U_g(t) = A_g(t)\sqrt{\frac{2P_{tg}}{k\rho}}. \tag{2.2}$$

Note that $P_{tg} = P_{sg} - V$, so that

$$\begin{aligned} U_g(t) &= A_g(t)\sqrt{\frac{2P_{tg}}{k\rho}} \\ &= A_g(t)\sqrt{\frac{2(P_{sg}-V)}{k\rho}} \\ &= A_g(t)\sqrt{\frac{2P_{sg}}{k\rho}}\sqrt{\left(1-\frac{V}{P_{sg}}\right)}. \end{aligned} \tag{2.3}$$

Assuming the pressure drop across the glottis is nearly as large as the sub-glottal pressure, i.e. $V \ll P_{sg}$, we can use the Taylor series approximation

$$\sqrt{(1-\frac{V}{P_{sg}})} \approx 1 - \frac{1}{2}\left(\frac{V}{P_{sg}}\right) \tag{2.4}$$

and equation 2.3 in equation 2.2, we obtain the following

$$C\frac{dV}{dt} + \frac{V}{R} + \frac{1}{L}\int V dt = U_g(t) = A_g(t)\sqrt{\frac{2P_{sg}}{k\rho}}\left(1-\frac{V}{2P_{sg}}\right)$$

$$C\frac{dV}{dt} + \frac{V}{R} + \frac{1}{L}\int V dt + V\left[A(t)\frac{1}{2P_{sg}}\sqrt{\frac{2P_{sg}}{k\rho}}\right] = A(t)\sqrt{\frac{2P_{sg}}{k\rho}}, \tag{2.5}$$

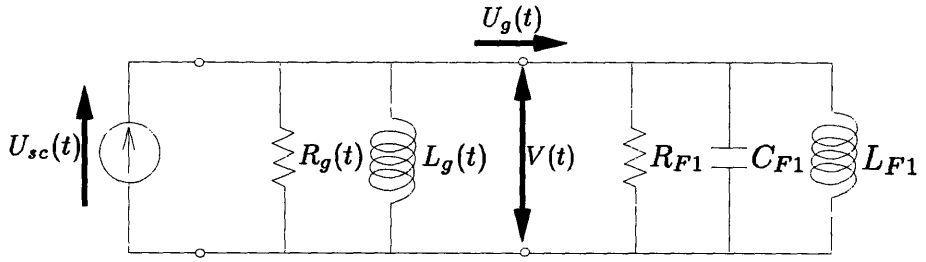

Figure 2-2: Norton equivalent circuit for single formant load showing time-varying elements, from [29]

and by substituting

$$\begin{aligned} U_{sc}(t) &= A(t)\sqrt{\frac{2P_{sg}(t)}{k\rho}}, \text{ and} \\ g_o(t) &= \frac{U_{sc}(t)}{P_{sg}} = A(t)\sqrt{\frac{2}{k\rho P_{sg}}}, \end{aligned}$$

we obtain

$$C\frac{dV}{dt} + \frac{V}{R} + \frac{1}{L}\int V dt + \frac{1}{2}Vg_o(t) = U_{sc}(t). \tag{2.6}$$

Differentiating equation 2.6 yields

$$C\frac{d^2V}{dt^2} + \frac{1}{R}\frac{dV}{dt} + \frac{V}{L} + \frac{1}{2}(V\dot{g}_o + \dot{V}g_o) = \dot{U}_{sc},$$

simplifying, we obtain

$$C\frac{d^2V}{dt^2} + \left[\frac{1}{R} + \frac{1}{2}g_o(t)\right]\frac{dV}{dt} + \left[\frac{1}{L} + \frac{1}{2}\dot{g}_o(t)\right]V = \dot{U}_{sc}(t). \tag{2.7}$$

Equation 2.7 shows a correspondence between R and $R_g(t) = 2/g_o(t)$ as resistances and L and $L_g(t) = 2/\dot{g}_o(t)$ as inductances. This leads to the Norton equivalent circuit in figure 2-2. We can therefore see that the effect of glottal interaction is to modulate the formant. To gain a better understanding of this formant modulation, Anantha-padmanabha and Fant suggest making the assumption that the glottal impedance is stationary so that a "pseudo-Laplace transform" of the transfer function from the volume-velocity to the output speech pressure can be calculated,

$$H(s,t) = \frac{V(s,t)}{U_{sc}(s)} = \frac{s/C}{s^2 + B_1(t)s + \Omega_1^2(t)}, \tag{2.8}$$

where the time-varying formant frequency and bandwidth

$$\Omega_1(t) = \Omega_0\sqrt{1 + \frac{1}{2}L\dot{g}_o(t)} \text{ and} \tag{2.9}$$

$$B_1(t) = B_0[1 + \frac{1}{2}Rg_o(t)] \tag{2.10}$$

are given in terms of the non-interactive frequency and bandwidth,

$$\Omega_0 = \sqrt{\frac{1}{LC}} \text{ and}$$

$$B_0 = \frac{1}{RC}.$$

Figure 2-3 shows the effects of this formant modulation. The following assumptions were made: $\rho = 1.275 * 10^{-3} g/cm^2$, $P_{sg} = 8cm\ H_2O$, $\Omega_0 = 524\ Hz$, and $B_0 = 35\ Hz$. The glottal area is given by a two part function,

$$A_g(t) = \begin{cases} A_{max}[0.5 - 0.5\cos(\frac{\pi t}{T_0})] & 0 < t < T_0 \\ A_{max}\cos[\frac{\pi(t-T_0)}{2*T_c}] & T_0 < t < T_0 + T_c \end{cases}$$

where A_{max} is the maximum glottal opening, $20mm^2$, T_0 is the time from glottal opening to maximum glottal area, $3ms$, and T_c is the time from T_0 to glottal closure, also $3ms$. This formant modulation occurs during the open phase, which is late in the decay of the previous glottal pulse. As the bandwidth increases, the response will decay much more rapidly, causing a truncation effect about halfway through each pitch period. Along with this decay comes an increase in formant frequency. While the formant frequency is shown mathematically to drop towards the very end of the open phase, the increase in bandwidth has normally almost completely damped out the signal by this point.

If we wish to use a linear time-invariant system to model the approximation to the time varying glottal impedance given in equation 2.8, we must adjust either the

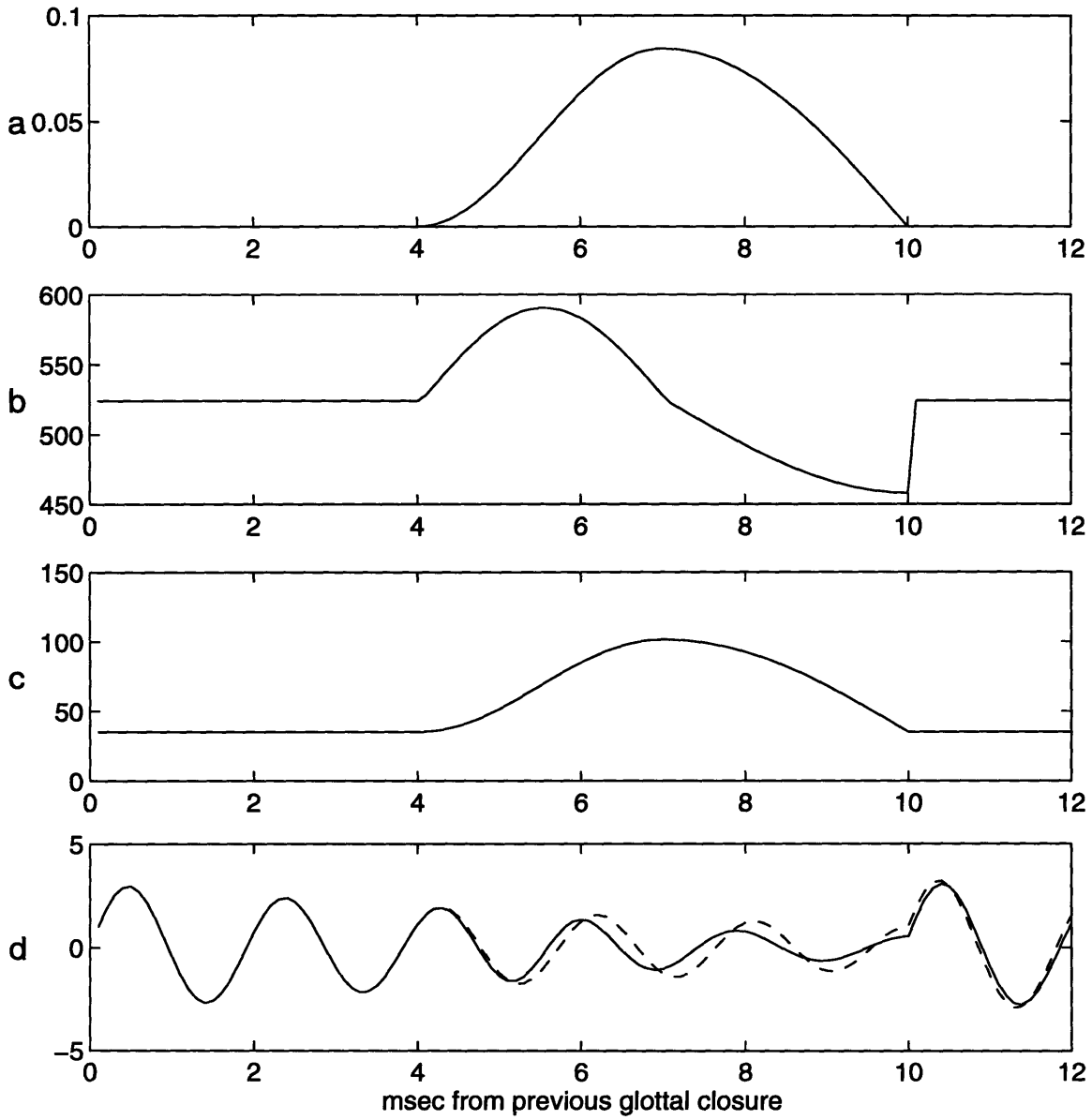

Figure 2-3: Source-filter interaction causes modulation of formant frequencies and bandwidths when the glottis is open. Although these effects are nonlinear, they can be approximated in a linear framework. This figure shows the function $g_o(t)$, which is proportional to the glottal area in (a), the formant frequency in Hz (b), the formant bandwidth in Hz (c), and speech waveforms generated with a time-invariant formant and with these formant modulations (d). Dashed : no formant modulation. Solid : formant modulation as shown in panels (b) and (c).

source or the filter to reintroduce this source-filter interaction. By holding the vocal tract filter fixed and mapping all source-filter interaction to the source, the system again becomes linear and time-invariant. Rewriting equation 2.8 in terms of a fixed vocal tract $H(s)$ and a time varying source, $U_g(s,t)$, the output becomes

$$V(s,t) = H(s)U_g(s,t), \tag{2.11}$$

where the time varying source $U_g(s,t)$ includes the non-interactive flow $U_{sc}(s)$ and the modulation in equation 2.8:

$$U_g(s,t) = U_{sc}(s)\frac{s^2 + B_0 s + \Omega_0^2}{s^2 + B_1(t)s + \Omega_1^2(t)}, \tag{2.12}$$

while the vocal tract filter $H(s)$ is just

$$H(s) = \frac{1}{s^2 + B_0 s + \Omega_0^2}. \tag{2.13}$$

Taking the inverse "pseudo-Laplace transform" of equation 2.12, we have

$$u_g(t) = u_{sc}(t) + f(t)e^{-0.5tB_1(t)}\cos(\Omega_1(t)t), \tag{2.14}$$

where $u_{sc}(t)$ contains the coarse structure of the flow, and the scale factor $f(t)$ is determined by the partial fraction expansion of equation 2.12. Equations 2.8 and 2.14 show the duality of the time domain ripple and the formant modulation in the frequency domain. It should be noted that while the glottal flow waveform is primarily effected by the first formant, all of the formants are approximately equally effected by this formant modulation [17]. The influence on the glottal flow due to higher formants is less because they tend to be of lower amplitude, and with their higher bandwidths, they have decayed even more by the time the glottis opens.

2.2 Feature Model

The feature model provides a parameterized version of the source, with parameters designed for their significance for speaker ID. In order to simplify the problem of representing the glottal flow derivative, we will break it up into two main parts, the coarse and fine structure of the flow. The coarse structure includes the large-scale portions of the flow, primarily the general shape. The fine structure includes the ripple and aspiration.

2.2.1 Coarse Structure

The coarse structure is dominated by the motion and size of the glottis and pulse skew due to loading of the source by the vocal tract. The features we want to capture through the coarse structure include the open quotient, the speed of opening and closing, and the relationship between the glottal pulse and the peak glottal flow. The open quotient is known to vary from speaker to speaker, and has been shown empirically to adjust the relative amplitudes of the first few harmonics [32]. Breathy voices tend to have larger open quotients, while pressed voices have smaller open quotients.

The relationship between the peak glottal flow and the amplitude of the glottal pulse indicate the efficiency of a speaker. As mentioned previously, the glottal pulse is the primary excitation for voiced speech. Thus it is the slope of the glottal flow at closure, rather than the peak glottal flow, that primarily determines the loudness of a speaker. Ripple can also play a role in efficiency, if the ripple is timed such that the supra-glottal pressure is at a maximum at the same time as the glottal flow. In this case, the ripple will tend to lessen the glottal flow, but not impact the rate of closure [14].

To model the coarse structure, we will use the Liljencrants-Fant (LF) model for the glottal flow derivative [14]. The LF model is described by the following equations:

$$E(t) = \frac{dU_g}{dt} = E_0 e^{\alpha t} \sin \omega_g t \tag{2.15}$$

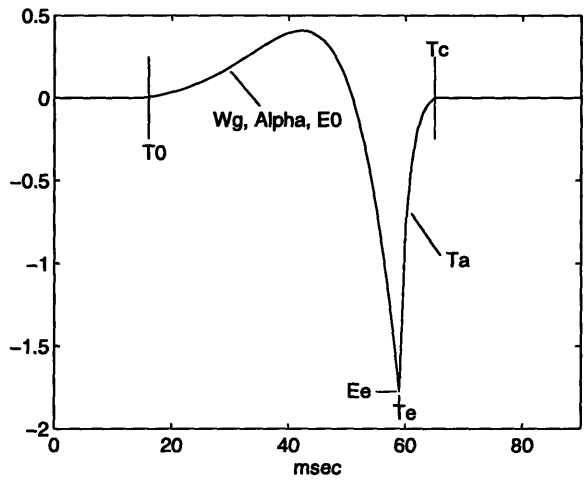

Figure 2-4: LF Model for the glottal flow derivative waveform

for the period from glottal opening (T_0) to the pitch pulse (T_e, time of excitation), at which time the return phase starts:

$$E(t) = \frac{-E_0}{\epsilon T_a}\left[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)}\right], \tag{2.16}$$

which continues until time T_c. See figure 2-4.

The model is considered a four parameter model. Three of the parameters describe the open phase; they are E_0, ω_g, and α, with one parameter describing the return phase, T_a. In order to ensure continuity between the open and return phases at the point T_e, ϵ is dependent on T_a. While the relationship between ϵ and T_a cannot be expressed in closed form, $\epsilon \approx 1/T_a$ for small values of T_e. Since these four waveshape parameters do not include the time of glottal opening, excitation, nor closure, the values of T_0, T_e, and T_c must be given. Generally, it is assumed that T_0 coincides with T_c from the previous pitch period, requiring only that the timing of T_e in relation to T_0 be known. This assumption results in no period for which the glottis is completely closed; however, a small T_a will result in flow derivative values essentially equal to zero, due to the exponential decay during the return phase.

The parameter T_a is probably the most important parameter in terms of human perception, as it controls the amount of spectral tilt present in the source. The return phase of the LF model is equivalent to a first order low-pass filter [15] with a corner

frequency of

$$F_a = 1/(2\pi T_a). \tag{2.17}$$

This equation illustrates the manner in which the parameter T_a controls the spectral tilt of the source, and thus the speech output. The parameter α determines how rounded the open phase is, while the parameter ω_g determines how rounded the left side of the pulse is. These parameters primarily influence the relationships between the first few harmonics of the source spectrum.

Ease of use dictated several changes to the above described LF model. First, the times T_0, T_e, and T_c are not given, and T_0 does not have to occur at the same time as the previous period's T_c. In order to express the model in a closed form, the assumption was made that $\epsilon = 1/T_a$. This requires that the value E_0 be different for the open and return phases, a simple closed form expression exists for this calculation. Also, the time variable is normalized during the open phase by the time difference between T_0 and T_e, which at time T_e gives the equation

$$E(t) = E_0 e^{\alpha} \sin \omega_g,$$

We thus have a seven parameter model to describe the glottal flow, with the four standard LF model parameters, and three indicating the timing of the pulse.

2.2.2 Fine Structure

In terms of the fine structure, we are interested in the timing of fine structure and how much is present. The two sources of fine structure used in this study are aspiration and ripple.

Aspiration and ripple due to source-filter interaction are departures from the ideal linear-system view of speech production, and can be roughly measured accordingly. If ripple and aspiration noise occur during the closed phase due to a constant glottal opening, the source will be nonzero during this region. Two ways in which aspiration noise can be modeled are through its spectral content and the energy of the noise as a

function of time. Since the source estimate was derived through inverse filtering, using a filter generated by linear prediction during the closed phase, any energy during the closed phase should be approximately white, which precludes attempting to model the spectral content of the noise. Due to the very short time period, typically on the order of three to five milliseconds, measuring the evolution of the energy of aspiration noise would be difficult, so we choose to simply calculate the energy during the closed phase.

There may be ripple present during the closed phase as well. Since ripple during the closed phase will result in small variations of the formant frequencies and bandwidths, we expect energy due to ripple to be in frequency bands around the formants. Ripple can thus be modeled similarly to aspiration noise, by its spectral content and the evolution of its energy. We want to estimate the ripple independent from the formants of the vocal tract, so we do not want to model the spectral content of the ripple. As for aspiration, estimating the time evolution of ripple would be difficult. We thus model the aspiration and ripple during the closed phase together by the amount of energy in the glottal flow derivative estimate during the closed phase.

The energy during the open phase that is not captured by the coarse model can be assumed to be due to ripple or aspiration. These two energy measures are particularly useful when compared. Complete glottal closure during the closed phase will result in significantly less energy present in the source during the closed phase than the ripple and aspiration energy present during the open phase. If the energy of the source during the closed phase is more comparable to the energy during the open phase, there is most likely a significant constant glottal opening.

Many speakers show no evidence of ripple during the closed phase, but the glottal flow derivative waveforms for these speakers do have noise during the closed phase, which we attribute to aspiration noise, possibly due to a very small opening which results in aspiration noise without significant ripple. For speakers with such a constant glottal opening, it is possible to estimate the aspiration noise separately from the ripple. Figure 2-5 shows an example of a source with these characteristics. The analysis window is chosen during the period when ripple is not present, as will be

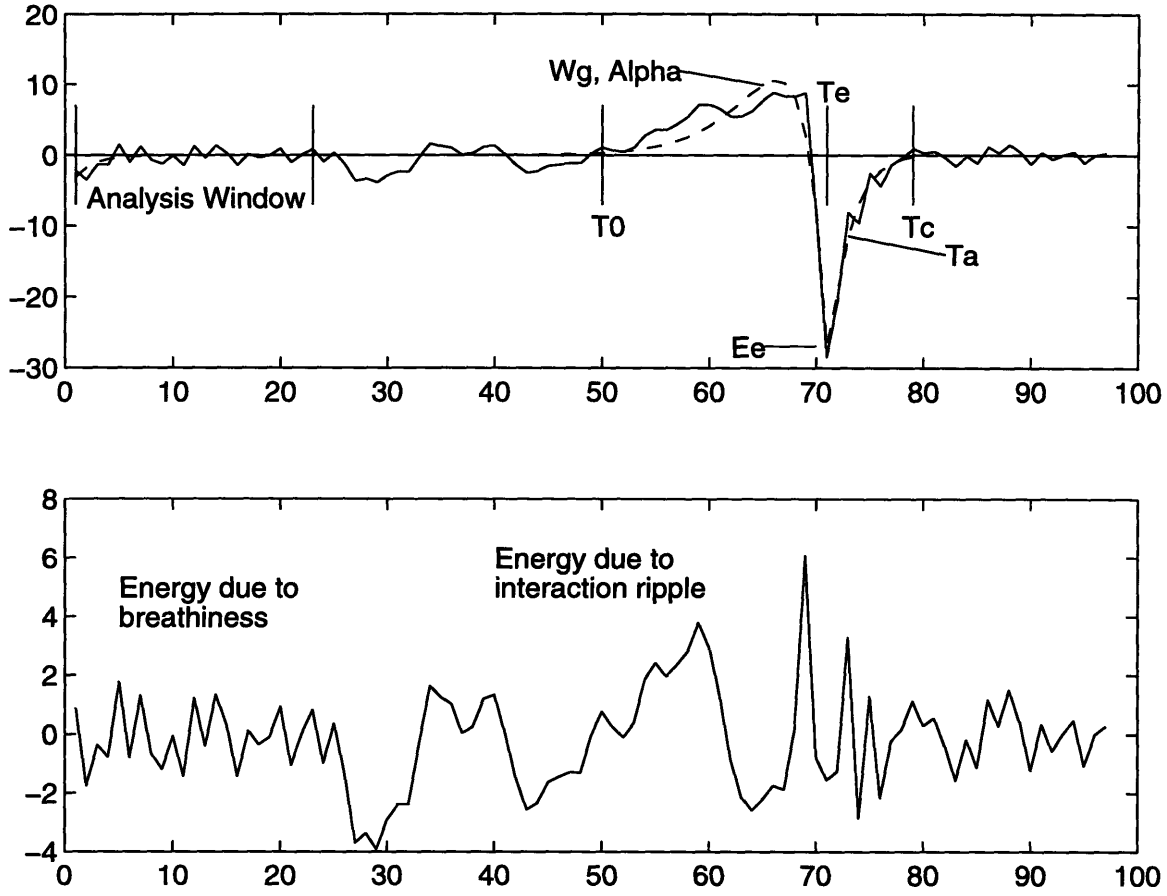

Figure 2-5: Often the glottal flow derivative exhibits a period containing a small amount of noise immediately after the return phase, followed by a period that shows ripple but no significant flow, followed by the standard glottal pulse and return phase. (a): Estimated Glottal Flow Derivative (solid) and overlaid LF model (dashed). (b): The error for the fitted LF model (see chapter 4), containing aspiration noise and ripple due to source-filter interaction

discussed further in section 3.1. After the analysis window, there is a period which seems to exhibit ripple, but as yet no significant glottal flow. After this period, we see the onset of glottal flow, with a sharp glottal pulse and a gradual return phase.

In particular, the normalized energy is calculated during the following periods:

1. The closed period bounded by the times T_c from the previous pitch period and T_0 from this pitch period,
2. The open period bounded by the times T_0 and T_e,
3. The closed phase as defined by the identified analysis window,
4. The open phase bounded by the end of the analysis window and T_e, and

5. The return phase starting at time T_e and ending at time T_c.

In addition to these energy measures, a direct measure of ripple is used. As discussed at the end of section 2.1.2, ripple can be equivalently represented in the time domain or as formant modulation in the frequency domain. The theory described above predicts that bandwidth modulation will be proportional to the glottal area, while frequency modulation will be proportional to the derivative of the glottal area. While not agreeing with theory, observation of calculated formant motion led us to model the modulation of the first formant frequency through the use of a parabola,

$$F(t) = (\bar{F}_1 + \Delta F) + Bt + Ct^2. \tag{2.18}$$

The average formant value, $\bar{F}_1$, represents vocal tract information, while the change in the formant value is due to source characteristics. In order to model the source separately from the vocal tract, we separate the average formant value ($\bar{F}_1$) from constant term. The three remaining parameters, ΔF, B, and C, model source-filter interaction and not filter characteristics.

2.3 Summary

This chapter laid out the theory describing the expected features of the glottal flow, as well as formant modulation, which will be used to determine the timing of glottal opening. The glottal flow will be skewed to the right due to loading of the source by the vocal tract, and will exhibit ripple due to nonlinear source-filter interaction. Aspiration noise is expected to be present during all times that the glottis is not completely closed.

Section 2.2 described the model features of the source over a pitch period. The general pulse shape is captured through a modified LF model, while aspiration noise and ripple will be measured through several energy measures as well as a parabola fit to the modulation of the first formant frequency.

In the next chapter we develop the procedures used to calculate the glottal flow

derivative waveform from speech.

Chapter 3

Calculation of the Glottal Flow Derivative Waveform Estimate

The theory for the production of voiced speech suggests that an accurate vocal tract estimate can be calculated during the glottal closed phase, when there is no source/vocal tract interaction. This estimate can then be used to inverse filter the speech signal during both the closed and open phases. Any source/vocal tract interaction is thus lumped into the glottal flow (or its derivative), the source for voiced speech, since the vocal tract is considered fixed.

3.1 Determination of the Closed Phase

The first and most difficult task in an analysis based on inverse filtering from a vocal tract estimate calculated during the closed phase is identification of the closed phase. A rough approximation of the beginning of the closed phase can be determined through inverse filtering the speech waveform. Since linear prediction matches the spectrum of the signal analyzed, inverse filtering a signal $S(z)$ with a filter $\hat{S}(z)$ determined by linear prediction will result in an approximately white signal:

$$\left| S(z) \frac{1}{\hat{S}(z)} \right| \approx 1.$$

For periodic speech signals, inverse filtering will result in impulses that occur at the point of primary excitation, the glottal pulse. The exact timing of these pitch pulses can be identified by finding the largest sample approximately every T_0 samples, where T_0 is the pitch period. This procedure is known as *peak picking*. The return phase shows that complete glottal closure does not occur until a short time after the glottal pulse, so additional processing is needed to find the onset of the closed phase.

Determination of glottal opening is much more difficult, since the glottal flow develops slowly, and glottal opening does not cause a significant excitation of the vocal tract. As discussed in section 2.1.2, formant modulation will occur when the glottis is open. By tracking the formants during a pitch period, the time at which the formants begin to move can be identified. This will be when the glottis begins to open.

To identify the closed phase, a two step procedure is therefore used:

1. Identify glottal pulses through peak picking of an initial whitening of the speech. This provides a frame for each pitch period in which to identify the closed phase.
2. Determine the closed phase as the period during which formant modulation does not occur. This formant modulation occurs due to source-filter interaction whenever the glottal opening is changing.

3.1.1 Initial Glottal Closure Estimate

In order to ease the analysis, pitch estimates and voicing probabilities are required as input to the system, along with the speech. The pitch estimates and voicing probabilities are generated with a sinusoidal-based pitch estimator [38], with one estimate every 10ms and an analysis window of length 30ms. Most any pitch estimator could be used in place of the sinusoidal pitch estimator. This pitch information is used to perform a pitch synchronous linear prediction. The covariance method of linear prediction is used, because it will generate a more accurate spectral match. No pre-emphasis is used, as pre-emphasis would result in a less perfect spectral match. The goal of this initial linear prediction is not an accurate model of the vocal tract, rather, the goal is an inverse filtered waveform amenable to peak picking. One measure of

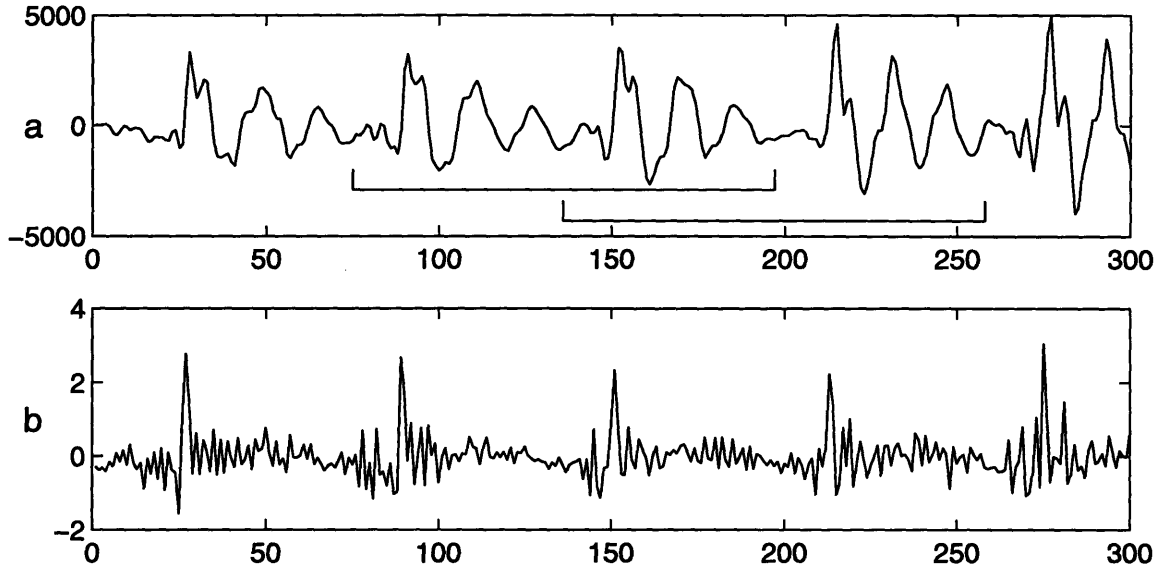

Figure 3-1: Inverse filtering speech using a pitch synchronous window and no pre-emphasis results in a waveform with sharp pulses occurring at the time of glottal closure. These pulses can be easily identified using a peak picking algorithm. (a): Speech waveform with a sequence of possible analysis windows. (b): Resulting inverse filtered waveform.

the ease of identifying peaks is the peak-to-RMS energy ratio, an indication of the height of the peak compared to the rest of the signal.

The size of the rectangular analysis window is two pitch periods, and the window shift is one pitch period. The location of the glottal pulse within this window is not controlled. Figure 3-1 shows two sequential analysis windows. This initial analysis is used to inverse filter the waveform. The resulting source estimate tends to be very impulse-like, easing identification of the glottal pulse.

The peaks of the inverse filtered waveform are identified as follows: The voicing probabilities taken as input to the system are used to identify voiced regions in the speech. Each voiced region will consist of one or more voiced phonemes, such as the entire word "man." In order to identify all the glottal pulses, we will first identify one pulse which we expect to identify with a good deal of accuracy. The remaining glottal pulses will be identified in small regions around where the pitch estimates predict they should occur.

For each voiced region, the largest peak is found; this is considered to be a glot-

tal pulse. The pitch information provided as input to the system is used to give an estimate of the location of the following glottal pulse. A small window around this estimated location is searched for the largest peak, whose location is considered to be the timing of the next glottal pulse. This is continued until the end of the voiced region, and then repeated for the voiced region before the initially identified glottal pulse location. The procedure is repeated for all remaining voiced region. The procedure is illustrated in figure 3-2.

Overshoots and ringing after the impulse are sometimes a problem in the inverse filtered waveform, which can cause incorrect identification of glottal closure. To help correct errors caused by overshoots and ringing, other peaks of similar amplitude are looked for before each identified glottal closure. The region searched is less than half the pitch period and less than half the period of the estimated first formant. This small search region ensures that the previous glottal pulse will not be accidentally found, and eliminates problems that could occur if the first formant were inaccurately estimated. These glottal closure estimates are only used to identify a beginning and end for each pitch period, so no further refining is performed.

3.1.2 Sliding Covariance Analysis

The glottal closure estimates provide a frame for each pitch period, since each closed phase must be entirely contained between two consecutive glottal closures. This frame enables identification of the closed phase based on changes which happen each period. The formant frequencies and bandwidths are expected to remain constant during the closed phase but will shift during the open phase. For voices in which the glottis never completely closes, such as breathy voices, a similar formant modulation will occur. During the nominally closed phase, the glottal opening should remain approximately constant, resulting in an effect on the formants of stable magnitude. Due to the nonlinear nature of the source-filter interaction, the formants will vary even with a constant glottal area as present during the closed phase of a breathy speaker. When the glottis begins to open, the formants will move from the relatively stable values they had during closed phase.

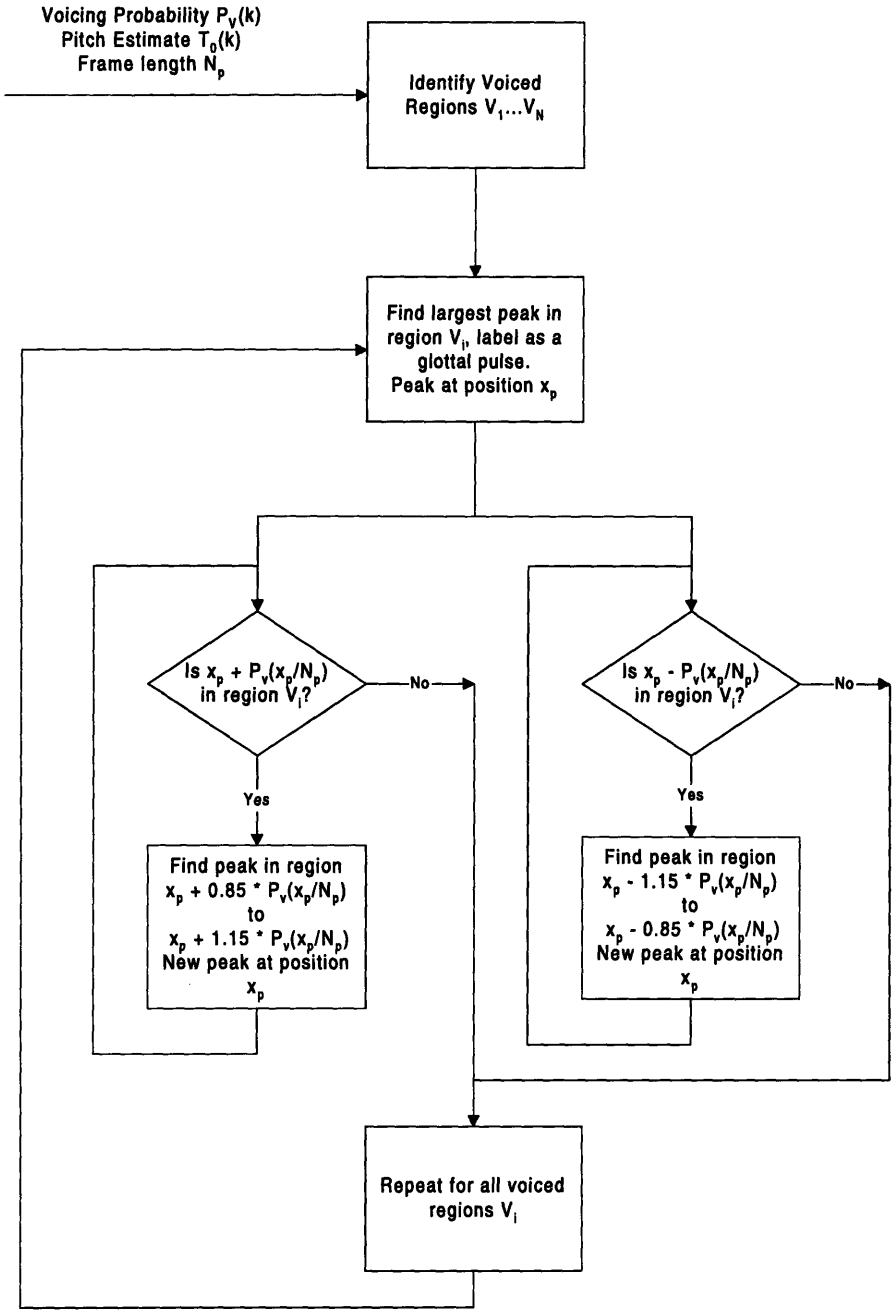

Figure 3-2: Flowchart of procedure to estimate glottal pulses.

To measure the formant frequencies and bandwidths during each pitch period, a sliding covariance based linear prediction analysis with a one sample shift is used. The size of the rectangular analysis window is constrained to be slightly larger than the prediction order, while still being several times smaller than the pitch period. In particular, the length of the analysis window is chosen for each frame to be

$$N_w = N_p/4,$$

with upper and lower bounds of

$$p + 3 \leq N_w \leq 2p,$$

where N_w is the size of the sliding covariance analysis window, N_p is the length of the pitch period as calculated by the time between the glottal pulses identified above, and p is the order of the linear prediction analysis, 14 for this study. Window lengths less than $p + 3$ cause occasional failure of the Cholesky decomposition, while using more than $2p$ points will not make the estimate significantly more accurate but will decrease the time resolution. The first analysis window begins immediately after the previous glottal pulse, while the last analysis window ends the sample before the next glottal pulse. There are thus a total of $N - N_w$ windows for each pitch period. This sliding covariance analysis gives one vocal tract estimate per sample in the pitch period. Formant tracking is performed in each pitch period on the formants calculated from the vocal tract estimates[1]. This provides estimates of each formant during both the

[1]The first four formants are tracked by their frequency using a viterbi search. The search space is the four lowest poles with bandwidth less than 500 Hz calculated by the sliding covariance analysis. The cost function is the variance of the formant track including the proposed pole to be added to the end of the track. Since a viterbi search can result in a single pole being assigned to multiple formant tracks, after the viterbi search is completed the path with the minimum variance is considered to be an actual formant track, and the poles used in that track are removed from the search space. Three new formants are found from the reduced search space, and again the formant track with the lowest variance is considered an actual formant track. This is repeated one more time to give a third formant track, with the remaining poles assigned to the last formant track. While the track of the first formant will generally have the least variance, this need not be the case, and the formant track identified in the first pass will not necessarily be the formant track with the lowest frequency. Due to this manner of searching, often the higher formant tracks will include poles that clearly do not belong, but are included in the track because a pole must be included for each sample and no better

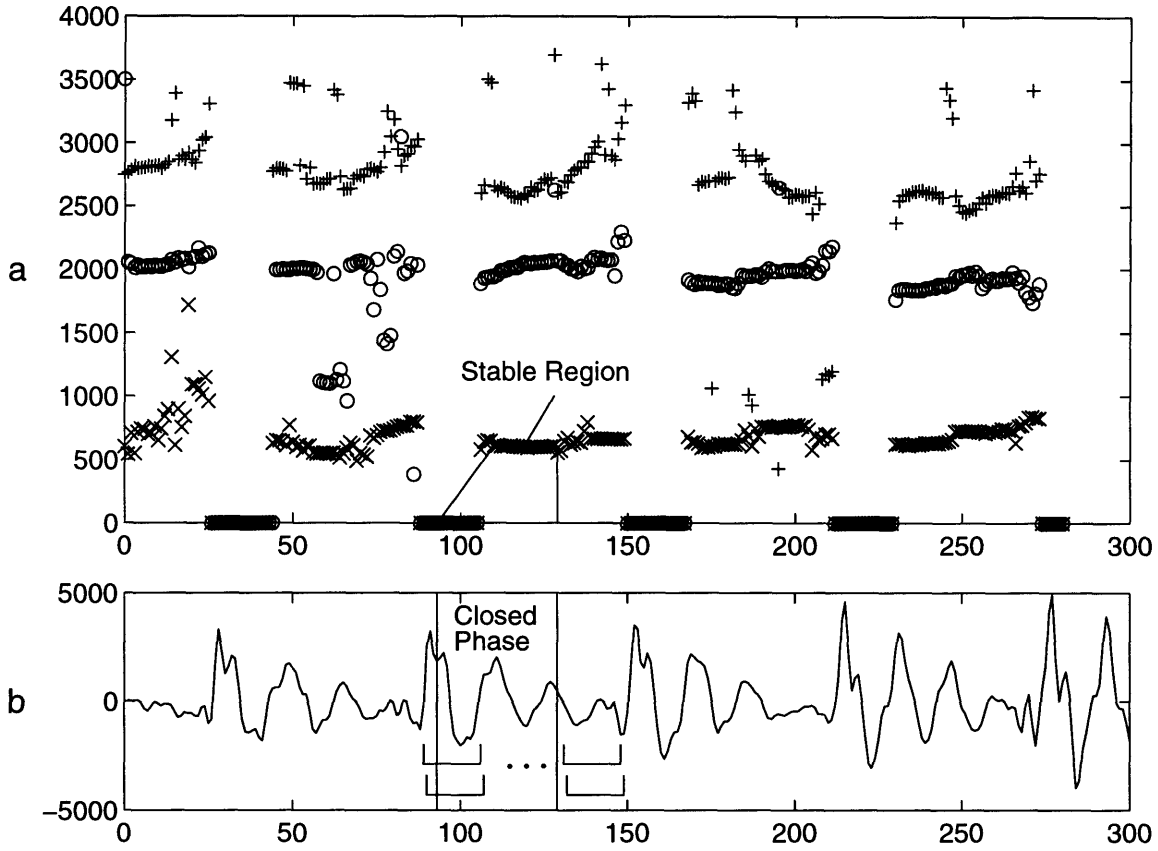

Figure 3-3: The closed phase is identified by the region in which the first formant frequency is stable. Panel (a) shows the formant frequency tracks over several pitch periods of the speech in (b). Panel (b) also shows the first and last two analysis windows for one frame. Each formant value is calculated from a small window of speech samples. The closed phase is defined as every speech sample in the windows used to calculate the formant values in the stable region. x : F1, o : F2, + : F3. Formant values are shown at the end of the corresponding analysis window. Formant values of 0 are displayed for regions in which no value was calculated.

closed and open phases, enabling identification of the time of glottal opening based on formant modulation. See figure 3-3 for the resulting formant track for the speech shown in figure 3-1.

While a mathematical framework for calculating the expected modulation of the formant frequencies and bandwidths was developed in section 2.1.2, we have found a large variety in the frequency and bandwidth changes that occur in the open phase. Also, due to different fixed glottal openings from speaker to speaker, the amount of

pole was available at that stage in the formant tracking. This can be seen in figure 3-3.

formant modulation that occurs during the closed phase will vary from speaker to speaker. This varying amount of formant modulation during the closed phase makes it difficult to set a threshold for an amount of formant modulation that indicates glottal opening. Because of these two problems, we have chosen to take a statistical approach to identifying the glottal opening. The approach taken is also a more practical approach, in that we want to estimate the vocal tract when the formant values are constant. The basic idea is to find a region during which the formant values vary minimally, while outside this region the formant values change considerably.

A small region of sequential formant samples is determined in which the formant modulation is minimal as defined by the sum of the absolute difference between successive formant estimates:

$$\min D = \sum_{i=n_0}^{n_0+4} |F(n_0) - F(n_0 - 1)| \ : \ 1 \leq n_0 < N - N_w - 5, \tag{3.1}$$

where D is the sum of absolute differences to be minimized, n_0 is the first sample of this small region, which is varied to minimize D, F are the formant values calculated for each sample in the pitch period, and N is the number of samples in the pitch period. The size of the initial stable region is five formant samples, which ensures meaningful statistics are available to extend the region.

Once an initial stable region is identified, the mean and standard deviation of the formants within this small region are calculated, and the region is grown based on the following criteria. If the next sample is less than two standard deviations from the mean, it is included in the stable region and the mean and standard deviation are recalculated before continuing on to test the next point. A slightly different algorithm is used to extend the window to the left. The final mean and standard deviation from extending the stable region to the right are kept constant, and the region is grown to the left until a sample is more than two of these standard deviations from this mean. The closed phase is considered to include every speech sample which was used to calculate the stable formant values. Since each formant value is calculated from N_w speech samples, the total length of the closed phase will be $n_2 - n_1 + N_w$ samples,

where n_1 is time of the first formant in the stable region and n_2 is the time of the last formant in the stable region.

There are two primary reasons for the different techniques used to identify the glottal opening and closure. First, after the region has been extended to the right to identify the glottal opening, the statistics have been estimated from sufficient data and extending the window to the left will not improve those estimates. More importantly, we have found that the glottal opening tends to result in sudden formant shifts, while gradual formant shifts are found when extending the region to the left towards glottal closure. This may be because the sub- and supra-glottal pressures are approximately equal during the return phase, which combined with the minimal flow results in little influence on the vocal tract estimate. If we attempted to update the statistics during a gradual change in the formant estimate, the statistics would likely incorporate this change, and glottal closure would not be identified. The flowcharts in figures 3-4 and 3-5 illustrate the techniques used to identify glottal opening and glottal closure.

Identifying a small initial stable region allows the algorithm to adapt to the variability of the formants for each frame. If there is more aspiration or ripple during the closed phase, the initial standard deviation calculated from this window will reflect the greater variability that will occur in the formant estimates due to the nonlinear source-filter interaction. When the glottis begins opening from its maximally closed position, the interaction will increase, and the standard deviation limits will be exceeded, indicating the glottis has begun to open.

3.1.3 Measuring Formant Motion

In the above discussion the specific parameter used for the formant estimates was not stated. According to the theory presented in section 2.1.2, all of the formants will undergo modulation of both their frequencies and bandwidths. The first formant shows these modulations clearer than other formants, in part because the energy of the first formant is greater and estimates of it tend to be less effected by noise. In general, both the formant frequencies and bandwidths tend to increase during the open phase, while they remain relatively constant during the closed phase.

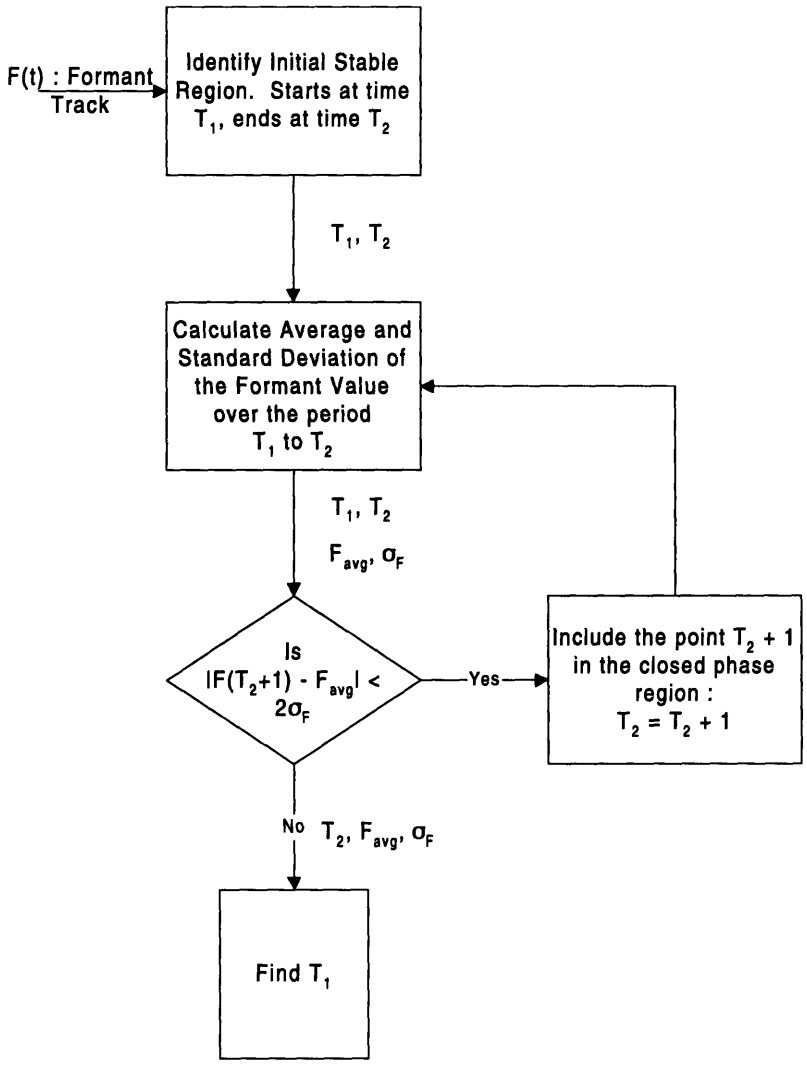

Figure 3-4: Glottal opening is identified by growing a small region in which the first formant frequency is stable until the next sample is greater than two standard deviations from the mean of the formants in the region. The procedure is illustrated in this flow chart.

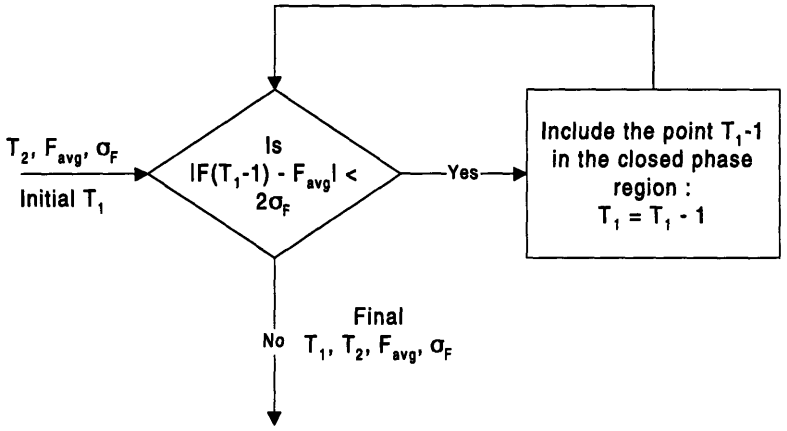

Figure 3-5: Glottal Closure is identified similarly to glottal opening, except that the mean and standard deviation of the formant values within the region are not updated as the region is grown. The procedure is illustrated in this flow chart.

Experiments showed that the best measure to use in determining formant modulation is the frequency of the first formant. As can be seen from figure 3-3, the first formant is more stable than higher formants during the closed phase and exhibits a more observable change at the start of the open phase. Also, the sliding covariance and formant tracker tend to make more errors for higher formants; examples of these errors can be seen in figure 3-3, such as the poles around 1000 Hz assigned to F3 between samples 150 and 200, since no poles in near F3 were calculated by the covariance analysis for these sample.

Tables 3.1 and 3.2 show a measure of signal to noise ratio (SNR) for various statistics which could be used to identify the closed phase. The SNR is calculated as the ratio of the average variance at the start of the open phase to the average variance during the closed phase. For example, for the frequency of the first formant, for each pitch period, we would have:

$$\bar{F}_1 = \frac{1}{T_o - T_c} \sum_{t=T_o}^{T_c-1} F_1(t) \tag{3.2}$$

$$\sigma_C^2 = \frac{1}{T_o - T_c} \sum_{t=T_o}^{T_c-1} (F_1(t) - \bar{F}_1)^2 \tag{3.3}$$

$$\sigma_O^2 = \frac{1}{5} \sum_{t=T_c}^{T_c+4} (F_1(t) - \bar{F}_1)^2, \tag{3.4}$$

where $F_1(t)$ is the frequency track of the first formant, $\bar{F}_1$ is the average frequency of the first formant during the closed phase, T_c is the time of glottal closing, T_o is the time of glottal opening, σ_C^2 is the variance of the closed phase, and σ_O^2 is the variance of the open phase. The variance of a signal is the average AC energy per sample in the signal, thus a ratio of variances is a signal to noise ratio in the standard sense. The variance during the closed phase can be viewed as background noise—this is the formant modulation that is measured when the glottis is maximally closed. The variance in the open phase is the signal of interest—the variance in this signal is used to identify the opening of the glottis. For all pitch periods across multiple speakers,

SNR Measure	Males	Females
F_1 Freq	161	155
F_1 BW	8.7	4.2
F_2 Freq	8.3	5.7
F_2 BW	1.2	0.9

Table 3.1: Average Signal to Noise Ratios for several potential measures used in identifying the glottal opening. The closed phase was identified using the first formant frequency.

SNR Measure	Males	Females
F_1 Freq	25.1	13.7
F_1 BW	6.3	2.4
F_2 Freq	42.7	59.4
F_2 BW	2.2	2.0

Table 3.2: Average Signal to Noise Ratios for several potential measures used in identifying the glottal opening. The closed phase was identified using the second formant frequency.

we have:

$$SNR = \frac{\frac{1}{\sum_{S=1}^{N_S} N_P(S)} \sum_{S=1}^{N_S} \sum_{P=1}^{N_P(S)} \sigma_O^2(S,P)}{\frac{1}{\sum_{S=1}^{N_S} N_P(S)} \sum_{S=1}^{N_S} \sum_{P=1}^{N_P(S)} \sigma_C^2(S,P)}, \tag{3.5}$$

where N_S is the number of speakers, $N_P()$ is the number of pitch periods for each speaker, $\sigma_O^2()$ is the open phase variance for each pitch period, and $\sigma_C^2()$ is the closed phase variance for each pitch period.

The closed phase was determined for table 3.1 using the frequency of the first formant as the measure of formant modulation, while for table 3.2 the frequency of the second formant was used. The key feature to notice in this data is that the SNR for the F1 frequency in the first table is higher than the SNR for the F2 frequency in the second table. This indicates that the change in F1 frequency at the boundary of the identified closed phase is more noticeable than the change in F2 frequency at the boundary. Separately marked closed phase timings were not available, so a more rigorous evaluation of the features was not possible.

3.1.4 High Pitch Speakers: Using Two Analysis Windows

For high pitch speakers, it is possible that the analysis technique will require too large a region in attempting to determine a closed phase. The various windows used result in the closed phase identified being at least 21 samples in length. In particular, the minimum length of the sliding covariance window is 17 samples, while the minimum size of the initial stable region is five sequential sliding covariance windows, which will cover a total of 21 samples. At a 10khz sampling rate, this corresponds to a minimum closed phase of 2.1 ms. A speaker with a fundamental frequency of 200 Hz and a 70% open quotient will have a closed phase of only 1.5 ms:

$$\frac{1}{200\,Hz}0.3 = 1.5\,ms.$$

Many female speakers will accordingly have closed phases which are less than 2.1 ms. To help solve this problem, we use a covariance analysis which is based on two windows, one in each of a pair of pitch periods.

Assuming that the rate of change of the vocal tract is dependent on time and not on the number of pitch periods, the vocal tract variation over two frames for a 200 Hz voice is approximately the same as one frame of a 100 Hz voice, since both last for 10 ms. By splitting the sliding covariance analysis window into two parts, each one need be slightly larger than half the desired linear prediction order, which results in a minimum identifiable closed phase size of 1.3 ms, five sequential windows each half the size of the standard minimum window length of 17 samples. Since this technique is more dependent on stability of both the vocal tract and the source across multiple pitch periods, it is only used when the pitch period is small and a closed phase close to the minimum duration is identified.

As shown in section 1.1.2, the covariance method of linear prediction is the solution to the equation

$$\mathbf{\Phi}\vec{\alpha} = \vec{\psi}, \tag{3.6}$$

where the $(i,j)^{th}$ term of $\mathbf{\Phi}$ is given by $\phi_{i,j}$, where

$$\phi_{i,j} = \sum_{n=0}^{N-1} s[n-i]s[n-j] \ : \ 1 \leq i,j \leq p \tag{3.7}$$

and the two vectors are given by

$$\begin{aligned} \vec{\alpha} &= [\alpha_i, \alpha_2, \ldots, \alpha_p]^T, \text{ and} \\ \vec{\psi} &= [\phi_{0,1}, \phi_{0,2}, \ldots, \phi_{0,p}]^T. \end{aligned}$$

Two windows of speech data can be used to calculate the matrix $\mathbf{\Phi}$ and the vector $\vec{\psi}$,

$$\phi_{i,j} = \sum_{n=N_1}^{N_1+N_{l_1}-1} s[n-i]s[n-j] + \sum_{n=N_2}^{N_2+N_{l_2}-1} s[n-i]s[n-j] \ : \ 1 \leq i,j \leq p, \tag{3.8}$$

where N_1 is the start of the first region, N_{l_1} is the length of the first region, N_2 is the start of the second region, and N_{l_2} is the length of the second region. The only change required to convert the standard covariance linear prediction procedure into a two window procedure is this change in the calculation of the matrix $\mathbf{\Phi}$. The properties of the matrix $\mathbf{\Phi}$ still hold as long as the windows are non-overlapping, allowing efficient solution by Cholesky decomposition.

3.2 From Closed Phase to Glottal Flow Derivative

Once the closed phase is determined, the vocal tract response is calculated, and then used to inverse filter the speech signal to generate the glottal flow derivative waveform.

3.2.1 Vocal Tract Response

The vocal tract response is calculated from a rectangularly windowed region of the speech signal bounded on the left by the glottal closure and on the right by the glottal opening, as determined in the preceding section. The vocal tract is estimated using a covariance based linear prediction, with an adaptive pre-emphasis. To determine the

pre-emphasis coefficient, a first-order autocorrelation linear prediction is performed on the analysis window, including the preceding samples required to initialize the covariance analysis. This filter is then used to pre-emphasize the data. We have found this adaptive pre-emphasis to work better than a fixed pre-emphasis filter.

3.2.2 Inverse Filtering

There is some uncertainty as to what region to inverse filter with a particular vocal tract response. This problem arises due to the fact that the vocal tract is estimated during the closed phase but must be used to inverse filter both the closed and open phases. At first, we used a given filter to whiten the closed phase and the following open phase. This can create a problem, since the difference equation implementing the inverse vocal tract filter is changed at the start of the analysis window, when there is significant energy in the speech signal, and thus significant energy in the inverse filter. This sudden change of filter artificially excites the formants, and sometimes results in a large output shift.

The decay of a linear filter with zero input only contains components at pole locations. For speech, we have

$$s[n] = e[n] + \sum_{i=1}^{p} a_i s[n-i].$$

Considering $e[n]$ to be zero (superposition allows us to add in the response to $e[n]$ later), we have

$$s[n] - \sum_{i=1}^{p} a_i s[n-i] = 0.$$

Difference equations are easily solved through the z-transform, giving

$$S(z) - \sum_{i=1}^{p} a_i \left(S(z)z^{-i} + \sum_{k=1}^{i} s[-k]z^{1-k} \right) = 0, \tag{3.9}$$

where the inner sum is due to the initial conditions. Rearranging in the form required

for partial fraction expansion, we have

$$\begin{aligned}\frac{S(z)}{z} &= \frac{\sum_{i=1}^{p} a_i s[-i] z^{p-i}}{\sum_{i=0}^{p} a_i z^{p-i}} \\ &= \frac{\sum_{i=1}^{p} a_i s[-i] z^{p-i}}{\prod_{i=1}^{p} (z - z_i)},\end{aligned} \tag{3.10}$$

where $a_0 = 1$, and z_i are the complex pole locations. The partial fraction expansion of equation 3.10 will generally be of the form

$$S(z) = \sum_{i=1}^{p} \frac{C_i z}{z - z_i}, \tag{3.11}$$

where the C_i's are due to the initial conditions. A slightly different form of equation 3.11 will result under the unusual condition of repeated poles. The inverse f-transform of equation 3.11 is of the form

$$s[n] = \sum_{i=1}^{p} C_i z_i^n u[n], \tag{3.12}$$

where $u[n]$ is the unit step function. Under the normal condition of complex pole locations z_i, poles will appear in complex conjugate pairs, with their responses combining to form a decaying sine wave. Equation 3.12 shows that the only possible output is a combination of decaying sine waves at the pole frequencies. Since the only possible outputs are at the pole frequencies, if the filter is suddenly changed, the energy in the filter must be redistributed to the new frequencies. Experiments have confirmed that this redistribution can cause excitation of some of the formants, as shown in figure 3-6.

In order to minimize the impact of changing filters, we change the filter at the end of the analysis region, rather than the beginning. This means that a given vocal tract response whitens the closed phase from which it was calculated, as well as the preceding open phase. The filter change occurs later in the decay of the speech signal, so the speech has less energy. Also, the beginning of the closed phase is sometimes determined before the end of the return phase. By changing filters at the end of the

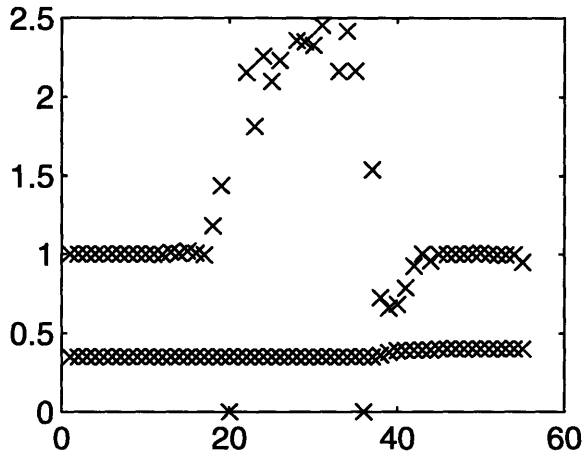

Figure 3-6: This plot shows the pole frequencies of the impulse response to a two pole system whose lower pole frequency changes at approximately sample 40. The frequencies are generated by a sliding covariance analysis. The higher pole has a larger bandwidth, causing the energy at that frequency to fall below the noise floor at approximately sample 20. When the filter is changed, the higher pole is excited due to the redistribution of energy that occurs when the characteristic modes of the response change.

identified closed phase, the glottal flow derivative waveform will generally be closer to zero, which results in a smaller jump in the output.

3.3 Examples

Figure 3-7 shows several examples of glottal flow derivatives. Two examples are shown for each of four speakers. The small pulses superimposed on the glottal flow derivative waveforms show the times of glottal closure and opening as identified through the modulation for the first formant, while the large pulses represent the initial estimates of the time of the glottal pulse. All the examples in the first column come from the vowel in the word "had". The plots in the second column come from different sounds. The first column demonstrates that glottal flow varies from one speaker to another for a particular sound. The two examples for each speaker demonstrate some of the speaker dependent characteristics of the flow, as well as the variety of flow for each speaker. These eight speech segments will also be used in chapters 4 and 5 to

demonstrate the algorithms discussed in those chapters.

The first speaker shows significant energy during the closed phase, and a sudden increase and drop-off of the glottal flow derivative. The second speaker typically exhibits very little energy during the closed phase. This speaker also tends to have a very impulse-like glottal pulse, the first example is atypical in this feature. The third speaker commonly exhibits a large amount of ripple during the open phase. The first example for this speaker is taken from the onset of the vowel, and demonstrates that superposition of the decay from previous pulses causes the ripple to build up over several periods. The fourth speaker has a large open quotient, as can be seen from the second example.

3.4 Summary

The estimation of the glottal flow derivative waveform is automatic and requires only information which can be directly calculated from the speech signal. Two innovative techniques are used: identifying the closed phase through formant motion calculated by a sliding covariance analysis, and a two-window covariance analysis used for high pitch speakers. By identifying statistically significant variations in the frequency of the estimated first formant, we are able to identify when the glottis finishes closing and when it begins opening. The formant motions are predicted by the theory of interaction between the glottal flow and the vocal tract.

A high fundamental frequency poses a problem for this algorithm, as it does for many speech algorithms. To help minimize the difficulty associated with high pitch speakers, a covariance based linear prediction was developed which has two disjoint windows. For a standard covariance analysis, the window size must be greater than or equal to the order of the analysis. The two window technique allows the sum of the length of the two windows to be greater than or equal to the analysis order. This enables the identification of a smaller closed phase for high pitch speakers.

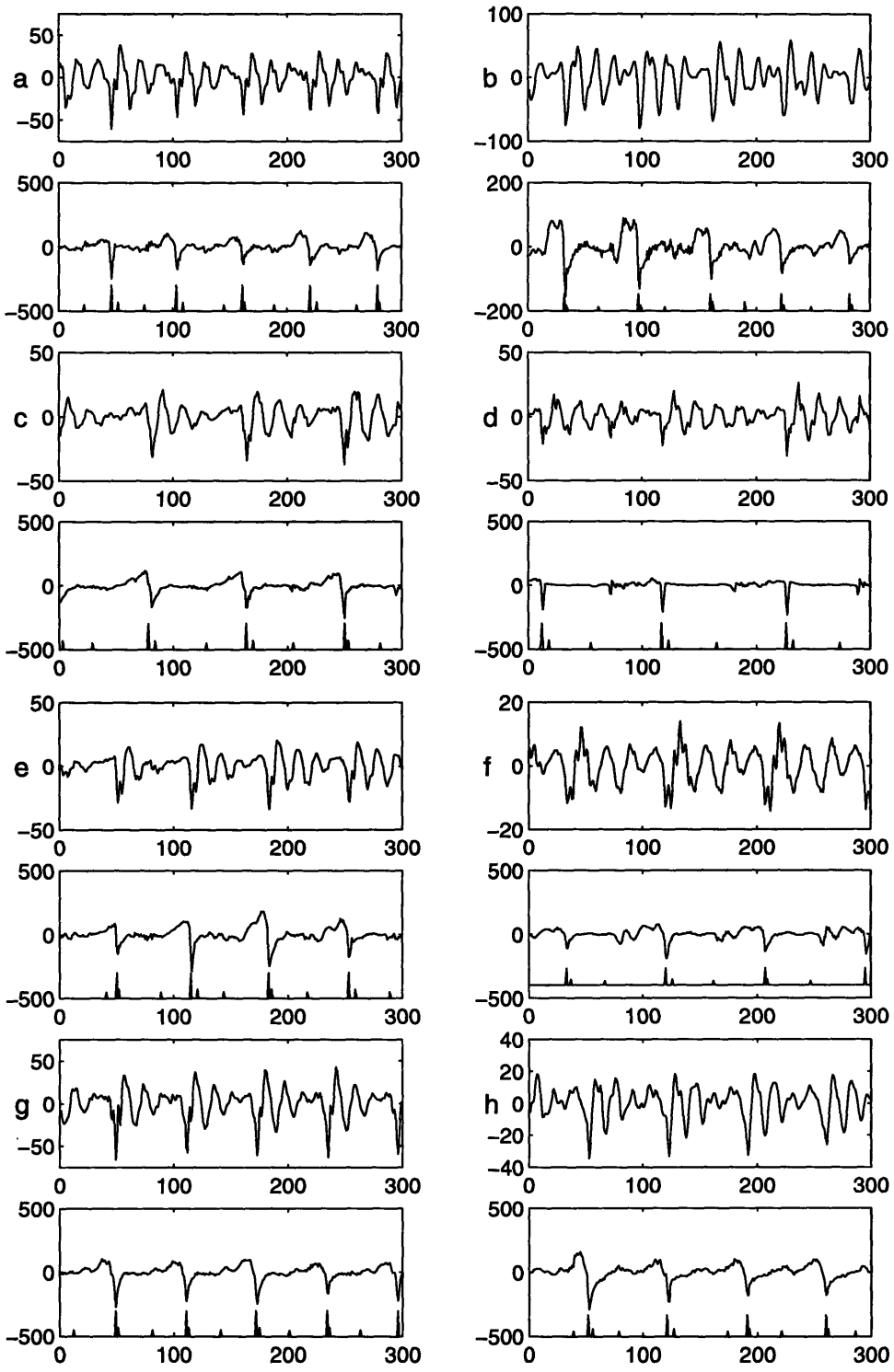

Figure 3-7: Several examples of estimated glottal flow derivatives. The speech signals are above the corresponding glottal flow derivative waveforms. Each row represents a speaker. All the examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels.

Chapter 4

Estimating Coarse Structure

Chapter 3 developed the techniques used to calculate the glottal flow derivative waveform from the speech signal. Now that we have the source waveform, we can estimate the parameters of a model describing the general shape of the waveform.

4.1 Formulation of the Estimation Problem

The coarse structure of the glottal flow derivative is captured using the LF model, described by the equation

$$E(t) = \frac{dU_g}{dt} = E_0 e^{\alpha t} \sin \omega_g t \tag{4.1}$$

for the period from glottal opening (T_0) to the pitch pulse (T_e, time of excitation), at which time the return phase starts:

$$E(t) = \frac{-E_0}{\epsilon T_a} \left[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)} \right], \tag{4.2}$$

which continues until time T_c. Figure 4-1 shows an example of the LF model. Due to the large dependence of E_0 on α as will be discussed below, the parameter E_e, the value of the waveform at time T_e, is estimated instead of E_0. To calculate E_0 from

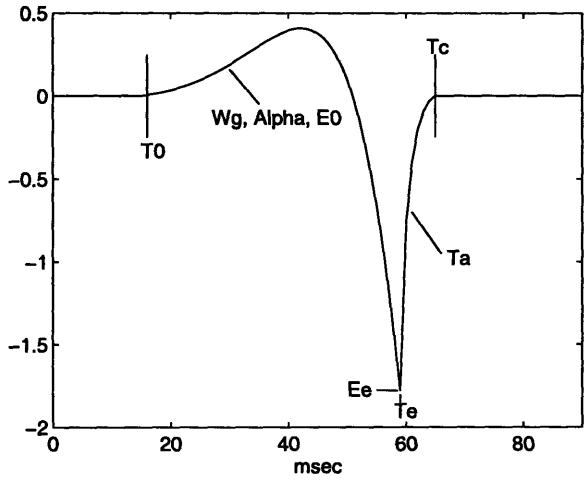

Figure 4-1: LF Model for the glottal flow derivative waveform. (Repeat of figure 2-4 for convenience)

E_e, the equation

$$E_0 = \frac{E_e}{e^{\alpha T_e} \sin \omega_g T_e} \tag{4.3}$$

is used.

The seven parameters to be estimated for each pitch period are described in table 4.1. A least squares minimization problem can be setup to fit the LF model to the glottal flow derivative waveform:

$$\begin{aligned} E(\vec{x}) \;=\; & \sum_{n=0}^{T_0} G^2[i] + \sum_{n=T_0+1}^{T_e} \left(G[n] - E_0 e^{\alpha n} \sin \omega_g n\right)^2 + \\ & \sum_{n=T_e+1}^{T_c} \left(G[n] - \frac{-E_0}{\epsilon T_a}\left[e^{-\epsilon(n-T_e)} - e^{-\epsilon(T_c-T_e)}\right]\right)^2 + \\ & \sum_{n=T_c+1}^{N} G^2[n], \end{aligned} \tag{4.4}$$

where the point $n = 0$ occurs after the end of the previous return phase, $n = N$ occurs before the next open phase, $\vec{x}$ is a vector of the seven parameters, and $G[n]$ is the glottal flow derivative waveform at sample n. The error E is a nonlinear function of the seven model parameters, so the problem must be solved iteratively using a nonlinear least-squares algorithm.

The standard Gauss-Newton and Levenberg-Marquardt methods for solving non-

T_0	The time of glottal opening
α	Determines the ratio of E_e to the height of the positive portion of the glottal flow derivative
ω_g	Determines the curvature of the left side of the glottal pulse, also determines how much time elapses between the zero crossing and T_e
T_e	The time of the maximum negative value of the waveform, called the glottal pulse
E_e	The value of the waveform at time T_e
T_a	An exponential time constant which determines how quickly the waveform returns to zero after time T_e
T_c	The time of glottal closure

Table 4.1: Description of the seven parameters of the LF model for the glottal flow derivative waveform.

linear least-squares problems do not work well when the minimum error E is large [30], which is often the case in fitting the LF model to the glottal flow derivative waveform, since ripple, not modeled by the LF model, will show up in E. After finding that the Levenberg-Marquardt routine in [43] does not work well for our particular problem, we switched to using a version of the NL2SOL algorithm for adaptive nonlinear least-squares regression, Association for Computing Machinery (ACM) algorithm 573 [30, 31]. The modified version we used has the addition of bounds to enable parameters to be limited to physically reasonable values.

4.2 The NL2SOL Algorithm

The NL2SOL algorithm uses two models of the error function to iteratively reduce the residue. The residue is defined as:

$$r_i(\vec{x}) = m_i(\vec{x}) - y_i, \tag{4.5}$$

where $\vec{x}$ is a vector of the parameters to be solved for, y_i is the data to be fitted, $m_i(\vec{x})$ is the value of the curve at point i using the parameters $\vec{x}$, and $r_i(\vec{x})$ is the

fitting error, or residue, which is to be minimized in a least squares sense:

$$\min f(\vec{x}) = \frac{1}{2}\sum_{i=1}^{N} r_i^2(\vec{x}) = \frac{1}{2}\vec{R(\vec{x})}^T\vec{R}(\vec{x}), \tag{4.6}$$

with

$$\vec{R(\vec{x})} = [r_1(\vec{x}), r_2(\vec{x}), \ldots, r_N(\vec{x})],$$

where $f(\vec{x})$ is the summed squared residue to be minimized. The specific value of $\vec{x}$ that minimizes equation 4.6 is written as $\vec{x}^*$. If $f(\vec{x}^*) = 0$, the fitted curve perfectly matches the data, and the global minimum has been found. For the typical case, $f(\vec{x})$ will be considered to be a local minimum when one of various convergence criteria are reached.

In order to iteratively minimize f, we need to know how to change the parameter vector $\vec{x}$ from its current value, which we will call $\vec{x}_k$. The Taylor series expansion of $f(\vec{x})$ around the point $\vec{x}_k$ is given by

$$f(\vec{x}) \approx \frac{1}{2}\vec{R}(\vec{x}_k)^T\vec{R}(\vec{x}_k) + (\vec{x} - \vec{x}_k)^T\nabla f(\vec{x}) + \frac{1}{2}(\vec{x} - \vec{x}_k)^T\nabla^2 f(\vec{x})(\vec{x} - \vec{x}_k) + \ldots, \tag{4.7}$$

where the gradient of $f(\vec{x})$ is given by

$$\nabla f(\vec{x}) = \boldsymbol{J}(\vec{x})^T\vec{R}(\vec{x}), \tag{4.8}$$

where the $(i,l)^{th}$ element of the Jacobian matrix $\boldsymbol{J}(\vec{x})$ of the vector $\vec{R}(\vec{x})$ is given by

$$j_{i,l}(\vec{x}) = \frac{\partial r_i(\vec{x})}{\partial \vec{x}_l}. \tag{4.9}$$

In other words, the $(i,l)^{th}$ element of $\boldsymbol{J}$ is the partial derivative of the residue $\vec{R}(\vec{x})$ at the point i with respect to the l^{th} element of the parameter vector $\vec{x}$. The second order gradient of $f(\vec{x})$, called the Hessian, is given by the equation

$$\boldsymbol{\nabla^2} f(\vec{x}) = \boldsymbol{J}(\vec{x})^T\boldsymbol{J}(\vec{x}) + \sum_{i=1}^{n} r_i(\vec{x})\nabla^2 r_i(\vec{x}). \tag{4.10}$$

The Taylor series approximation to $f(\vec{x})$ is a linear function of the powers of $\vec{x}$, which can be minimized explicitly. Since the Taylor series is just an approximation if a finite number of terms are used, the minimum of $f(\vec{x})$ must be found iteratively:

1. Start with an initial guess for $\vec{x}^*$, $\vec{x}_0$.
2. Calculate the Taylor series expansion of $f(\vec{x})$ around the point $\vec{x}_k$, where k is the current iteration number.
3. Choose $\vec{x}_{k+1}$ as the parameter vector which minimizes the Taylor series.
4. If the difference between $\vec{x}_k$ and $\vec{x}_{k+1}$ is small, or the value of $f(\vec{x})$ is small at $\vec{x}_{k+1}$, consider $\vec{x}^*$ to equal $\vec{x}_{k+1}$. If not, return to step 2 to refine the estimate of $\vec{x}^*$.

Using just the first order term of the Taylor series makes the assumption that $f(\vec{x})$ can be adequately modeled by an affine function, giving the Newton method. Including the second term makes the assumption that $f(\vec{x})$ can be modeled by a quadratic. In general, the second partial derivatives of $r_i(\vec{x})$ will not be available, so the second term must be approximated if we wish to use it. The Jacobian of $f(\vec{x})$ will give a good approximation to the Hessian of $f(\vec{x})$ through the first term in equation 4.10. Just using the first term of the Hessian of $f(\vec{x})$ gives the Gauss-Newton method.

The second term in the Hessian of $f(\vec{x})$ can be approximated by the first difference of the Jacobian, giving

$$\boldsymbol{S}_{k+1}\Delta\vec{x}_k = \boldsymbol{J}_{k+1}^T\vec{R}_{k+1} - \boldsymbol{J}_k^T\vec{R}_{k+1}, \tag{4.11}$$

where $\boldsymbol{S}_{k+1}$ is the new approximation to the second term of the Hessian of $f(\vec{x})$ and $\Delta\vec{x}_k = \vec{x}_{k+1} - \vec{x}_k$. By rearranging equation 4.11 in a non-valid manner for the purposes of understanding the value $\boldsymbol{S}_{k+1}$, we have

$$\boldsymbol{S}_{k+1} = \frac{\boldsymbol{J}_{k+1}^T - \boldsymbol{J}_k^T}{\Delta\vec{x}_k}\vec{R}_{k+1}. \tag{4.12}$$

Comparing equation 4.12 to the second term of equation 4.10, we see that the Hessian of $\vec{R}$ is approximated through use of the change in the Jacobian from iteration k to

iteration $k+1$. As $\Delta\vec{x}_k$ approaches 0, this approximation approaches the actual value in equation 4.10. An equation exists to calculate $\boldsymbol{S}_{k+1}$ from $\boldsymbol{S}_k$. $\boldsymbol{S}_0$ is initialized to zero.

The NL2SOL algorithm uses two methods to minimize $f(\vec{x})$, the standard Gauss-Newton method and the Gauss-Newton method refined by using $\boldsymbol{S}_k$ to more accurately approximate the Hessian of $f(\vec{x})$. It has been found that the Gauss-Newton method predicts $f(\vec{x}_{k+1})$ better than the Hessian method for small values of k. Accordingly, the authors of the NL2SOL algorithm have devised a technique for choosing between the two methods. Also, they implement a "trust region" which indicates the region around $\vec{x}_k$ in which they have confidence of the current model for $f(\vec{x})$. If a proposed step is outside the trust region, the step is changed to the closest point in the trust region to the proposed step. The algorithm is considered adaptive because the calculation of the trust region and the selection of which model to use for a given step are based on the previous and current steps in the iteration.

4.3 Using the NL2SOL Algorithm

For our application of the NL2SOL algorithm to model the glottal flow derivative waveform with the LF model, the parameter vector $\vec{x}$ consists of the seven model parameters, the vector $\vec{R}$ is the difference between the model and the waveform, with one element of $\vec{R}$ for each sample between the previous glottal closure and the subsequent glottal opening. The implementation of the algorithm takes as input the vector $\vec{R}$ and the matrix $\boldsymbol{J}$. Calculation of $\boldsymbol{J}$ requires evaluation of the partial derivatives of the LF model equations. The 14 partial derivative equations (seven for the open phase and seven for the return phase) were derived and are included in the analysis software.

4.3.1 Difficulties with the NL2SOL Algorithm

We have encountered several difficulties in using this algorithm, some of which will be described here. The most fundamental problem is that of identifying the times T_0, T_e,

and T_c, due to discontinuities of the partial derivatives at these points. The partial derivatives will be discontinuous at these points because of the piecewise nature of the LF model. If the partial derivatives are discontinuous, $f(\vec{x})$ will not be adequately modeled by the first two terms of the Taylor series expansion. The result of this inadequate modeling is that the NL2SOL algorithm will have a harder time converging to the correct solution, and will be more likely to find a local minimum that is not the global minimum.

The problems caused by poor modeling of $f(\vec{x})$ can be reduced by starting the iteration with an accurate estimate of these parameters. The partial derivatives of the parameter T_e will have the largest discontinuity, due to the pulse-like shape of the waveform at time T_e. Luckily, the time T_e is easily estimated by identifying the largest negative sample during the pitch period. In general, we find that the parameter T_e is varied less than a sample from its initial estimate.

The partial derivatives will be zero before the time T_0 and after the time T_c, since the model function equals zero for these regions. This indicates to the NL2SOL algorithm that no change in parameters will reduce the error in these regions. In order that the algorithm be given the opportunity to adjust the parameters for these regions, and possibly include them in the pulse shape, the initial values for the parameters T_0 and T_c are each moved five samples further away from the center of the pulse than they would normally be set.

The LF model is perhaps over-specified when described by the seven parameters, which can cause problems in their estimation. For example, E_0 and α can be traded off for each other at high values of α. A large α will result in a very impulse-like waveform. There is little difference between a waveform that is larger because it is slightly more impulse-like and one that is larger everywhere, since most of the energy is in the impulse already, as illustrated in figure 4-2. Similarly, T_a and T_c can be confused, since small values of T_a make T_c difficult to estimate through a least-squares algorithm, since the exponentially decaying tail of the return phase will become very close to 0 (A similar effect can happen with large values of α and T_0).

To solve the confusion between E_0 and α, we estimate E_e, the energy at time T_e.

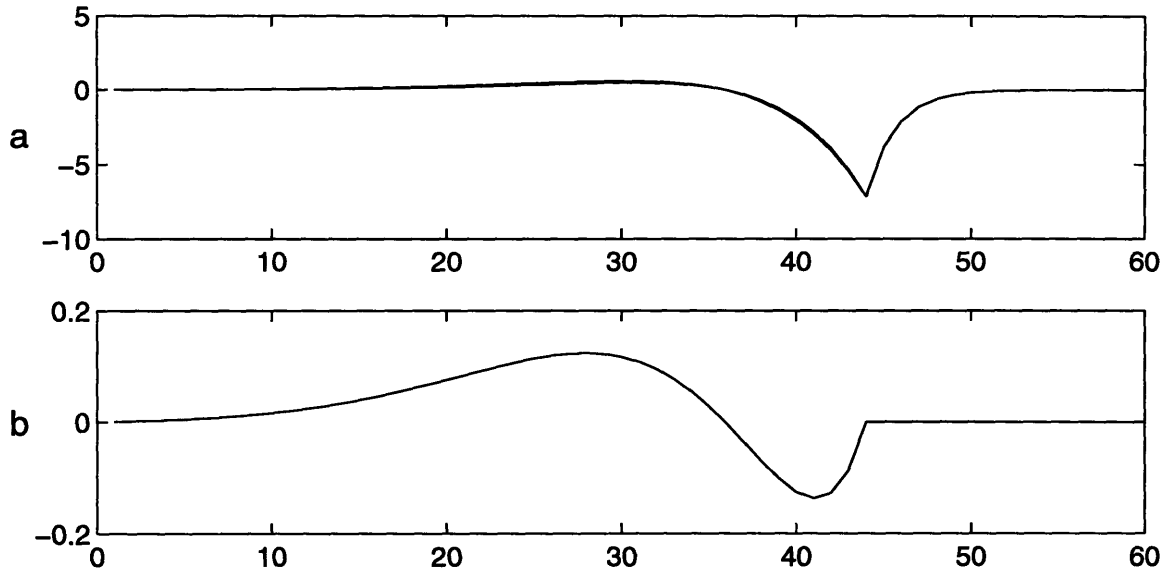

Figure 4-2: The parameters α and E_0 of the LF model for the glottal flow derivative waveform can be traded off for large values of α as demonstrated in this figure. A 10% increase in α and a 50% decrease in E_0 result in a squared error change of only 0.2% for this example. (a): Two superimposed LF waveforms, one with $\alpha = 7$ and $E_0 = 100$, the other with $\alpha = 7.7$ and $E_0 = 50$. (b): Difference between the two waveforms in (a).

This has two benefits, an accurate estimate is easy to obtain, since it is the largest negative value of the data during the pitch period. Also, E_e is essentially unrelated to α. To work around the problem of inaccurate estimates of T_0 and T_c depending on values of α and T_a, we recalculate T_0 and T_c from the model waveform after the NL2SOL algorithm estimates the parameter values. These times are set at the time at which the predicted model is less than one percent of E_e.

In order to ensure that the NL2SOL algorithm estimates physically reasonable parameter values, we set bounds on the parameters. For example, if the value ω_g is less than π, the model will have no negative samples, since the sine term will never go below zero.

4.3.2 Discarding Data

Once the model has been estimated, parameter values that are too close to their bounds, or too low a value for E_e will cause the data for that frame to be considered

unreliable and discarded before further analysis. This was found to be important for improving speaker identification scores as will be further explained in chapter 6. The particular bounds used were not chosen in a rigorous manner. Informal experiments indicated that the initial bounds choose were reasonable in the context of speaker identification.

4.3.3 Initialization

The curve fitting algorithm must be initialized with parameters close to the proper values, else a local minimum that is not the global minimum may be identified. The parameters T_e and E_e are easily calculated, and are used to identify the location of the waveform and its scale. The other five parameters are initialized with the value calculated for the previous frame, with T_0 decreased and T_c increased as described above. For the first pitch period in an utterance, the parameters are initialized with preset reasonable values.

4.4 Other Possible Approaches

There are two primary alternate approaches to determine the parameter values. One alternative approach is to model the glottal flow derivative in the frequency domain rather than the time domain. There are several problems associated with such an approach. First, there is no closed form solution for the frequency response of the LF model, so each set of trial parameters would require an FFT to convert from the time domain representation of the model to the frequency domain representation, making the problem computationally unfeasible. Another problem is that the observed waveforms tend to have their high-frequency energy distributed throughout the period, likely due to aspiration noise. Fitting in the frequency domain would result in reduced temporal resolution, clumping the distributed high frequency energy into a slightly sharper pulse, making the task of estimating aspiration more difficult.

Fant has described a more direct method of identifying the parameter T_a through frequency domain characteristics [15]. The effect of T_a is to low-pass filter the pulse

with a first order filter with a corner frequency

$$F_a = 1/(2\pi T_a). \tag{4.13}$$

This can be used to estimate T_a through its effect on the initial energy in a formant. Simple estimators like this one are not available for the other parameters.

4.5 Examples

Figure 4-3 shows the LF model of the coarse structure extracted from the same examples shown in figure 3-7 of section 3.3. The plots are in sets of two, the top is the estimated glottal flow derivative waveform, while the lower plot is the modeled coarse structure of the estimated glottal flow derivative waveform. Each row shows two examples for a particular speaker.

The first speaker generally exhibits a long return phase, and positive and negative extremes of approximately the same amplitude. The closed phase is fairly short, this is captured as a large open quotient. The second speaker shows a mush longer closed phase and smaller open quotient, as well as a lower pitch, which we do not model. The glottal pulse is of much larger amplitude than the positive portion of the glottal flow derivative. This speaker seems to often have a very short return phase. The third speaker has a pulse shape that is somewhat of an average of the first two speakers. The open quotient is closer to 50%, and the glottal pulse seems to be slightly larger than the positive portion of the waveform. The fourth speaker exhibits a much more gradual flow. The return phase is sometimes quite long, while the positive portion of the waveform develops slowly. These two factors lead to a short closed phase and large open quotient.

4.6 Summary

A nonlinear least-squares algorithm is used to fit the LF model to the glottal flow derivative waveform for each pitch period. Steps must be taken in order to ensure

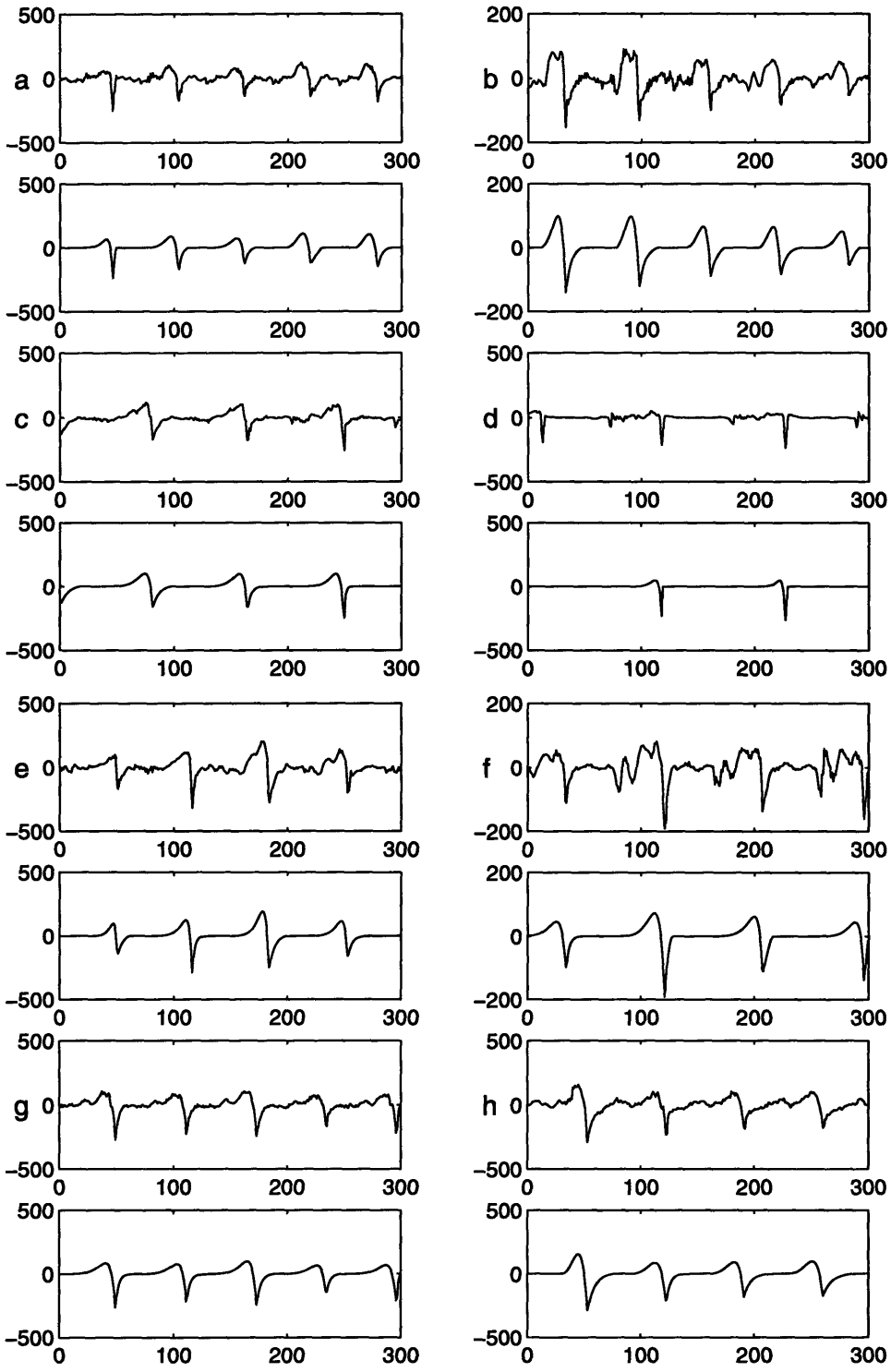

Figure 4-3: Several examples of the LF model for the coarse structure in the estimated glottal flow derivative. The glottal flow derivatives are shown above the corresponding model waveforms. Each row represents a speaker. All the examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels.

that the curve fitting is performed in a manner that yields meaningful results. Large dependencies between several pairs of parameters required that some parameters be changed, while other parameters are recalculated after the curve fitting is complete. Once the coarse structure of the glottal flow derivative has been modeled, it can be subtracted from the waveform, leaving the fine structure. Modeling of the fine structure is discussed in the next chapter.

Chapter 5

Estimating Fine Structure

In the previous chapter we estimated the coarse structure of the glottal flow derivative waveform. Subtracting the estimated coarse structure from the glottal flow derivative waveform yields the fine structure. We look at two sources of fine structure, aspiration noise and ripple due to source-filter interaction. This chapter discusses the techniques used to model the fine structure of the source.

In general, we are interested in the magnitude and timing of fine structure. For example, we are interested in when there is aspiration noise and how much aspiration noise is present. In addition to magnitude and timing measures, the extent of formant modulation can be modeled as separate information.

5.1 Modeling Ripple Through Formant Modulation

As shown in section 2.1.2, the bandwidth and frequency modulation of the formants due to source-filter interaction are related to the glottal area and the derivative of the glottal area, respectively. We model only frequency modulation in this study. Bandwidth modulation was not modeled due to time constraints.

The modulation of the first formant frequency is modeled using a parabola. Setting up a least sum of squares minimization of the error between the parabola and the

data, we have

$$e[n] = \left(A + Bn + Cn^2\right) - F_1[n] \tag{5.1}$$

$$\min E = \sum_{n=0}^{N} e^2[n], \tag{5.2}$$

where $n = 0$ is the start of the region to be modeled, $n = N$ is the end of the region, A, B, and C are the parameters to be estimated, and $F_1[n]$ is the frequency of the first formant. The formant estimates from the sliding covariance analysis tend to be noisy during the open phase. Because of this, we use a robust linear regression algorithm called PROGRESS [48].

In order to increase the robustness of least squares regression, we replace the summation by a median, giving a *least median of squares* (LMS) estimator,

$$\min E = \text{med}_{n=0}^{N} e^2[n], \tag{5.3}$$

where $\text{med}_{n=0}^{N}$ indicates the median value of the error samples from $e^2[0]$ to $e^2[N]$. Half of the samples of $F_1[n]$ have an squared error less than E as calculated in equation 5.3. The solution minimizing E can be seen to be the curve that most closely matches half of the data. One solution to this minimization problem, as presented in [48], is to take all subsets of p different observations of $F_1[n]$, where p is the dimension of the model to be fit to the data, 3 in this case. For each subset, the curve that passes through the three points is calculated as a trial solution, and the median given in equation 5.3 is calculated. The trial solution with the minimum median is the LMS estimate.

The LMS estimator can be shown to be very robust in the presence of outliers, but requires a larger number of samples to accurately fit the model to the data than traditional least squares regression. To increase the accuracy of the fit with a limited amount of data, the PROGRESS algorithm performs a *re-weighted least squares* (RLS)

estimation, as given by

$$\min \sum_{n=0}^{N} w_n \left((A + Bn + Cn^2) - F_1[n] \right)^2, \tag{5.4}$$

where the w_n are weights designed to lessen the influence of outlier points in the data. The outlier points are determined by an estimate of the average error of the LMS fit. In particular, the statistic

$$\hat{\sigma} = C_1 \sqrt{\text{med } (e^2[n])}, \tag{5.5}$$

is used, where $e[n]$ is the modeling error for each sample, and C_1 is a constant related to statistical modeling of the error. The value $\hat{\sigma}$ is the standard deviation of the error for the best half of the data. If we wish to discard outliers, the weights w_n can be chosen as

$$w_n = \begin{cases} 1 & \text{if } |e[n]/\hat{\sigma}| \leq 2.5 \\ 0 & \text{if } |e[n]/\hat{\sigma}| > 2.5 \end{cases} \tag{5.6}$$

where the bound 2.5 is chosen somewhat arbitrarily. Alternatively, a smooth function could be developed to lessen the impact of outliers on the fit.

Now that we have a method for robustly estimating the formant modulation, we must determine over what region the formant modulation should be estimated. The most obvious region to choose includes the samples after the closed phase and before the glottal pulse. Occasionally a single noisy estimate of the first formant will cause an early identification of glottal opening. To avoid modeling a region which includes formant values which belong in the closed phase, the start of the open phase is identified as the first of five sequential samples which are outside the two standard deviation bound set for identifying glottal opening.

5.2 Time Domain Fine Structure

The primary information extracted from the time-domain fine structure is the magnitude of the effects causing fine structure. The amount of fine structure present indicates how much the volume velocity flow through the glottis is varying due to

ripple and the amount of aspiration noise as determined by the configuration of the glottal opening. The five energy measures used in calculating time domain energy in the source waveform were described in section 2.2.2. The five energy measures are calculated as the energy of the fine structure during the appropriate period normalized by the total energy in the estimated glottal flow derivative waveform for that pitch period. The total energy is given by

$$E_{tot} = \sum_{n=T_{c-1}}^{T_c} G^2[n] \tag{5.7}$$

where $G[n]$ is the glottal flow derivative waveform, T_{c-1} is the end of the previous return phase, and T_c is the glottal closure of the current pitch period. As an example of the energy measures, the energy of the fine structure during the open phase as determined by the LF model parameters is calculated as

$$E_o = \frac{1}{E_{tot}} \sum_{n=T_0}^{T_e-1} (G[n] - LF[n])^2 \tag{5.8}$$

where $LF[n]$ is the model of the coarse structure. The energy of the fine structure during the other four periods is similarly calculated.

5.3 Examples

We continue to use the same speech segments shown in figure 3-7 of section 3.3 to demonstrate the fine structure in speech. This structure is much more difficult to observe from the speech waveform than the coarse structure shown in figure 4-3. Figure 5-1 shows the residue from which the energy measures of time domain ripple are calculated. The residue figures are scaled to make the features more visible; the numbers labeling the y-axes of the speech and fine structure can not be compared. Some of the examples exhibit clear ripple, while others seem to just show noise. Some of the examples do show a significant difference in energy between the closed and open phases, which is indicative of a complete glottal closure, while those examples

that have similar energy during the closed and open phases are likely indicative of incomplete glottal closure.

5.4 Summary

Modulation of the first formant frequency is estimated by fitting a parabola to the covariance formant track. The fit is accomplished using a robust least squares algorithm. We do not propose that a parabola is an accurate representation of the formant modulation, but rather it is a simple model intended to show that formant modulation is speaker dependent, as predicted by its dependence on the time evolution of the glottal area. The magnitude of the combined aspiration noise and ripple are modeled by calculating the normalized energy during five regions of the pitch period. The application of the model parameters to speaker identification is discussed in the next chapter.

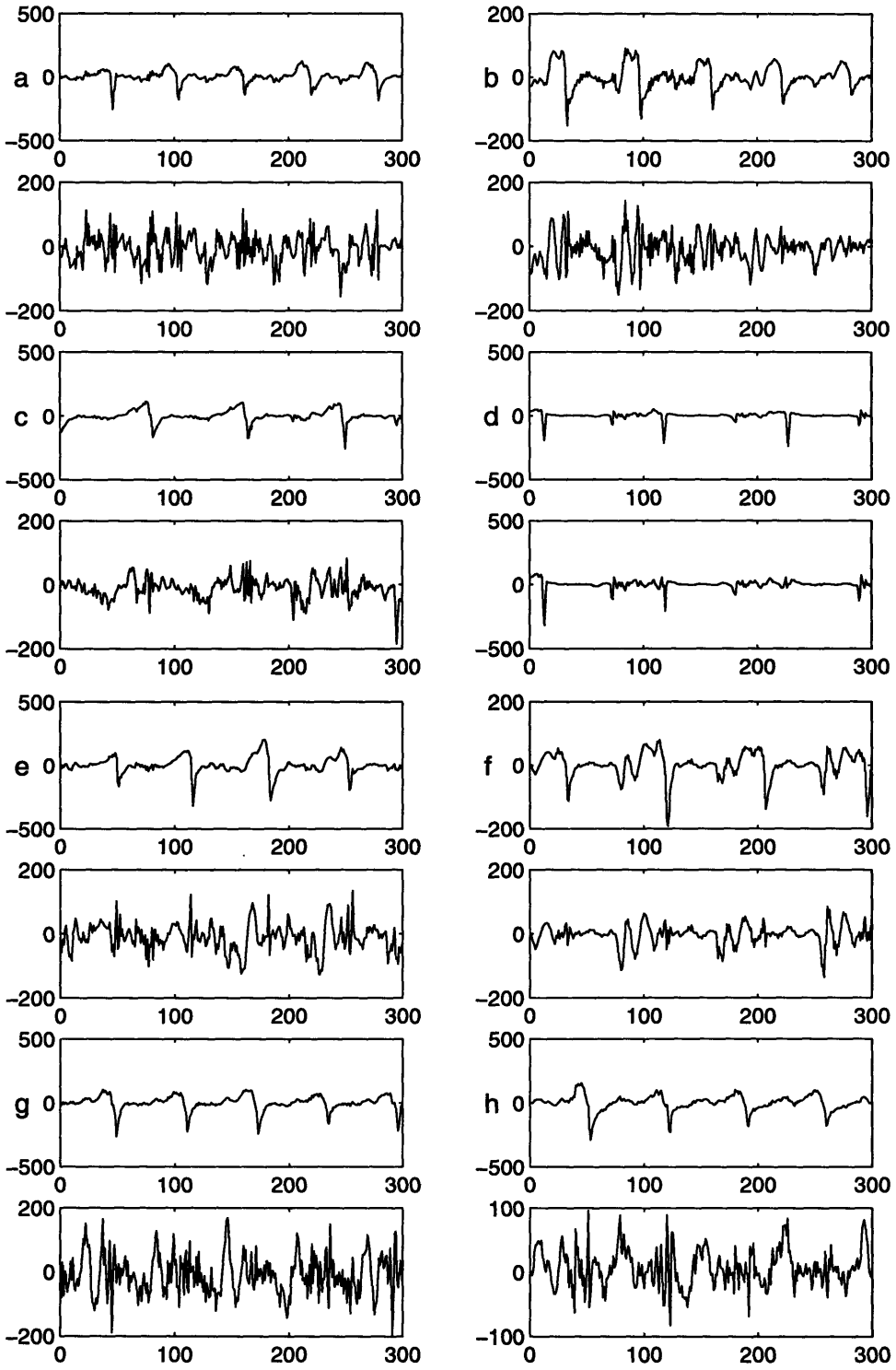

Figure 5-1: Several examples of the LF model for the coarse structure in the estimated glottal flow derivative. The glottal flow derivatives are shown above the corresponding model waveforms. Each row represents a speaker. All the examples in the first column are from the vowel in the word "had," while the examples in the second column come from various vowels.

Chapter 6

Speaker Identification Experiments

Previous chapters discussed estimation of the glottal flow derivative from speech and modeling the coarse and fine structure of this source waveform. We now discuss the application of the parameters of these models to speaker identification.

6.1 Background

We first begin with a discussion of the general topic of speaker identification. In speaker identification, we wish to take a spoken utterance, extract features from the utterance, and compare these feature values against speaker models that were previously generated to determine the identity of the speaker. We use the Reynolds Gaussian mixture model (GMM) speaker identification (SID) system in this thesis [46, 47].

The Reynolds GMM system models the probability distribution of the feature vector as a sum of weighted Gaussians. For our speech source features, we have found that a mixture of 16 Gaussians gives the highest accuracy unless otherwise noted. The feature vector has one element per parameter for each frame. A fixed frame rate might be used, or each frame can correspond to a pitch period, as in this thesis. The Gaussians are n dimensional, where n is the number of features in the feature vector. The parameters of the mixture model are estimated using the Expectation Maximization (EM) algorithm.

To identify the speaker of an utterance, the frame probabilities are multiplied for each potential speaker, and the speaker model with the highest probability is considered to indicate the correct speaker. In order to determine how well the system works, we calculate the percentage of correctly identified speakers, which we call the accuracy or score of the system for a particular set of features. The task of training speaker models is called *training*, while the task of identifying the speakers of a number of sentences is called *testing*.

6.2 Difficulties with the Reynolds GMM SID System

By using a mixture of Gaussians to model the feature vectors, we are making the assumption that the feature vectors can be adequately modeled by Gaussian distributions. This tends to be a good assumption for parameters that are directly calculated from the waveform, such as mel-cepstral coefficients, but is often not a good assumption to make for waveform models as in our glottal flow derivative model. Two primary areas of difficulty have arisen, outliers and singularities.

A feature might repeatedly take on the same value, due to difficulties in modeling, poor waveform estimation, or insufficient data. For example, the end of the return phase can occur one sample after the time T_e. It is not possible to estimate a value for T_c less than one sample after T_e, despite the fact that this may occur, either due to a very rapid glottal closure or poor modeling of the vocal tract. Another example is large values of the exponential constant α. The parameter α is limited to a value of 50 by the bounds given to the NL2SOL algorithm. Values larger than this are not meaningful, as the pulse is already extremely impulse-like. Any frame for which the estimated α is greater than or equal to 50 will return a value of 50 for α. The bounds help in improving estimates of the other parameters and in ensuring convergence, but can cause problems in training.

Outliers occur when an unusual feature vector is calculated. In general, outliers

occur due to poor estimation of the glottal flow derivative or unsuccessful modeling by the NL2SOL or PROGRESS algorithms. Outlier feature vectors commonly occur when the glottal flow derivative does not resemble an LF model, and when the timing of the glottal pulse is not correctly identified, either due to incorrect pitch estimates provided to the analysis system, or errors in the initial peak picking algorithm.

Singularities can cause problems in modeling because they will tend to "grab" Gaussians, reducing the number of models available for modeling meaningful feature values. The modeling routines of the Reynolds SID system identify and remove outlier vectors, but this is not possible during testing. Any outliers that occur during testing will have unpredictable results, but in general, a single outlier vector will cause a sentence whose speaker would otherwise be identified correctly to indicate a different speaker.

To reduce the problems of singularities and outlier vectors, a significant portion of the feature vectors are discarded before training and testing. If any element of a feature vector is equal to the bounds used in the NL2SOL algorithm, the vector is discarded. Also, if the value for E_e is below a certain threshold, the frame is discarded. This restriction is placed on the data because the estimation is often poor when there is very little energy in the waveform. While discarding data might generally be undesirable, we have found that it increases the accuracy of the speaker identification system by approximately 15%. Only about one third of the actual speech is used for training and testing, since no unvoiced frames are used, and some of the voiced frames are discarded.

6.3 Using Source Features for SID

The parameters used for SID are not all the same as those used in estimating the coarse model. For example, the parameter T_0, indicating the first sample of the open phase, will grow continually larger as we move further into an utterance. This large upward trend makes T_0 essentially useless for SID. To avoid this problem, instead of using the times T_0, T_e, and T_c, we calculate the length of the return phase as $T_c - T_e$

and two open quotient-type parameters, the first is

$$OQ = \frac{T_e - T_0}{T_e - T_{e-1}}, \tag{6.1}$$

where T_{e-1} is the time of the glottal pulse for the previous pitch period. The second parameter is actually a closed quotient as determined by the closed phase identified using the sliding covariance analysis,

$$CQ = \frac{n_o - n_c}{T_e - T_{e-1}}, \tag{6.2}$$

where n_o is the time of glottal opening, and n_c is the time of glottal closure, both for the current pitch period. The waveshape parameters α, ω_g, E_e, and T_a are included as calculated during modeling, giving a total of seven coarse structure parameters.

We now present speaker identification results for a number of different tests. Male and female sets are handled separately, as the large differences in anatomy result in cross-sex errors being very rare. For the first set of experiments, the data comes from the TIMIT database, which contains 10 sentences of read speech for each speaker, recorded in a quiet-room with a Sennheizer microphone. The recording methods result in a high quality database. The male subset contains 112 speakers, while the female subset contains 56 speakers. For each speaker, eight of the sentences are used for training and two are used for two independent tests.

Tests were conducted with the following sets of features:

1. The seven LF model parameters,
2. The five energy parameters,
3. The seven LF and five energy parameters,
4. The three formant modulation parameters,
5. The 14 LPC derived cepstral coefficients, and
6. The 12 source parameters and 14 cepstral parameters.

The LPC cepstrum consists of the first 14 coefficients of the real cepstrum as

Features		Male	Female
Coarse:	7 LF	58.3%	68.2%
Fine:	5 energy	39.5%	41.8%
Source:	12 LF & energy	69.1%	73.6%
Fine:	3 FM	7.6%	16.4%
Filter:	14 LPC Cepstrum	91.0%	93.6%
Combined	LF, energy, cep	93.7%	92.6%

Table 6.1: Speaker identification results for various combinations of the source parameters

calculated by the recursion

$$c_i = -\alpha_i - \frac{1}{i}\sum_{k=1}^{i-1}(i-k)\alpha_k c_{i-k} \tag{6.3}$$

where c_i are the 14 cepstral coefficients, and c_0 is not calculated [27]. The α_k's used are the estimated vocal tract parameters calculated using the covariance method of linear prediction over the closed phase. The recursion assumes a minimum-phase filter given by the α_k's. Any maximum-phase poles are flipped inside the unit circle before the cepstral coefficients are calculated.

The results in table 6.1 clearly show that the three categories of source parameters all contain significant speaker-dependent information. The source features contain information not present in the 14 cepstral parameters which model the vocal tract, as shown by the increase to 93.7% accuracy[1] for the combination of source and vocal tract data for the male subset. Outliers caused the reduction in score when source parameters are added to the vocal tract parameters for females. The 3 FM parameters show some speaker dependence, as their scores of approximately 8% and 15% correct are well above chance (less than 1% for both cases). Including the three formant modulation parameters with the other data, however, lowered the scores significantly, due to the large number of outliers in the formant modulation data.

[1]In comparing accuracy rates near 100%, it is generally more instructive to compare the relative reduction in error rate. For the male subset, the error was reduced from 9% to 6.3%, a 30% reduction in error.

Features	Male	Female
Modeled GFD	41.1%	51.8%
GFD	95.1%	95.5%

Table 6.2: Speaker identification results for mel-cepstral representations of the Glottal Flow Derivative (GFD) waveform and the modeled GFD waveform.

As a secondary measure of information in the glottal flow derivative waveform, we calculated the mel-cepstra of these waveforms and used these 23 coefficients as the features for SID. Both the glottal flow derivative waveform and the modeled waveform were processed in this manner. The results are shown in table 6.2. We note that the 7 LF parameters shown in the first row of table 6.1 better represent the modeled glottal flow derivative than the 23 cepstral parameters. The modeled glottal flow derivative contains significantly less information than the glottal flow derivative, because modeling significantly reduces the amount of vocal tract information present.

6.4 SID for Degraded Speech

The Reynolds SID system has been shown to be 100% accurate for a large number of speakers using the high quality speech in the TIMIT database. The true value of source information for SID is thus in improving recognition scores for degraded speech. To test how well the source information works on degraded speech, we use a subset of 20 male speakers and a subset of 20 female speakers from the NTIMIT database[28]. The NTIMIT database is the TIMIT database transmitted through a telephone handset and over long-distance phone lines. The 20 male and 20 female speakers used are speakers that the Reynolds SID system performs poorly on. For these tests, we use the mel-cepstral representation of the speech signal and the mel-cepstral representation of the source waveform. Results are shown in table 6.3.

For the tests in which speech and source mel-cepstral data were combined, the feature vectors were merged prior to training and testing. The mel-cepstral feature vectors each contain 23 parameters. The best results were achieved by training on a

Features	Male	Female
Speech	40.0%	52.5%
GFD	25.0%	22.5%
modeled GFD	12.5%	27.5%
Speech & GFD	50.0%	50.0%
Speech & modeled GFD	45.0%	47.5%
Speech & GFD with 32 Gaussians	57.5%	52.5%
Speech & modeled GFD with 32 Gaussians	60.0%	55.0%

Table 6.3: Speaker identification results for mel-cepstral representations of the speech signal, the Glottal Flow Derivative (GFD) waveform, the modeled GFD waveform, and combinations of the speech and source mel-cepstral data. All of the data was generated from the NTIMIT database.

23 element vector rather than the 46 element vector that would result by combining the two vectors into one. Using this approach, each feature vector contains either speech or source information. We believe that the training routines assigned some of the 16 Gaussians to model the speech signal, while some were used to model the source signal. A further increase in accuracy was achieved by increasing the number of Gaussians to 32. Making this change provides sufficient Gaussians to model each of the two sets of feature vectors.

The scores for the source information calculated from the NTIMIT database are significantly worse than those calculated using the TIMIT database. In an attempt to demonstrate the potential increase in accuracy that could be achieved by improving the estimation of the source in degraded environments, we repeat the above procedure using the NTIMIT speech waveforms and modeled glottal flow waveforms calculated from TIMIT data. Using 32 Gaussians, we achieved scores of 60% for the males and 57.5% for the females. One possible explanation for the small increase in scores is that the TIMIT and NTIMIT data differ enough that it is more difficult to model the two streams in combination.

6.5 Summary

The speaker identification tests show that the models of the coarse and fine structure of the glottal flow derivative contain significant speaker dependent information, and that the algorithms used to estimate parameters of these models work in a robust manner. Difficulties were encountered in modeling the parameters with a mixture of Gaussians, which results in lower scores than if these problems were solved, and restricts combinations of parameters with other data. Combining mel-cepstral representations of the source and speech waveforms transmitted over telephone lines demonstrates that the source information can be used to improve on current speaker identification systems.

Chapter 7

Conclusions

7.1 Summary of Findings

The goal of this thesis was to develop automatic techniques for reliably estimating and modeling the glottal flow derivative waveform from speech, and to determine the importance of the glottal flow derivative for speaker identification. The volume velocity airflow through the glottis, called the *glottal flow*, is the source for voiced speech. Incorporating the radiation effect into the source gives the glottal flow derivative waveform. Previous studies have shown the importance of the glottal flow for improving naturalness in speech synthesizers and for use as correlates to voice types such as loud, angry, breathy, etc.

There is strong evidence to suggest that the glottal flow should be speaker dependent. Videos of vocal fold vibration demonstrate a wide variety of vocal fold vibration patterns for different speakers. Some folds open or close in a zipper-like fashion, while others open or close along their entire length at the same time. For some speakers the glottis never closes, while for others a long period of complete closure is observed. The glottal flow derivative is determined by a nonlinear function of the glottal area, so information about the vocal fold vibration for a speaker will therefore be included in the glottal flow derivative. By modeling the coarse structure of the glottal flow derivative waveform, we expect to capture some of this speaker dependent information.

The glottal flow derivative will also contain vocal tract information, as seen through the theory of source-filter interaction presented in chapter 2. In order to determine the importance of the source for speaker identification, vocal tract information must be removed from the model parameters. Some of the source-filter interaction will result in information that should be included with source information. For example, the formant bandwidth and frequency modulation were shown to be dependent on the glottal area and its derivative. Similarly, the amount of fine structure during the closed phase as compared to the open phase will indicate how completely a speaker closes his or her glottis.

The glottal flow derivative was estimated using an inverse filter estimated during the closed phase. The timing of glottal closure and glottal opening were determined by formant frequency modulation calculated using a sliding covariance analysis with a one sample shift. A statistical technique was used to identify the glottal closure and opening. The statistical technique eliminates dependence on a particular type of frequency modulation, and allows the algorithm to adapt to the amount of formant modulation during the closed phase, which is dependent on the degree of glottal closure. A two-window covariance technique was developed to improve time resolution for high pitch speakers.

The LF model for the glottal flow derivative was used to model the coarse structure of the glottal flow. As modified for use in this thesis, it is a seven parameter model, with distinct closed, open, and return phases. The shape of the open phase is determined by two parameters, while one is used to control an exponential return phase. The parameters of this model were determined for each pitch period using the NL2SOL algorithm for nonlinear least-squares regression.

The fine structure of the glottal flow was modeled through the frequency modulation of the first formant and five energy measures. The energy measures estimate the amount of aspiration noise and ripple during various periods. The relationship between the energy of the fine structure during the open phase to the energy during the closed phase will indicate how completely the glottis closes. The frequency modulation of the first formant was modeled by a parabola using the PROGRESS

algorithm for robust least-squares regression. While a parabola is not a theoretically meaningful measure of frequency modulation, we have used it to demonstrate that the frequency modulation can be reasonably measured and is speaker dependent, as predicted by its dependence on the derivative of the glottal area.

The speaker identification experiments show that the algorithms described in this paper work reasonably, and that the source parameters estimated contain significant speaker dependent information. To achieve high SID scores, the source must contain speaker dependent information, the source must be modeled in a manner that captures this information, and the estimation must determine the parameters of these models in a consistent manner. If the algorithms were not consistent, the parameters calculated would not contain statistically significant information. Without a separate source of parameter values to compare against, the speaker identification experiments are a good indication that the parameters are estimated in a meaningful manner.

All aspects of the source model have been shown to contain speaker dependent information. The coarse structure parameters contain the most information, the time domain energy measures of fine structure less information, and the frequency modulation of the first formant contains the least amount of speaker dependent information, though still resulting in speaker identification scores well above chance. Difficulties in modeling the source parameters using a Gaussian Mixture Model likely reduced the scores. Integrating the source parameters with more traditional measures also proved difficult. By combining mel-cepstral representations of the speech waveform and the glottal flow derivative waveform, we demonstrated that source features can be used to improve the scores of current speaker identification systems in degraded environments.

7.2 Suggestions for Future Work

The results of this thesis demonstrate that the source for voiced speech is important in speaker identification. Many areas of this thesis require further study, to improve upon the results achieved so far, to show more conclusively that the source and vocal

tract can be separated and that they contain independent information, and to extend the analysis techniques to non-ideal environments.

Source Modeling

There are several ways in which the glottal flow derivative could be modeled more completely. In terms of the fine structure, the aspiration and ripple components could be separated, perhaps using a noise/harmonic model. Informal experiments have shown this is a difficult separation to achieve, in part because the ripple has a wide bandwidth due to the nonlinear manner in which it is created, and in part because the specifics of superposition will cause the ripple to vary from pitch period to pitch period. Also, the time evolution of aspiration noise and ripple could be modeled. It has been shown [25] through perceptual evaluation of synthetic speech that the timing of noise is important to its perceived integration with the speech as aspiration noise, rather than as a separate noise source.

The formant modulation due to ripple could be modeled more theoretically than with a simple parabola fit to the frequency modulation. This would include an improved model for the formant frequency modulation and the addition of a model for the formant bandwidth modulation. The ripple is determined by the supra-glottal pressure variations, the glottal area, and the derivative of the glottal-area. How these three physical features are best modeled through the time- and frequency-domain representations of ripple needs to be better understood from both a theoretical and empirical standpoint. It also remains to be seen whether higher formants would yield additional information or if the ripple due to higher formants contains redundant information (which could possibly be used to calculate a more robust estimate of the ripple).

The temporal change of the glottal flow derivative waveform is not included in our SID experiments. Changes in the glottal flow from period to period will indicate when glottal stops are used, how sudden the onset of voicing is, and the inter-period variability of the vocal fold vibration. Asymmetries in the vocal folds will result in a less stable pattern of vibration [52], which would be captured through the temporal

change of the glottal flow derivative.

Parameter Estimation and Statistical Modeling

There are several improvements to be made in the modeling algorithms for both estimating the glottal flow derivative parameters and using these parameters for SID. A nonlinear least-squares algorithm that is designed to handle piecewise functions should enable more accurate estimation of the times T_0, T_e, and T_c, which should in turn result in better estimates of the remaining parameters. Similarly, a statistical model that is better suited to the waveform models used should result in higher SID scores. Integration with traditional vocal tract measures is also a problem. The source data is calculated on a pitch period basis, while mel-cepstra are traditionally calculated using a fixed window size and shift. Even when the vocal tract information is calculated on a pitch period basis, integration of source and filter information may not increase scores, due to modeling problems, as demonstrated by the drop in female ID rate from 93.6% to 92.7% when source information was added to LPC derived cepstral coefficients.

It is difficult to demonstrate conclusively that we have achieved a nearly complete separation of the source and filter in our estimation procedure. One indication of both the importance of source information and the separation of vocal tract information is that SID scores for lpc-cepstrum parameters calculated during the closed phase are lower than for lpc-cepstrum parameters calculated during the open and closed phase. The lower scores indicate that information that is being captured during the entire pitch period is not present during the closed phase. We attribute this information to the source, which is not being modeled by the vocal tract estimates calculated during the closed phase.

The ripple present in the glottal flow derivative waveform will contain significant formant information. For this reason, the ripple must be modeled in a way that excludes vocal tract information. Pulse skew, due to loading of the flow by the vocal tract, will also contain vocal tract information, although this has not been mentioned previously. Fant states in [13] that a narrow pharynx will result in a

higher characteristic impedance of the back end of the vocal tract, which will result in more loading of the glottal flow than a wide pharynx. The amount of pulse skew will therefore contain information about the diameter of the pharynx. Pulse skew will also be determined by the speed of glottal closure, so a technique is needed to separate these two causes of pulse skew, in order to identify the portion that is independent of the vocal tract.

Estimation and Modeling of Degraded Speech

The problem of estimation and modeling in degraded environments, such as telephone speech, has not been carefully studied. The current algorithms do not work well on the NTIMIT database. Despite this difficulty, we have found that combining speech and source information for the NTIMIT database can significantly increase SID scores. We feel that the key difficulty in estimating the glottal flow derivative from degraded speech is that the current algorithm is a time-domain algorithm, which requires phase coherence. Simple phase compensation experiments gave promising results, but more detailed study is needed. By estimating a linear filter to approximate the difference between the TIMIT and NTIMIT form of an utterance, and inverse filtering the degraded speech with this filter, we were able to achieve glottal flow derivative waveforms that looked more similar to the waveforms estimated from TIMIT speech. A frequency domain approach to modeling the glottal flow may give a solution to the problem of estimation for degraded speech. Another potential solution is to separately estimate closed phases for different formants, and estimate each formant during the time when it indicates the glottis is closed.

Multiple Pulses and Other Unusual Examples

The examples presented in this thesis were chosen to illustrate certain points, but are still typical examples. We now present four examples of more unusual situations. Figure 7-1 shows these four examples. The small pulses superimposed on the glottal flow derivative waveforms show the times of glottal closure and opening as identified

through the modulation for the first formant, while the large pulses represent the initial estimates of the time of the glottal pulse.

The first two examples show multiple points of excitation. The secondary excitation is particularly easy to see in the speech signal for the second example, as the second formant is clearly re-excited at the second pulse. We have found that such multiple pulses occur primarily for speakers with very complete glottal closure. We propose that these pulses result from the large energy in the vocal tract suddenly being dissipated through the glottis. As the glottis opens, the vocal tract filter changes, and the formants are re-excited as demonstrated in figure 3-6 of section 3.2.2. This reasoning for secondary pulses may in part explain the improved SID scores achieved by measuring energy onset times in formant bands using the Teager operator as seen in [44].

The third example shows a case of very large ripple. The fourth example illustrates that the closed phase is occasionally identified incorrectly. The closed phase following the more impulse-like pulse is seen to have the glottal opening estimate occur later than it should. This results in some source information being included in the vocal tract filter, and thus a more white glottal pulse after inverse filtering.

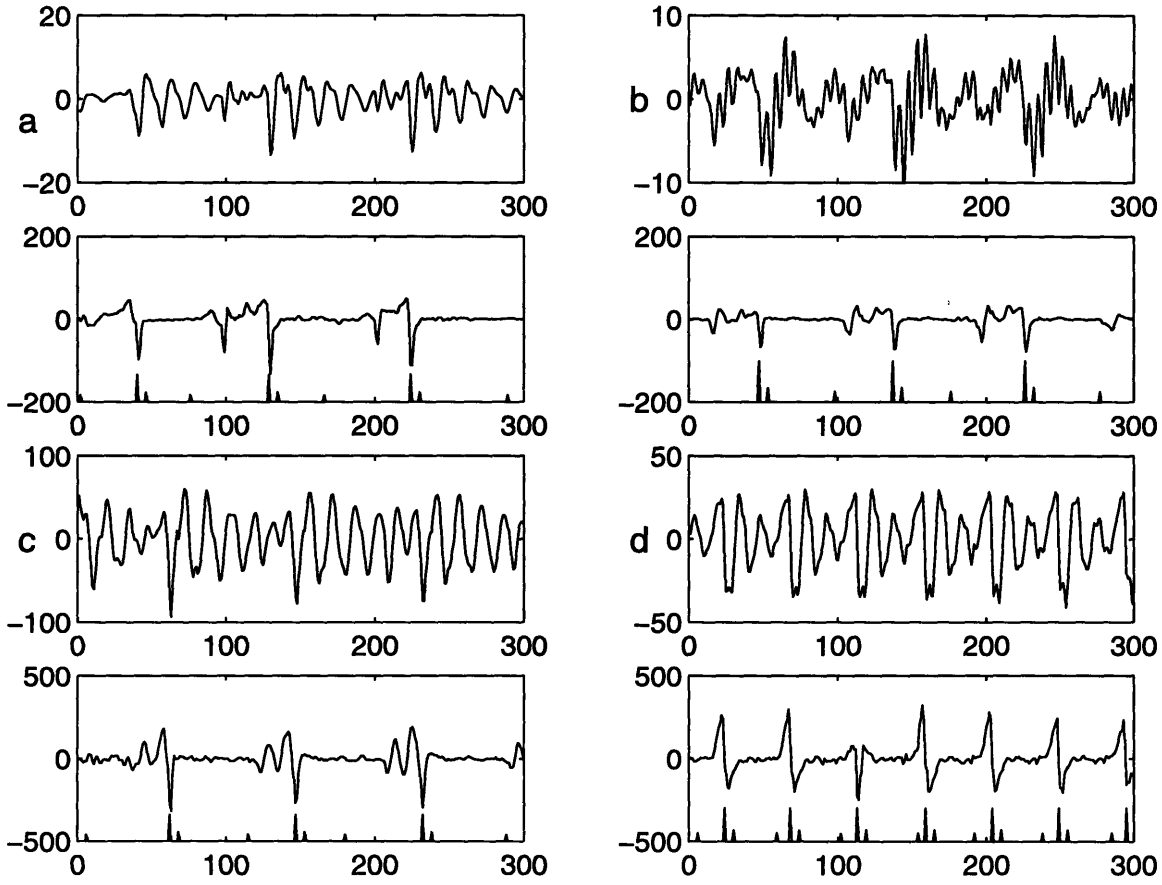

Figure 7-1: Four examples of unusual glottal flow derivatives. In each case, the speech waveform is shown above the glottal flow derivative waveform. Superimposed on the glottal flow derivative waveform are small pulses indicating the timing of the glottal opening, closure, and pulse. Panels (a) and (b) show evidence of multiple pulses, likely due to the sudden onset of ripple. Panel (c) shows a case of a large amount of ripple, while panel (d) shows an error in identification of the closed phase and the resultant incorrect glottal flow derivative waveform.

Bibliography

[1] T. V. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Communications*, pages 167–184, 1982.

[2] Tirupattur V. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):309–319, August 1979.

[3] Mats Båvegård and Gunnar Fant. Notes on glottal source interaction ripple. *STL-QPSR*, (4):63–77, 1994.

[4] D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acousitc Society of America*, 90(5):2394–2410, November 1991.

[5] D.G. Childers and Chieteuk Ahn. Modeling the glottal volume-velocity waveform for three voice types. *Journal of the Acoustic Society of America*, 97(1):505–519, January 1995.

[6] D.G. Childers, J.C. Principe, and Y.T. Ting. Adaptive WRLS-VFF for speech analysis. *IEEE transactions on speech and audio processing*, 3(3):209–2113, May 1995.

[7] Donald G. Childers and Chun-Fan Wong. Measuring and modeling vocal source-tract interaction. *IEEE Transactions on Biomendial Engineering*, 41(7):663–671, July 1994.

[8] Kathleen E. Cummings and Mark A. Clements. Analysis of glottal waveforms across stress styles. In *ICASSP*, pages 369–372, 1990.

[9] Kathleen E. Cummings and Mark A. Clements. Glottal models for digital speech processing: A historial survey and new results. *Digital Signal Processing*, 5:21–42, 1995.

[10] G. Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

[11] G. Fant and T.V. Ananthapadmanabha. Truncation and superposition. *STL-QPSR*, (2-3):1–17, 1982.

[12] G. Fant and J. Liljencrants. Perception of vowels with truncatied intraperiod decay envelopes. *STL-QPSR*, (1):79–84, 1979.

[13] Gunnar Fant. Temporal fine structure of formant damping and excitation. In *Proceedings of the ASA*, pages 161–165, 1979.

[14] Gunnar Fant. Glottal flow: models and interaction. *Journal of Phonetics*, 14:393–399, 1986.

[15] Gunnar Fant. Some problems in voice source analysis. *Speech Communication*, 13:7–22, 1993.

[16] Gunnar Fant, Qi guang Lin, and Christer Gobl. Notes on glottal flow interaction. *STL-QPSR*, (2-3):21–45, 1985.

[17] Gunnar Fant and Qiguang Lin. Glottal source – vocal tract acoustic interaction. *STL-QPSR*, (1):13–27, 1987.

[18] Gunnar Fant and Qiguang Lin. Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR*, (2-3):1–21, 1988.

[19] Hiroya Fujisaki and Mats Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *ICASSP*, pages 1605–1608, 1986.

[20] Hiroya Fukisaki and Mats Ljungqvist. Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. In *ICASSP*, pages 637–640, 1987.

[21] J. Gauffin, N. Binh, T. V. Ananthapadmanabha, and G. Fant. Glottal geometry and volume velocity waveform. In *Proc. Research Conf. Voice Physiology*, 1981.

[22] Christer Gobl. Voice source dynamics in connected speech. *STL-QPSR*, (1):123–159, 1988.

[23] Helen Hanson. *Glottal Characteristics of Female Speakers.* PhD thesis, Harvard University, May 1995.

[24] Jialong He, Li Liu, and Günther Palm. On the use of features from prediction residual signals in speaker identification. In *EUROSPEECH*, pages 313–316, 1995.

[25] Dik J. Hermes. Synthesis of breathy vowels: Some research methods. *Speech Communication*, 10(5-6):497–502, December 1991.

[26] Eva B. Holmberg, Robert E. Hillman, and Joseph S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustic Society of America*, 84(2):511–529, August 1988.

[27] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[28] Charles R. Jankowski, Ashok Kalyanswamy, Sara Basson, and Judity Spitz. Ntimit: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *ICASSP*, pages 109–112, Albuquerque, 1990. IEEE.

[29] Charles R. Jankowski Jr. *Fine Structure Features for Speaker Identification.* PhD thesis, Massachusetts Institute of Technology, 1996.

[30] John E. Dennis Jr., David M. Gay, and Roy E. Welsch. An adaptive nonlinear least-squares algorith. *ACM Transactions on Mathematical Software*, 7(3):348–368, September 1981.

[31] John E. Dennis Jr., David M. Gay, and Roy E. Welsch. Algorithm 573 NL2SOL–an adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7(3):369–383, September 1981.

[32] Dennis H. Klatt and Laura C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustic Society of America*, 87(2):820–857, February 1990.

[33] Ashok K. Krishnamurthy and Donald G. Childers. Two-channel speech analysis. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, 34(4):730–743, August 1986.

[34] Bell Telephone Laboratories. High speed motion pictures of the human vocal cords. Motion Picture.

[35] Qiguang Lin. Nonlinear interaction in voice production. *STL-QPSR*, (1):1–12, 1987.

[36] Changxue Ma, Yves Kamp, and Lei F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–264, April 1994.

[37] John Makhoul. Linear prediction: A tutorial review. In *Proceedings of the IEEE*, volume 63, pages 561–580, April 1975.

[38] R. J. McAulay and T. F. Quatieri. Pitch estimation and voicing detection based on a sinusoidal model. In *ICASSP*, pages 249–252, 1990.

[39] Yoshiaki Miyoshi, Kazuharu Yamato, Riichiro Mizoguchi, Masuzo Yanagida, and Osamu Kakusho. Analysis of speech signals of short pitch period by a sample-selective linear prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(9):1233–1239, September 1987.

[40] Randall B. Monsen and A. Maynard Engebretson. Study of variations in the male and female glottal wave. *Journal of the Acoustic Society of America*, 62(4):981–993, October 1977.

[41] Burhan F. Necioğlu, Mark A. Clements, and Thomas P. Barnwell III. Objectively measured descriptors applied to speaker characterization. In *ICASSP*, pages 483–486, 1996.

[42] Janet B. Pierrehumbert. A preliminary study of the consequences of intonation for the voice source. *STL-QPSR*, (4):23–36, 1989.

[43] William H. Press, Sual A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: the Art of Scientific Computing.* Cambridge University Press, 1992.

[44] 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.

[45] Lawrence R. Rabiner and Ronald W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Inc., 1978.

[46] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, August 1995.

[47] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions of Speech and Audio Processing*, 3:72–83, January 1995.

[48] Peter J. Rousseeuw and Annick M. Leroy. *Robust regression and outlier detection.* John Wiley & Sons, 1987.

[49] K. N. Stevens. *Introduction to Communication Sciences and Disorders*, chapter Scientific substrates of speech production. Singular, San Diego, 1994.

[50] H. Strik and L. Boves. On the relation between voice source parameter and prosodic features in connected speech. *Speech Communication*, 11:167–174, 1992.

[51] Philipee Thévenaz and Heinz Hügli. Usefulness of the LPC-residue in text-independent speaker verification. *Speech Communication*, 17(1-2):145–157, August 1995.

[52] Ingo Titze. What's in a voice. *New Scientist*, pages 38–42, 23 September 1995.

[53] Robert L. Whitehad, Dale E. Metz, and Brenda H. Whitehead. Vibratory patterns of the vocal folds during pulse register phonation. *Journal of the Acoustic Society of America*, 75(4):1293–1297, April 1984.

[54] David Y. Wong, John D. Markel, and Augustine H. Gray Jr. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):350–355, August 1979.

[55] B. Yegnanarayana and P. Satyanarayana Murthy. Source-system windowing for speech analysis and synthesis. *IEEE Transactions on Speech and Audio Processing*, 4(2):133–137, March 1996.