Analysis of Nonmodal Glottal Event Patterns with Application to Automatic Speaker Recognition

by

Nicolas Malyska

B.S., Electrical Engineering
B.S., Computer Engineering
University of Florida, 2000

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2004

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN HEALTH SCIENCE AND TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2008

Signature of Author ....
Harvard-MIT Division of Health Sciences and Technology
December 7, 2007

Certified by ....
Thomas F. Quatieri
Senior Member of Technical Staff; MIT Lincoln Laboratory
Faculty of MIT Speech and Hearing Bioscience and Technology Program
Thesis Supervisor

Accepted by ....
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

This work was sponsored by the Air Force Research Laboratory, under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Analysis of Nonmodal Glottal Event Patterns with Application to Automatic Speaker Recognition

by

Nicolas Malyska

Submitted to the Division of Health Sciences and Technology
on December 7, 2007 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Health Science and Technology

Abstract

Regions of phonation exhibiting nonmodal characteristics are likely to contain information about speaker identity, language, dialect, and vocal-fold health. As a basis for testing such dependencies, we develop a representation of patterns in the relative timing and height of nonmodal glottal pulses. To extract the timing and height of candidate pulses, we investigate a variety of inverse-filtering schemes including maximum-entropy deconvolution that minimizes predictability of a signal and minimum-entropy deconvolution that maximizes pulse-likeness. Hybrid formulations of these methods are also considered.

We then derive a theoretical framework for understanding frequency- and time-domain properties of a pulse sequence, a process that sheds light on the transformation of nonmodal pulse trains into useful parameters. In the frequency domain, we introduce the first comprehensive mathematical derivation of the effect of deterministic and stochastic source perturbation on the short-time spectrum. We also propose a pitch representation of nonmodality that provides an alternative viewpoint on the frequency content that does not rely on Fourier bases. In developing time-domain properties, we use projected low-dimensional histograms of feature vectors derived from pulse timing and height parameters. For these features, we have found clusters of distinct pulse patterns, reflecting a wide variety of glottal-pulse phenomena including near-modal phonation, shimmer and jitter, diplophonia and triplophonia, and aperiodicity. Using temporal relationships between successive feature vectors, an algorithm by which to separate these different classes of glottal-pulse characteristics has also been developed.

We have used our glottal-pulse-pattern representation to automatically test for one signal dependency: speaker dependence of glottal-pulse sequences. This choice is motivated by differences observed between talkers in our separated feature space. Using an automatic speaker verification experiment, we investigate tradeoffs in speaker dependency for short-time pulse patterns, reflecting local irregularity, as well as long-time patterns related to higher-level cyclic variations. Results, using speakers with a broad array of modal and nonmodal behaviors, indicate a high accuracy in speaker recognition performance, complementary to the use of conventional mel-cepstral features. These results suggest that there is rich structure to the source excitation that provides information about a particular speaker's identity.

Thesis Supervisor: Thomas F. Quatieri

Title: Senior Member of Technical Staff; MIT Lincoln Laboratory

Faculty of MIT Speech and Hearing Bioscience and Technology Program

Acknowledgement

A large number of people helped me on my journey towards researching and writing this thesis. First I would like to thank my advisor, Tom, for all of his encouragement, insight, and guidance. I feel as if I have just returned from a trek through the jungle, dirty and insect-ravished with my boots wet, but with plenty of stories to tell. Tom is a great advisor for many reasons, but I think I learned the most from the way he always pushed me to navigate some of the windiest paths, and for this I am grateful. Tom has taught me that many "long solved" research problems have aspects that deserve further thought and innovation. He has a knack for listening to a student describe a large number of topics that they have been working on, and selecting one or two ideas that are likely to lead to success. Thanks Tom for a great five years; I hope that with students and with others I advise, I can be as patient and supportive as you have been.

I would also like to thank my committee, Lou Braida, Stefanie Shattuck-Hufnagel, and TJ Hazen, for its help and encouragement. I have been thrilled by the number of insightful questions, good ideas, and suggestions for future work that I have received. Lou has supported me throughout my entire graduate-school journey as a professor, mentor, and colleague, and I appreciated his guidance as my committee chair. Thanks to Stefanie for long discussions about how we might apply my work to prosody research and other areas. Her wise words have made me confident that there is a future for my research in solving real-world speech-science problems. Many thanks to TJ, who provided a much-appreciated fresh perspective on my methods and how to make them relevant for the speech-technology community. TJ also provided very detailed notes on my thesis draft and defense presentation with numerous good suggestions, helping me immensely during the final few weeks of the process. I look forward to working with him as a colleague in my future at MIT Lincoln Laboratory.

Thanks very much to Group 62 at Lincoln Laboratory for allowing me to pursue my research and to follow the many paths that it took. I would like to especially thank Cliff Weinstein, our group leader, who has built an environment in this group where I feel free to research what intrigues me and still be part of the team. The entire Group 62 team, and especially Doug Sturim, Bob Dunn, Mike Brandstein, Wade Shen, Robert Granville, Doug Reynolds, Kevin Brady, Bill Campbell, and Joe Campbell have made themselves available as friends and colleagues from the beginning, with doors open, always ready for an interesting discussion. Thanks to all of you, and I am ecstatic to have the opportunity to continue to work with you in the coming years.

To my student labmates, Daryush Mehta, Tian Wang, Nancy Chen, Zahi Karam, Dan Rudoy, and Barry Jacobson, I give great thanks and best wishes! I am very happy that I am not required to list every great insight, idea, or question that these colleagues have contributed to this thesis. From discussions about time-varying filtering, to a refresher course on the physiology of the ear, and afternoons thinking through the physics of snowboarding, I think we have tackled many important topics with wide-reaching implications. I look forward to watching each of you continue towards reaching your goals.

There are many other people that I would like to thank. Janet Slifka was very kind for allowing us access to her data containing nonmodal regions, which provided much of the inspiration for this work. Also, I have enjoyed working with Dr. Robert Hillman, who has provided me with a clinical perspective and has helped me see how my work can be helpful to the medical

community. The MIT Voice-Quality Study Group as a whole also deserves my thanks, as they have helped cement voice-quality research as a growing field that takes minds from many different disciplines to understand.

Finally, I would like to very enthusiastically thank my family, who has been very supportive of this effort. I especially appreciate my father's subtle encouragement and my mother's steadfast confidence in me. Thanks also to my brother, William, who always seems to call me just when I need a study break. My wonderful girlfriend Dani has made many sacrifices of time and hugs during the thesis-writing process, and to her I give my love, my thanks, and the majority of my newly rediscovered free time. Also, getting to this point would have been much less fun without my other "family," Mike, Matt, Kelly, and Brian, who have seen my life evolve since our first lunch together in the freshman year of high school. May we together continue to be irregular, impulsive excitation functions.

Remembering Gram and Brian:

Wish you were here to share this with me

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The distinct opening and closing gesture of the vocal folds and its effect on measured speech signals has been a motivator of research in many aspects of speech production. It is the fundamental repeating unit that researchers associate with the percept of voice pitch, it is a major landmark in visual representations like spectrograms, and its properties are thought to indicate voice qualities like breathiness and roughness. Although the details are not well agreed upon, each opening and closing cycle of the glottis, or *glottal cycle*, provides a burst of excitation to the remainder of the speech-production system. At this point, we will be intentionally vague about what exactly is this excitation—later, we will describe in more detail the representation that we use in this thesis. As we shall discuss, the glottal cycle can be viewed as contributing a series of discrete excitation impulses, called *glottal events*, to the speech-production system.

In this thesis, we focus on relationships between neighboring glottal-events. Phonation where these events are not evenly spaced or do not have identical heights—elements of so-called *nonmodal* phonation (see Figure 1-1)—remains an area of intense study. It is a domain where researchers active in linguistics [22, 59], voice perception [20], voice physiology [24, 27, 33, 53], and signal processing [46, 56] participate. Our approach is to create a set of parameters describing series of glottal events in general, without attempting to describe the patterns in terms of currently used, and often conflicting, definitions. Although the specific application focus is automatic speaker recognition with normal speakers, the intent is to form a technology that can be adapted to many other applications. To this end, we explore the theory governing the analysis of speech generated using a nonmodal impulse excitation model.

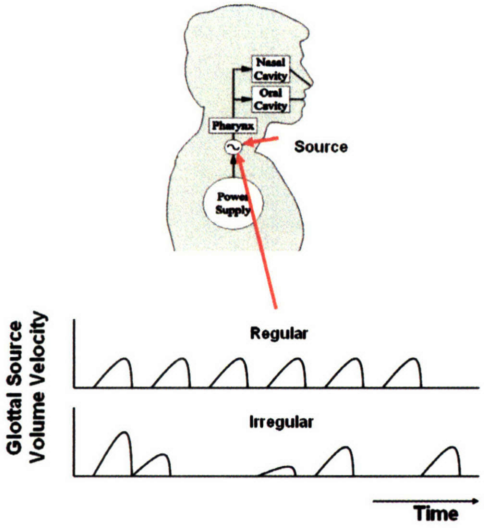


Figure 1-1. Illustration of regular and nonmodal phonation originating at the glottal source.

While the idea of extracting instances of glottal excitation is not new, an algorithm has not been developed specifically to accomplish this task for nonmodal speech. Current algorithms for characterizing phonation often assume an underlying fundamental frequency as well as pulse amplitudes and pulse shapes that are similar between neighboring glottal cycles. We propose to develop an algorithm that avoids these assumptions and is instead able to capture arbitrary patterns.

1.1 Problem

Models of phonatory excitation as a series of impulses have been pursued for much of the history of speech processing. Typically, these representations have focused on a *near-periodic* impulsive source. The excitation mechanisms of real speech, however, are likely better represented in general as impulses exhibiting non-uniform spacing and amplitudes.

The underlying causes of these non-uniform impulse-like excitations are varied and not well defined. At the level of the individual glottal cycle, multiple nonuniform impulses may represent events such as the opening and closing of the glottis. At a higher level, glottal cycles with nonuniform amplitudes and timings are known to be common in both normal and disordered speech and potentially may provide information about speaker identity, language, and dialect, and the state of the glottal source.

The motivation for our work is that nonmodal glottal events occur often in speech but are not well understood. Additionally, a better representation of nonmodal excitation has applications in many areas including automatic speaker, language, dialect, and voice-disorder recognition where nonmodality is not currently exploited. It is also important in speech analysis/synthesis, modification, and coding where these sorts of behaviors challenge current speech signal processing systems.

Nonmodal phenomena in the source include both random pulse variations, such as those found in jitter (timing variation) and shimmer (amplitude variation) as well as deterministic variations exhibiting distinct patterns in their timings and amplitudes. A well-known example of such variation from glottal cycle to glottal cycle is diplophonia, characterized by a repeating pattern of alternating large and small amplitudes and/or periods. Figure 1-2 shows examples for several of the different timing and amplitude patterns that underlie different phonation behaviors.

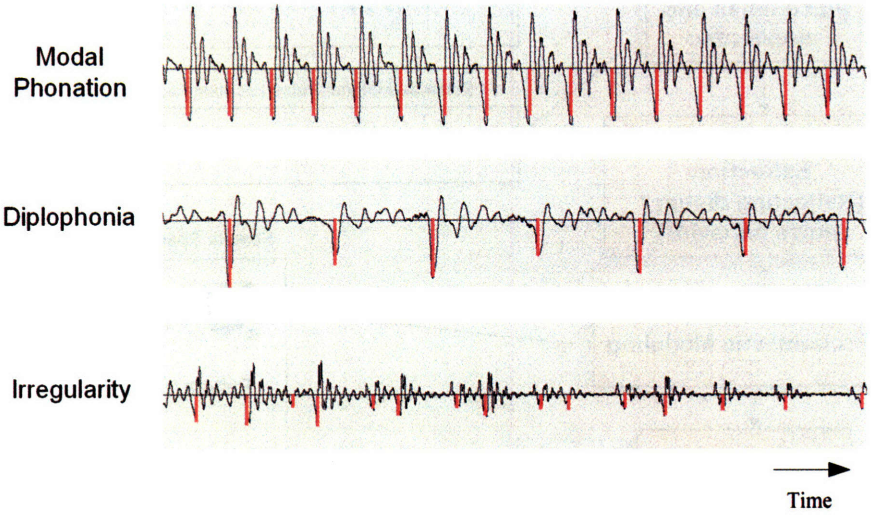


Figure 1-2. Illustration of different timing and amplitude patterns for different phonation behaviors.

Exploiting changes from one glottal cycle to the next is not a new concept. In clinical voice analysis, for example, shimmer and jitter have been used to characterize changes in the heights and amplitudes from period to period. As discussed by Titze [72], these measures can be thought of as operations on an underlying *perturbation function*, which characterizes deviations from each preceding pitch period as a function of time. Such methods are designed primarily to work with small variations rather than with larger perturbations. In a recent comparison of several different automatic perturbation-analysis techniques, for example, measures of jitter were found to become more unreliable for highly-aperiodic stimuli [58]. Additionally, such measures are intended to analyze random patterns in the voice rather than deterministic patterns like diplophonia and triplophonia. Information about period-to-period structure in most perturbation measures is averaged out [46].

1.2 Approach

The objectives of this thesis are illustrated in Figure 1-3. Our goal is to develop an automatic feature-extraction method that is able to characterize a wide range of glottal-excitation phenomena including near-modal phonation, shimmer and jitter, diplophonia, and aperiodicity. In order to do this, we find ratios of neighboring pitch periods and amplitudes, the approach being to characterize differences in neighboring patterns of pulses. When analyzed with our feature-extraction methods, each particular kind of phonation creates a distinct cluster. Our methods capture *both* random and deterministic irregularities that are found in the speech signal. Additionally, our approach allows these different kinds of phenomena to be separated from one another if desired. Our contributions are a general framework for analyzing nonmodal regions in speech and the techniques needed to visualize the data or provide it to automatic systems for efficient processing.

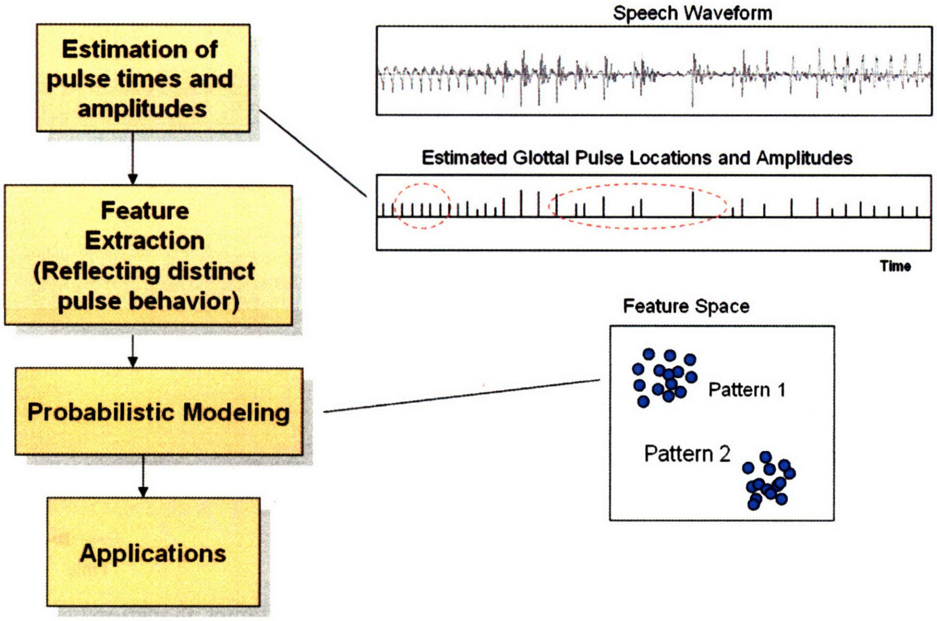

Estimation of pulse times and amplitudes
Feature Extraction (Reflecting distinct pulse behavior)
Probabilistic Modeling
Applications
Speech Waveform
Estimated Glottal Pulse Locations and Amplitudes
Time
Feature Space
Pattern 1
Pattern 2

Figure 1-3. Block diagram of the goals of this thesis.

In order to extract glottal events from the acoustic signal, we use several different inverse-filtering techniques that are extensions to linear prediction. These techniques model the acoustic signal as the result of an impulse-like excitation signal exciting a system consisting of the combination of the glottal-pulse shape, vocal-tract filtering, and radiation characteristic.

We have used our representation to automatically test for one signal dependency: the speaker dependence of irregular pulse patterns. We present experiments using clean speech databases and our pulse-pattern features which are able to achieve promising speaker-recognition results.

1.3 Outline and Thesis Summary

The thesis begins in Chapter 2 with a discussion clarifying the concept of discrete instants of glottal excitation, which we call *glottal events*. Here, we define glottal events and provide several motivating examples from real speech showing that the speech excitation can be abstracted into a series of events. We then discuss scales at which events can be seen including within individual glottal cycles as well as between different cycles, and we motivate the idea that these events are different between speakers. Finally, we describe how glottal events fit into the *source-filter* representation of speech, a viewpoint that we will use throughout the thesis.

This discussion of glottal events is followed by three chapters describing the theory behind the analysis of nonmodal glottal events. Chapter 3 investigates the spectral representation of series of nonmodal events, which relate to conventional spectral-based analysis approaches such as the mel-cepstrum that are used later in the thesis. It also gives insight into short-time spectral measurement throughout the thesis. Chapter 4 discusses the notion of time-varying fundamental-frequency as it relates to nonmodal sequences of events, proving that the pitch underlying a given excitation sequence is not unique. This is an alternate perspective in the frequency domain to the short-time Fourier transform view of Chapter 3. Chapter 5 then investigates the time-domain representation of sequences of impulses that *does* uniquely define an excitation sequence. Here, we will also describe an algorithm by which different patterns underlying an observed sequence may be automatically identified in a waveform.

Applying the theory described in the first chapters of the thesis to practical speech problems depends on our ability to extract glottal events from natural utterances. Chapter 6 addresses this problem, describing the application of three inverse-filtering approaches, linear prediction, minimum-entropy deconvolution, and a hybrid of the two, to glottal-event extraction. Through a series of experiments using both natural and synthetic speech, we present evidence that our methods are able to reliably extract meaningful glottal events.

Built upon the techniques and findings of the earlier chapters, Chapter 7 is concerned with the speaker dependence of nonmodal phonation, concentrating on automatic speaker recognition. We present evidence that our glottal-event-pattern features are able to extract significant speaker-dependent information from real speech data. We also show that the performance of the best-performing methods is not significantly degraded between sessions, and that our features provide complimentary information to the spectrally-based mel-cepstrum.

The thesis includes four appendices. Appendix A is a discussion of voice-quality measurements using our techniques that is aimed at a non-specialist audience including linguists, clinicians, and speech scientists. The goal is to distill the technologies that we have developed into their practical importance to these communities. We briefly discuss the general problem of voice-quality measurement and use a series of examples to show how the tools in this thesis can improve upon current practices. Appendix B describes the implementation details of the glottal-event extraction algorithms including block diagrams and code segments to complement the description in Chapter 6. Appendix C includes the raw data for the speaker-recognition experiments presented in this thesis in order to aid future researchers in replicating and interpreting the results. Finally, Appendix D lists the specific patterns used to study the time-domain glottal event features in Chapter 5.

Chapter 2

Glottal Events

In this chapter, we clarify the concept that the speech-excitation function may be abstracted as occurring at discrete instants of time. In Section 2.1, we begin by providing motivation for this viewpoint using definitions from the literature and both real and synthetic speech examples. Then in Section 2.2, we provide a summary of work in the literature describing the importance of particular patterns of the discrete excitation points. Finally in Section 2.3, we describe the impulse-excited representation of speech that will be used in the remainder of this thesis, relating it to our motivating examples.

2.1 Motivation

The motivation of this chapter is that the excitation function for phonation can be modeled as a series of discrete events. Such a view is consistent with representations of speech posed in the literature as well as observations of natural speech. Even though the speech-production process actually varies continuously in time, we adopt an abstraction that the system produces discontinuities that are well-described as discrete excitations. In this thesis, we quantify sequences of these *glottal events* that are extracted from the acoustic signal and use them to analyze differences between voice qualities.

Classically, phonation has been described as a periodic process [16], the repeating unit of which is the *glottal cycle*. The area of the *glottis*, the space between the vocal folds, is at a minimum when the folds are closed. The folds then open, and the glottal area increases and reaches a maximum. Finally, as the folds close, the area between the vocal folds decreases, eventually returning to its starting value. Schematically, the periodic nature of the glottal cycle is shown in Figure 2-1, with each solid line corresponding to a subsequent cycle, occurring with period *T* from the previous one and with each cycle having the same amplitude. Conventionally, we associate the boundary between one glottal cycle and the next with the closing of the glottis. Another term for phonation exhibiting glottal cycles with uniform timing and amplitudes is *modal* [37].

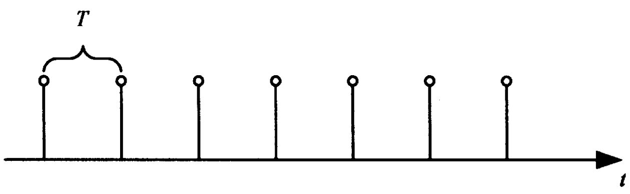


Figure 2-1. Schematic of the classical periodic description of phonation. Each solid line represents an abstraction of the closing time in each glottal cycle.

Glottal cycles may also occur with *nonuniform* timings and amplitudes, a property known as *nonmodality*. This behavior is shown schematically in Figure 2-2. Modality and nonmodality do not form discrete categories, but rather describe a continuum of behaviors. As we will explore later in the thesis, deviations from modality take on many different forms, having deterministic and random elements, and also occur to different degrees, from slight deviations from completely modal behavior to more drastic irregularity.

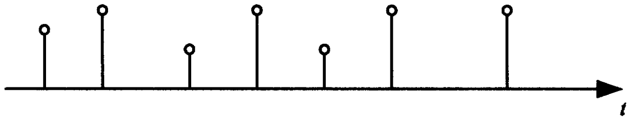


Figure 2-2. Schematic of nonmodal phonation, having amplitudes and periods which vary from one glottal cycle to the next. Each solid line represents an abstraction of the closing time in the glottal cycle.

In addition to the nonmodal variation *between* glottal cycles that we have shown in Figure 2-2, it is also common to see points of significant excitation *within* a single glottal cycle. As shown schematically in Figure 2-3, these events may occur due to glottal openings, ripples, vortices, and possibly other phenomena. In the figure, solid lines represent an abstraction of the closure point associated with each glottal cycle and each dotted line represent additional glottal events occurring within each period.

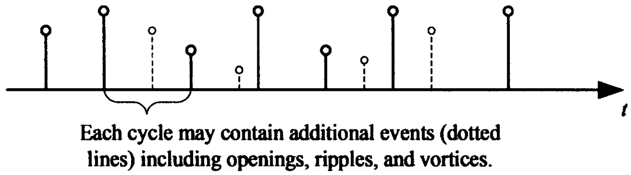

t
Each cycle may contain additional events (dotted lines) including openings, ripples, and vortices.

Figure 2-3. Schematic of events occuring within a glottal cycle. Each solid line represents an abstraction of the closing time for each glottal cycle and each dotted line represents abstractions of additional events within a single cycle.

2.1.1 Glottal Events as Points of Rapid Change

Defining discrete events in a glottal cycle quantitatively has been persued in the literature. One particular example is found in Ananthapadmanabha and Yegnanarayana's theory of epoch extraction [3]. Here, the significant discrete excitation events are defined in terms of discontinuities in the derivatives of the waveforms. In particular, they describe that "the point of discontinuity of the lowest ordered derivative will be regarded as an epoch." They state that the term "discontinuity of the lowest ordered derivative" refers to the case where $f^{(n+1)}(t)$ is discontinuous while $f^{(n)}(t)$ is continuous. Therefore, each epoch has an *order* equal to $(n+1)$ associated with it, which denotes how many derivatives need to be taken before a discontinuity is seen. Ananthapadmanabha and Yegnanarayana show analytically that *lower-order events tend to dominate the spectral content of the waveform*, more so at frequencies further away from DC [3]. Figure 2-4 shows an example from [3] describing the locations and orders of events for four different waveforms.

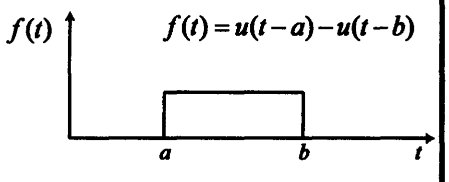
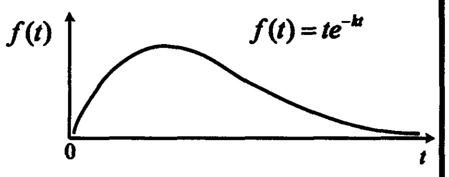
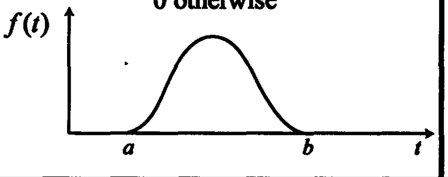
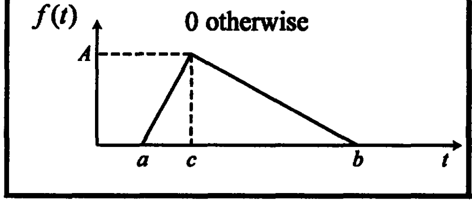
Function	Epoch Times	Order of Epochs
$f(t)=u(t-a)-u(t-b)$ 	$t_i = a,b$	0 due to a discontinuous signal: $f(t_i^-) \neq f(t_i^+)$
$f(t)=te^{-kt}$ 	$t_i = 0$	1 due to a discontinuous first derivative: $f(t_i^-) = f(t_i^+)$ $f^1(t_i^-) \neq f^1(t_i^+)$
$f(t)=5-5\cos(2\pi(t-a)/(b-a))$ for $a \le t \le b$ 0 otherwise 	$t_i = a,b$	2 due to a discontinuous second derivative: $f(t_i^-) = f(t_i^+)$ $f^1(t_i^-) = f^1(t_i^+)$ $f^2(t_i^-) \neq f^2(t_i^+)$
$f(t)=k_1(t-a) \quad a \le t \le c$ $A-k_2(t-c) \quad c \le t \le b$ 0 otherwise 	$t_i = a,b,c$	1 due to a discontinuous first derivative: $f(t_i^-) = f(t_i^+)$ $f^1(t_i^-) \neq f^1(t_i^+)$

Figure 2-4. Diagram adapted from [3] describing several glottal-flow-like functions and their glottal events (called epochs) defined in terms of discontinuities in the derivatives.

The above definition of glottal events based on the derivatives of the excitation waveform helps us begin to understand quantitatively why a continuous speech waveform may be viewed as excited at a series of discrete times. The most abrupt changes (equivalently lowest-order events) in a waveform will dominate its spectral energy, especially in the higher frequencies such as those that give a "complete" excitation of the vocal tract.

In order to better understand the nature of glottal events, we now show several examples using synthetic stimuli. These examples show that rapid changes in the source waveform, vertical striations in the spectrogram, and a subset of the peaks in the acoustic waveform are all associated with glottal events. We also observe that different waveshapes exhibit different glottal-event structures. In Figure 2-5, we see the effects of a periodic rectangular source waveform. The top two panels show the effect on the spectrogram and acoustic waveform for the synthetic vowel /ɑ/.

The bottom two panels show the source waveform itself and the accompanying spectrogram. Observe that both the onset and offset of the waveform yield peaks in the acoustic signal and vertical striations in the spectrograms. In contrast, Figure 2-6 depicts the analysis of a sawtooth waveform. Here, the offset of each triangle dominates the excitation, although there is some evidence of an increase in excitation at the onset as well, due to the discontinuity in the derivative.

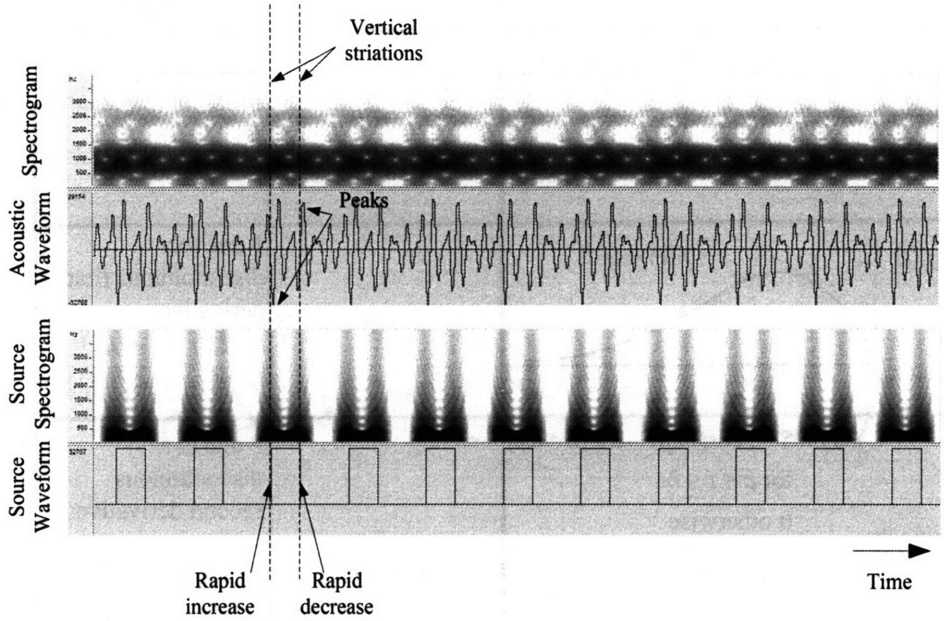


Figure 2-5. Analysis of the properties of a periodic rectangular waveform with a fundamental frequency of 125 Hz. Vertical striations in the wideband spectrogram of both the source and radiated synthetic acoustic signal, strong peaks in the acoustic waveform, and rapid increases and decreases in the source waveform all occur at the instants of glottal events. The spectrograms use 4-ms Hamming windows.

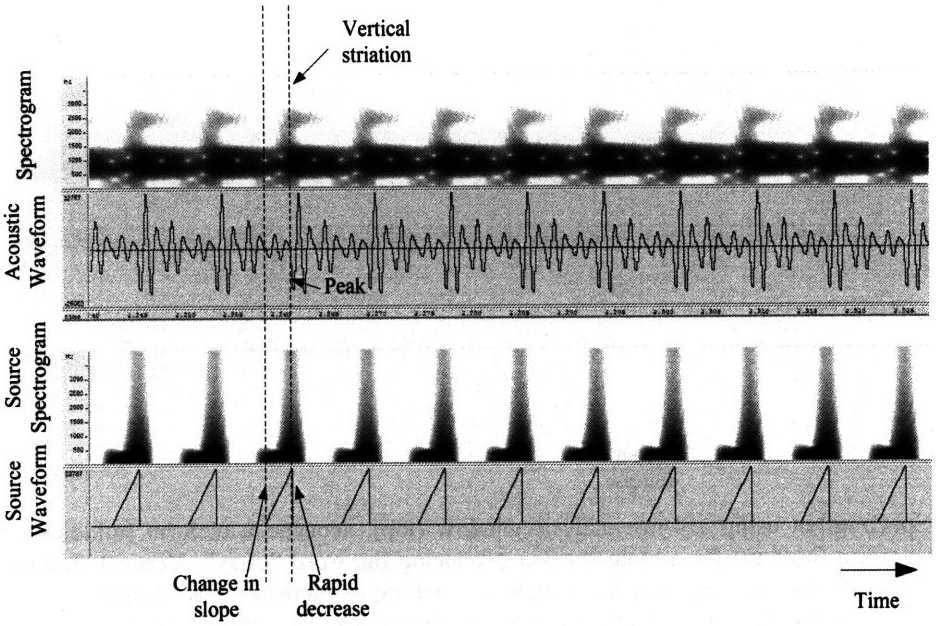


Figure 2-6. Analysis of the properties of a periodic sawtooth waveform with a fundamental frequency of 125 Hz. Vertical striations in the wideband spectrograms of both the source and radiated synthetic acoustic signal, strong peaks in the acoustic waveform, and rapid increases and decreases in the source waveform all occur at the instants of glottal events. The spectrograms use 4-ms Hamming windows.

2.1.2 Motivating Examples from Real Speech

Before we delve into the specifics of the model used in the remainder of this thesis, we present several examples further illustrating the notion that the real speech excitation can be abstracted into a series of discrete glottal events. In Figure 2-7, taken from a male speaker, for example, we see a section of voiced speech from the word "carry" with spectrogram, acoustic waveform, and electroglottograph (EGG) measurements. We use EGG data as a rough indicator of changes in the glottal source, keeping in mind that the two are not equivalent. EGG shows contact between the vocal folds, while the source itself is a more complicated aeroacoustic phenomenon. Each of these representations shows evidence of rapid change, manifesting as discrete points in the excitation. In the spectrogram, we see vertical striations, in the time-domain signal we see a series of strong peaks, and in the EGG signal we see evidence of rapid increases and decreases in the impedance corresponding roughly to spreading apart and closing of the glottis.

We see more evidence of these rapid changes in the excitation source also in Figure 2-8, taken from a female speaker. Again, vertical striations in the short-time spectrogram, sharp peaks in the acoustic waveform, and rapid changes in the EGG waveform are all present and agree roughly in time. The goal of these pictures is to give a flavor for the kind of phenomenon that inspires this thesis. At this point, it is sufficient for the reader to appreciate that the speech-production system may be abstracted as having instants of excitation, which manifest across different kinds of measurements.

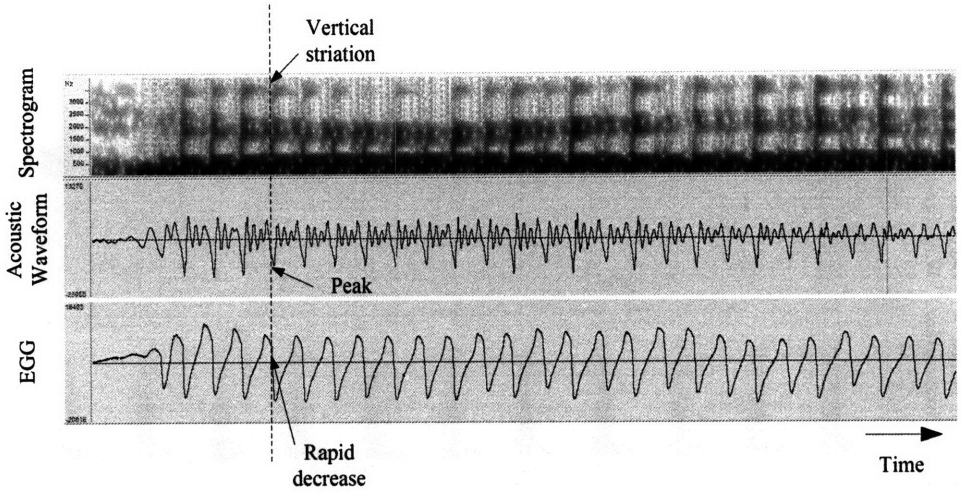


Figure 2-7. Spectrogram using a 4-ms analysis window (top), acoustic waveform (middle), and EGG impedance waveform (bottom) for a male speaker producing the word "carry". Vertical striations in the spectrogram, strong peaks in the acoustic waveform, and rapid increases and decreases in the EGG waveform are consistent with discrete excitation events in the speech-production mechanisms.

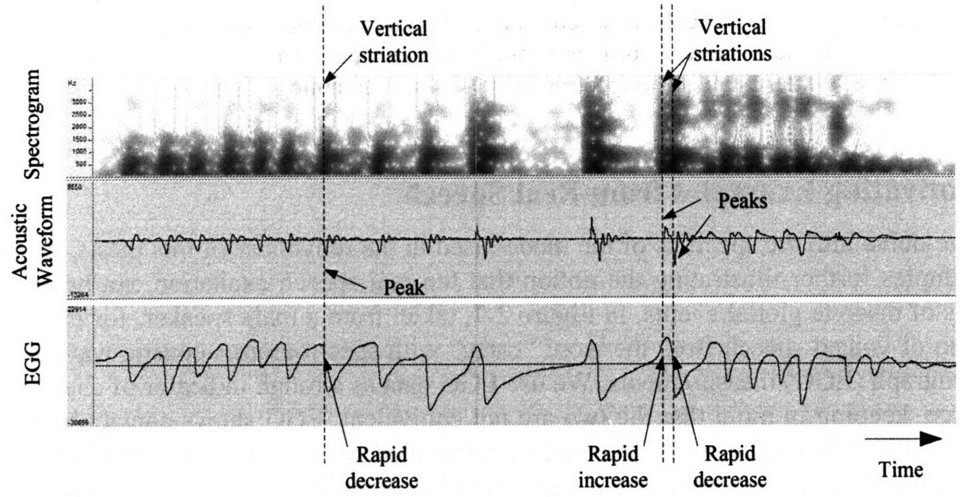


Figure 2-8. Spectrogram using a 4-ms analysis window (top), acoustic waveform (middle), and EGG impedance waveform (bottom) for a female speaker producing the word "like". Vertical striations in the spectrogram, strong peaks in the acoustic waveform, and rapid increases and decreases in the EGG waveform are consistent with discrete excitation events in the speech-production mechanisms.

2.1.3 Speaker-Dependence in Impulsive Patterns

This thesis hypothesizes that there exist differences in sequences of discrete glottal events between different speakers. As we have discussed earlier in this chapter, these differences occur on different scales including within a glottal cycle and between neighboring glottal cycles. We will show motivating examples of both kinds of phenomena in this section.

Within-Cycle Events

Impulsive events may occur at multiple times within a single glottal cycle. One example of this phenomenon is shown in Figure 2-9 where a male speaker is producing the /ɛ/ in "carry." Here, the glottal events manifest in the acoustic signal as sets of two vertical striations in the spectrogram. The first of these likely corresponds to an opening gesture of the glottis as seen in the rising EGG impedance, and the second is related to a closing gesture of the glottis evidenced by the falling EGG impedance. For comparison, Figure 2-10 shows phonation from the same word produced by a second male speaker, this time with only one distinct event per glottal cycle apparent in the acoustic signal. In contrast to the first case, there is only one distinct striation per cycle.

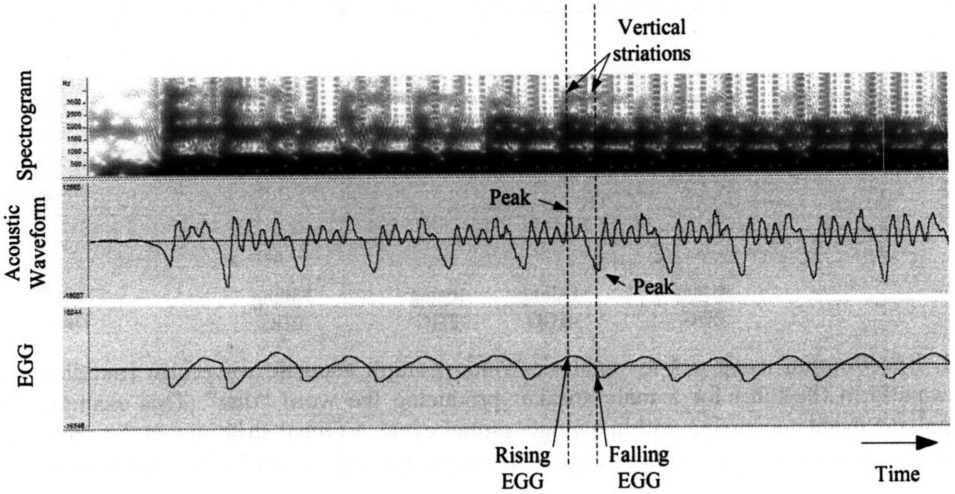


Figure 2-9. Spectrogram using a 4-ms analysis window (top), acoustic waveform (middle), and EGG impedance waveform (bottom) for a male speaker producing the word "carry". This example illustrates multiple impulsive events within a single glottal cycle. Vertical striations in the spectrogram are grouped in sets of two, one corresponding roughly to glottal opening (rising EGG) and the other to closure (falling EGG).

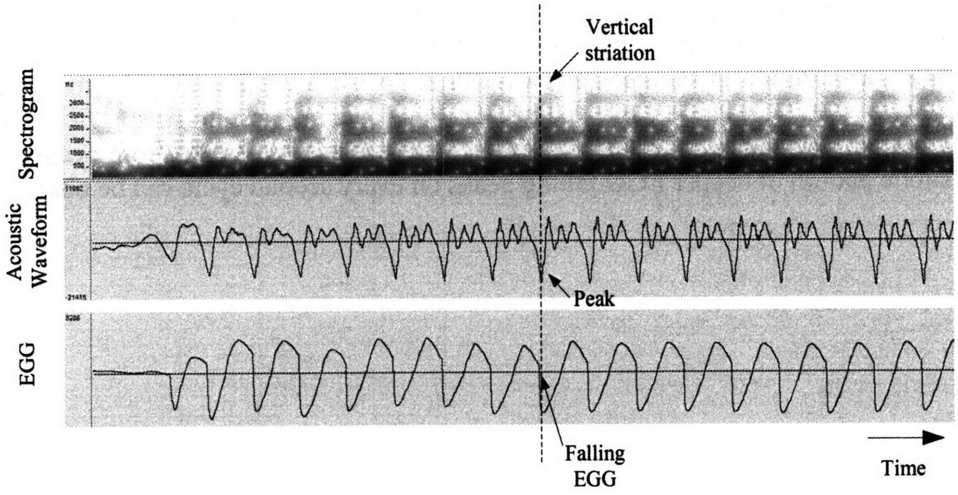


Figure 2-10. Spectrogram using a 4-ms analysis window (top), acoustic waveform (middle), and EGG impedance waveform (bottom) for a male speaker producing the word "carry". This example illustrates one glottal event per glottal cycle manifesting as a vertical striation in the spectrogram, corresponding to glottal closure (falling EGG).

Another example is shown in Figure 2-11 for the word "that" produced by a male speaker: In this figure, we see clear pairs of vertical striations in the spectrogram indicating strong within-cycle events in the source. The changes are also partially visible in the acoustic waveform as positive and negative prominences. The first of each pair is likely associated with an opening gesture of the glottis as seen in the rising EGG impedance, and the second is related to a closing gesture of the glottis evidenced by the falling EGG impedance.

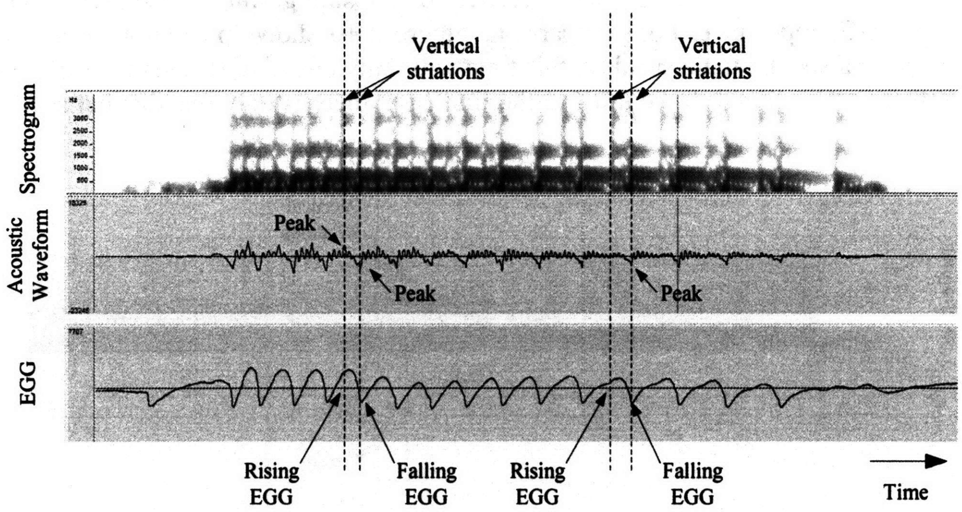


Figure 2-11. Spectrogram using a 4-ms analysis window (top), acoustic waveform (middle), and EGG impedance waveform (bottom) for a male speaker producing the word "that". This example illustrates multiple impulsive events occurring within a single glottal cycle. Vertical striations in the spectrogram are grouped in sets of two, one corresponding roughly to glottal opening (rising EGG) and the other to closure (falling EGG).

Between-Cycle Events

Impulsive glottal events may also take on patterns *between* neighboring glottal cycles. In Figure 2-12, we see an example of this phenomenon for a male speaker. Observe that there is a pattern of glottal events from one cycle to the next, with every other cycle having a glottal event with a different quality as evidenced by the weaker vertical striation in the spectrogram and the slightly weaker prominence in the acoustic waveform. Every second cycle actually appears to have a set of two weaker striations, one likely due to an opening gesture (rising EGG) and the other due to glottal closing (falling EGG). The EGG stays relatively similar from cycle to cycle, although a subtle pattern of deeper EGG falling peaks on every second cycle can be observed.

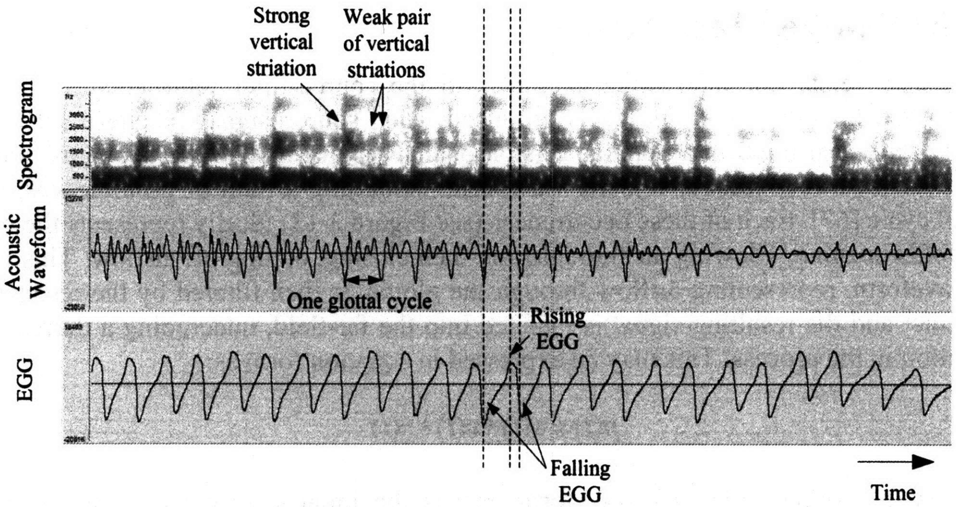


Figure 2-12. Spectrogram using a 4-ms analysis window (top), acoustic waveform (middle), and EGG impedance waveform (bottom) for a male speaker producing the word "carry". This example illustrates patterns of glottal events between glottal cycles. The glottal events alternate between weak and strong every cycle. Additionally, every weak cycle consists of a pair of two glottal events, one due to opening and one to closing.

2.2 Previous Work with Nonmodality

Nonuniform glottal events occurring across successive cycles have been studied extensively in the literature. Recall that there is a continuum along which all natural voices should fall, ranging from just short of modal to strongly nonmodal. In terms of this continuum, a single speaker may change as they speak, both over extended regions of speech as well as in short intermittent sections. Such changes can be important for linguistic and prosodic reasons [22] and may also indicate disorders of the voice [75].

Different aspects of nonmodal speech have been explored in the literature. It has been shown, for example, that different speakers vary in how frequently they use nonmodal voicing. Dilley, Shattuck-Hufnagel, and Ostendorf report a group of five speakers who differ in the frequency of overall word-initial nonmodality they express from 13 percent to 44 percent [15]. There is also evidence that certain acoustic characteristics of pulse patterns vary characteristically between speakers [20, 59]. Evidence for repeating glottal pulse patterns has also been shown using nonacoustic sensors detecting movement of the vocal folds.

In the typical speech of healthy speakers, there exist several common classes of cycle-to-cycle nonuniformity in glottal events. In their work discussing occurrences of a subset of nonmodal phonation in normal speakers, Redi and Shattuck-Hufnagel discuss a scheme using four possible categories: (1) aperiodicity or "irregularity in duration of glottal pulses from period to period," (2) creak or "prolonged low fundamental frequency accompanied by almost total damping", (3) diplophonia or "regular alteration in shape, duration, or amplitude of glottal periods", and (4) glottal squeak or "a sudden shift to relatively high sustained f_o, which [is] usually very low amplitude." Speakers have been shown to vary consistently in the amount they express each of these four categories [59]. For a complete review of nonmodal pulse characteristics in healthy speakers see [22].

2.3 The Source-Filter Model

We will now relate our motivating examples to the source-filter representation of speech production [16]. In this model, speech is produced when source signals excite a set of filters representing the vocal tract. In a popular version of this type of model, the phonation source of speech is described as a volume-velocity waveform, exhibiting a train of prominences, one for each glottal cycle [67]. Each of these occurrences (see Figure 2-13) ideally represents the result of the vocal folds opening, allowing air through, and then closing, cutting off airflow. This *volume-velocity* waveform, representing airflow through the glottis, is then filtered by the resonances of the vocal tract and the resulting signal is radiated into the far-field, undergoing a derivative-like transformation in the process. This may be expressed in equation form as:

$$p(t) = s(t) * v(t) * r(t)$$

where $s(t)$ is the source-excitation waveform, $v(t)$ is the impulse response of the vocal tract resonances, and $r(t)$ is the radiation characteristic.

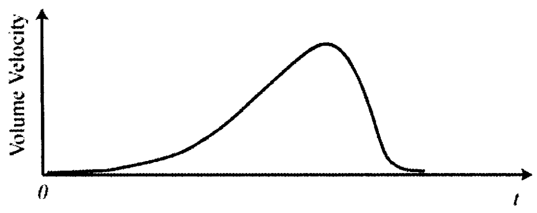


Figure 2-13. Schematic view of a glottal volume-velocity waveform, adapted from [54].

Though this model is often assumed, it is important to realize that it is a very simple abstraction of the physical system. The ideal source waveform, $s(t)$, that we have described likely cannot be measured anywhere in the three-dimensional physical system. Abstracting a complex system such as this to a single one-dimensional source is common in lumped-element modeling of acoustics and has been able to yield many insights into the speech-production process. In the words of Fant [16], "the glottis represents a high impedance termination of the vocal tract, and it is thus possible and convenient to define the voice source by the pulsating airflow through the glottis."

For the purposes of this thesis, one of the problems with the model is that it is often implied to have a periodically-varying source-excitation waveform, $s(t)$, and textbook shape as in Figure 2-13. As we have described, this is not always the case. In this section, we describe two different perspectives on representing the source waveform, a parametric approach and a filtered impulse excitation approach.

2.3.1 Modeling the Glottal Cycle

One perspective on modeling the source component of the source-filter model is to arrive at a set of *parameters* to describe each glottal cycle. The resulting speech waveform is then simply the sum of a set of such parameterized cycles. There exists a significant body of literature specifically dealing with ways in which to parameterize the glottal cycle [36, 37, 62, 67]. Two of the more popular models are the Liljencrants-Fant (LF) model [17] and Klatt's KLGLOTT88 model [37]. Both of these models attempt to model the shape of one volume-velocity prominence.

The LF model is popular in the literature and parametrically describes the shape of a single glottal cycle. Important parameters of this model are the time of the glottal opening, the time of the maximum negative value of the glottal flow derivative, and the time of glottal closure. Several additional parameters describe the specific shape of the waveform. Researchers who have used this representation to describe the glottal-source include [54] and [18].

The Klatt model [37] is based on the relationships shown by Rosenberg [62]. The model uses a modulated sinusoid to obtain the shape of the volume velocity for each glottal cycle. The parameters are the time of closure, open quotient, and optionally a speed quotient, which shapes the steepness of the waveform during opening. This model also allows a low-pass filter controlled by a parameter called *tilt* to model the smoothness at the glottal-pulse closure. Researchers including [34] and [48] have used the Klatt model to parametrically describe the glottal cycle.

Although it is convenient to describe the source as a train of discrete bursts of air, each with a shape that can be described with a small set of parameters, the physical reality is not always manifested in this way. There are many situations when the folds, for example, do not fully close to cut off the airflow, or do close but only partially at different points in the cycle. As we have discussed, though, the speech excitation tends to be well described as a series of events representing points of rapid change in the source. The behaviors of the phonation source mechanisms that we have described, which are common in many normal and disordered speakers, are the motivation for using a more general impulse-excited model as described in the next section.

2.3.2 A Filtered-Impulse Abstraction of Glottal Events

Earlier in this chapter, we presented the idea that speech can be abstracted as a series of discrete glottal excitation events. In this section, we will present a rationale for modeling these events in the source as a train of filtered impulses, each with a certain amplitude and timing.

This thesis takes the perspective that dealing with a wide range of excitation characteristics requires a more general approach to modeling the glottal source than a parametric model constrained to a limited range of glottal cycle shapes as described above. One approach is to model the speech source as a series of (in general) nonmodal impulses exciting a filter. This filter captures both the local effects of the vocal tract as well as elements of the source shape. The assumption is that details of the glottal shape are difficult to estimate, but that this underlying impulsive representation captures meaningful events in the glottal cycle. For the purposes of this thesis, we hypothesize that these impulses represent differences between different speakers.

Related to our earlier discussion of glottal events, we can see that impulses have much in common with the notion of a glottal event. In fact the notion that the underlying excitation source is impulse-like is not new. The idea has been central to early synthesis engines including KLATTSYN [35] and others, as well as to analysis techniques such as linear prediction. Atal [6] also uses such a model in his analysis-by-synthesis approach. Impulses and glottal events are similar in that they are both assigned a discrete time and amplitude. Both also have the property that they can excite the vocal tract, providing spectral components at upper frequencies. They also both cause peaks in the time-signal and vertical striations in the spectrogram. The biggest leap in mapping glottal events to impulses is that glottal events do not impart flat spectral excitation in the same way that impulses do. To address this issue, we build the shape of the spectrum into the filter component of our source filter model as first mentioned above and discussed in more detail below.

The model we use represents speech as the sum of filtered discrete-time impulses. Each impulse has associated with it a certain time of occurrence and amplitude. In order to make such a train of impulses speech-like, the model imposes a certain impulse response to each impulse in the train, representing contributions due to the physiological and aeroacoustic mechanisms of the source as well as the vocal-tract resonances and the radiation characteristic. All of the filtering elements combined convolutionally are called the ***composite impulse response***. This system is shown schematically in Figure 2-14 where each glottal event is associated with a particular composite-impulse response. Two example responses are shown along with how they fit in the resulting acoustic waveform.

Unlike strict parametric models, the underlying *meaning* of the impulses is in terms of the more abstract glottal events instead of particular characteristics of the physical glottal mechanisms. The characteristics of each individual glottal cycle are not separated out, allowing multiple impulses to represent one cycle. This ambiguity is imposed with intent, to allow the model to represent a wider variety of phenomena.

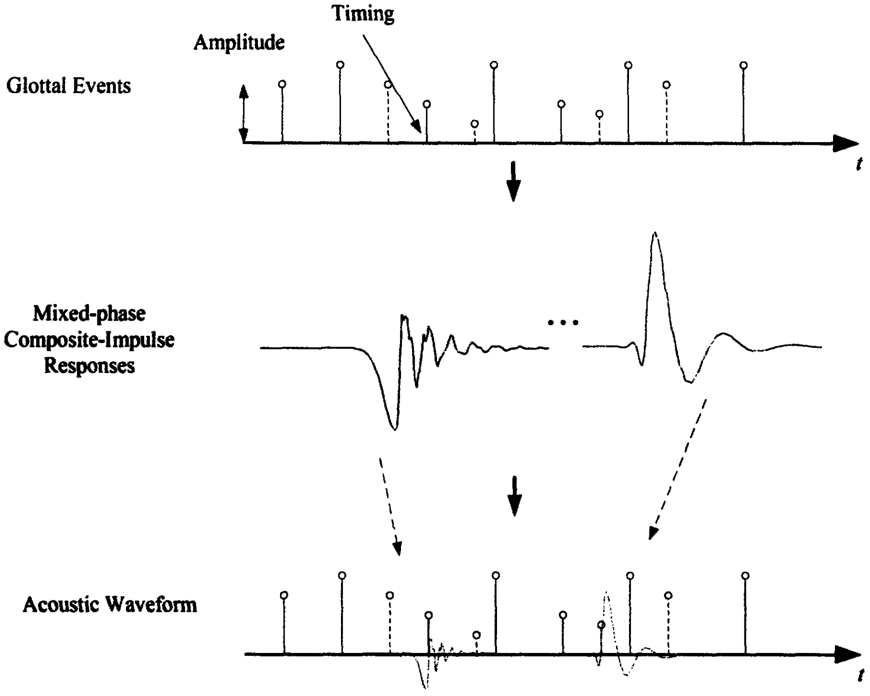


Figure 2-14. A schematic of the filtered-impulse representation of the voice. Each glottal event has associated with it a certain composite impulse response waveform. The acoustic waveform is the result of summing each glottal-event impulse with its associated composite impulse response.

2.4 Conclusion

In this chapter, we have defined glottal events in terms of a source-filter model of speech production. We have shown that, contrary to the periodic textbook representation of phonation, real speech may exhibit nonuniform timings and amplitudes between neighboring glottal cycles. Additionally, multiple points of excitation may occur *within* a single glottal cycle, imparting a local structure.

Series of these discrete events are used in the rest of this thesis to represent natural and synthetic speech. As we have discussed, the events underlying this model may often be associated with meaningful characteristics of the glottal cycle, such as openings and closings. When we estimate the locations of glottal events in later chapters, we will reinforce these interpretations.

Chapter 3

Spectral Representation of Nonmodality

In the last chapter, we described a glottal-event representation of the speech excitation source. We will now embark on an understanding of the theory of how sequences of these events affect *spectral*, *pitch-track*, and *time-domain* representations. The current chapter addresses how deterministic patterns as well as stochastic timing and amplitude variations affect the spectral representation of glottal-event trains. The next chapter describes the pitch of a nonmodal train of glottal events. Finally, Chapter 5 presents a new technique for analyzing the timing and amplitude relationships between neighboring glottal events.

The spectral, fundamental-frequency, and time-domain representations for a periodic source excitation are well known. The theory that we develop is important because it extends the understanding of periodic events to nonmodal sequences, which, as we will see, have special properties that do not follow intuitively from our understanding of the periodic source. We will use these properties later in the thesis when we extract and analyze sequences of glottal events from real speech.

This chapter on a spectral theory [42] of nonmodality is outlined as follows[1]. In Section 3.1, we provide motivation and background for developing a spectral theory. In Section 3.2, we describe our framework for introducing nonmodality in deterministic impulse trains, including both multiplicative and additive disturbances in the frequency domain, and give examples that illustrate our models. In Section 3.3, we investigate a stochastic framework for random perturbations and present empirical studies in this framework. Finally, Section 3.4 provides a brief look into extensions to real speech cases.

3.1 Motivation and Background

Our investigation of the spectral representation of impulse *patterns* arises from a desire to understand the spectral-temporal properties of nonmodality in speech. As we have discussed, the underlying impulsive excitation of speech can exhibit repeating deterministic patterns as well as perturbations that appear random in nature. The deterministic aspects involve repeating patterns of excitation timings and amplitudes. Examples include diplophonia (a 2-long pattern of alternating large and small pulse periods and/or amplitudes) and triplophonia (3-long pattern) [8, 20]. The stochastic elements of nonmodality are characterized by randomness in the timing and amplitude. For small variations, randomness in timing is often called jitter, and randomness in

[1] **A large part of this chapter is © 2008 IEEE. Reprinted, with permission, from N. Malyska and T. F. Quatieri, "Spectral representations of nonmodal phonation,"** ***IEEE Transactions on Audio, Speech, and Language Processing*****, to appear January, 2008.**

amplitude is called shimmer. Large perturbations, however, can also occur, and there exists a continuum of this dimension of nonmodality from minor to significant. The two aspects of nonmodality mentioned here are not mutually exclusive, and it is common to find speech signals with *both* structured variation and random variation from an underlying pattern.

Harmonic speech spectra can be quite sensitive to aberrations in periodicity of an impulsive model of the glottal source. Even very small perturbations can take the form of short-time spectral changes that mislead the viewer in terms of signal composition. For example, jitter and shimmer in periodicity due to vocal-fold instability can introduce the appearance of "noise" contributions that are not present in the source. Significant spectral modifications can also be present with timing and amplitude perturbations due to nonmodal types of vibration such as diplophonia and triplophonia. A spectral characterization of nonmodality then is important in a variety of speech applications such as feature-extraction techniques for recognition and for typical speech signal-processing approaches including linear prediction and sinusoidal analysis-synthesis.

An example of short-time spectral sensitivity to timing perturbations for a periodic train of impulses is shown in Figure 3-1. Here one impulse of a periodic impulse train of 10.1-ms period is shifted by one sample (0.1 ms), and its short-time spectrum is obtained with a 40-ms Hamming-window. Observe that even this relatively minor change of a one-sample shift over a long window contributes a noticeable difference in the harmonic structure in that the mid-frequency harmonics are attenuated and high-frequency harmonics have the appearance of being shifted half the fundamental frequency.

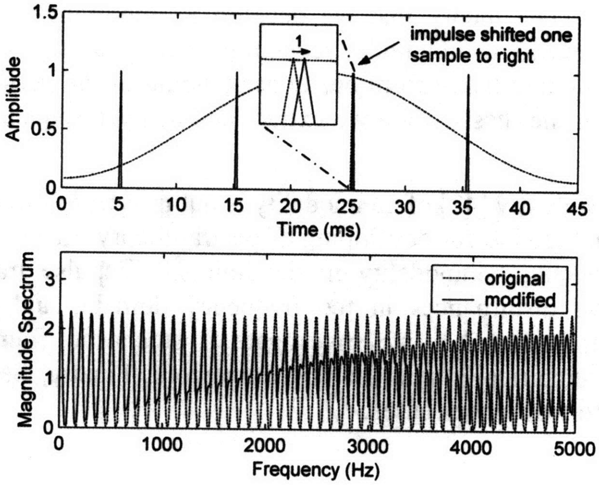


Figure 3-1. Time-frequency illustration of one-sample shift in a periodic impulse train. (Upper) Segment of perturbed impulse train (solid) superimposed with the original (dashed); (Bottom) Fourier-transform magnitude of Hamming windowed waveforms. Period is 10.1 ms and sampling rate is 10000 samples/s. A one-sample shift equals 0.1 ms.

The objective of this chapter is to formulate a general framework for how manipulations like the one in the example of Figure 3-1, where impulse timings as well as heights are modified, alter short-time spectral speech source characteristics. We will do this both from a *deterministic* perspective in which perturbations are performed in a cyclic fashion and from a *stochastic* perspective in which perturbations occur randomly. Our cyclic perturbations encompass repeating diplophonic and triplophonic patterns alluded to earlier, as well as generalizations of these repeating patterns, while our random perturbations encompass jitter, shimmer, and their larger counterparts. Using our general frameworks, we will derive the spectrum for deterministic and stochastic impulse sequences that typically occur in speech production. The spectral representations are relevant to a number of speech-processing areas where nonmodality can lead to misinterpretation of aharmonic structure, peaks and nulls in the spectral envelope, and additive noise contributions.

Aspects of our work are motivated in part by previous observations and analysis of harmonic spectral modification due to jitter, shimmer, and additive noise ([28], [48], and [63]). Hillenbrand was one of the first to experimentally consider the effect of small random perturbations of pitch and amplitude on short-time spectral characteristics. With random pitch perturbations, he observed a breakdown of "harmonic organization" in high-frequency regions, while with random amplitude changes he observed similar properties but with less harmonic breakdown. Hillenbrand found that, consistent with this harmonic breakdown, was the inability of harmonic-to-noise measurements to distinguish noise from random jitter and shimmer.

A number of researchers built on the work of Hillenbrand. Murphy [48], for example, provided a Fourier series-based description of "cyclic" amplitude and pitch perturbations of periodic signals and a Fourier-transform-based description of random perturbations. With these Fourier representations, Murphy was able to predict some of Hillenbrand's experimental observations. In more recent work, Schoentgen [63] gives a different perspective in terms of a modulation model of shimmer and jitter, decomposing a periodic sequence into a harmonic series of sinusoids which are randomly frequency-modulated or randomly amplitude-modulated. In this approach, the amplitude and/or phase of each harmonic in the original signal is modulated, resulting in sidebands that provide new frequencies and that may interact with the harmonics themselves.

In these previous developments, important insights were obtained on the effects of perturbations on the spectrum, either experimentally or through modeling approaches. However, a quantitative framework to describe the spectrum of arbitrary cyclic pulse patterns, such as Murphy's "cyclic jitter" and "cyclic shimmer", has not yet been derived. Additionally, although some properties of the power spectral density resulting from stochastic perturbations are described experimentally by Murphy and Hillenbrand, neither they nor Schoentgen analytically derive an expression for the power spectral density. In this chapter, we provide a more general approach that includes the possibility of arbitrary perturbations, encompassing jitter and shimmer, as well as other aspects of the source. In the deterministic case, corresponding to structured patterns in nonmodality, the key is a *filtering interpretation of an underlying harmonic series* that results in alterations of harmonics and the evolution of new "sub-harmonics". As part of this development, we introduce the use of a sequence of doublets to represent the movement of impulses as a means for spectral modification. In the stochastic case, involving random aspects of nonmodality, the key is a model of *how the autocorrelation function of an underlying periodic signal is modified* by random amplitude and timing changes on the periodic signal.

3.2 Spectra for Deterministic Impulse Sequences

In this section, we will derive the general frequency spectrum for a deterministic sequence of impulses. In particular, we will discuss (1) construction of pulse sequences, (2) spectra of these sequences, and (3) the effect of an analysis window.

3.2.1 Framework

The first step in our derivation of spectral representations of impulse sequences is to formulate the sequence of interest. For now, we will ignore the effect of an analysis window.

In the time domain, a series of impulses can be viewed as a set of scaled and time-displaced versions of the unit sample $\delta[n]$. In general, a sequence of impulses, each with amplitude C_k and position η_k, can be expressed as

$$y[n]=\sum_k C_k\delta[n-\eta_k]. \tag{3.1}$$

Such a representation can be used for any deterministic sequence of impulses. We will begin our derivation with the spectrum of this general sequence of impulses and then address the case of repeating patterns of impulses that relate to certain phenomena commonly seen in nonmodal speech.

The spectrum of the sequence in (3.1) is

$$Y(\omega)=\sum_k C_k e^{-j\omega n_k}. \tag{3.2}$$

While we can express the frequency spectrum for any sequence of pulses by (3.2), the process of phonation tends not to be discussed in terms of the locations of individual excitation pulses. Instead, it is more conventional to view phonation as a *repeating* pattern of impulses. With perfectly modal speech, for example, we often talk about its fundamental frequency, which is inversely proportional to the spacing between adjacent glottal pulses.

One method to obtain a periodically-repeating pulse pattern is simply to take a sequence of pulses and convolve it with a repeating impulse train with unity amplitude and period *P*, denoted $x[n]$. Throughout this chapter, we will call *P* the *pattern-period*:

$$\begin{aligned} &y[n]=\sum_k A_k\delta[n-n_k]*x[n] \quad 0\le n_k\le P-1 \\ &\text{where } x[n]=\sum_k \delta[n-kP] \end{aligned} \tag{3.3}$$

and where A_k is the amplitude of the impulse at time n_k. For simplicity, we use the notation $m[n]$ to refer to the sequence $y[n]$ over one period of the repeating impulse pattern:

$$\begin{aligned} &m[n]=\sum_k A_k\delta[n-n_k] \qquad 0\le n_k\le P-1 \\ &\text{where its spectrum is given by } M(\omega)=\sum_k A_k e^{-j\omega n_k}. \end{aligned} \tag{3.4}$$

In the frequency domain, the convolution shown in (3.3) is represented as the product

$$\begin{aligned} Y(\omega)&=\sum_k A_k e^{-j\omega n_k}\left(\sum_k \frac{2\pi}{P}\delta(\omega-k2\pi/P)\right) \quad 0\le n_k\le P-1 \\ &=M(\omega)X(\omega) \end{aligned}. \tag{3.5}$$

The interpretation of this expression is that the spectrum of any arbitrary pulse pattern is equal to a uniform line spectrum occurring at $\omega=2\pi k/P$ modulated by the sum of a set of weighted complex exponentials, $M(\omega)$. Alternatively, we can interpret this process as a sampling of $M(\omega)$ in the frequency domain. The magnitude of this product is depicted schematically in Figure 3-2.

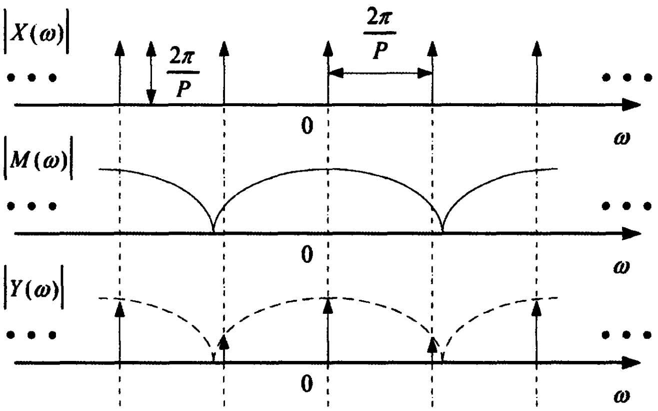


Figure 3-2. Schematic for the process of generating the spectrum of a repeating impulse pattern. The envelope, $M(\omega)$, derived from one period of the repeating pattern, is multiplied, i.e. sampled, by a line spectrum, $X(\omega)$, to yield the spectrum of the repeating pattern.

We have thus far described two approaches that, together, can be used to construct a pattern of impulses—the first can build any general sequence of impulses while the second provides a straightforward way to deal with repeating sequences. We will now examine an example of this process.

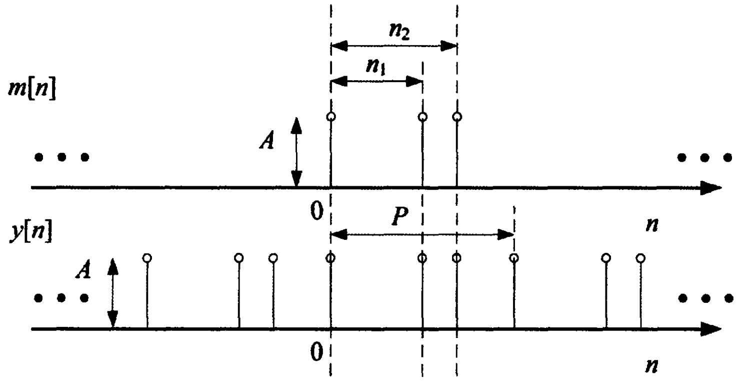


Figure 3-3. Process of building a pattern of impulses by repeating one period of the pattern.

We begin by constructing a basic timing pattern involving a pattern of three pulses with constant amplitude A, one at the origin, one at n_1, and one at n_2 and with pattern-period P. As shown in Figure 3-3, this pattern is built up by first describing one period of the pattern, $m[n]$, described by

$$m[n] = A\delta[n] + A\delta[n - n_1] + A\delta[n - n_2]. \tag{3.6}$$

This individual period is then repeated in time at the pattern-period P, resulting in the impulse sequence, $y[n]$. In the frequency domain, the envelope $M(\omega)$ is represented by the function

$$M(\omega) = A + Ae^{-j\omega n_1} + Ae^{-j\omega n_2}. \tag{3.7}$$

The effect is to impose a complex frequency envelope. An example of such a magnitude function is shown in Figure 3-4. Observe that the envelope is different from the flat envelope expected for a perfectly periodic impulse case—it has numerous peaks and valleys.

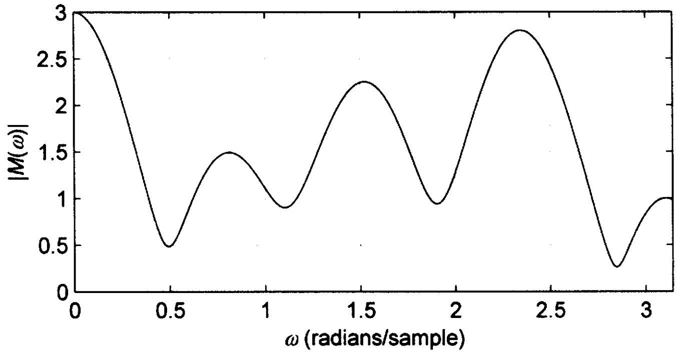


Figure 3-4. Spectral magnitude imposed by the pattern described in (3.6) shown on a linear-amplitude scale. The values $A=1$, $n_1=5$, and $n_2=8$ are used.

Generally, the effect of an impulse pattern may be interpreted as a set of spectral zeros, which effectively create peaks and dips in the source spectrum. We can view $M(\omega)$ as the complex spectrum of a filter which shapes an underlying line spectrum in the frequency domain.

3.2.2 Spectral Effects of Modified Impulse Amplitude and Timing

Another perspective on the spectral effects of a nonmodal impulse sequence is as a modification of the spectrum of a known impulse train, in contrast to the filtering of an underlying reference harmonic series of the previous section. The input, $y_{in}[n]$, to the modification process is a general sequence of impulses, each with amplitude B_k and position ς_k, described by the equation

$$y_{in}[n]=\sum_k B_k\delta[n-\varsigma_k]. \tag{3.8}$$

We can modify each of these impulses by shifting it to the left or right or by changing its amplitude. An equivalent operation is to add a term that cancels the original impulse and adds an impulse with the desired timing and amplitude. We call this term a *doublet* and denote it $q[n]$. Each doublet and its Fourier transform has the form

$$q[n]=-a_{start}\delta[n-n_{start}]+a_0\delta[n-n_{start}-n_0] \tag{3.9}$$

$$Q(\omega)=\left(-a_{start}e^{-j\omega n_{start}}+a_0e^{-j\omega(n_{start}+n_0)}\right)$$

where a_{start} and n_{start} are the amplitude and time of the original impulse and a_0 and $n_{start}+n_0$ are the amplitude and time of the new impulse.

The process of modifying $y_{in}[n]$ with $q[n]$ to form a new sequence $y[n]$ is shown schematically in Figure 3-5. The effect of performing a shift and scaling in the time domain becomes a simple addition of the doublet spectrum to the original spectrum $Y_{in}(\omega)$:

$$Y(\omega)=Y_{in}(\omega)+Q(\omega). \tag{3.10}$$

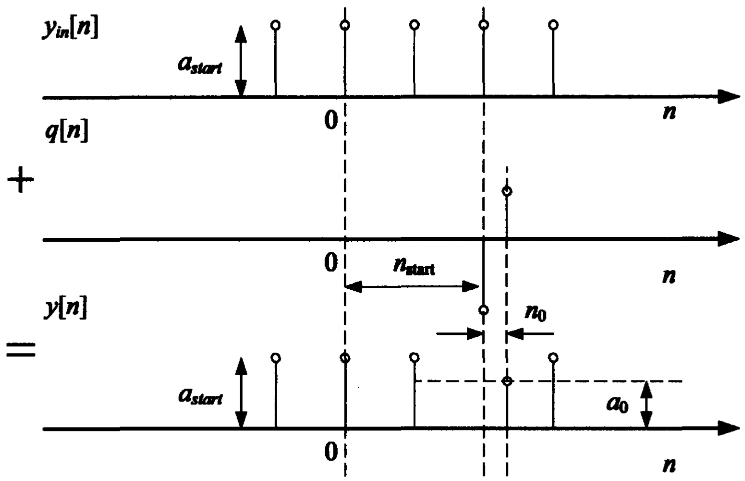


Figure 3-5. Depiction of the addition of a doublet as a means to move and scale an impulse.

Transforming a periodic impulse train into an arbitrary repeating impulse pattern is a special case of modifying a general impulse train. That is, we can add an infinite sum of doublets to the periodic impulse sequence in order to generate the desired pattern. Here, we will demonstrate this process.

To begin, we assume that the input to be modified, $y_{in}[n]$, is periodic with period P_{in}, having the time-domain representation and spectrum:

$$y_{in}[n]=\sum_k \delta[n-kP_{in}] \tag{3.11}$$

$$Y_{in}(\omega)=\sum_k \frac{2\pi}{P_{in}}\delta(\omega-k2\pi/P_{in}).$$

If we wish to move and/or scale every K^{th} impulse in this sequence by the same amount, we can do so by adding a repeating sequence of doublets with period $P=KP_{in}$. We build the repeating sequence of doublets by convolving one period of the doublet, $q[n]$, by a repeating impulse train with period P. The resulting sequence $r[n]$ is

$$r[n]=q[n]*\sum_k \delta[n-kP]. \tag{3.12}$$

Adding $r[n]$ to $y_{in}[n]$ we obtain

$$y[n]=r[n]+y_{in}[n]=q[n]*\sum_k \delta[n-kP]+\sum_k \delta[n-kP_{in}]. \tag{3.13}$$

In order to simplify (3.13) so that it is a function of only the pattern-period P, it can be shown that

$$\sum_k \delta[n-kP_{in}]=v[n]*\sum_k \delta[n-kP] \tag{3.14}$$

where

$$v[n]=\sum_{k=0}^{K-1}\delta[n-kP_{in}]\,. \tag{3.15}$$

That is, a periodic impulse train with period P_{in} in the time domain and harmonics at $2\pi/P_{in}$ in the frequency domain can be viewed as a periodically-repeating sequence with the longer pattern period $P = KP_{in}$.

Using (3.15), the frequency-domain equivalency of $y_{in}[n]$ is written as

$$Y_{in}(\omega)=\sum_k \frac{2\pi}{P_{in}}\delta(\omega-k\frac{2\pi}{P_{in}})=V(\omega)\left(\sum_k \frac{2\pi}{P}\delta(\omega-k2\pi/P)\right) \tag{3.16}$$

where

$$V(\omega)=\sum_{k=0}^{K-1} e^{-j\omega kP_{in}}\,. \tag{3.17}$$

This equivalency says that a periodic set of impulses can always be written in such a way that it has a set of *subharmonics* between *harmonics*, corresponding to a periodic impulse sequence interpreted as having a larger pattern period. The idea of an introduction of subharmonics by timing perturbations was explored by Murphy in his investigation of spectral correlates to perturbation [48].

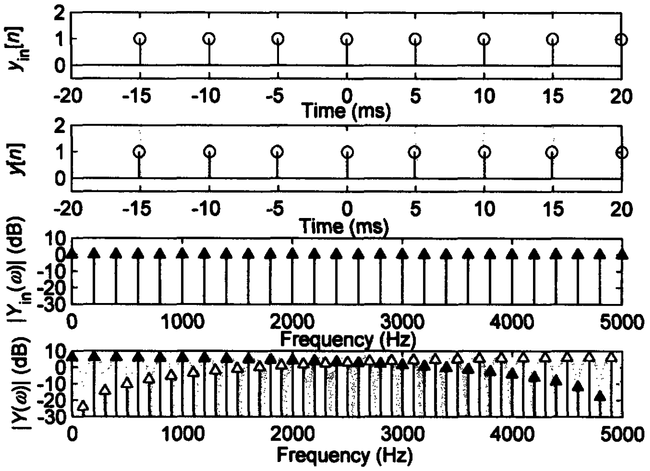


Figure 3-6. Timing modification of a periodic impulse train with period 5 ms by shifting every second impulse one sample, i.e. 0.1 ms, to the left. The grey contour in the bottom panel that shapes the line spectrum is equal to $|M(\omega)|$.

Using the equivalency given in (3.14), we can rewrite (3.13) as

$$y[n]=(q[n]+v[n])*\sum_k \delta[n-kP]\,. \tag{3.18}$$

In the frequency domain, this yields the product:

$$Y(\omega) = \left(Q(\omega) + V(\omega)\right)\left(\sum_k \frac{2\pi}{P}\delta(\omega - k2\pi/P)\right). \tag{3.19}$$

The sum $q[n]+v[n]$ is a case of the function $m[n]$ discussed earlier, which is the sequence over one period of the repeating impulse pattern in time. Recall that $M(\omega)$, the spectrum of $m[n]$, corresponds to the complex spectral envelope which is applied to an underlying periodic line spectrum with spacing $2\pi/P$. In Figure 3-6, Figure 3-7, and Figure 3-8 we give three examples of modification—timing only, amplitude only, and a combination of timing and amplitude—to a periodic impulse train with period 5 ms. As modifications are introduced through the introduction of doublets, $M(\omega)$ changes, becoming compressed or expanded in frequency and/or having dips that become nonzero in the case of amplitude variations. These changes allow the subharmonic frequencies to arise. In particular, we see for these examples how $M(\omega)$ (grey curve) shapes the relative magnitudes of the harmonics (filled triangles) and subharmonics (unfilled triangles).

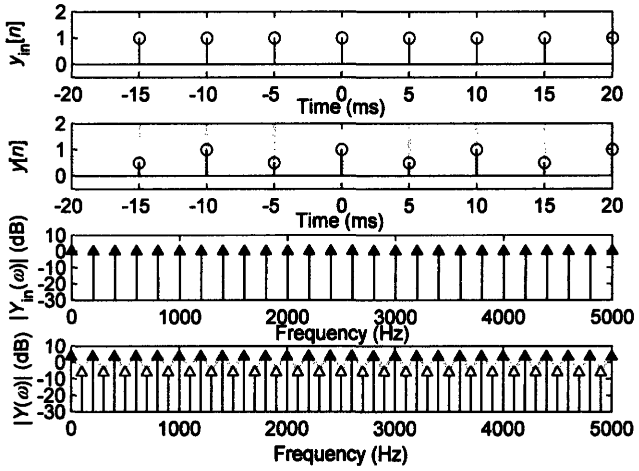


Figure 3-7. Amplitude modification of a periodic impulse train with period 5 ms by scaling every second impulse by 0.5. The grey contour in the bottom panel that shapes the line spectrum is equal to $|M(\omega)|$.

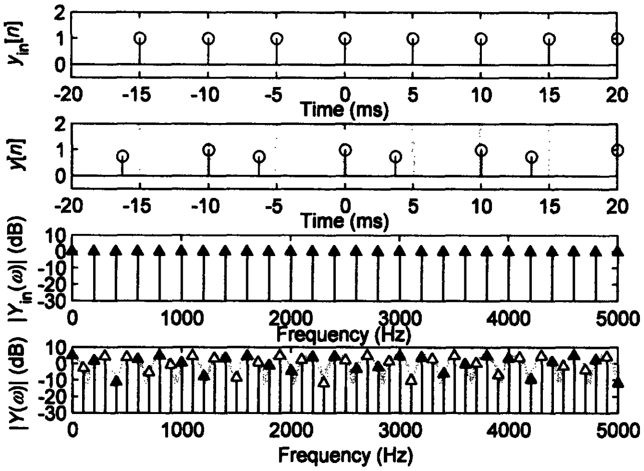


Figure 3-8. Timing and amplitude modification of a periodic impulse train with period 5 ms by shifting every second impulse to the left by 1.3 ms and scaling every second impulse by 0.75. The grey contour in the bottom panel that shapes the line spectrum is equal to $|M(\omega)|$.

For the special case of 2-long (short period, long period, short period…) timing patterns such as the one shown in Fig. 6, $M(\omega)$ has the form

$$\begin{aligned}|M(\omega)| &= |Q(\omega)+V(\omega)| = \left|1+e^{-j\omega(P_{in}+n_0)}\right| \\ &= 2\cos(\omega(P_{in}+n_0)/2)\,.\end{aligned} \tag{3.20}$$

For this special case, we see that the locations of the nulls in $M(\omega)$ are sensitive to the shift of the second impulse, n_0. We can observe this spectral sensitivity in Figure 3-9, which illustrates the effect of moving every second impulse in 0.1 ms increments. We can see that each of these small shifts moves the nulls of $M(\omega)$ enough to drastically change the relative magnitudes of the harmonic and subharmonic components.

Additionally, the examples in Figure 3-9 show different phenomena that can occur due to timing shifts. $Y_1(\omega)$ demonstrates "harmonic shifting," where the prominent line components switch to the subharmonic frequencies, here above 2500 Hz. At larger shifts, such as in $Y_3(\omega)$ and $Y_4(\omega)$, regions where either the harmonics or subharmonics are prominent alternate. Finally, at regions where these sections interface, we can get the appearance of multiple pitches, based on the distances between the line components. For example in $Y_5(\omega)$, we see prominent components spaced by 100 Hz around ω= 500 Hz and spaced by 200 Hz around ω= 1000 Hz.

For comparison, Figure 3-10 shows how changing the amplitude scaling of every second impulse from 0.9 to 0.1 in increments of 0.2 affects the spectrum. Here we see that as the amplitude of the second impulse is reduced, the subharmonic components increase in magnitude. The spectral changes are not qualitatively as drastic as for the timing-shift case.

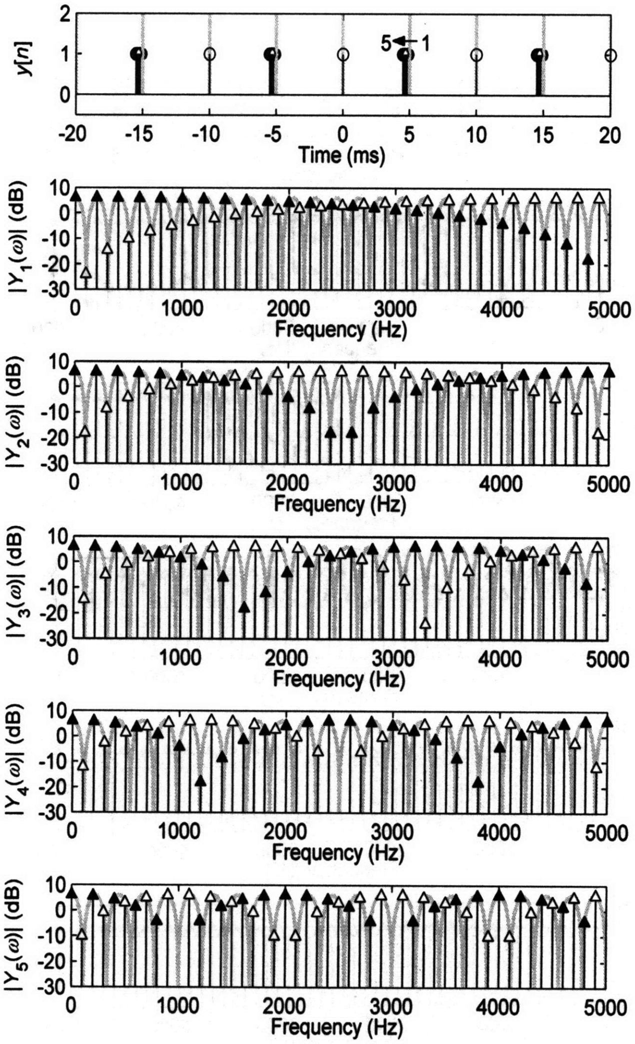


Figure 3-9. Sweep of deterministic timing patterns from every other impulse shifted by 0.1 ms (2nd panel from top) to every other impulse shifted by 0.5 ms (bottom panel) in 0.1 ms increments. The top panel indicates the movement of every second impulse relative to modal (grey vertical lines). Filled triangles indicate the heights of the harmonic components, unfilled triangles indicate the heights of the subharmonic components, and the grey contours that shape the line spectra are equal to $|M(\omega)|$. Observe that the spectrum is sensitive to the shift used.

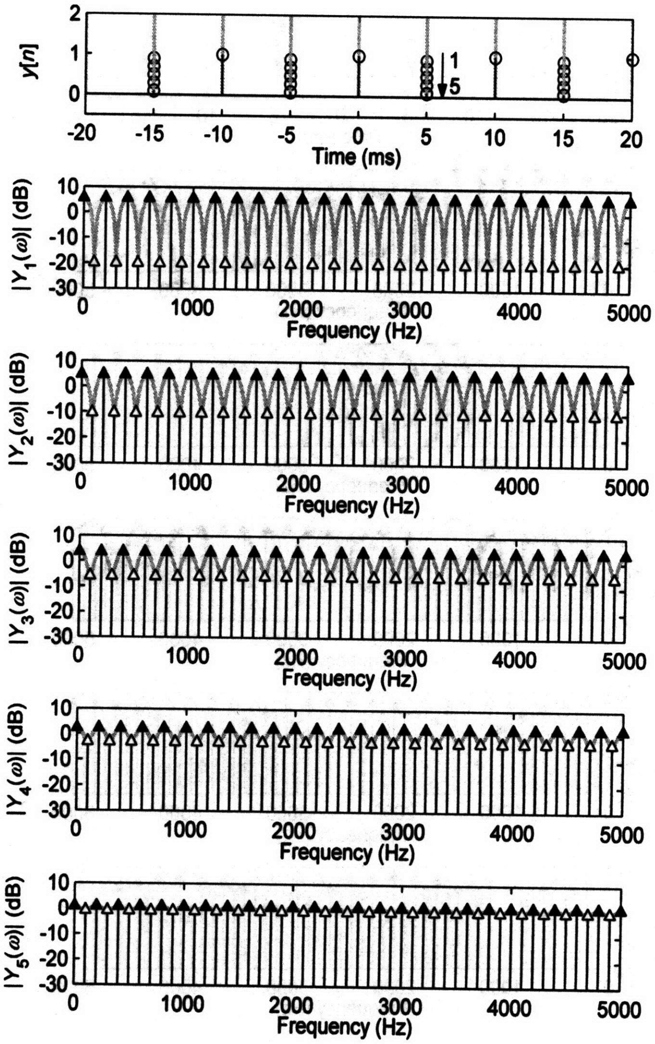


Figure 3-10. Sweep of deterministic amplitude patterns from every other impulse scaled by 0.9 (2nd panel from top) to every other impulse scaled by 0.1 (bottom panel) in increments of 0.2. The top panel indicates the decrease in amplitude of every second impulse. Filled triangles indicate the heights of the harmonic components, unfilled triangles indicate the heights of the subharmonic components, and the grey contours that shape the line spectra are equal to $|M(\omega)|$.

3.2.3 Effects of Windowing on Periodically-Repeating Patterns

Having developed a filtering interpretation of cyclic patterns, we now explore the effect of a short-time window on the spectrum of general perturbations as described by (3.5). It is well-known that windowing in the time domain by the sequence $w[n]$ leads to convolution in the frequency domain:

$$\hat{Y}(\omega) = \frac{1}{2\pi} Y(\omega) * W(\omega) . \qquad (3.21)$$

For a repeating impulse pattern with pattern-period P, we can substitute our expression for $Y(\omega)$, (3.5), to obtain:

$$\hat{Y}(\omega) = \sum_k \left(\frac{1}{P} M(k2\pi/P) W(\omega - k2\pi/P) \right). \tag{3.22}$$

$\hat{Y}(\omega)$ changes depending on the window shape and position, as well as on the length. For changes in position, one way to evaluate fluctuations in this function is to calculate the sample variance of the windowed spectral magnitudes over each of the P different possible window positions:

$$\sigma_{\hat{Y}}^2 = \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \mathrm{var}\left(\left| \hat{Y}(\omega) \right| \right) d\omega. \tag{3.23}$$

We can approximate this integral of the variance using a summation and the discrete Fourier transform (DFT) with length L:

$$\sigma_{\hat{Y}}^2 \approx \frac{1}{L} \sum_{n=0}^{L-1} \mathrm{var}\left(\left| \hat{Y}(n2\pi/L) \right| \right). \tag{3.24}$$

A simulation to calculate this metric as a function of the window length for several different diplophonic impulse patterns, each having an underlying period of 5 ms and a 10-kHz sampling rate, was conducted. This is a special case of the more general set of possible patterns, intended as a useful example. The results using DFT length $L = 8192$ are shown in Figure 3-11, which sweeps over different timing patterns, and Figure 3-12, which sweeps over different amplitude patterns.

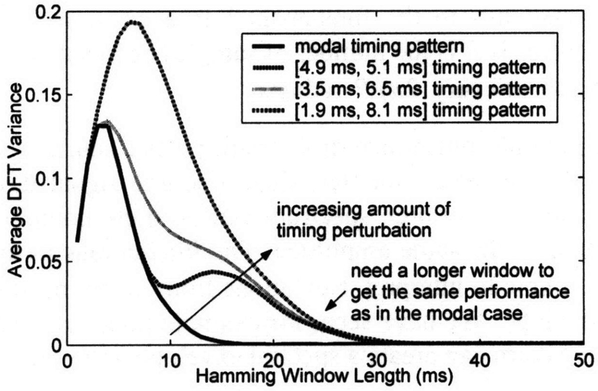


Figure 3-11. Average DFT variance for timing perturbation sweep in 2-long pattern. DFT length $L = 8192$ is used. The notation [first, second] in the legend denotes the periods between the first and second impulses and the second impulse and the start of the next cycle in the 2-long pattern. The sampling rate is 10 kHz. A one-sample shift equals 0.1 ms.

These figures indicate that increasing the amount of perturbation, both in time and amplitude, increases the length of the analysis window necessary to get a spectral representation with a given variance. This jump in variance is particularly apparent in the timing-variation case, where a one-sample shift in one of the samples of $m[n]$ yields a large jump in the average variance in windows between about 10 and 30 ms. A small (10-percent) variation in the amplitude of one of the impulses, on the other hand, causes only a minor fluctuation in the average variance profile, while a 50-percent and 90-percent amplitude variation do yield a large change.

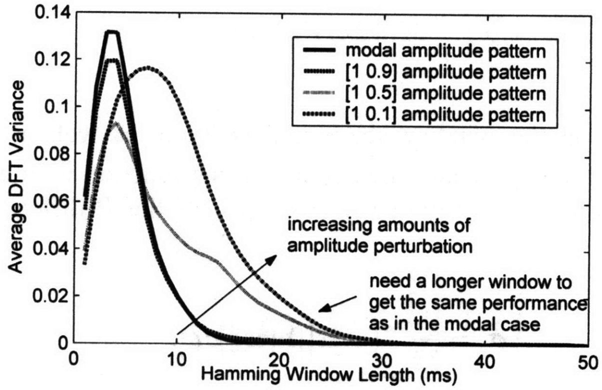


Figure 3-12. Average DFT variance for amplitude perturbation sweep in 2-long pattern. DFT length $L = 8192$ is used. The notation [first, second] in the legend denotes the scaling of the first and second impulse in the 2-long pattern. The sampling rate is 10 kHz.

3.2.4 Harmonic Line Spectra as a Special Case of More General Impulse Behavior

The above analysis has explored how deterministically-structured aspects of nonmodal pulses are reflected in the frequency domain. This perspective helps us to interpret line structure in the excitation component of actual speech spectra in a way that differs from our more standard view.

Conventionally, we expect that the spectrum of a segment of stationary voiced speech will consist of a flat harmonic line spectrum, shaped by the glottal-flow spectrum and by the spectrum of the vocal-tract transfer function. The spacing of the components of the flat underlying harmonic line spectrum corresponds to the fundamental frequency, which reflects the rate at which glottal source pulses occur. If P is the number of samples between neighboring pulses, then we expect a harmonic line to occur every $2\pi/P$ radians.

As we have seen in our periodic pulse-pattern derivations, the source spectrum (without the glottal-shaping contribution) is generally not flat when a source signal deviates from an ideal periodic impulse train. An impulsive-source spectrum can be quite complicated even if all the impulses in the $m[n]$ pattern have the same amplitude. The fluctuations in the spectrum of $m[n]$, $M(\omega)$, in most cases are not simple to interpret. Deviations from modal need not be large in order to significantly affect the envelope. We have seen, for example, that even shifting a single pulse by one sample in an impulse pattern can create a spectral envelope with large peaks and dips.

Additionally, the spacing of the lines in the spectra of nonmodal speech is, in general, *not* related to the apparent rate at which glottal pulses occur. This is an important point because it is often assumed that this is the case. Instead, for periodic patterns, the spacing of the line spectrum reflects the rate at which the *pattern repeats*. The "extra" harmonic lines that occur due to repeating patterns are often referred to in the literature as *subharmonics*. This term, however, can be deceiving. It implies that there is a constant period between time-domain impulses, which is not the case in general. Another common implication is that the spectral lines with large magnitude belong to the "real" harmonics reflecting the period between impulses in the time domain. As we have seen in the modification section, this is not the case in general. For example, in the one-sample shift case, we saw subharmonics having larger magnitude than any of the harmonics at the higher frequencies.

In summary, we have seen that the spacing between spectral lines is related to the rate at which an impulse glottal pattern occurs and not an apparent rate of individual glottal pulses. This *local* pattern shapes a flat underlying spectral line structure. If there is no repeating pattern, then the spectrum will not necessarily exhibit a line structure but will take on a stochastic-like appearance as described in the following section.

3.3 Stochastic Variation in Timing and Amplitude to a Periodic Impulse Train

3.3.1 Stochastic Variation in Timing

Random Process Specifications

The first case we consider is a periodic impulse train with the addition of random timing variation. In particular, we are interested in a perfectly-periodic series of impulses, each with height A and spacing T. Each of the impulses shall be independently perturbed in time by a random discrete number of samples, $\boldsymbol{\lambda}$, with a probability-density function $p_{\boldsymbol{\lambda}}[\lambda]$. We can interpret this random variable as modeling the amount of jitter on each glottal pulse. We call the resulting random process $\boldsymbol{x}[m]$, depicted graphically in Figure 3-13:

$$\boldsymbol{x}[m]=\sum_k A\delta[m-kT-\boldsymbol{\lambda}_k]. \tag{3.25}$$

The subscript k on $\boldsymbol{\lambda}$ indicates the random shift associated with a particular impulse k, and the bold-italic font indicates that $\boldsymbol{\lambda}$ is a random variable. We set $p_{\boldsymbol{\lambda}}[\lambda]$ to 0 for all $\lambda > T/2$ and $\lambda < -T/2$, guaranteeing that the distributions do not overlap.

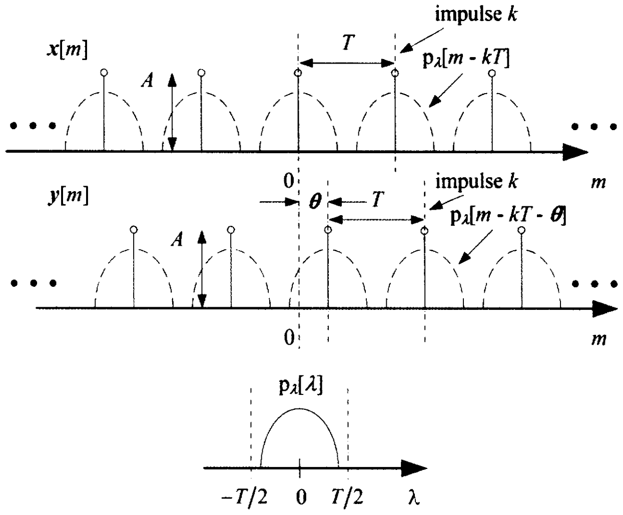


Figure 3-13. Schematic representation of the random processes of impulses perturbed randomly in time. The dashed distributions each represent the extent to which a given impulse can be shifted. θ indicates the random offset of the entire impulse train.

We also randomly offset the sequence of jittered impulses by θ samples, capturing the idea that glottal pulses will not typically fall on the origin. θ is distributed with constant probability $1/T$ over any interval of length T and with probability zero elsewhere. The resulting random process is denoted $y[m]$:

$$y[m] = x[m-\theta] = \sum_k A\delta\left[m - kT - \lambda_k - \theta\right]. \tag{3.26}$$

The random process is shown schematically in Figure 3-13 with the range of possible impulse locations due to the random perturbations, λ_k, drawn with dashed lines.

Computing Autocorrelation

Our derivation begins by writing out the definition of the autocorrelation function of random process $y[m]$ as a function of the sample n and the autocorrelation lag τ written as

$$\mathrm{R}_{yy}[\tau,m] = \mathrm{E}\left[y[m]\,y[m+\tau]\right]. \tag{3.27}$$

Writing this as a double expectation, we obtain

$$\mathrm{R}_{yy}[\tau,m] = \mathrm{E}_{\theta}\left[\mathrm{E}_{x}\left[x[m-\theta]\,x[m-\theta+\tau] \mid \theta\right]\right]. \tag{3.28}$$

In words, this expression finds the expected value of $x[m-\theta]x[m-\theta+\tau]$, averaging over all allowed functions $x[m]$, or sequences λ_k, for each possible offset θ. It then takes the expectation of this conditional expectation over all offset values. The subscript on the expectation operator indicates which random variable is being averaged over.

In calculating the inner expectation of (3.28), the autocorrelation of $\boldsymbol{x}[m]$ for a given offset, we make the substitution $\boldsymbol{u} = m - \boldsymbol{\theta}$. This substitution allows $\boldsymbol{x}[m-\boldsymbol{\theta}]$ to be the function of only the variable $\boldsymbol{u}$ instead of both m and $\boldsymbol{\theta}$:

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}[m-\boldsymbol{\theta}]\boldsymbol{x}[m-\boldsymbol{\theta}+\tau] \mid \boldsymbol{\theta}\right]\\ &= \mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}[u]\boldsymbol{x}[u+\tau] \mid \boldsymbol{u} = m-\boldsymbol{\theta}\right].\end{aligned} \tag{3.29}$$

Using the definition of expectation, we obtain an expression for the inner expectation:

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}[u]\boldsymbol{x}[u+\tau] \mid \boldsymbol{u}\right]\\ &= \sum_{\lambda_0}\sum_{\lambda_1}\cdots\left(x[u]x[u+\tau] \mid \boldsymbol{\lambda}_0 = \lambda_0, \boldsymbol{\lambda}_1 = \lambda_1, \cdots\right)\mathrm{P}\left(\bigcap_k\left(\boldsymbol{\lambda}_k = \lambda_k\right)\right)\end{aligned} \tag{3.30}$$

where P denotes probability. This expression finds the value of the autocorrelation given each of the infinite possible trains, multiplies each possible autocorrelation by the probability of its occurring, and sums over all of these possibilities. We note that there is an infinite set of sums due to the fact that there are, in general, an infinite number of possible pulse trains.

We can simplify this sum by enumerating only the cases for which the product $\boldsymbol{x}[m-\boldsymbol{\theta}]\boldsymbol{x}[m-\boldsymbol{\theta}+\tau]$ is nonzero. This product equals zero except when $\boldsymbol{x}[m-\boldsymbol{\theta}]$ has impulses at both the point u and also at a point τ samples to the right of u, in which case the autocorrelation is equal to the product of the two amplitudes, A^2. We must include every possible pair of impulses and find the probability that one of these impulses occurs at u when the other occurs at $u+\tau$. All of the cases for this occurrence are combined by the union symbol. The resulting simplification of (3.30) is

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}[u]\boldsymbol{x}[u+\tau] \mid \boldsymbol{u}\right]\\ &= A^2\mathrm{P}\left(\bigcup_q\bigcup_r\left(\boldsymbol{\lambda}_q = u - qT\right)\cap\left(\boldsymbol{\lambda}_r = u+\tau-rT\right)\right).\end{aligned} \tag{3.31}$$

Additionally, recall that only one impulse may occur at any particular point u—we do not allow the jitter imposed on one impulse to exceed the range $-T/2 < \lambda < T/2$ and overlap the neighboring impulse's range. This allows us to write (3.31) as the sum

$$\begin{aligned}&\mathrm{E}_{\boldsymbol{x}}\left[\boldsymbol{x}[u]\boldsymbol{x}[u+\tau] \mid \boldsymbol{u}\right]\\ &= \delta[\tau]\sum_k A^2\mathrm{p}_{\lambda}[u-kT]\\ &+\sum_k\sum_{r\neq k} A^2\mathrm{p}_{\lambda}[u-kT]\,\mathrm{p}_{\lambda}[u+\tau-rT].\end{aligned} \tag{3.32}$$

The first term of this expression is a scaled impulse at $\tau = 0$, representing the fact that when $q = r$ in (3.31), both indices are referring to the same impulse, implying that $\lambda_q = \lambda_r$ and $\tau = 0$. The second term covers all the other probabilities that an impulse will occur at u concurrently with another impulse occurring at $u+\tau$. Each jitter value, $\boldsymbol{\lambda}_k$, is assumed to be independent of the others, allowing us to write the intersection in this term as a product.

We now have an expression for the inner expectation, but need to substitute it into (3.28). We first rewrite (3.28) in terms of the variable u:

$$\begin{aligned}\mathrm{R}_{yy}[\tau,m] &= \mathrm{E}_{\theta}\left[\mathrm{E}_{x}\left[x[m-\theta]x[m-\theta+\tau] \mid \theta\right]\right]\\ &= \mathrm{E}_{r\le(m-\theta)<r+T}\left[\mathrm{E}_{x}\left[x[u]x[u+\tau] \mid u=m-\theta\right]\right]\\ &= \mathrm{E}_{r\le u<r+T}\left[\mathrm{E}_{x}\left[x[u]x[u+\tau] \mid u\right]\right]\end{aligned} \tag{3.33}$$

where r is any integer. Recall that we have defined θ, and therefore u, to be uniformly distributed over any interval of length T.

We may now choose an arbitrary range of T contiguous values of u over which to evaluate the expected value of (3.32). We will see that a convenient choice is $-T/2 \le u < T/2$. Using this range of u, we may ignore all terms including $A^2\mathrm{p}_{\lambda}[u-kT]$ in (3.32) for which k is not equal to zero. We can now substitute (3.32) into (3.33) resulting in

$$\begin{aligned}\mathrm{R}_{yy}[\tau] &= \mathrm{E}_{-T/2\le u<T/2}\left[A^2\mathrm{p}_{\lambda}[u]\delta[\tau]\right]\\ &+\mathrm{E}_{-T/2\le u<T/2}\left[\sum_{r\neq 0} A^2\mathrm{p}_{\lambda}[u]\mathrm{p}_{\lambda}[u+\tau-rT]\right].\end{aligned} \tag{3.34}$$

The resulting autocorrelation, depicted schematically in Figure 3-14, is:

$$\mathrm{R}_{yy}[\tau] = \frac{A^2}{T}\delta[\tau] + A^2\varphi[\tau] * \sum_{r}\delta[\tau-rT] - A^2\varphi[\tau], \tag{3.35}$$

where

$$\varphi[\tau] = \frac{1}{T}\sum_{u=-T/2}^{T/2-1} \mathrm{p}_{\lambda}[u]\,\mathrm{p}_{\lambda}[u+\tau] \tag{3.36}$$

from which we observe that $\varphi[\tau]=0$ for $|\tau|>T$.

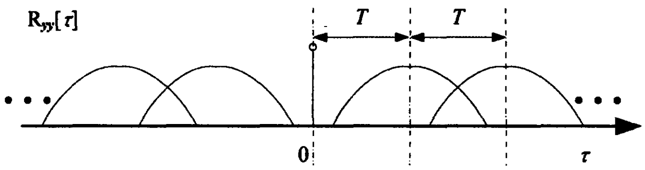


Figure 3-14. Illustration of the autocorrelation function, $\mathrm{R}_{yy}[\tau]$, of a periodic impulse train when each impulse is randomly perturbed. Each lobe is a shifted rendition of $\varphi[\tau]$ as described in (3.35).

By a similar argument to the above, we can show that the mean of $y[m]$ is equal to

$$\mathrm{E}[y[m]] = \mathrm{E}_{r\le u<r+T}\left[\mathrm{E}_{x}\left[x[u] \mid u\right]\right] = A/T. \tag{3.37}$$

Since the values of both $\mathrm{R}_{yy}[\tau,m]=\mathrm{R}_{yy}[\tau]$ and $\mathrm{E}[y[m]]=\text{constant}$ do not depend on *m*, this process has both a stationary first and second-order moment and is, therefore, wide-sense stationary. This property is important since the power-spectrum is only meaningful for a wide-sense stationary process.

Power Spectral Density

We can compute the power spectral density as the sum of three different components. Each of these corresponds to the Fourier transform of one of the terms in the autocorrelation (3.35) as shown in Table 3-1. From the table, component A, which is periodic in the autocorrelation domain, contributes a line-spectrum component to the power spectral density. Each spectral line is found at the location $\omega=2\pi k/T$, and $\Phi(\omega)$, the Fourier transform of $\varphi[\tau]$, shapes the line-components. For the shape of the lobes depicted in Figure 3-14, we obtain a low-pass characteristic. The cutoff-frequency of this low-pass effect becomes higher as the lobes in the autocorrelation domain become narrower. We will refer to this term as the *low-pass line spectrum.*

Table 3-1. Corresponding components of the autocorrelation and power spectral density

	Autocorrelation	Power Spectral Density
A	$A^2\varphi[\tau]*\sum_r \delta[\tau-rT]$	$A^2\Phi(\omega)\sum_k \frac{2\pi}{T}\delta(\omega-k2\pi/T$
B	$\frac{A^2}{T}\delta[\tau]$	A^2/T
C	$-A^2\varphi[\tau]$	$-A^2\Phi(\omega)$

Components of the autocorrelation function and power spectral density. Each power-spectral density component is found by taking the Fourier transform of the component in the autocorrelation domain. Component *A* is known as the low-pass line spectrum and the sum of components *B* and *C* are called the high-pass noise floor.

Component B of the autocorrelation yields a continuous flat contribution, A^2/T, in the power spectral density. As the distribution $\mathrm{p}_\lambda[\lambda]$ widens, the line-spectrum components will approach this noise floor. Likewise, component C of the autocorrelation contributes a continuous element, $A^2\Phi(\omega)$, to the power spectral density but has a low-pass shape instead of being flat. When this low-pass component is subtracted from the flat noise floor, it effectively yields a noise floor with a high-pass characteristic. We will call this difference the *high-pass noise floor*. The zero at $\omega=0$ in the resulting term comes about since $\Phi(0)=1/T$, which follows from (3.36). The sum of these three components results in $\mathrm{S}_{yy}(\omega)$, depicted schematically in Figure 3-15:

$$\mathrm{S}_{yy}(\omega)=A^2\Phi(\omega)\sum_k\frac{2\pi}{T}\delta(\omega-k2\pi/T)+\left(A^2/T-A^2\Phi(\omega)\right). \qquad (3.38)$$

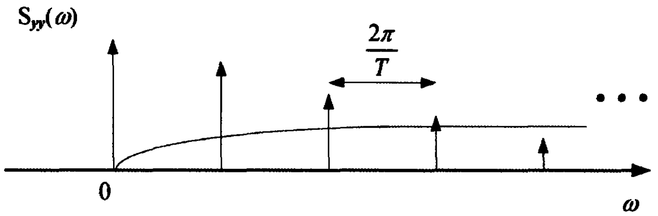


Figure 3-15. Schematic of the power-spectral density of a periodic impulse train perturbed randomly in time. Note that the dirac-delta functions are not drawn to scale with reference to the continuous noise floor.

In summary, we expect two major components in the power spectral density of a jittered impulse sequence—a high-pass noise floor and a low-pass line spectrum. We can confirm this empirically by computing the power-spectral density of an impulse train with random perturbation. Figure 3-16 shows the impulse train, autocorrelation function, and power-spectral density of an impulse train perturbed by the discrete distributions illustrated. The maximum perturbation for the case using a wide distribution (thick line) is ±0.5 ms while the maximum perturbation for a narrow distribution (thin line) is ±0.1 ms. Observe that the prediction of a low-pass harmonic spectrum and high-pass noise spectrum is consistent with the figure. As the amount of perturbation grows smaller, the higher-frequency harmonics become more prominent. One difference between the derivation and the figure is that the figure was generated using windowed signals, a technique known as the Welch average modified-periodogram method. It is known that this technique leads to a biased estimate of the PSD and approaches the true PSD as the window length increases (see for example [49] pp.733-737).

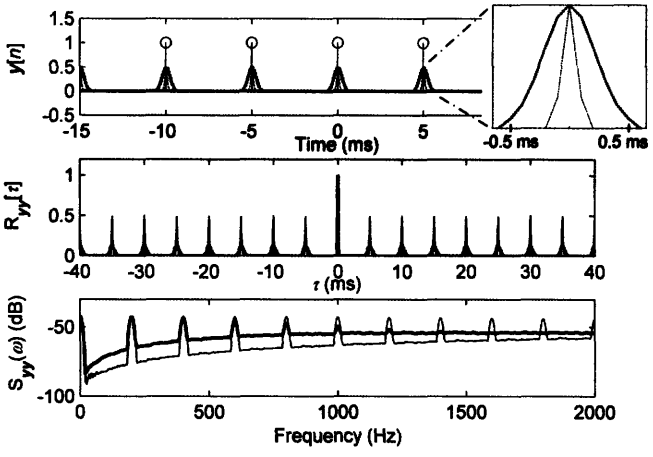


Figure 3-16. This figure shows a 5-ms-period periodic impulse train with two different normalized pulse shift distributions superimposed, as detailed in the inset. The thick line corresponds to a maximum deviation of ±0.5 ms; the narrow line to ±0.1 ms. 10 seconds of synthesized signal were used. Analysis was performed with the Welch average modified-periodogram method using a 1024-point Hamming window, with 512 points of overlap.

3.3.2 Stochastic Variation in Amplitude

We now modify the derivation for timing variation to include amplitude variation. We assume that the amplitude and timing perturbations are independent from one another. As in our previous random process (3.26) the mean impulse amplitude is A and we add amplitude perturbation, α, having a continuous distribution $p_\alpha(\alpha)$ and zero mean.

Using a derivation similar to the timing-perturbation case, it can be shown that we obtain the autocorrelation function shown in (3.39) and the PSD in (3.40). These expressions handle any combination of *both* timing and amplitude perturbations.

$$\mathrm{R}_{yy}[\tau] = \frac{\left(A^2 + \sigma_a^2\right)}{T}\delta[\tau] + A^2\varphi[\tau] * \sum_r \delta[\tau - rT] - A^2\varphi[\tau] \tag{3.39}$$

$$\begin{aligned}\mathrm{S}_{yy}(\omega) &= A^2\Phi(\omega)\sum_k \frac{2\pi}{T}\delta(\omega - k2\pi/T) \\ &+\left(\left(A^2 + \sigma_a^2\right)/T - A^2\Phi(\omega)\right)\end{aligned} . \tag{3.40}$$

We can conclude from the derived PSD that the effect of random amplitude perturbation adds linearly in the PSD domain to the PSD previously derived for the time-perturbation-only case. In the terms of traditional speech-perturbation analysis, this may be stated that the effect of shimmer adds linearly to the jittered PSD. This additional effect is always in the form of a flat additive noise floor having energy σ_a^2/T.

For the random-amplitude perturbation only case, we have $\varphi[\tau] = \delta[\tau]/T$ and $\Phi(\omega) = 1/T$. This yields a PSD of

$$A^2/T\sum_k \frac{2\pi}{T}\delta(\omega - k2\pi/T) + \sigma_a^2/T . \tag{3.41}$$

The power-spectral density for an amplitude-perturbation-only case with uniformly-distributed perturbation from 0.1 to 1.9 is shown in Figure 3-17.

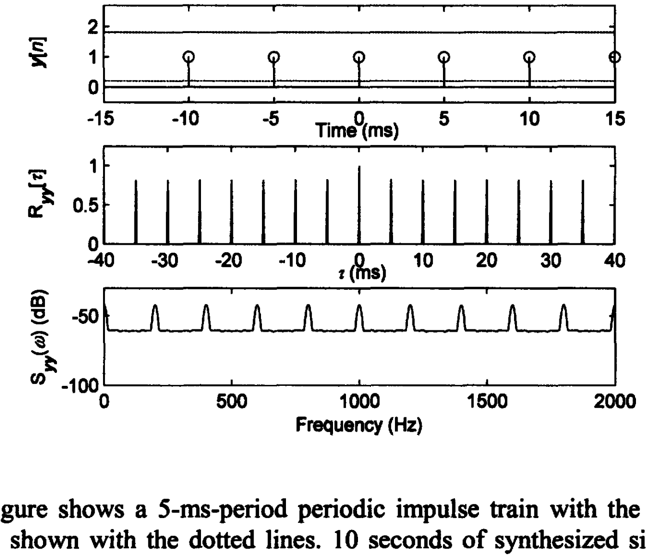


Figure 3-17. This figure shows a 5-ms-period periodic impulse train with the minimum and maximum amplitude deviations shown with the dotted lines. 10 seconds of synthesized signal were used. Analysis was performed with the Welch average modified-periodogram method using a 1024-point Hamming window, with 512 points of overlap.

3.4 Extension to Real Speech Cases

Thus far we have focused on nonmodal impulse trains. In this section, we will argue that our spectral models of nonmodality are applicable to real speech signals.

We model the generation of natural speech using the source-filter model of speech production. This model consists of a volume-velocity *source* waveform, filtered by both an all-pole vocal tract filter and a radiation characteristic at the mouth to produce an acoustic pressure signal. As an additional step, each source pulse can be modeled as a pure impulse source convolved with a mixed-phase *source response* [56]. The input to this model is a series of impulses, each characterized by a time of occurrence and amplitude. Speech is generated according to this model by filtering the impulses by vocal tract and source responses.

Based on our derivation of the spectrum of nonmodal impulse trains, we can derive the spectrum of any statically-filtered impulse train simply by performing a multiplication in the frequency domain. This process shapes the line spectrum. Natural speech also contains time-varying filters which are important but beyond the scope of this chapter. Another issue that is beyond this chapter but very relevant is the influence of time-varying pitch contours.

In addition to the linear filtering of the speech-production system, the effects of the analysis window are also different than on the impulse excitation alone. We have shown that windowing of an impulse train can be described by (3.22). With the addition of the vocal tract transfer function, denoted by $H(\omega)$, this expression is extended to become

$$\hat{Y}(\omega) = \sum_k \left(\frac{1}{P} M(k2\pi/P) H(k2\pi/P) W(\omega - k2\pi/P) \right). \tag{3.42}$$

One of the lessons of our development is that the spectrum is sensitive to the exact locations and amplitudes of source impulses. Real speech is complicated in that it contains combinations of random and deterministic impulse patterns. It also contains time-varying vocal tract and source filtering as mentioned above, as well as aspiration and frication noise components.

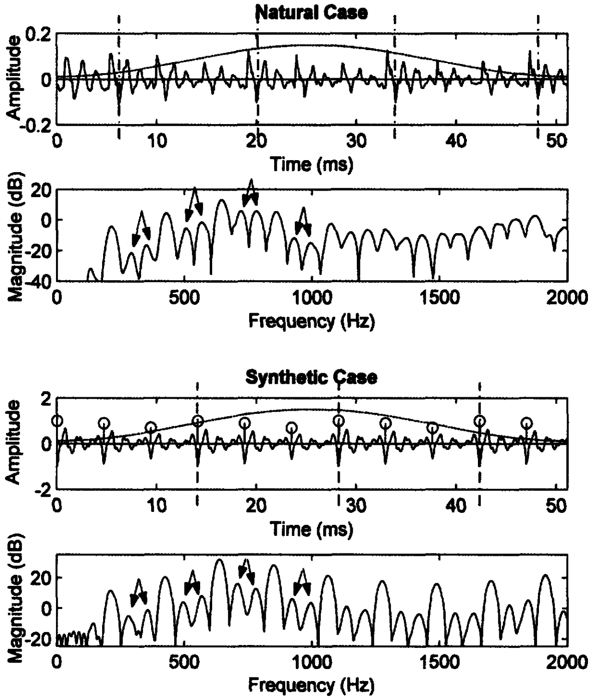


Figure 3-18. Waveform, and short-time spectrum of a section of natural and synthetic nonmodal speech displaying triplophonia—a repeating pattern of three pulses. Vertical dashed lines highlight each cycle of repetition. Observe the pairs of depressed harmonics in both the natural and synthetic cases. The Hamming window is 51.2-ms long and the sampling rate is 10000 samples/s.

We can show, nevertheless, qualitative agreement between certain aspects of spectra of natural speech and our model spectra. Figure 3-18 shows the waveform and short-time spectrum for a section of triplophonic speech produced by a female talker from the TIMIT database. The speech sound shown is /æ/ from the word "that." Recall that triplophonia is a repeating pattern of three pulses—here these pulses decrease in amplitude over the period of the pattern. We use the impulses shown in the third panel of Figure 3-18, together with the KLGLOTT88 glottal-pulse model [37] and a Klatt-like formant synthesizer similar to KLSYN88 [37], to model this behavior. Here, the natural and synthetic utterances have similar behaviors in the lower frequencies—up to about 1200 Hz. Both spectral line components exhibit a series of two subharmonics between each pair of harmonics. Higher frequencies suggest a raised noise floor in the natural speech due to timing and other time-varying fluctuations.

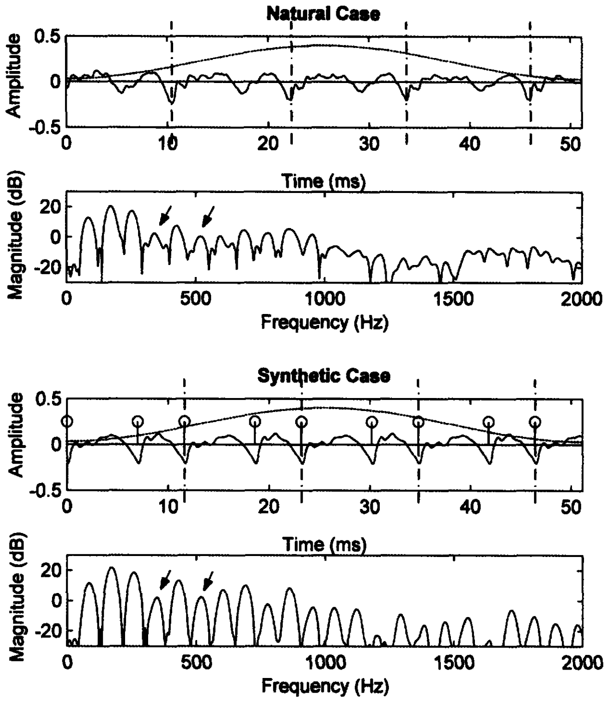


Figure 3-19. Waveform and short-time spectrum of a section of nonmodal speech displaying diplophonia—one period followed by a shorter one. Vertical dashed lines highlight each cycle of repetition. Observe the harmonics with lowered amplitude in both the natural and synthetic cases. The Hamming window is 51.2-ms long and the sampling rate is 10000 samples/s.

In Figure 3-19, we observe a case of diplophonia produced by a female speaker from a database of TIMIT sentences recorded in clean conditions. As in the previous case, the speech sound shown is /æ/ from the word "that." Here, we have a repeating pattern of two impulses, one period being shorter than the other. We use the impulses shown in the third panel of Figure 3-19 to model this behavior. The natural and synthetic speech have similar behaviors in the lower frequencies. The harmonics follow a similar pattern of relative amplitudes, the depressed harmonics indicated by the arrows being especially clear examples.

To help illustrate the stochastic spectral model, we present Figure 3-20, which shows the estimated power-spectral density for 500 ms of a sustained vowel /ɑ/ produced by a male speaker. With the caveat that other influences such as changes in pitch and additive wideband speech noise can affect the power-spectrum, we note that there are clear similarities between the synthesized case and the real-speech case. In the synthesized case, we introduced a random perturbation equal to 0.6-percent of one period. Observe that the higher-frequency harmonics of both the natural and perturbed synthetic speech grow smaller relative to the noise floor.

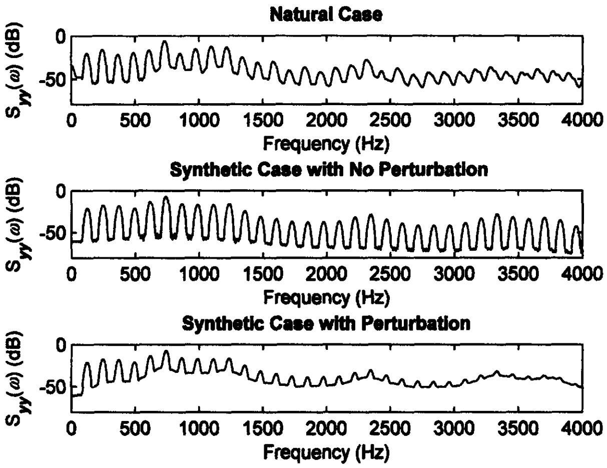


Figure 3-20. Comparison of the estimated PSD for 500 ms of natural speech (top), synthetic modal speech (middle), and synthetic speech with ±0.5-ms random perturbation (bottom). Analysis was performed with the Welch average modified-periodogram method using a 512-point Hamming window, with 256 points of overlap. Both the natural and synthetic case with perturbation have a raised "noise floor" as frequency increases.

3.5 Conclusion

In this chapter, we have derived the spectra of both deterministic and stochastic nonmodal impulse trains. We have also argued that such relations can be applied to the analysis of natural speech. The principal contribution of this chapter is an analytical connection between temporal patterns of nonmodality in speech and their spectral characteristics. For deterministic patterns, we have shown that there is a complex envelope function that shapes the underlying line components of speech. This filtering interpretation showed the alterations of harmonics and the evolution of new "sub-harmonics". Using doublet sequences, we also investigated spectral sensitivity, showing that small changes in the locations and amplitudes of impulses can significantly alter the shape of the spectrum. For the stochastic case, we have shown that random impulse timing and amplitude variations lead to a continuous high-pass "noise floor" PSD component summed with a low-pass series of harmonic lines. We used in this case an autocorrelation-based derivation to find the spectral density. There may be other ways to approach this problem, such as using the theory of point-processes [7].

Chapter 4

The Fundamental Frequency of Nonuniform Glottal Events

In the last chapter, we presented a view of the spectrum resulting from nonuniform impulse trains. Another common perspective for describing the dynamics of pulses is using the concept of pitch and its acoustical correlate, the fundamental frequency. For a periodic waveform, the fundamental frequency is usually described as a constant, equal to the reciprocal of the period of the glottal cycle in time and also equal to the spacing of the harmonically-related peaks in the spectrum. For slight changes in the period of the glottal cycles, the fundamental frequency becomes a time-varying function. As the spacing of the cycles grows smaller, the fundamental increases, and as the spacing of the cycles grows larger, the fundamental decreases. In the short-time frequency domain, the local fundamental frequency is approximately equal to the spacing of the peaks.

It would be useful to be able to apply this time-varying fundamental frequency description to speech having large nonuniformities in its underlying glottal cycles. Pitch and fundamental frequency are common ways to describe dynamics of the source, and previous researchers have made it a goal to determine this fundamental frequency for aperiodic signals [25]. How one can describe irregular phonation using a fundamental-frequency description, however, is not clear. The term fundamental frequency *implies harmonicity*, but, as we have seen in the last chapter, such harmonicity is not present in the spectrum of nonuniform glottal cycles over short times and with regard to the standard Fourier transform. *In this chapter, our goal is to model the fundamental frequency of such nonmodal sequences of glottal events.*

In order to arrive at a fundamental-frequency description, we introduce a time-varying harmonically-related sinusoid representation of the speech source. Such a model for speech excitation has been proposed by Schoentgen [63] in which the frequencies and amplitudes of a set of harmonically-related sinusoids are modulated in order to obtain perturbations of speech excitation signals. This domain of time-varying frequencies is very different than the domain of stationary Fourier basis functions discussed in the last chapter.

In this chapter, we show an important relationship between an impulse train and the instantaneous frequency of its harmonic components. This relation helps us to better understand the meaning of the term fundamental frequency by showing that there exists an infinite number of possible instantaneous frequency tracks that underlie a given impulse train. As an offshoot of this work, this property also contributes to understanding certain modifications of speech with nonuniform glottal events that occur by sinewave analysis and synthesis [44], specifically the addition, deletion, and movement of pulses. The representation discussed in this chapter is relevant to sinewave analysis-synthesis since it uses the same model.

Section 4.1 of this chapter provides conditions on the fundamental sinewave phase to maintain the impulsiveness of harmonic signals. Section 4.2 then focuses on the problem of uniqueness and shows that different frequency and phase trajectories can result in the same impulse train, for both modal and nonmodal cases. The converse problem, where measurements of the underlying sinewave representation can yield different impulse trains, is also introduced. Examples are given illustrating both the issue of uniqueness and its converse. Section 4.3 describes theoretical elements related to filtering the sinusoidal representation of impulsive excitation. This section is important since we have represented real speech as a series of impulsive glottal-excitation events filtered by the vocal tract. Finally, in Section 4.4, we extend our results to its implications for one application, sinusoidal analysis-synthesis.

4.1 Condition for Impulsiveness

This section presents conditions that relate a series of continuous-time impulses to a sinusoidal model consisting of harmonically-related sinusoidal components.

4.1.1 Definitions

The class of signals that we consider in this chapter involves harmonically-related instantaneous frequencies. At a given time, *t*, each harmonic, denoted by the index *k*, has the instantaneous frequency

$$\dot{\theta}_k(t) = k\dot{\theta}_o(t) \tag{4.1}$$

where $\dot{\theta}_o(t)$ is termed the *instantaneous fundamental frequency*. Using this definition, we can find the instantaneous phase of each harmonic component by integrating (4.1), yielding

$$\theta_k(t) = \int_{-\infty}^{t} k\dot{\theta}_o(\tau)d\tau = k\int_{-\infty}^{t} \dot{\theta}_o(\tau)d\tau = k\theta_o(t). \tag{4.2}$$

The output signal, *s*(*t*), created using the sum of an infinite number of such components, each having amplitude C_k, is

$$s(t) = \sum_{k=1}^{\infty} C_k \cos(\theta_k(t)) = \sum_{k=1}^{\infty} C_k \cos(k\theta_o(t)). \tag{4.3}$$

If we further restrict all C_k to be equal to *C*, we obtain

$$s(t) = C\sum_{k=1}^{\infty} \cos(k\theta_o(t)). \tag{4.4}$$

For a constant fundamental frequency, $\dot{\theta}_o(t) = f_o$, it is clear that (4.4) gives a periodic series of impulses. With time-varying fundamental frequency, on the other hand, it is not clear if (4.4) yields a series of impulses and, if so, where the impulses lie in time. In the following section, we show that indeed (4.4) is always an impulse train (generally nonmodal) and we will provide a condition for where in time the impulses occur.

4.1.2 Condition

We now show that for any phase $\theta_o(t) \neq 2\pi m$, where m is an integer, our time function (4.4) is guaranteed to be zero. We will thus show that (4.4) always yields a train of impulses, with m being the positive or negative index for the mth impulse in the series. That is, our objective is to show:

$$s(t) = C\sum_{k=1}^{\infty}\cos(k\theta_o(t)) = 0 \quad \text{for} \quad \theta_o(t) \neq 2\pi m \tag{4.5}$$

where m is any integer. This can be considered a condition of impulsiveness.

Equation (4.4) can be interpreted as a mapping of each possible pair of values $(\theta_o(t), C)$ at a given time t to an output signal $s(t)$. In other words,

$$s(t) = f(\theta_o(t), C). \tag{4.6}$$

It is well known [50] that the periodic impulse train

$$s(t) = \alpha \sum_{m=-\infty}^{+\infty} \delta(t - m/f_o), \tag{4.7}$$

where f_o is the fundamental frequency and α is a nonzero scale factor, can be written as a Fourier series

$$s(t) = C\sum_{k=1}^{\infty}\cos(k2\pi f_o t), \tag{4.8}$$

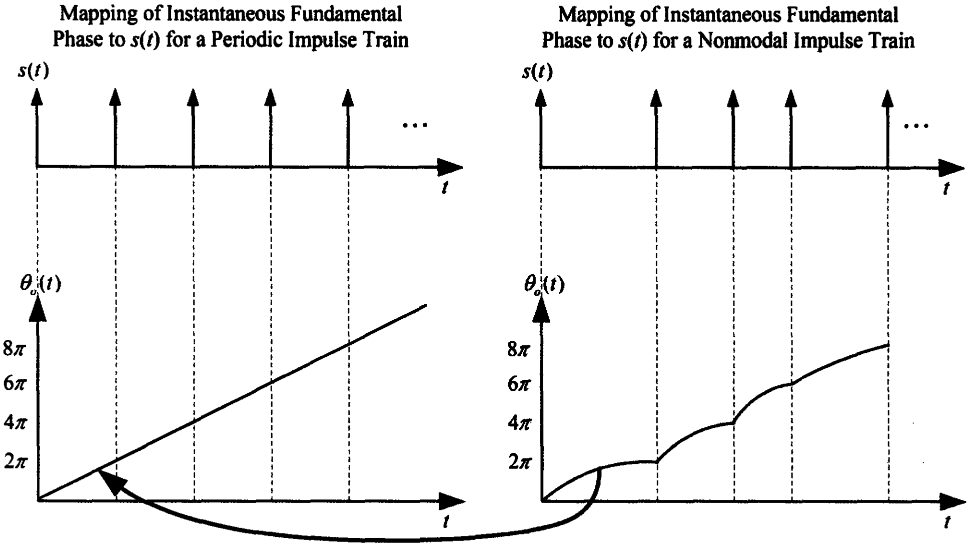
where C is a scale factor proportional to α.[2]

This specific function captures the relationship between any $s(t)$, $\theta_o(t)$, and C because the term $2\pi f_o t$ covers all possible values of phase. We see that $s(t)$ is only nonzero when the phase is equal to a multiple of 2π. Therefore, the function described by (4.6) may be rewritten

$$s(t) = f(\theta_o(t), C) \propto \begin{Bmatrix} 0 & \text{for } \theta_o(t) \neq 2\pi m \text{ where m} \subseteq \mathbb{Z} \\ \delta(0) & \text{for } \theta_o(t) = 2\pi m \text{ where m} \subseteq \mathbb{Z} \end{Bmatrix} \tag{4.9}$$

where $\mathbb{Z}$ indicates the set of all integers. Another way to think about the mapping in (4.9) is that for $\theta_o(t)$ not equal to a multiple of 2π, one can always find a corresponding phase in (4.8) that gives a zero sum. This mapping is shown in Figure 4-1.

[2] The Fourier series also includes a constant DC offset term equal to $C/2$ which will be ignored in the derivation since it is not dependent on the instantaneous phase.



Mapping of Instantaneous Fundamental Phase to $s(t)$ for a Periodic Impulse Train

Mapping of Instantaneous Fundamental Phase to $s(t)$ for a Nonmodal Impulse Train

For $\theta_o(t)$ not equal to a multiple of 2π, one can always find a corresponding phase in the periodic impulse train case that gives a zero sum

Figure 4-1. Mapping of instantaneous fundamental phase in a nonmodal impulse train to phase in a periodic impulse train. An impulse occurs only if $\theta_o(t)$ is a multiple of 2π. Otherwise, the value of $s(t)$ will be zero. The thick arrow indicates the process of finding a corresponding phase in the periodic impulse train case that gives a zero sum.

4.2 Uniqueness

4.2.1 Finding the Instantaneous Frequency from an Observed Impulse Train

Given the above result, we now strive to find a set of functions $\dot{\theta}_o(t)$ that yield the phase $\theta_o(t) = 2\pi m$ where $m \subseteq \mathbb{Z}$ at the times of an observed set of impulses (and only at these points). Let us write the function for $s(t)$ in terms of the instantaneous frequency by combining (4.2) and (4.4):

$$s(t) = C\sum_{k=1}^{\infty}\cos(k\int_{-\infty}^{t}\dot{\theta}_o(t)dt). \tag{4.10}$$

Suppose that an impulse occurs at known time $t = t_a$ so that $\theta_o(t)$ at time t_a is equal to an integer multiple of 2π. We thus rewrite $s(t)$ with reference to this time instant:

$$s(t) = C\sum_{k=1}^{\infty}\cos(k\int_{t_a}^{t}\dot{\theta}_o(t)dt) \quad \text{for } t > t_a. \tag{4.11}$$

In order to simplify our expressions and not allow negative frequencies, we constrain $\dot{\theta}_o(t)$ to be positive for all t. When $s(t)$ becomes nonzero after time t_a, we have reached the next impulse at a time t_b, and the following equation must hold:

$$\int_{t_a}^{t_b} \dot{\theta}_o(t)dt = 2\pi \quad \text{for } t_b > t_a \tag{4.12}$$

because at t_b we must have another multiple of 2π and no such multiple can occur within (t_a, t_b). Given this relationship, it is interesting to consider the value of the instantaneous frequency in this region if we constrain it to be constant, $2\pi f_o$. Given that there are impulses at times t_a and t_b, yielding *period* $T_o = t_b - t_a$, we obtain:

$$\begin{gathered} 2\pi f_o(t_b - t_a) = 2\pi \\ f_o = 1/T_o \end{gathered} . \tag{4.13}$$

This matches the conventional definition of the fundamental frequency being the reciprocal of the period.

We may also allow $\dot{\theta}_o(t)$ to vary with time, exhibiting frequency modulation. If we choose t_b ahead of time and solve for all the possible $\dot{\theta}_o(t)$ functions that yield an impulse at time t_b, we see that there are an infinite number and that they are constrained only by:

$$\int_{t_a}^{t_b} \dot{\theta}_o(t)dt = 2\pi \quad \text{where} \quad \dot{\theta}_o(t) > 0 \quad \text{for} \quad t_a < t < t_b . \tag{4.14}$$

One can model the time-varying $\dot{\theta}_o(t)$ between impulses *a* and *b* in a number of ways. One typical way might be to assume that it is the sum of a constant component with period $(t_b - t_a)$ and a time-varying component:

$$\dot{\theta}_o(t) = 2\pi/(t_b - t_a) + \dot{\theta}_{FM}(t) \quad \text{for} \quad t_a < t < t_b . \tag{4.15}$$

If this is the case, then we can solve for $\dot{\theta}_{FM}(t)$ resulting in:

$$\int_{t_a}^{t_b} \dot{\theta}_{FM}(t)dt = 0 \quad \text{where} \quad \dot{\theta}_{FM}(t) > -2\pi/(t_b - t_a) \quad \text{for} \quad t_a < t < t_b . \tag{4.16}$$

The lower limit on the value of $\dot{\theta}_{FM}(t)$ ensures that $\dot{\theta}_o(t)$ remains positive for all *t* when the frequency-modulated part is summed with the constant-frequency component.

4.2.2 Constructing Instantaneous Frequency for an Observed Impulse Train

Using the constraints set by (4.15) and (4.16), we can construct many different valid instantaneous fundamental-frequency tracks. This section describes this process and presents several examples as illustrations.

An intuitive way to construct an instantaneous fundamental-frequency track for an observed pulse train is to simply make the frequency constant between each set of impulses, equal to the reciprocal of the time between impulses. That is, we can set the frequency-modulated part of (4.15) to zero, yielding

$$\dot{\theta}_o(t) = 2\pi/(t_b - t_a) \quad \text{for} \quad t_a < t < t_b, \tag{4.17}$$

which satisfies constraint (4.12). Figure 4-2 shows a schematic of a fundamental-frequency track obtained in this way for a series of nonmodal impulses.

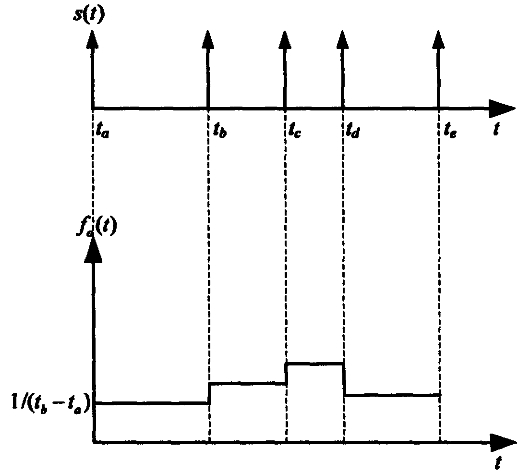
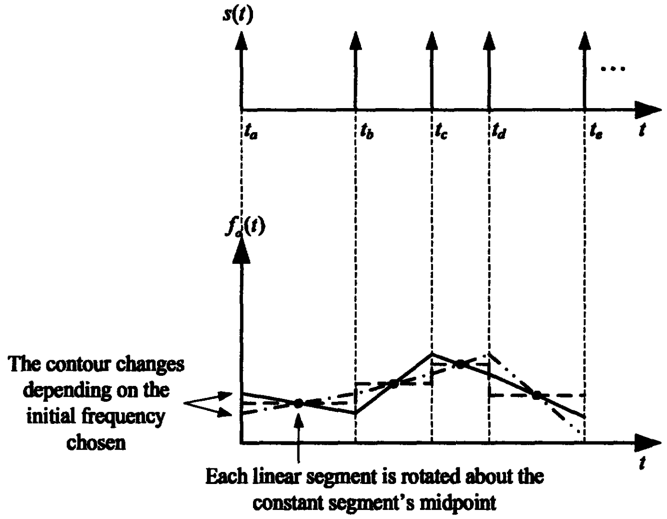


Figure 4-2. Construction of the piecewise-constant instantaneous fundamental-frequency track for a given impulse train $s(t)$. The fundamental frequency between every pair of impulses is equal to the reciprocal of the time between the impulses.

Though constructing a pitch contour using piecewise constant segments is intuitive, it is not the only possible fundamental frequency track satisfying (4.12). Pitch is often thought of as a continuous contour rather than the series of disjoint segments constructed using the piecewise-constant approach. We can construct such a contour by using linear segments which adhere to the constraint in (4.16) and are continuous. As illustrated in Figure 4-3, a straightforward way to build them is to begin with the piecewise-constant segments found using (4.17). As shown, each of these segments can then be rotated about its midpoint, which is the way to satisfy (4.16) using a linear track. This is because the FM part integrates to 2π, there being equal positive and negative area relative to the constant (4.17). As illustrated, the continuity constraint can be met in many different ways, depending on what fundamental frequency is chosen at the first impulse. As can be seen in the figure between t_c and t_d, the direction of change can be different depending on the contour chosen.



s(t)
...
t_a
t_b
t_c
t_d
t_e
t
$f_0(t)$
The contour changes depending on the initial frequency chosen
Each linear segment is rotated about the constant segment's midpoint
t

Figure 4-3. Construction of continuous linear instantaneous fundamental-frequency tracks for a given impulse train $s(t)$. There are multiple possible tracks, each passing through the midpoint of each segment of the piecewise constant track shown in Figure 4-2. The contour changes depending on the initial frequency chosen.

Instantaneous-frequency contours other than the constant and linear examples above may also be chosen. Segments may have a higher order such as quadratic or cubic, for example. We may additionally place continuity constraints on the derivatives of the functions, for example specifying that the first and second derivatives of the frequency contours are continuous at all times. These situations will not be explicitly explored in this thesis, although they follow readily from combining (4.15) and (4.16) with the equations described by conventional interpolation methods.

We can approximately synthesize the impulse trains resulting from different instantaneous fundamental frequency contours by using a limited set of harmonics. An example of the ambiguity of sinusoidal tracks corresponding to a nonmodal impulse train using the first 4000 harmonics is shown in Figure 4-4. Here we demonstrate that a piecewise constant instantaneous fundamental frequency (left) and a linearly-varying contour (right) produce the same sequence of impulses. In Figure 4-5, we also constructed a "ragged" fundamental-frequency track (right) fitting the constraint in (4.16) resulting in the same sequence of impulses as a constant contour (left).

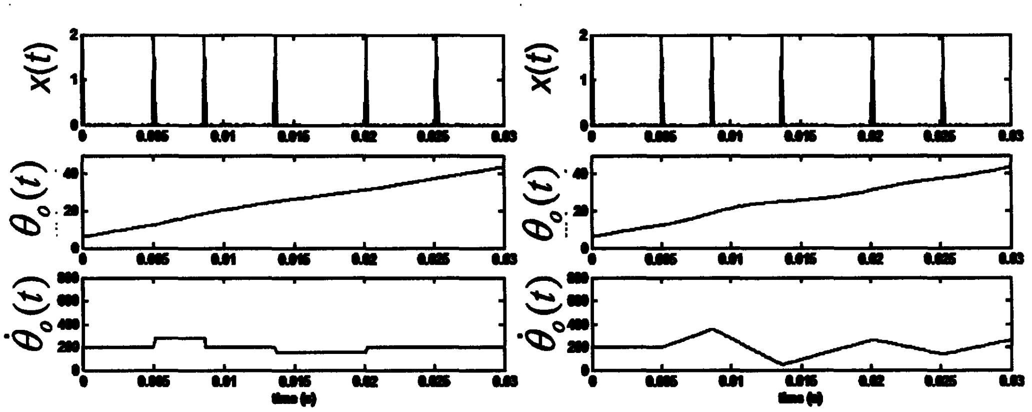


Figure 4-4. Multiple harmonically-related instantaneous frequencies (bottom panels) can yield the same nonmodal impulse train (top panels). A piecewise-constant instantaneous fundamental frequency (left) and piecewise-linear track (right) are compared.

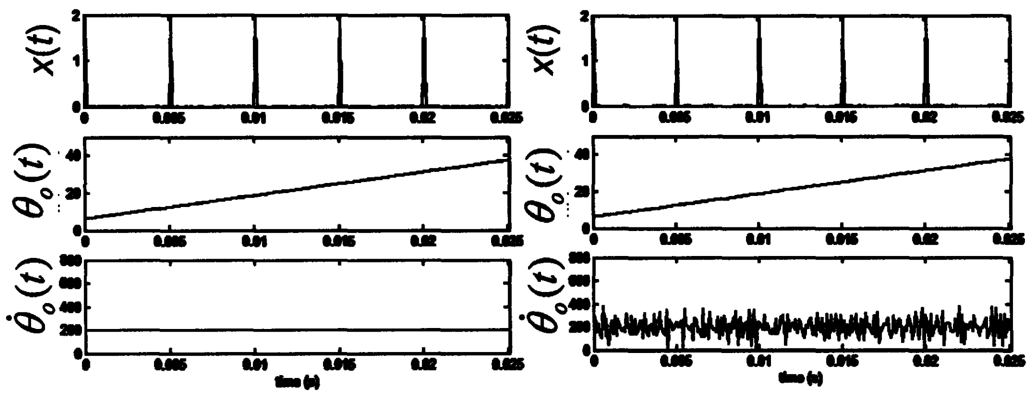


Figure 4-5. Multiple harmonically-related instantaneous frequencies (bottom panels) can yield the same impulse train (top panels). A constant instantaneous fundamental frequency (left) and "ragged" track (right) are compared.

4.2.3 Finding the Impulse Train Given the Pitch

Let us suppose that instead of the discrete times of a series of impulses, we instead have a pitch contour for a nonmodal/modal passage. This situation may come about, for example, through the use of frame-wise pitch estimation algorithms. Such pitch tracks are popular in the literature and are even derived for aperiodic phonation sources [25]. The implication is that such pitch tracks can uniquely represent the behavior at the source. In this section, however, we will show that even if we are able to derive the true value of the instantaneous fundamental frequency at a set of discrete times, there exist multiple valid sets of impulses that fit those values.

As we have shown, if the phases and frequencies are harmonically related, we get an impulse each time the phase of the fundamental crosses a multiple of 2π. We have shown an integral constraint for the instantaneous fundamental frequency between two impulses. We can equivalently examine the converse case where we only know the instantaneous phase at several points and wish to find the time instants of the impulses. For clarity of explanation, we initially assume that we know these phases, and will later discuss the case where we have measured the instantaneous fundamental frequency.

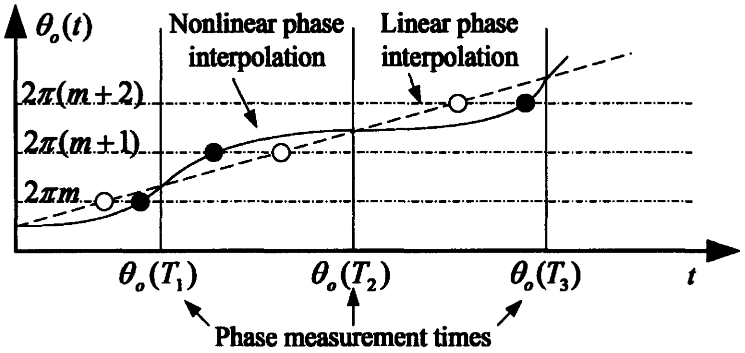


Figure 4-6. Comparison of impulse locations determined by nonlinear-phase interpolation or linear-phase interpolation.

Depending on how we interpolate the phases between a set of known phases, we will get impulses at varying points. Figure 4-6 illustrates this situation, with the times of impulses using a linear phase interpolation denoted with unfilled dots and the times of impulses using a nonlinear phase interpolation denoted with filled dots. Let us initially assume that we have a set of correctly-unwrapped and monotonically-increasing phase values. The monotonically-increasing phase constraint restricts the frequency to only positive values. Also, each sinusoid is constrained to be harmonically related to the fundamental at each time. If the difference in $\theta_o(t)$ exceeds 2π between two fundamental phase measurements, at least one impulse will occur between the measurements, and its location can change depending on the interpolation technique used. We know from the condition derived in Section 4.1, that the output will be zero when $\theta_o(t) \neq 2\pi m$ and nonzero only when $\theta_o(t) = 2\pi m$. If the phase value at each point is only known modulo 2π, the phase goes through multiples of 2π an indeterminate number of times. This means that both the number and times of the impulses between the phase measurements are ambiguous.

In a variety of practical cases, interpolation is done between measurements of the instantaneous frequency rather than the phase. An example is in the work of Milenkovic, where parabolic interpolation is done for a set of consecutive pitch-period estimates to "obtain a sensitive jitter measurement without requiring that the speech signal be sampled at a very high rate" [46]. Figure 4-7 illustrates an example of this interpolation process, here using simple piecewise constant sections. In this figure, we first show the phase function and impulse times derived from the true instantaneous-frequency track. We then uniformly sample this true frequency and use piecewise-constant sections centered on each frequency sample, and derive the corresponding phase function and impulse times. As shown, the times of the true impulses do not correspond with the times of the impulses derived using the interpolation method.

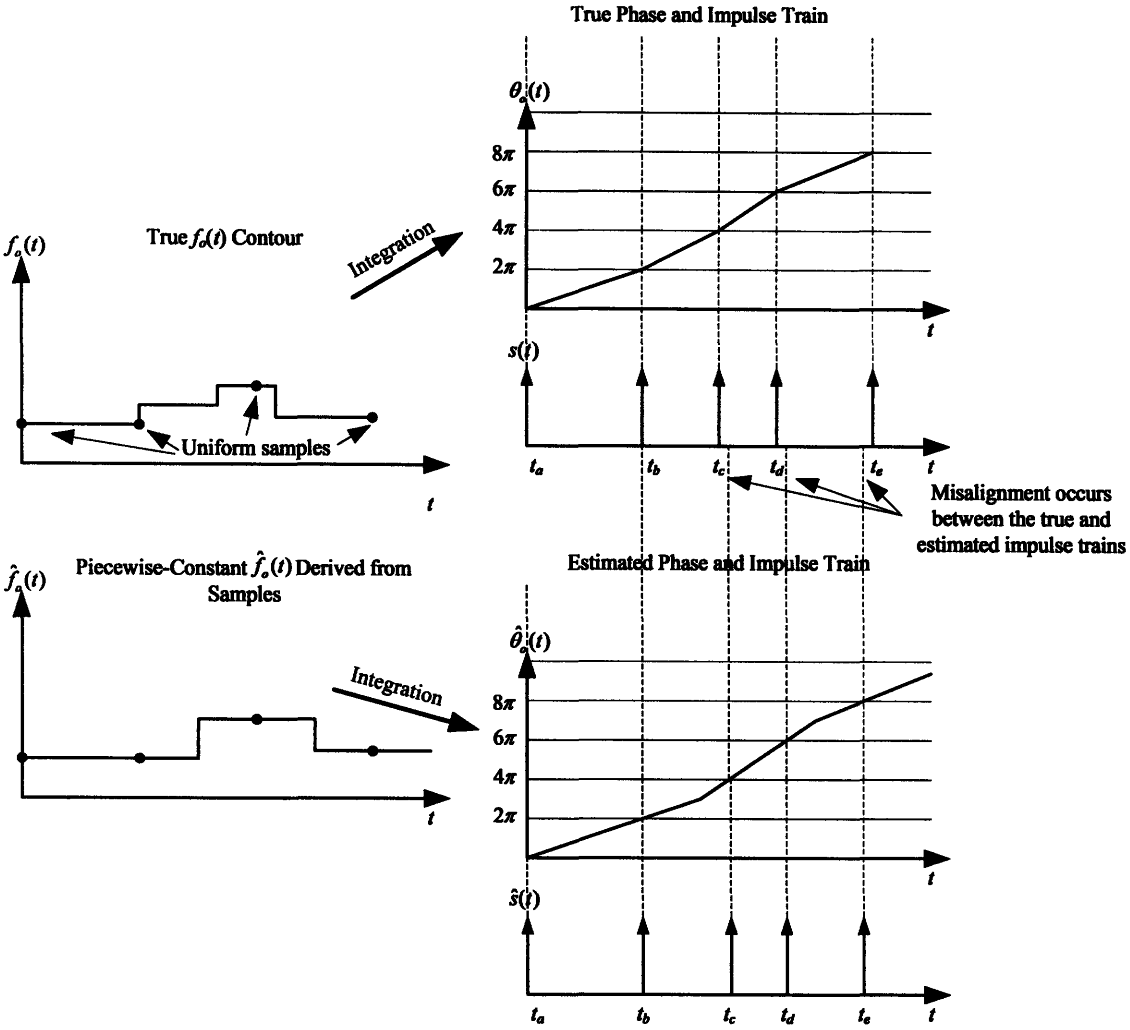


Figure 4-7. Illustration that interpolating uniform samples of the true instantaneous frequency does not always yield the true impulse locations. At the top of the figure, we see the true frequency contour and its corresponding instantaneous-phase and impulse train. The result of sampling and interpolating the true frequency contour is shown at the bottom of the figure. The estimated impulse train does not equal the true impulse train.

Even though we have only shown an example using simple piecewise-constant frequency interpolation, we could use any desired interpolation scheme. The details of such approaches, including which represent the underlying impulse train more accurately, are beyond this thesis and the subject of future work. As we have illustrated in Figure 4-6 and Figure 4-7, each possible frequency interpolation will potentially yield a different set of impulses because of resulting different phase functions.

4.3 Effects of Filtering

To this point, we have discussed the sinewave representation of ideal impulse trains, describing their dynamics in time with a single instantaneous-frequency parameter. Practical systems excited by impulses, however, are always bandlimited or otherwise filtered. Two examples are the application of a lowpass channel and filtering of the impulsive source by the resonances of the vocal tract. It is intuitive that we cannot use such processes to reduce the ambiguity in the instantaneous fundamental frequency used to generate the signal. An illustration of this fact can be seen in Figure 4-8, which illustrates that multiple sinusoidal representations yield the same impulsive source signal, from which the true sinusoidal tracks used cannot be recovered. As shown, this is because the impulse signal that enters the filter is exactly the same for each of infinite possible sinusoidal representation.

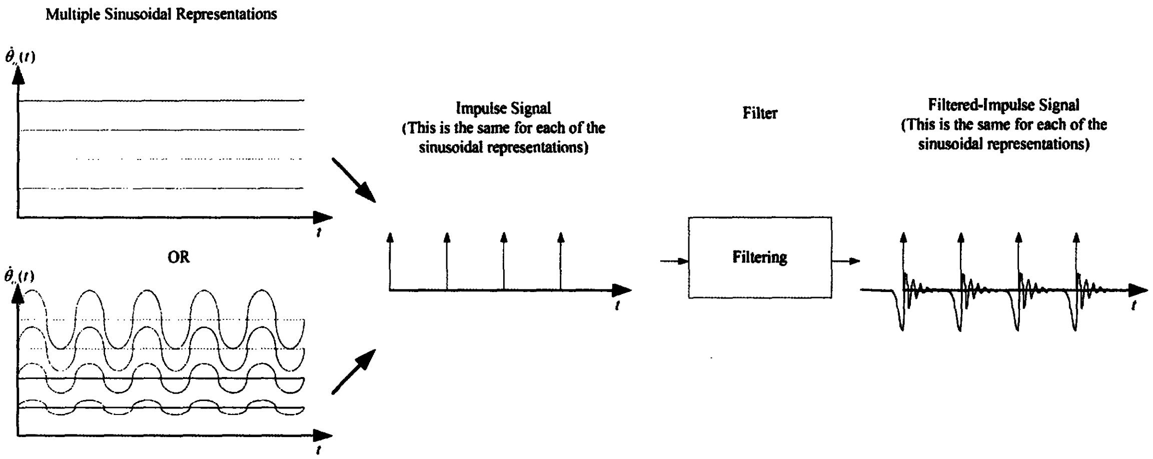

Multiple Sinusoidal Representations
$\dot{\theta}_n(t)$
t
OR
$\dot{\theta}_n(t)$
t
Impulse Signal
(This is the same for each of the sinusoidal representations)
t
Filter
Filtering
Filtered-Impulse Signal
(This is the same for each of the sinusoidal representations)
t

Figure 4-8. Schematic illustration that multiple sinusoidal representations may lead to the same impulse signal, which yields the same signal after filtering.

As we have discussed, it is impossible to recover which of the infinite valid sinusoidal representations was used to create an observed filtered-impulse signal. Here, we present an apparent paradox that seems to contradict this fact. This paradox comes about when we consider using an ideal low-pass filter to extract the instantaneous fundamental frequency. We will find that even though filtering sometimes yields a seemingly unambiguous sinusoidal representation, the underlying fundamental-frequency track remains ambiguous.

This strategy is shown schematically in Figure 4-9. Here, we consider that one of two possible sets of instantaneous frequency tracks are used and we wish to determine the correct one. One set, shown with solid lines, consists of a set of constant harmonically-related elements, and the other, shown with dashed lines, consists of sinusoidally-varying harmonically-related contours varying at frequency f_{FM}. In order to determine the correct set of tracks, we place the cutoff frequency of a lowpass filter above the first harmonic and below the second harmonic of each set as shown. The paradox comes about because while bandlimiting apparently can choose the first sinusoid only, indicating whether it is a constant sinusoid or a time-varying signal, we know that there is a unique time signal resulting from low-pass filtering an impulse train. This well-known signal in time is simply the impulse response of the low-pass filter convolved with the impulse train. Thus, while we expect to be able to extract the lowest sinusoid of each harmonic set by bandlimiting, we also see that this is not possible.

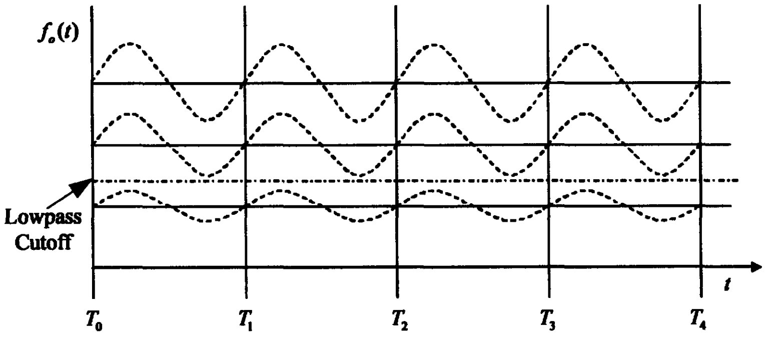


Figure 4-9. Schematic drawing of the setup leading to the bandlimiting paradox. The vertical lines indicate the times of impulses. The solid horizontal lines are the constant sinusoidal tracks. The dotted curves are one possible set of time-varying sinusoidal tracks leading to the observed impulse train. The dash-dot horizontal line indicates the proposed low-pass cutoff frequency.

The resolution to this paradox comes about by realizing that any frequency-modulated sinusoid actually has infinite bandwidth (see [50] pp.479-487), not the limited bandwidth it appears to have if just looking at its time-varying sinusoidal track. Linear filtering with an infinitely-wide time impulse response assumes one spectrum over all time. The (infinite-time) spectrum of each FM sinusoid is not bandlimited. An interesting consequence of this is that each time-varying sinusoid making up the excitation function will have infinite bandwidth, and these different bands will interact with one another.

From an FM point of view, the spectrum has each of the "carrier" components as well as a Bessel expansion of a time-varying sinusoidal component having spacing f_{FM} [12]. The sum will still be a flat line spectrum as shown schematically in Figure 4-10. This spectrum is the same as "seen" by the low-pass filter in the no-frequency-modulation case. One can equivalently say that once the sidebands due to FM sum together, they cannot be decomposed (separated) in the spectrum.

Thus, we see that two different sets of $f_o(t)$ tracks yield the same bandlimited output signal. Because this bandlimited signal is the same as the no frequency-modulation case, a naive observer might infer that the steady-tone output signal is from a system with no fundamental-frequency variation. However, as we have shown, this is a faulty conclusion to draw. The implication is that we cannot rely on measurements of a filtered speech signal to provide us with information about which possible fundamental-frequency track created it.

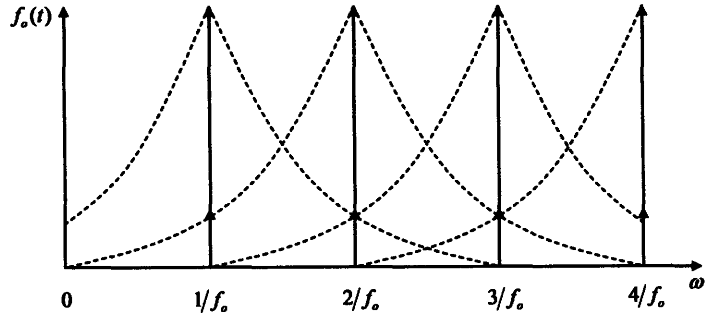


Figure 4-10. Illustration that once sidebands are summed together, they cannot be recovered.

4.4 Implication for Sinusoidal Analysis-Synthesis

As an offshoot of this work, the properties we have explored also contribute to understanding certain modifications of speech with nonuniform glottal events that occur by sinewave analysis and synthesis, specifically the addition, deletion, and movement of pulses. An example of such a phenomemon is shown in Figure 4-11 using standard sinewave analysis/synthesis [44]. Sinewave analysis-synthesis is the process of finding a finite set of sinusoids with time-varying frequency and amplitude to represent a speech signal. The sum of these sinusoids is then used to resynthesize the signal. This approach is common in many modern speech-technology applications including speech coding and voice modification.

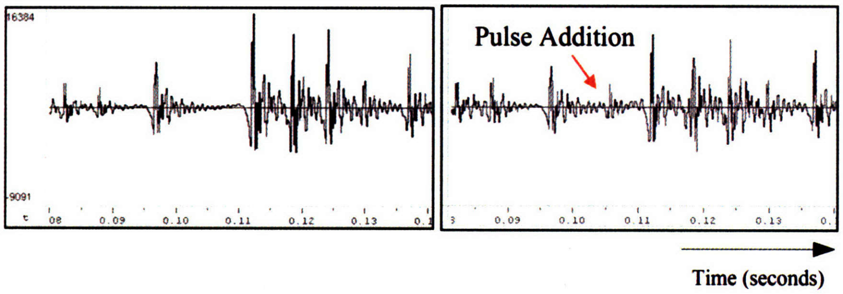


Figure 4-11. Generation of an unwanted pulse in sinewave analysis/synthesis. (left) original; (right) synthesis.

It is surprising that the sinewave synthesis of nonmodal impulses can result in effects that are more similar to the modification of the underlying pulse train through which pulses which are added, deleted, or moved than to distortion of the acoustic signal such as pulse smearing. Ultimately, we seek then to understand how a pulse can be added, deleted, or moved in synthesis relative to its position in the source signal.

For many applications, it is of interest to provide insight into why standard sinewave analysis/synthesis can add, delete, and move pulses of nonmodal speech, as we saw illustrated in Figure 4-11. We have in this chapter investigated the simplifying case of idealized impulses and have argued in Section 4.2.3, for harmonically-related instantaneous frequencies, how their locations can *move* in synthesis. With similar arguments, we can show that *addition* and *deletion* of impulses are also possible with resynthesis from a sampled phase function due to the assumptions placed on the phase contours in the sinusoidal model.

Developing a formal argument for the modification of underlying pulses in actual speech will require a generalization to the relations presented in this chapter and specifically will require addressing two properties of real speech: First, standard sinusoidal analysis does not always result in frequency harmonicity or continuity such as we have described in our model. In nonmodal speech regions, sinewave analysis via peak-picking in the frequency domain can yield what appear to be frequencies that are erratic spectrally and temporally, as we saw in Chapter 3. An important consequence is that, although impulses are generated approximately in synthesis, the resulting frequency tracks are broken and sporadic and appear to have little meaning physiologically. The second property to address is that natural speech is filtered by bandlimiting and vocal-tract filtering, as we outlined in Section 4.3. We must address how such manipulations can be represented while keeping a sinusoidal representation.

4.5 Conclusions

In Chapter 3 we described how spectral representations are affected by nonmodality in sequences of glottal events. In the current chapter, we have discussed the application of another conventional descriptor of the speech source, the fundamental frequency, to glottal events. In particular, we have described the process of representing series of modal and nonmodal glottal events in terms of a time-varying fundamental frequency. This is accomplished using a set of harmonically-related sinusoids with time-varying frequency.

We have shown that the fundamental-frequency contours for sequences of events are not unique. Rather, there are an infinite number of possible underlying tracks, subject to a set of constraints that we have derived. We have also shown the converse property that having samples of the true underlying fundamental frequency or phase is not sufficient to uniquely describe the underlying pulse train. This result is relevant to practical problems because it calls into question the common goal of extracting a ***unique*** fundamental frequency contour for voiced speech, especially for highly-nonmodal regions. Our work shows that the fundamental frequency that should be assigned to a given region is still an open question. Therefore, although it is common in the literature to describe changes in the source with fundamental frequency, we strive in this thesis to find a more unique description of glottal-event dynamics. To address this issue, the next chapter will explore explicit representations of patterns of neighboring glottal events in the time domain.

Chapter 5

Time-Domain Features for the Representation of Sequences of Glottal Events

In the previous two chapters, we have discussed two representations of sequences of glottal events. In the first, we presented the theory of spectra for these sequences. We have shown that this spectral representation is extremely sensitive to changes in a specific nonmodal pattern. We have also explored a representation using harmonically-related sinusoids. We have shown that different multiple sinusoidal tracks are valid for a single observed impulse train. In developing features to represent glottal events, it is not clear that either high sensitivity or ambiguity in representation are good characteristics. Instead, we seek an alternative representation that depends on the event times and amplitudes themselves since these are unambiguous parameters. We also seek a representation for which similar patterns are close to each other and dissimilar pulse patterns are far away from each other, according to a meaningful criterion.

In this chapter, we describe a time-domain method for representing the local characteristics of sequences of glottal events. Our technique finds relationships between the amplitudes and timings of sets of multiple neighboring impulses. We also present an algorithm by which repeating deterministic patterns of pulses may be separated in the feature space and extracted for further analysis. Deterministic patterns of glottal event timings and amplitudes are an important property of nonmodal phonation, and our algorithm is useful for characterizing the frequency of occurrence of such patterns. This separation technique will be used later in the thesis to analyze natural speech as a means to better understanding the speaker-dependence of glottal-event patterns.

The chapter begins in Section 5.1 with a description of previous work in time-domain representations of glottal-event patterns. In Section 5.2, we then describe the method for feature extraction that we will use in the remainder of this thesis. Section 5.3 presents visualizations that come about from the event-pattern characterization, demonstrating that different classes of event patterns become separate in this domain. In Section 5.4, we discuss the effects of constant and linearly-varying event amplitudes and timings on the feature space, highlighting the normalization properties of our time-domain approach. Finally, in Section 5.5, we describe the algorithm by which different deterministic patterns of glottal events may be separated based on their continuity across time. We will present this algorithm and examine its application to synthetic speech, as a prelude to its use with real speech later in the thesis.

We will see in later chapters that the features devised here are useful for automatic speaker recognition. We also have developed the feature space so that it has intuitive meaning, hopefully aiding its eventual use in speech science and clinical applications.

5.1 Previous Work in Time-Domain Glottal Event Patterns

5.1.1 Clinical Pulse Perturbation Measures

Characterizing sequences of nonmodal glottal cycles is a common practice in the domain of clinical voice analysis, where acoustic measures are integrated into clinical practice. These methods analyze speech automatically and provide the clinician with a numerical value related to the severity of the voice. Although there are many variations available in the literature, a large number of clinically-used objective acoustic measurements for pathological voices fall loosely into one of two different categories—vocal-fold perturbation measures and glottal noise measures [53]. Perturbation measures, such as *shimmer* and *jitter*, measure the irregularity of the speech signal from one pitch period to the next, usually in the time domain.

Jitter is defined as a "short term (cycle-to-cycle) perturbation in the fundamental frequency of the voice" [72]. Several automatic and hand-assisted techniques exist to measure jitter in sustained phonation [30, 46]. Shimmer is cycle-to-cycle change in the amplitude of successive voice periods [72]. As with jitter, both automatic and hand-assisted techniques have been created to measure shimmer in sustained phonation [31, 46]. There are multiple approaches to measuring shimmer and jitter acoustically [9, 72], one of the early and most popular is that of Milenkovic [46], which operates directly on the acoustic far-field pressure waveform.

In contrast to glottal-perturbation measures, glottal noise measures, such as harmonic-to-noise ratio (HNR), quantify the energy in harmonic components of a signal relative to energy in the inharmonic components [46, 53]. Perturbation measures explicitly deal with the relationships between neighboring glottal pulses and thus are the most relevant to this thesis.

5.1.2 Previous Work in Glottal-Pulse Pattern Features

The use of patterns between neighboring glottal cycles as features has been reported in the literature by several authors [54, 57]. The pulse-landmark feature that appears closest to the one that we will describe in this chapter is found in [57]. Here, an algorithm captures patterns of timing between points in three neighboring glottal cycles, with the relative times used as features. This technique is made to fit a frame-by-frame analysis, though, effectively smoothing the measurements of the glottal cycles and sampling them every 10 ms.

Visualizations similar to the ones we will derive in this chapter also exist. Fourcin uses a 2-dimensional plot to display glottal-cycle timing between neighboring periods derived from EGG measurements [19]. Additionally, Hirson and Duckworth [29] have used two-dimensional graphics to examine time-varying irregular glottal excitation.

5.2 Extracting Features

In this section, we will describe the process by which we characterize the pattern underlying a series of glottal events. These features assume that we have obtained the impulse representation of glottal events introduced in Chapter 2. The method to accomplish this will be detailed in the next chapter. For now, we focus on representing the event train once it is extracted. As shown in Figure 5-1, recall that each event is represented by a time of occurrence and amplitude. Given a sequence of these glottal events, we generate feature vectors using sets of neighboring impulse times and amplitudes.

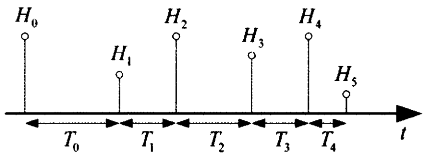
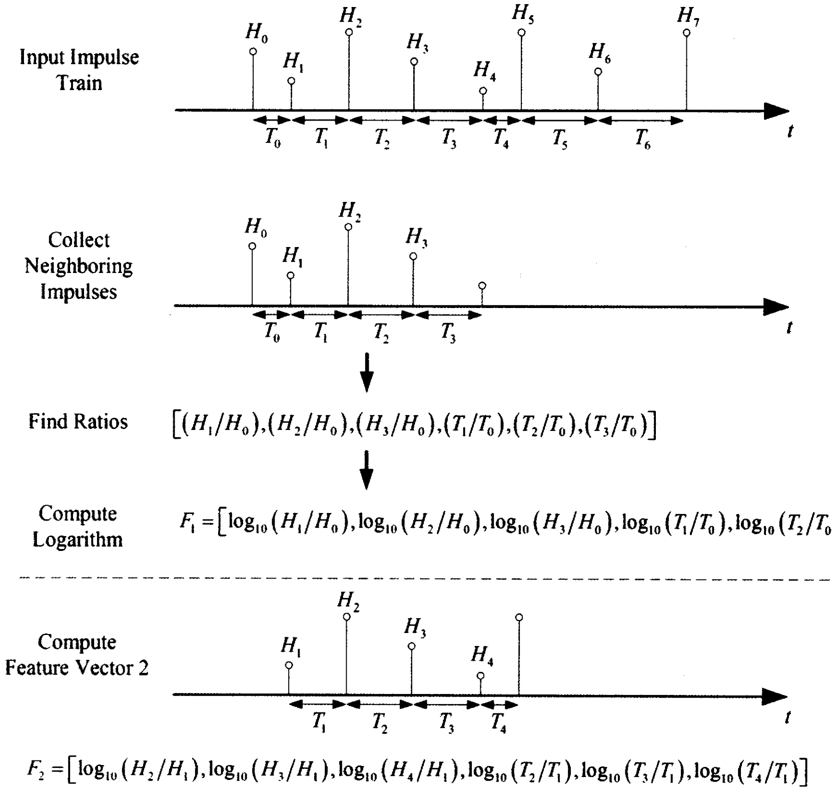


Figure 5-1. Schematic of the heights and times used to compute amplitude- and timing-ratio features.

The feature-creation process is illustrated in Figure 5-2. In this example, we choose to extract three timing features and three amplitude features per feature vector, yielding a total of six elements per vector. As shown, the first step is to collect amplitude and timing differences for a set of neighboring impulses to be analyzed. Next, each of the timing differences is normalized by finding its ratio with the first timing difference in the vector, and each of the amplitudes is normalized by finding its ratio with the first amplitude. This step is meant to help eliminate the effects of different absolute timings and amplitudes, due to differences in overall pitch or loudness, for example. Instead, the features are designed to concentrate on local patterns of variation. This is similar in spirit to the normalization over sequences of glottal cycles in the calculation of jitter and shimmer discussed in [72].



Input Impulse Train
H_0 H_1 H_2 H_3 H_4 H_5 H_6 H_7
T_0 T_1 T_2 T_3 T_4 T_5 T_6
t
Collect Neighboring Impulses
H_0 H_1 H_2 H_3
T_0 T_1 T_2 T_3
t
Find Ratios
$[(H_1/H_0),(H_2/H_0),(H_3/H_0),(T_1/T_0),(T_2/T_0),(T_3/T_0)]$
Compute Logarithm
$F_1=[\log_{10}(H_1/H_0),\log_{10}(H_2/H_0),\log_{10}(H_3/H_0),\log_{10}(T_1/T_0),\log_{10}(T_2/T_0),\log_{10}(T_3/T_0)]$
Compute Feature Vector 2
H_1 H_2 H_3 H_4
T_1 T_2 T_3 T_4
t
$F_2=[\log_{10}(H_2/H_1),\log_{10}(H_3/H_1),\log_{10}(H_4/H_1),\log_{10}(T_2/T_1),\log_{10}(T_3/T_1),\log_{10}(T_4/T_1)]$

Figure 5-2. Schematic of the feature-extraction process using a series of glottal event measurements. The example shown yields feature vectors each with six total features, three timing and three amplitude.

The next step is to compute the base-10 logarithm of each ratio. Taking the logarithm allows the feature space to become more compact and also introduces symmetry about the origin for reciprocal ratios, a property we will later find useful. For example, $\log_{10}(H_1/H_0) = -\log_{10}(H_0/H_1)$ and the midpoint of these two values is the origin.

The six-element feature vector obtained using the steps above, with three amplitude ratios and three timing ratios is given by

$$F_1 = \left[\log_{10}(H_1/H_0), \log_{10}(H_2/H_0), \log_{10}(H_3/H_0), \log_{10}(T_1/T_0), \log_{10}(T_2/T_0), \log_{10}(T_3/T_0)\right].$$

Also, as illustrated in Figure 5-2, a new feature vector is computed beginning at each incoming impulse. Using our six-element example, for instance, the next feature vector in the series is

$$F_2 = \left[\log_{10}(H_2/H_1), \log_{10}(H_3/H_1), \log_{10}(H_4/H_1), \log_{10}(T_2/T_1), \log_{10}(T_3/T_1), \log_{10}(T_4/T_1)\right].$$

The length of the feature vector used may be changed depending on the application.

5.3 Representation of Nonmodal Sequences of Glottal Events

Nonmodal phonation consists of many deterministic and stochastic behaviors. The goal of this section is to show the features resulting from each of several such behaviors. Through the synthetic examples in this section, we will show that different deterministic patterns of nonmodal events are represented differently in the feature space, and that these representations form sets of tight clusters. We will also see that the degree of random perturbation is characterized by the amount of dispersion in the feature space.

Using synthetic sequences of impulses, we can visualize how the feature space represents different patterns of pulses. In the following examples, we generate different pulse patterns by perturbing a modal 125-Hz pulse train. We refer to deterministic patterns based on their order of repetition, for example with the term *1-long* referring to a series of equal event periods or amplitudes, *2-long* denoting a series of periods or amplitudes that alternate every other event, and *3-long* denoting a series of periods or amplitudes that repeat every three events. The word *period* is used to mean the interval of time between two particular neighboring events. Specific 2-long patterns are referred to using the notation [T_1, T_2], where T_1 is the first period and T_2 is the second period of the repeating pattern. Likewise, specific 3-long patterns use the notation [T_1, T_2, T_3], where T_1, T_2, and T_3 are the periods of the repeating pattern.

The specific 2-long and 3-long timing patterns used in this chapter are detailed in Appendix D and were designed to represent a variety of different ratios. A small amount (±0.125 ms) of random timing perturbation was also added to each of the deterministic patterns in order to create natural variation. The signal generated for each particular pattern used consists of five seconds of continuous impulses. Signals exhibiting irregular events were built by uniformly perturbing the timing of each impulse in a periodic 125-Hz impulse train over a certain interval. Twenty-eight different signals, detailed in Appendix D, were generated using this approach, each using a different amount of perturbation. The feature-space visualizations that are displayed in this chapter are 2-dimensional histograms, with the axes representing the values of the first two timing features and the shade describing the number of occurrences in each bin.

For a near-periodic impulse train, we obtain the plot shown in Figure 5-3. We see that this kind of pattern clusters around the origin of our feature space.

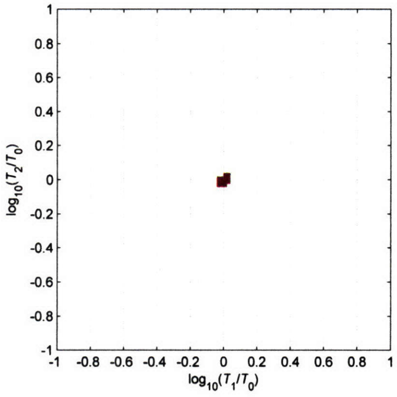


Figure 5-3. Timing-feature space for a 125-Hz near-periodic event train. ±0.125 ms of random timing perturbation has been added to create natural variation.

For 2-long patterns, we obtain the feature space shown in Figure 5-4. As shown in the top-left panel, each individual diplophonic pattern, here [12 ms, 4 ms], forms two distinct clusters in the feature space. When additional 2-long patterns are used, here [12 ms, 4 ms], [5.75 ms, 10.25 ms], and [8.5 ms, 7.25 ms], we obtain several clusters, as plotted in the top-right panel. Finally, if we use many different diplophonic patterns, we obtain the space shown in the bottom panel, a line along the *x*-axis. The widening on the right side of the line is due to the effect of the absolute ±0.125 ms jitter on the ratios when small periods are present. Toward the right, for example, the pattern [2 ms, 14 ms] has a cluster, and at the origin, [8 ms, 8 ms] has a cluster. The ratio of 2±0.125 ms to 2±0.125 ms exhibits much more variation than the ratio of 8±0.125 ms to 8±0.125 ms, which leads to spreading in the plot. This spreading effect unintentionally occurs more on the right side of the plot than on the left side due to the particular 2-long patterns we used (see Appendix D).

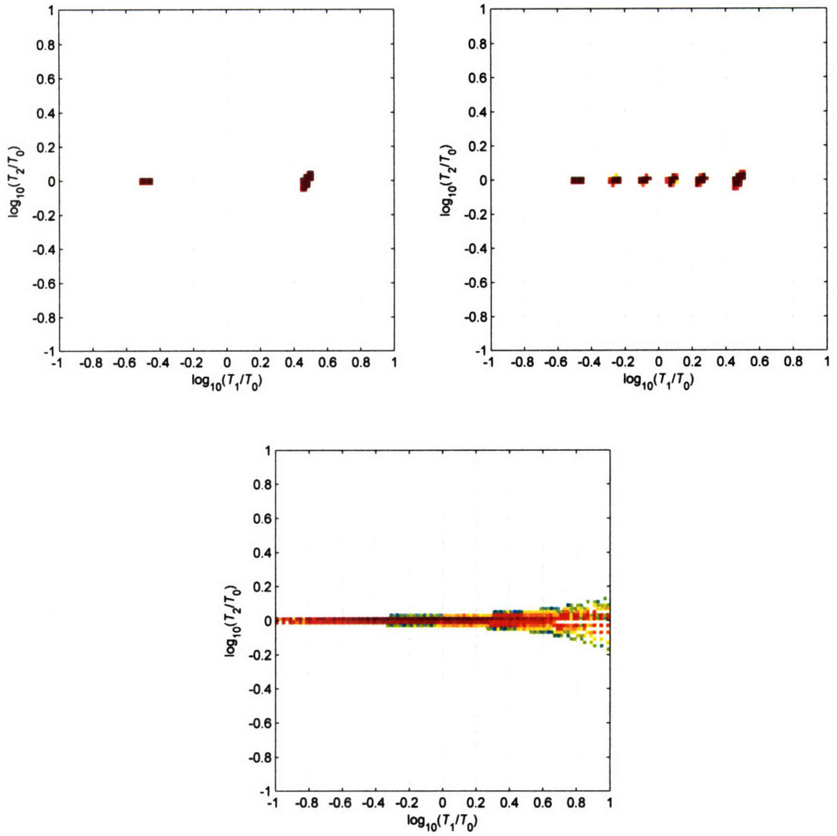


Figure 5-4. Timing-feature space for 2-long event patterns. The top-left panel shows the space for one particular 2-long pattern, [12 ms, 4 ms]; the top-right panel shows the space for three 2-long patterns, [12 ms, 4 ms], [5.75 ms, 10.25 ms], and [8.5 ms, 7.25 ms]; the bottom panel shows the space for a wide range of 2-long patterns. ±0.125 ms of random timing perturbation has been added in each case to create natural variation.

For 3-long patterns, we obtain the results shown in Figure 5-5. As we can observe in the top-left pane, for each individual pulse pattern, here [10.5 ms, 7.38 ms, 4.13 ms], we obtain a series of three distinct clusters. When additional 3-long patterns are shown, here [10.5 ms, 7.38 ms, 4.13 ms], [2.25 ms, 10 ms, 13.5 ms], and [3.5 ms, 11.25 ms, 9.25 ms], we see additional distinct sets of clusters. If we look at many different possible triplophonic patterns, we see that these clusters can be found across the *x*-*y* plane. It can be shown that any perfect triplophonic pattern falls on a single plane in the higher-dimensional feature space. As with the 2-long case, spread can be observed in the clusters due to the amplified effect of jitter in clusters containing small periods.

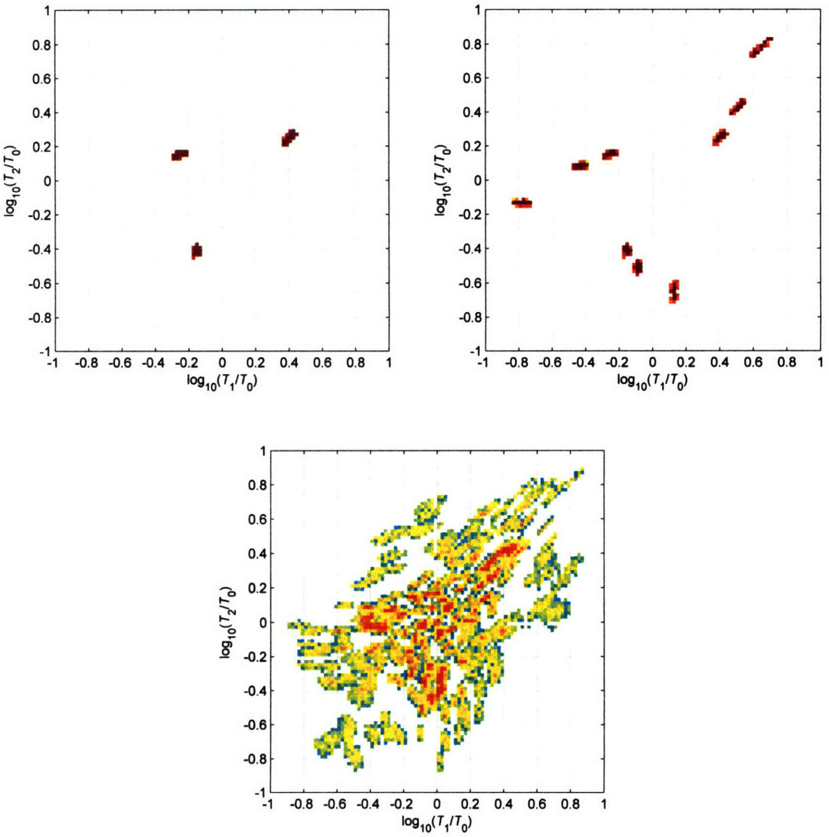


Figure 5-5. Timing-feature space for 3-long event patterns. The top-left panel shows the space for one particular 3-long pattern, [10.5 ms, 7.38 ms, 4.13 ms]; the top-right panel shows the space for three 3-long patterns, [10.5 ms, 7.38 ms, 4.13 ms], [2.25 ms, 10 ms, 13.5 ms], and [3.5 ms, 11.25 ms, 9.25 ms]; the bottom panel shows the space for a wide range of 3-long patterns. ±0.125 ms of random timing perturbation has been added in each case to create natural variation.

For different degrees of timing irregularity projected down to the first two timing dimensions, we obtain the plots shown in Figure 5-6. Observe that these figures no longer appear like the distinct clusters shown above. The points are dispersed over a large area, with greater dispersion occurring as the amount of irregularity increases.

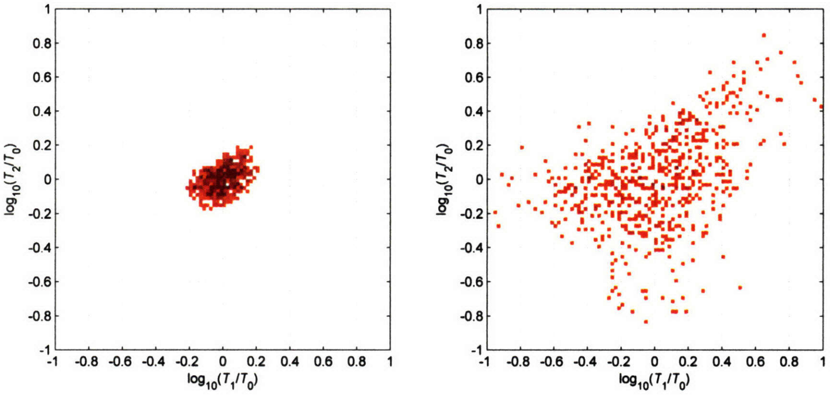


Figure 5-6. Timing feature space for different amounts of timing irregularity. The left plane shows the space for ±1 ms irregularity; the right panel shows the space for ±3.5 ms irregularity.

5.4 Representation of Constant and Linearly-Varying Sequences

As the features we have described are based on ratios, they impart a form of normalization on the relationships between neighboring glottal events. When these relationships are simply scaled by a constant, the features extracted remain the same. For instance, scaling each of the amplitude or periods in a feature vector F_m by a constant yields an identical feature vector, $\hat{F}_m$:

$$\begin{aligned}\hat{F}_m &= \left[\log_{10}\left(\alpha T_{m+1}/\alpha T_m\right), \log_{10}\left(\alpha T_{m+2}/\alpha T_m\right), \ldots\right] \\ &= \left[\log_{10}\left(T_{m+1}/T_m\right), \log_{10}\left(T_{m+2}/T_m\right), \ldots\right] \\ &= F_m\end{aligned} \quad . \tag{5.1}$$

This property allows the features to reflect relative patterns instead of absolute pitch or amplitude.

The relationships between neighboring periods or amplitudes may also change linearly from event to event. We can write this change as a function of the original timing intervals, T_m, or amplitudes, H_m, and β, which describes the scaling from one event to the next:

$$\begin{aligned}\hat{T}_m &= \beta^m T_m && \text{for linearly changing event periods} \\ \hat{H}_m &= \beta^m H_m && \text{for linearly changing event amplitudes}\end{aligned} \tag{5.2}$$

where the hat over the variable, as with $\hat{T}_m$, denotes the linearly-changing period or amplitude. The feature vectors $\hat{F}_m$ resulting from the linear change can be computed as follows:

$$\begin{aligned}\hat{F}_m &= \left[\log_{10}\left(\beta^{m+1} T_{m+1}/\beta^m T_m\right), \log_{10}\left(\beta^{m+2} T_{m+2}/\beta^m T_m\right), \ldots\right] \\ &= \left[\left(\log_{10}(\beta) + \log_{10}\left(T_{m+1}/T_m\right)\right), \left(\log_{10}\left(\beta^2\right) + \log_{10}\left(T_{m+2}/T_m\right)\right), \ldots\right] \\ &= F_m + \left[\log_{10}(\beta), 2\log_{10}(\beta), 3\log_{10}(\beta), \ldots\right]\end{aligned} \tag{5.3}$$

The same result holds if amplitudes are used instead of periods. The effect of linear change is therefore an additive offset term with the form shown. With this term, the feature-space geometry should remain as we have shown in the plots of Section 5.3, but centered at a point $\left[\log_{10}(\beta), 2\log_{10}(\beta), 3\log_{10}(\beta), \ldots\right]$ instead of at the origin.

Our discussion of the representation of constant and linearly-varying event sequences forms the basis for modeling more complicated "intonation" contours, such as those found in natural speech. A first step for future work may be representing intonation with a piecewise-linear approximation, allowing us to use the results derived above over each linear region. The effects of quadratic and higher order curves may also be derived using similar strategies.

5.5 Separation of Different Pulse Phenomena

As we have shown, our feature space displays different qualities for different kinds of pulse patterns. Figure 5-7 illustrates, however, that these patterns may overlap when modal speech, diplophonia, triplophonia, and irregularity are viewed in the feature-space plot. In this section, we describe and test an algorithm for separating different repeating patterns from this combined space. We will exploit the property that features located next to one in another in time tend to reflect similar pulse patterns.

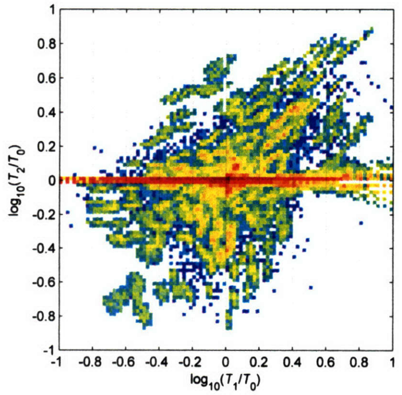


Figure 5-7. Timing feature space for a combination of modal, diplophonic, triplophonic, and irregular impulse patterns.

5.5.1 Formulation

In the separation algorithm, we are trying to separate feature vectors that are taken from a long repeating pattern of glottal-events. Figure 5-8 shows a 2-long example of such a pattern. An important property of any *N*-long repeating pattern is that we can reconstruct the entire pattern given only a single feature vector of length $N-1$. As an example, for the 2-long pattern of periods shown in Figure 5-8, we only need a single feature in order to reconstruct the entire pattern. In other words, a feature vector of length $N-1$ represents an *N*-long pattern. Once we have this feature vector, we can find every other feature vector in the sequence. For example, given only F_1 in Figure 5-8, we can predict F_2, F_3, F_4, and beyond.

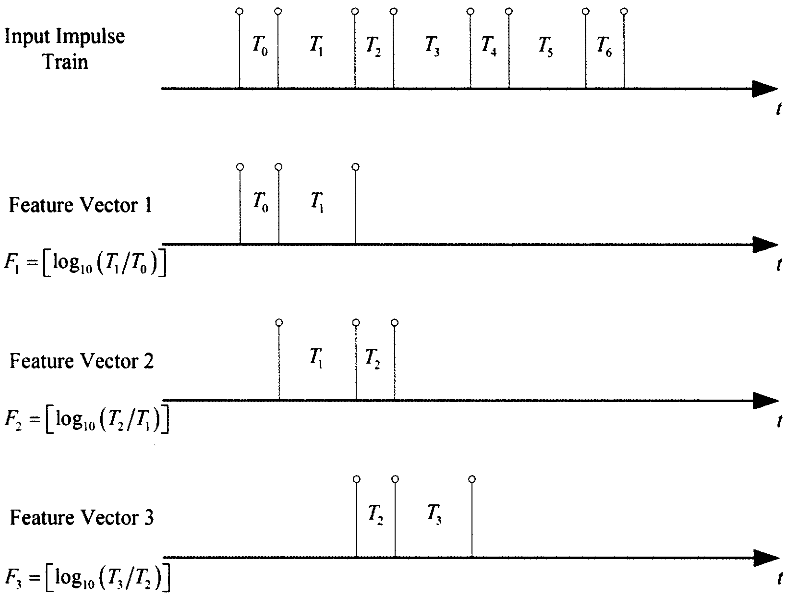


Figure 5-8. Schematic of the timing-feature-extraction process for a diplophonic pattern. As shown, each feature is created using the ratios of neighboring periods.

If we are within a section of phonation with an N-long repeating pattern, then we should be able to calculate the value of the next feature vector using only the first $N-1$ elements of the current feature vector. As shown in Figure 5-9 for an example 2-long pattern, the next period (timing interval) in a sequence of glottal events is the same as the first period of the current 2-long sequence. The implementation of predicting the next $N-1$ long feature vector from the current one involves only three steps. First, we append the value zero to the end of the current feature vector. Then we subtract the value of the first feature in the vector from the other features. Finally, we remove the first element, leaving $N-1$ features in the remaining vector. For the special case of a 1-long sequence, we predict that the next feature will be equal to the current feature. We denote this function $g(\,)$:

$$\begin{aligned} &g(F_m[1:N-1])=[F_m[2:N-1],0]-F_m[1] \quad \text{for } N>1 \\ &g(F_m[1])=F_m[1] \quad \text{for } N=1 \end{aligned}. \tag{5.4}$$

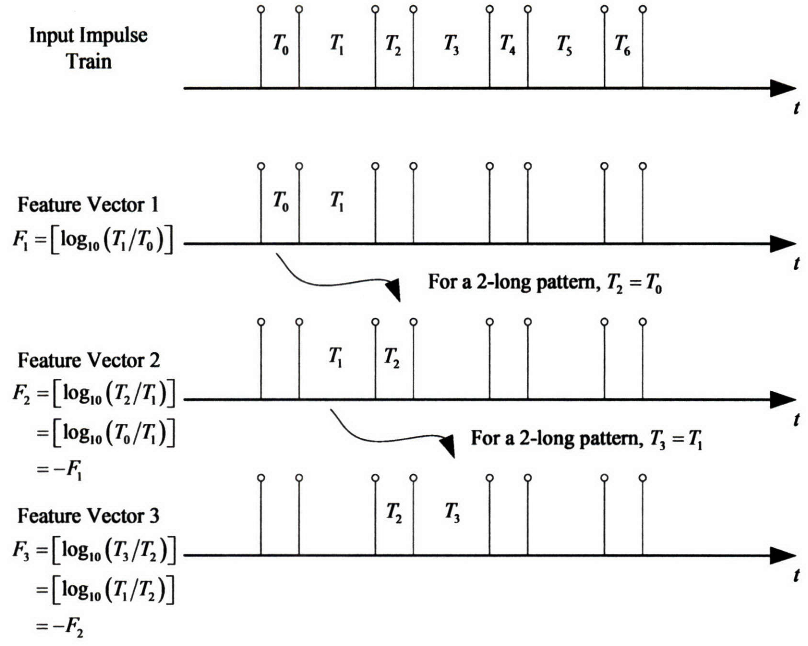

Input Impulse Train
T_0
T_1
T_2
T_3
T_4
T_5
T_6
t
Feature Vector 1
$F_1 = \left[\log_{10}(T_1/T_0)\right]$
T_0
T_1
t
For a 2-long pattern, $T_2 = T_0$
Feature Vector 2
$F_2 = \left[\log_{10}(T_2/T_1)\right]$
$= \left[\log_{10}(T_0/T_1)\right]$
$= -F_1$
T_1
T_2
t
For a 2-long pattern, $T_3 = T_1$
Feature Vector 3
$F_3 = \left[\log_{10}(T_3/T_2)\right]$
$= \left[\log_{10}(T_1/T_2)\right]$
$= -F_2$
T_2
T_3
t

Figure 5-9. Schematic showing that each consecutive feature vector for a 2-long pattern may be calculated from a single feature vector.

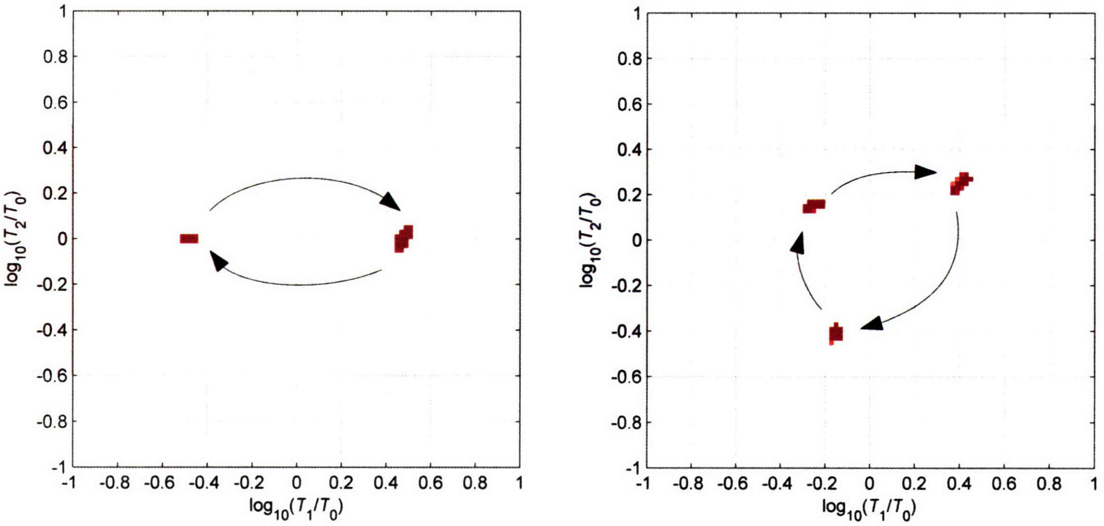


Figure 5-10. Illustration of a repeating pattern as a trajectory between a set of clusters. The trajectory on the left arises from the 2-long pattern, [12 ms, 4 ms], and the trajectory on the right arises from the 3-long pattern, [10.5 ms, 7.38 ms, 4.13 ms].

An illustration of the relationships between subsequent feature vectors is shown using the feature histograms in Figure 5-10 for a 2-long and 3-long pattern. The process of moving from one feature vector to the next can be visualized as a trajectory from one cluster to the next. The separation algorithm uses this property, finding series of contiguous feature vectors where *each feature vector may be calculated by the preceding feature vector*. If we are within a repeating pattern with period greater than N, then the calculation of the next feature will be incorrect at some point. This is shown for a 3-long underlying pattern in Figure 5-11. We quantify how well we have predicted each subsequent feature vector using the Euclidean distance,

$$\begin{array}{l} \left|F_{m+1}[1:N-1]-F_m[1:N-1]\right| \text{ for } N>1 \\ \left|F_{m+1}[1]-F_m[1]\right| \text{ for } N=1 \end{array} . \tag{5.5}$$

Observe in Figure 5-11 that it is possible to predict a subset of the consecutive features (here F_2) even if the underlying pattern has an order that is greater than that of the one being separated. We will allow for this behavior in our separation algorithm.

Mathematically, we can describe the first $N-1$ elements of each feature vector, F_{m+1}, as being related by the function $g(\,)$ to the first $N-1$ elements of the previous feature vector, F_m, with an additional prediction-error vector, ε:

$$F_{m+1}[1:N-1]=g(F_m[1:N-1])+\varepsilon . \tag{5.6}$$

When the sequence of features is drawn from an underlying perfect N-long pattern, the magnitude of the prediction-error vector will be zero for every extracted feature. We will use the term *prediction error* to refer to the magnitude of the vector ε.

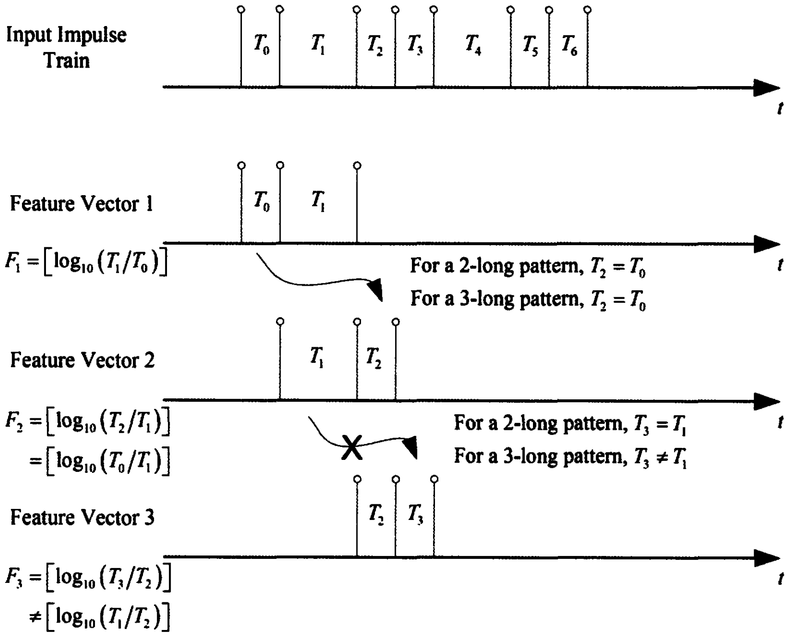

Input Impulse Train
T_0 T_1 T_2 T_3 T_4 T_5 T_6
t
Feature Vector 1
T_0 T_1
$F_1=[\log_{10}(T_1/T_0)]$
For a 2-long pattern, $T_2 = T_0$
For a 3-long pattern, $T_2 = T_0$
t
Feature Vector 2
T_1 T_2
$F_2=[\log_{10}(T_2/T_1)]$
$=[\log_{10}(T_0/T_1)]$
For a 2-long pattern, $T_3 = T_1$
For a 3-long pattern, $T_3 \neq T_1$
t
Feature Vector 3
T_2 T_3
$F_3=[\log_{10}(T_3/T_2)]$
$\neq[\log_{10}(T_1/T_2)]$
t

Figure 5-11. Schematic showing that each consecutive feature vector for a 3-long repeating pattern is not correctly calculated using a single-element feature vector. Because we incorrectly predict the next feature, we know that we do not have a 2-long pattern.

It is important to realize that repeating patterns of length N also repeat at orders which are integer multiples of N. For instance, a periodic pattern may also be interpreted as a 2-long or a 3-long pattern. Therefore, using our algorithm, each feature vector may be deemed to be repeating at more than one rate. One way to handle such cases is to attribute each feature to the lowest-order pattern that it is attributed to. A periodic sequence of events, for example, is deemed to be periodic and not 2-long or 3-long.

Any feature vector that is not attributed to a repeating pattern is assigned to a *residual* set of features which is not categorized as having any deterministic pattern. This residual space contains irregularly-occurring glottal events as well as short sequences of other patterns that do not repeat enough to be assigned to one of the patterns.

5.5.2 Algorithm

The specific algorithm used for separating feature vectors in terms of their underlying event patterns consists of the following steps:

(1) Compute the feature vectors.

(2) Determine the prediction error sequence for each pattern order, *N*. This sequence is computed by finding the Euclidean distance between the first *N*-1 elements of a feature vector and the first *N*-1 elements of the following feature vector. In our examples, we calculate the prediction error assuming 1-long, 2-long, and 3-long patterns.

(3) Find all features for which the prediction error is less than the specified *error threshold.* Unless otherwise specified, we use an error threshold of 0.1 in this thesis.

(4) Identify sets of contiguous features that are able to predict the next feature. Keep only those sets containing as many or more repeats as specified by the *minimum number of repeats* parameter. Unless otherwise specified, we require six repeats in this thesis.

(5) Extend each of the identified contiguous regions to include the feature following the last one of the sequence, since it is within the error of the previous feature.

(6) Identify each feature as possessing the lowest-order pattern it is attributed to. If the feature is not attributed to any deterministic pattern, assign it to the *residual* category.

Figure 5-12 illustrates schematically the algorithm operating on a set of feature vectors, where the feature index refers to each consecutive feature. The prediction errors are shown for 3-long, 2-long and 1-long patterns. We have circled features for which at least the minimum number of contiguous features are below the prediction-error threshold. As shown, the lowest-order pattern having the required number of repeats below the error threshold is identified to be the correct one.

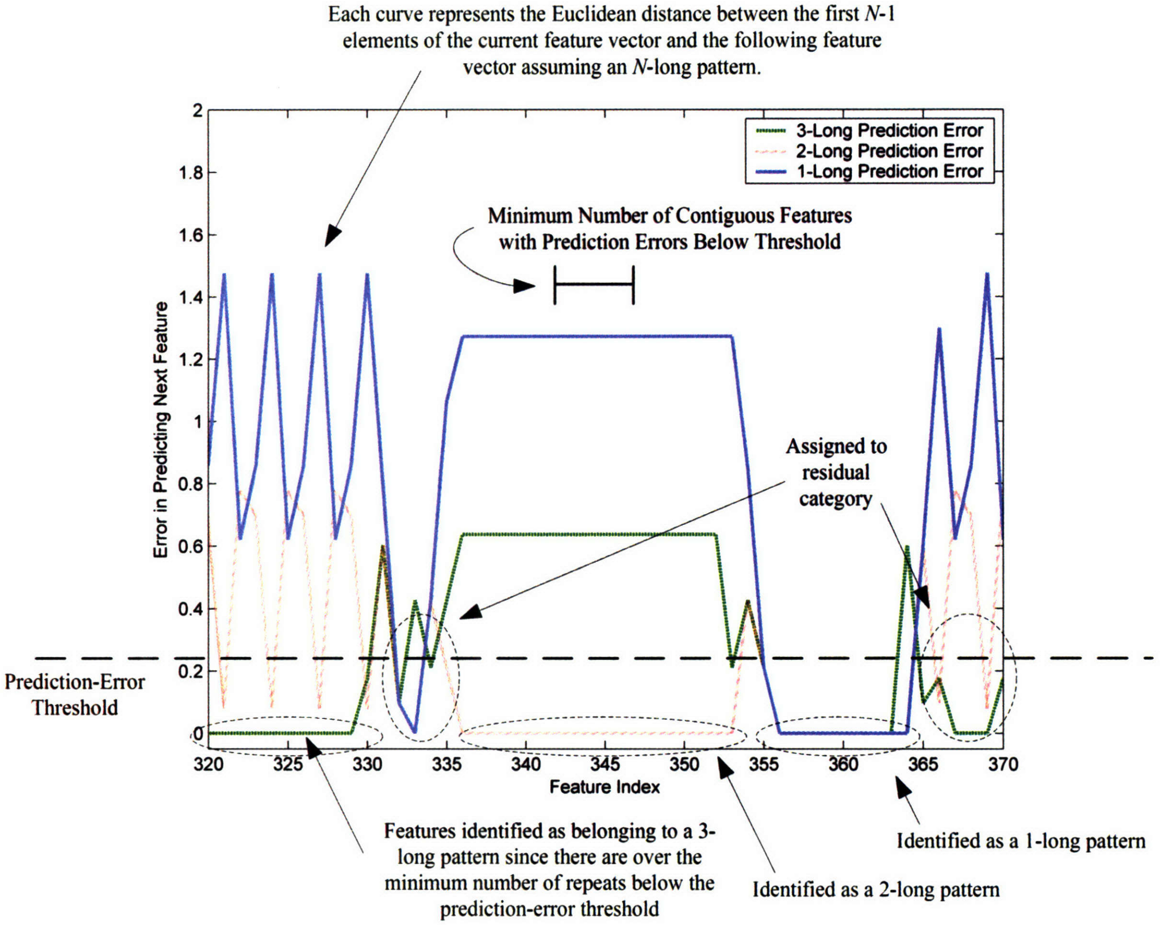


Figure 5-12. Schematic diagram of the pattern-separation algorithm acting on a sequence of time-domain glottal-event features. The two parameters of the algorithm, the prediction-error threshold and the minimum number of repeats, are illustrated. As shown, the lowest-order pattern having the required number of repeats below the error threshold is identified to be the correct one.

5.5.3 Synthetic Examples

Using the separation algorithm, we are able to partially separate out different impulse patterns. From the combined synthetic feature space shown in Figure 5-7, for example, we obtain the separated spaces shown in Figure 5-13 using a minimum of six consecutive repeats below a prediction-error threshold of 0.1. Each of the resulting spaces approximately captures the different behaviors seen in the earlier deterministic-pattern figures. There is some error, especially at the origin for both 2-long and 3-long patterns. This error is due partially to the random variation we added to the pulse positions and to the amount that we have allowed consecutive features to differ from on another.

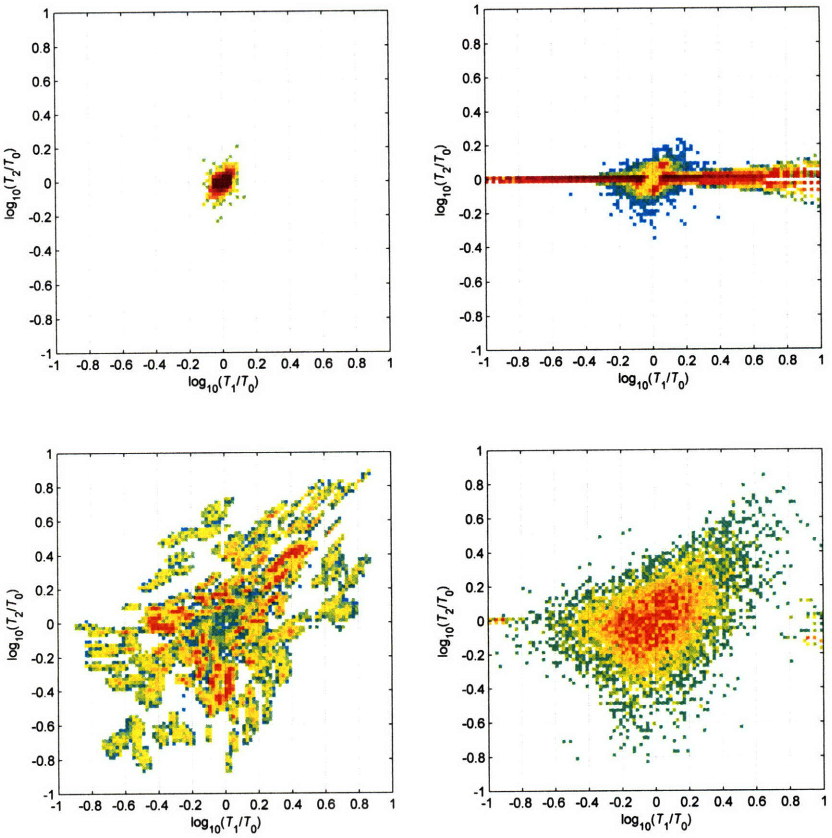


Figure 5-13. Automatic separation of the timing-feature space into features representing its modal, diplophonic, triplophonic, and irregular components of Figure 5-7. A feature prediction error threshold of 0.1 for a minimum of six consecutive features is required for identification as a particular pattern.

We can quantify the percentage of features correctly classified as belonging to each category as a function of the two algorithm parameters, the prediction threshold and the minimum number of repeats. Figure 5-14 shows the results of sweeping the prediction-error threshold for a fixed minimum six repeats. The results indicate that, for a low threshold, deterministic patterns will not be detected because the algorithm will not allow for small natural variations, and many features will be identified as being irregular. As the threshold is increased, the deterministic patterns begin to be correctly identified and the irregular sequences begin to be identified as belonging to one of the deterministic categories. We can see this effect clearly near the origin for the separated 2-long patterns in Figure 5-13. At large enough thresholds many of even the deterministic patterns are incorrectly identified as modal since so much prediction error is allowed. Figure 5-15 sweeps the minimum number of repeats for a fixed prediction-error threshold of 0.1. In this case, the primary effect is on the correct detection of irregular segments. For few repeats, many irregular sequences are mistaken to be a deterministic pattern. As more repeats are required, random changes in the period are mistaken less often for deterministic patterns.

When we perform separation in Chapter 7, we will use a prediction-error threshold of 0.1 and a minimum of six repeats. This choice is somewhat arbitrary, but is motivated by the tradeoffs we have seen for synthetic stimuli and informal observations of the effect of the separation parameters on real speech.

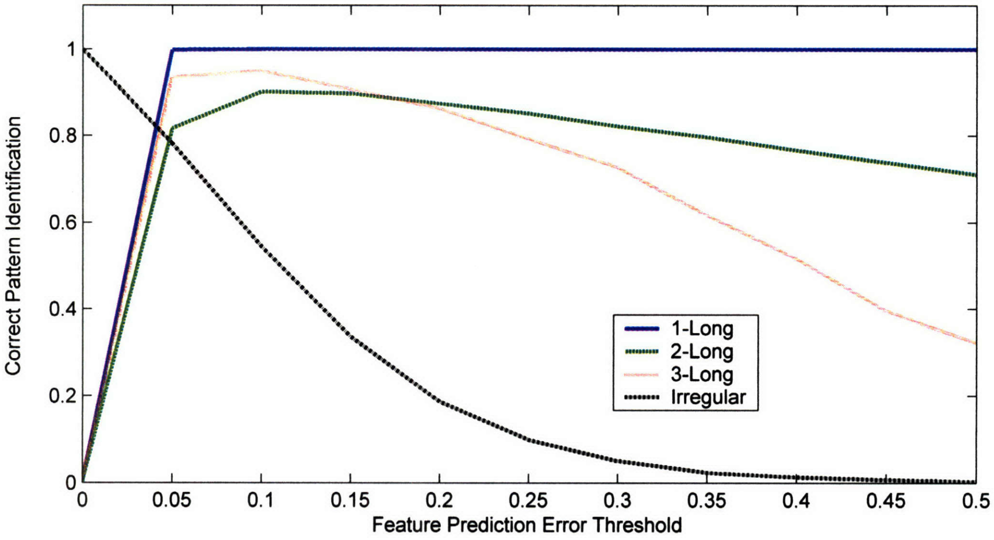


Figure 5-14. Pattern identification rate for a sweep of the prediction error threshold.

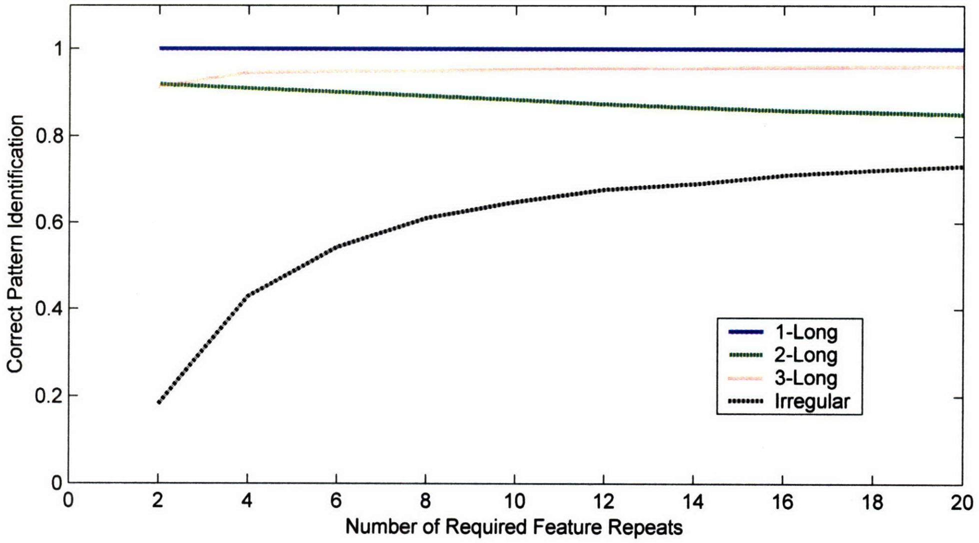


Figure 5-15. Pattern identification rate for a sweep of the minimum number of contiguous features with prediction errors below threshold.

5.6 Conclusion

In this chapter, we have presented a time-domain representation for patterns of glottal events. The features we have devised attempt to represent the underlying relationships between neighboring events in a region of speech and attempt to normalize out the local amplitude and timing scale. Instead, only relative timing and amplitude patterns are captured. We have shown that the feature space defined by this representation allows for separation of different event patterns based on the number of repetitions they contain (1-long, 2-long, etc) as well as the specific sequence embodied in the repeating pattern.

In Chapter 7, we will explore one particular application of the features we have derived here, which is automatic speaker recognition. We will see evidence that these features carry speaker-dependent information and that they can provide complimentary information when fused with spectrally-based features.

Chapter 6

Automatic Extraction of Glottal Events from Speech

In the preceding chapters, we have described a theory for speech represented as a series of filtered impulses. In this chapter, we describe methods by which such a representation may be extracted from natural speech. This chapter additionally explores the meanings of these underlying impulses in terms of the physiology and acoustics of speech.

The model that we have presented for the generation of voiced speech in Chapter 2 is a series of pure impulses, the *glottal events*, each convolved with a mixed-phase *composite impulse response* to produce an acoustic pressure signal. Our primary goals are to (1) estimate the composite impulse response for a given speech region and (2) generate the source impulse sequence using an inverse filter derived from the composite impulse response. Our motivation for obtaining this representation is the use of timings and amplitudes of the impulsive glottal event sequence in automatic speech, speaker, language, and dialect recognition systems.

In particular, we examine the application of an algorithm called minimum-entropy deconvolution (MED) [74] to the problem of determining the composite impulse response and associated source impulse sequence in sections with nonmodal glottal events. The use of MED on near-modal speech to derive a pulse-like residual has been studied previously with promising results [21, 41, 76]. We build upon the previous work by showing MED's application to deriving an impulse-like signal from highly-nonmodal synthetic and natural phonation and to estimating composite impulse responses for such signals. In addition, we propose a hybrid method that combines MED with conventional linear prediction. Evidence is presented that this hybrid method has benefit over MED alone for composite impulse-response estimation by being more robust to certain effects of short-time windowing.

The chapter is outlined as follows. We begin by describing three different methods of deconvolution, with linear-prediction described in Section 6.1, minimum-entropy deconvolution in Section 6.2, and the hybrid method in Section 6.3. Section 6.4 then discusses the method used to transform the outputs of these techniques into discrete impulses representing the glottal events. In the final two sections of the chapter, we evaluate the performance of our techniques in terms of how well the composite impulse response is estimated in Section 6.5 and the validity of the extracted glottal events in Section 6.6.

6.1 Linear Prediction-Based Deconvolution

A classical way to obtain the speech source from a measured acoustic signal is using the linear-prediction residual. The goal of this set of inverse-filtering techniques is to deconvolve the response of the vocal tract from a measured acoustic pressure signal outside the mouth [4, 54, 56]. The basic principle of linear-prediction-based inverse filtering is that the vocal tract acts as an all-pole minimum-phase linear system. The filter coefficients of such a system can be estimated approximately using autocorrelation- or covariance-based linear prediction over a window of speech. Under ideal conditions, this process yields a residual signal which has spikes at the points of an underlying impulsive excitation.

Some authors iterate upon the autocorrelation-based method, recalculating the residual during the closed-phase of the phonation cycle using the covariance method of linear prediction. This method uses an estimate of the times at which the glottis opens and closes which are estimated during the initial autocorrelation-based pass [54] or using other methods [18].

Inverse filtering for arbitrary underlying glottal-event heights and timings remains an unsolved problem. Hedelin and Huber [25] have specifically attempted to apply a version of inverse filtering to nonmodal speech, but use the results to estimate an underlying "pitch," not individual glottal events. Most existing linear-prediction-based algorithms incorporate a fundamental-frequency estimation step. This component allows for the adjustment of analysis window size and aids in the estimation of each instant of glottal closure, assuming that it occurs one pitch period from the previous one. As we saw in Chapter 4, an unambiguous fundamental frequency does not exist for a given sequence of glottal events. Also, for irregular sequences of glottal events, the spectrum loses much of its harmonic structure. This finding, discussed in Chapter 3, may impede the ability of conventional algorithms to estimate the fundamental frequency as well as alter the underlying formant structure.

6.1.1 Formulation

Autocorrelation-based linear prediction is formulated as follows. We estimate or "predict" the current observed signal at time *n* using a weighted sum of the previous signals. This is equivalent to assuming an all-pole model.

$$\hat{y}[n] = \sum_{k=1}^{N} \alpha_k y[n-k] \tag{6.1}$$

We then solve for the values of the parameter α_k by minimizing the mean-square error of the resulting signal, i.e., we compute:

$$\arg\min_{\alpha_k}\left(E\left\{e^2[n]\right\}\right) \text{ where } e[n] = (y[n] - \hat{y}[n])$$

where $\hat{y}[n]$ is given in (6.1).

A key assumption of the model underlying linear prediction is that the system is all-pole. It additionally always yields a solution that is minimum phase when the autocorrelation method is used [56]. As we shall see, this minimum-phase assumption may negatively affect pulse extraction by incorrectly estimating the poles brought about by the shape of the source waveform.

6.1.2 Algorithm

Unless otherwise noted, the algorithm to perform linear prediction in this thesis is as follows. Readers interested in implementation details may refer to Appendix B for a more complete description.

(1) Apply a 20-ms Hamming window at a 10-ms frame interval.

(2) Find the linear-prediction coefficients for this 20-ms window of speech using the autocorrelation method.

(3) Predict the current sample based on the previous *N* samples where *N* is the filter order. The coefficients used to do this are updated every 10 ms.

(4) Subtract out the estimate of the current sample from the measured value, resulting in a residual signal.

6.1.3 Examples using Nonmodal Phonation

An example of autocorrelation-based linear prediction operating on synthetic speech is shown in Figure 6-1. The three waveforms plotted are the acoustic waveform of the vowel /ɑ/, the source volume velocity, and the linear-prediction residual. In order to model natural speech, we have approximated the radiation characteristic using a first-order high-pass filter. As we can observe, the residual we obtain with linear prediction is not a set of impulses. The same behavior is seen for natural speech in Figure 6-2, where the residual appears even less impulse-like. In later sections, we will compare these responses with that of other deconvolution techniques that we develop.

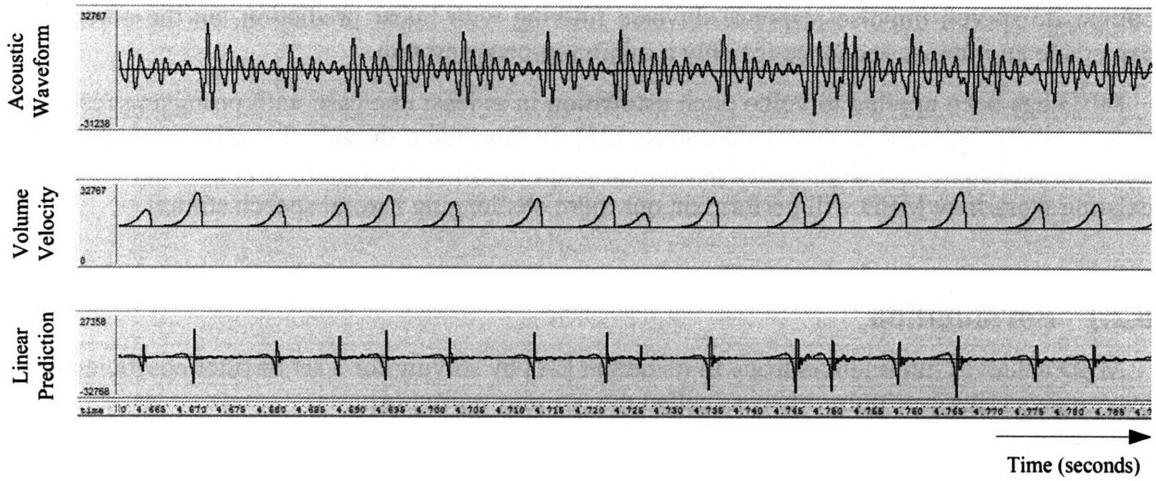


Figure 6-1. Application of order-15 linear prediction to synthetic vowel /ɑ/ with irregular glottal event timings and amplitudes. The bottom panel show the application of the algorithm to the acoustic waveform and corresponding volume-velocity source in the top two panels.

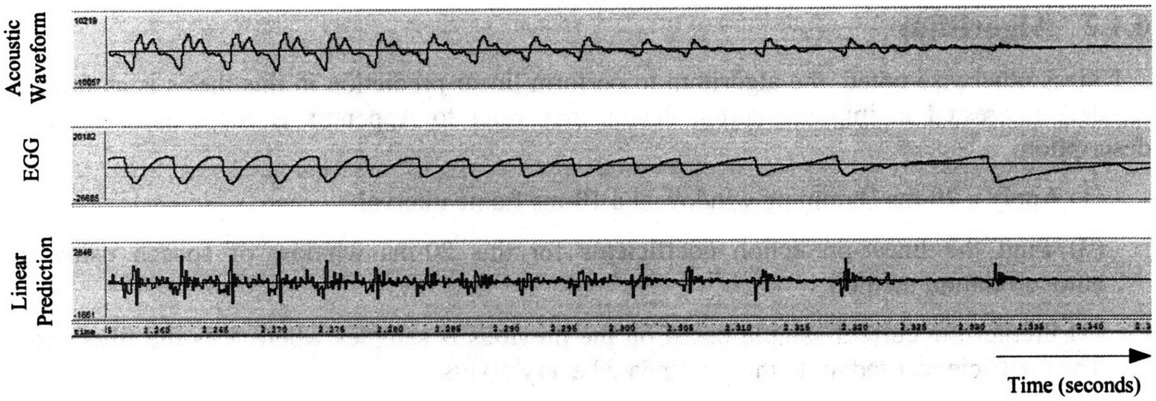


Figure 6-2. Output of order-15 linear prediction for a natural speech signal.

6.2 Minimum-Entropy Deconvolution

In this thesis, we have represented phonation as the result of convolving a source impulse sequence with a set of composite impulse responses. A reasonable approach to the decomposition problem, then, is to design an inverse-filtering method that yields an impulse-like residual. An alternative to linear prediction is a technique from the geophysical literature called minimum-entropy deconvolution (MED) [74]. This method contrasts linear prediction primarily in two ways. First, while linear prediction minimizes the predictability in the residual, MED attempts to minimize disorder. In effect, it creates a filter that generates the most "impulse-like" residual for a given input. The second way in which MED and linear prediction differ is that the MED filter is able to generate pure impulses from a mixed-phase system such as the one associated with the composite speech impulse response. Inverse filtering with linear prediction, on the other hand, can create an impulse sequence for only a minimum-phase system.

MED has been applied to voice pitch estimation in at least one case with promising results on both synthetic and natural sustained vowels [21]. In this study, though, the only nonmodal effect that was looked at was slight pulse-period variation in synthetic stimuli. It is thus not clear from existing work how MED will perform on our more challenging natural speech stimuli.

6.2.1 Formulation

MED achieves minimum entropy in its output [21] by solving for a set of filter coefficients that maximize a criteria of pulse-likeness called the *varimax norm* [74]:

$$V = \sum_{j=1}^{N} x^4[j] \Bigg/ \left(\sum_{j=1}^{N} x^2[j] \right)^2 \tag{6.2}$$

where $x[j]$ is the input signal and N is the length of the analysis window. The varimax norm is an instance of a normalized 4th-order moment, *kurtosis*, and is larger for signals with a small number of sharp pulses and smaller for signals with less structure. The problem of maximizing the varimax norm of the residual is nonlinear and requires an iterative approach as detailed in [74] and [21].

While using a single long analysis window with MED has been performed in the literature [21], there are few attempts to perform MED in a frame-wise manner. Five hurdles that face such efforts are:

(1) Compensation for a delay term introduced by the maximum-phase component of the filter

(2) Normalization of the residual to provide consistent output from frame to frame

(3) Choice of window

(4) Finding global solutions

(5) Implementation of block convolution.

The first problem, compensation for the delay term, comes about because MED generates a set of inverse-filter coefficients that yield an impulse-like residual but does not specify the correct delay term. We assume that the underlying system has the all-pole mixed-phase model:

$$H(z) = \frac{1}{\prod_{k}(1 - a_k z^{-k})\prod_{l}(1 - b_l z^{l})} \text{ where } |a_k| < 1 \text{ and } |b_l| < 1 . \tag{6.3}$$

MED yields an FIR filter that inverse filters the system in (6.3). Because (6.3) is noncausal in general, the correct inverse filter is also noncausal. We can calculate the correct amount to delay the MED coefficients by first interpreting the inverse filter as causal with the first element at the origin. The resulting filter will be a delayed version of the correct inverse filter:

$$G_{MED}(z) = z^{-n_0}\hat{H}^{-1}(z) = z^{-n_0}\prod_{k}(1 - \hat{a}_k z^{-k})\prod_{l}(1 - \hat{b}_l z^{l}) \text{ where } |\hat{a}_k| < 1 \text{ and } |\hat{b}_l| < 1 \tag{6.4}$$

where n_0 is equal to the number of zeros outside of the unit circle. In order to remove the delay term z^{-n_0}, we must apply a compensating term of z^{n_0} to the causal transfer function, equivalently shifting the impulse response to the left by n_0 samples.

The second problem that arises is that the MED output requires normalization. The coefficients that are output may be scaled arbitrarily because this scaling does not affect the value of the varimax norm described by (6.2). This behavior differs from linear prediction in that the inverse filter derived from linear prediction always has a value of 1 in its first bin, scaling it in a consistent way. There are many possible approaches to dealing with this scaling problem and it is a subject for future research. The approach that we chose was to normalize the energy in each output frame to a local energy estimate using the acoustic speech signal. The goal was to provide an output waveform with similar local energy characteristics to the input, although admittedly this choice makes the amplitude envelope of pulses in the residual dependent upon the vocal-tract filter. A vocal tract with high losses such as during sounds like nasals, for example, may lead to lower amplitude impulses than a vowel produced with the same source.

MED is also sensitive to the window used. For consistency with linear prediction, we have used a 20-ms window with a 10-ms frame rate throughout this thesis. This choice is somewhat arbitrary, but the intent is to be consistent. Window shape is another difficult choice. Steep edges tend to yield impulsiveness in the residual because they present a discontinuity to MED. Again, for consistency with linear prediction, we chose to use a Hamming window.

Another problem that comes about when implementing MED is that the MED algorithm does not yield a global maximum. Depending on the initial values used, it finds different solutions. Other authors have noted this problem. Our solution is to randomly initialize the coefficient values a number of times and take the solution yielding the maximum varimax-norm value. Finding the global maximum may take many of these monte-carlo trials, so we allow for some suboptimal solutions. A practical number of trials is 10, empirically found to converge to a global maximum for many cases, although some experiments use 100 trials where noted.

The final issue that arises is synthesis of the residual signal using the processed blocks. In this thesis, we used an overlap-save method of block convolution [49], with the implementation described in Appendix B, although an overlap-add approach could also have been devised. The overlap-save approach allowed us to compute the local residual *first* and select the valid portion from the resulting signal. This made it relatively simple to compensate for the delay term from noncausal filtering as previously described. Obvious edge effects resulting from block convolution, such as doubled pulses, were not observed between neighboring frames, although this topic should be analyzed more rigorously in the future.

6.2.2 Algorithm

In this thesis, unless otherwise noted, MED is performed as follows. Readers interested in implementation details may refer to Appendix B for a more complete description.

(1) Apply 20-ms Hamming window at a 10-ms frame interval.

(2) Find the MED coefficients for this 20-ms window of speech using ten iterations and ten monte-carlo trials

(3) Calculate the number of poles outside the unit circle—in order to yield a stable inverse filter, these must be made non-causal.

(4) Apply the MED-derived inverse filter, positioned with the leftmost coefficient in the zeroth bin.

(5) Shift the output of the MED-derived filter to the left by a number of samples equal to the number of poles outside the unit circle.

(6) Use ("save") only the middle 10-ms worth of samples. This is the overlap-save method of block convolution [49].

(7) Normalize the output to an estimate of the local acoustic energy.

(8) Adjust the polarity of the residual in each frame so that the largest value is positive. The polarity of the residual resulting from MED is ambiguous.

6.2.3 Examples using Nonmodal Phonation

MED may be applied to continuous speech using analysis with overlapping windows. As discussed in the previous section, we process the speech using 20-ms Hamming windows with 50-percent overlap. Creation of the residual from MED filters is performed using 20-ms windows at a 10-ms frame rate using overlap-and-save synthesis.

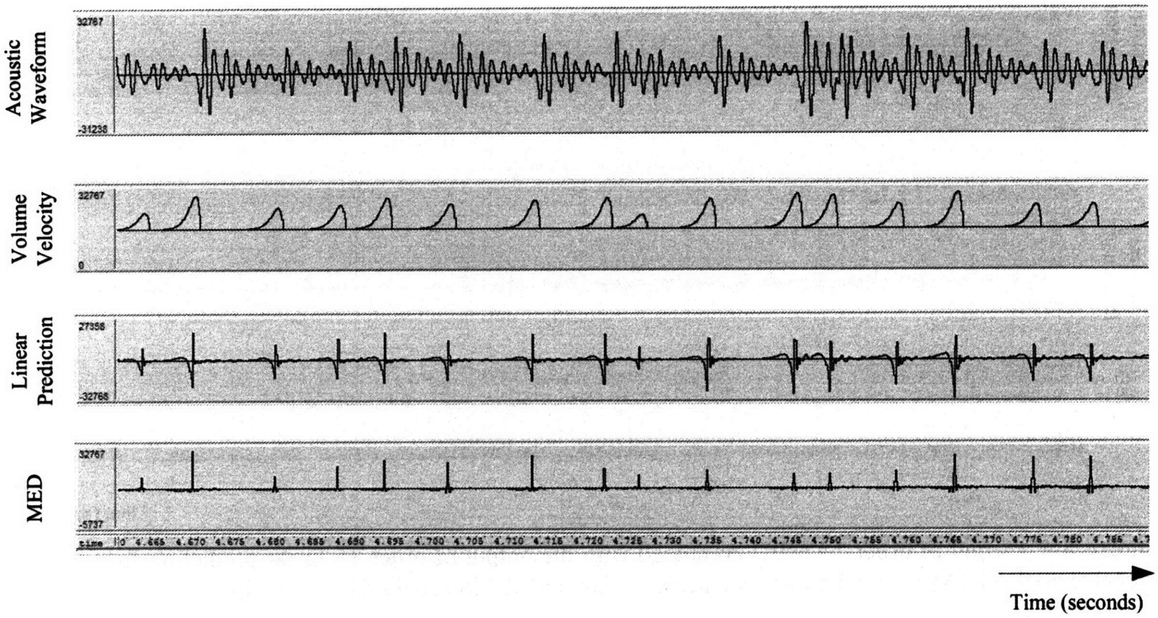


Figure 6-3. Application of order-15 MED and linear prediction to synthetic phonation with irregular glottal event timings and amplitudes. The bottom two panels show the application of the algorithms to the acoustic waveform and corresponding volume-velocity source in the top two panels. The MED residual is more impulse-like than the linear-prediction residual.

Figure 6-3 shows the results of applying MED to synthetic nonmodal phonation. The top two panes depict the synthetic glottal events and the resulting acoustic signal. This particular case contains source pulses with nonuniform interpulse timings. The vowel /ɑ/ was created using three formants and a first-difference radiation characteristic. The two remaining panes show the results of using a 15th-order linear prediction and a 15th-order MED to obtain a residual. The output of linear prediction is less pulse-like than the output of MED, with activity to the left of the responses to glottal closure. In contrast, MED yields a qualitatively more impulse-like residual; there is less activity to the sides of the main pulses. In both examples, the most impulse-like activity corresponds to near the closing point of the modeled synthetic volume-velocity source. Figure 6-4 repeats this comparison for a natural nonmodal utterance. We see that MED is again qualitatively more impulse-like.

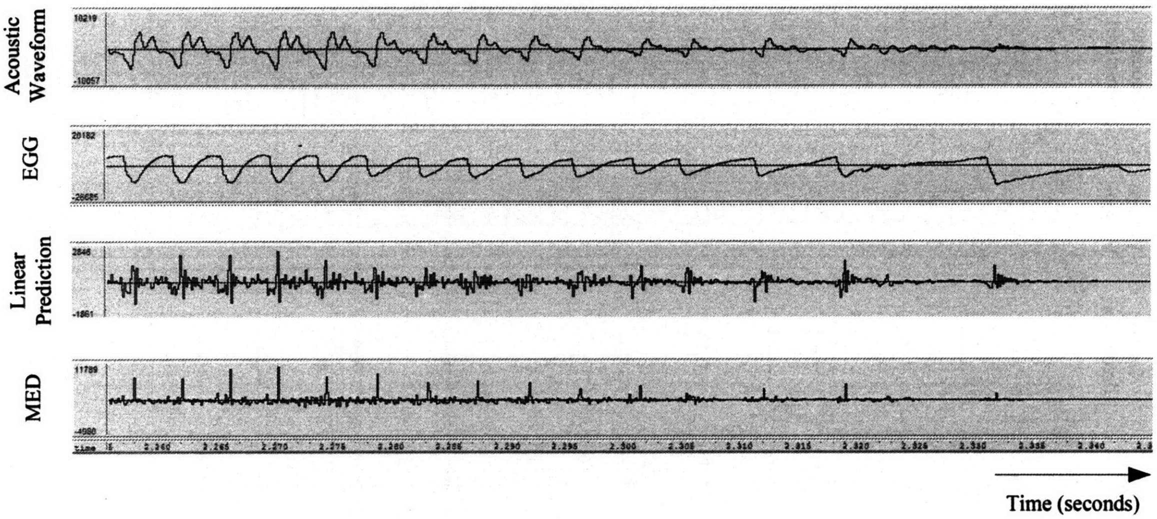


Figure 6-4. Output of order-15 MED compared with the output of order-15 linear prediction for a natural speech signal. The MED residual is more impulse-like than the linear-prediction residual.

Using the varimax norm as an objective measure of impulse-likeness, we find that MED is the most impulse-like, having the highest varimax norm, compared with linear prediction and the acoustic waveform for both the natural and synthetic cases. For the 94-ms segment of natural nonmodal phonation in Figure 6-4, MED residual has varimax norm of 0.0571 versus 0.0128 for linear prediction and 0.0065 for the original acoustic signal. For the 130-ms synthetic case in Figure 6-3, MED has a varimax norm of 0.0673 versus 0.0188 for linear prediction and 0.0036 for the original acoustic signal.

6.3 Hybrid Linear-Prediction/MED Approach

We have shown that MED produces a higher varimax norm than linear prediction, but this result is not surprising. The composite impulse response we aim to recover is represented well as the output of a mixed-phase system [43, 56]. Linear prediction effectively "flips" maximum-phase pole estimates to their minimum-phase reciprocal locations. In this section, we present a modification of standard linear prediction capable of obtaining an improved mean-squared error fit to mixed-phase impulse responses over linear prediction. The goal is to combine the well-behaved spectral-fitting of linear prediction with the impulse-like residual of MED.

6.3.1 Formulation

The least-mean-square residual criteria upon which linear prediction is based cannot be used to determine which location for each pole—inside or outside the unit circle—is "better." Instead we need an additional criterion. We propose to use the varimax norm criteria from MED to choose among the possible configurations. In this way, a filter is created that yields a linear-prediction-based residual with the maximum pulse-likeness possible. We will show that this combination allows us to estimate realistic composite impulse response shapes, especially in cases when MED does not. For a specified order, the solution we use is to (1) try every possible configuration of zeros inside and outside the unit circle, (2) find the residual resulting from each configuration, and (3) keep the configuration yielding the maximum varimax norm. The complexity of this technique grows exponentially with increasing linear-prediction order, but is manageable for typical orders. In this chapter, we show results for an order-15 case.

6.3.2 Algorithm

The hybrid method has the same shift, normalization, windowing, and other issues as described for the MED algorithm. The algorithm described as follows is used in this thesis to perform the hybrid method unless otherwise noted. Readers interested in implementation details may refer to Appendix B for a more complete description.

(1) Apply a 20-ms Hamming window at a 10-ms frame interval.

(2) Find the linear-prediction coefficients for this 20-ms window of speech

(3) Try every combination of poles—inside and outside the unit circle. Record the varimax norm for each residual.

(4) Use the configuration of poles with the highest varimax norm.

(5) Calculate the number of poles outside the unit circle.

(6) Apply the hybrid-derived inverse filter.

(7) Shift the output of the hybrid-derived filter to the left by a number of samples equal to the number of poles outside the unit circle.

(8) Use ("save") only the middle 10-ms worth of samples. This is the overlap-save method of block convolution.

(9) Normalize the output to an estimate of the local acoustic energy.

(10) Adjust the polarity of the residual in each frame so that the largest value is positive. The polarity of the residual resulting from the hybrid method is ambiguous.

6.3.3 Examples using Nonmodal Phonation

Figure 6-6 compares linear prediction, MED, and the new hybrid method for the same segment of natural speech as was used in Figure 6-4. We can see that the hybrid method, having a varimax norm of 0.0226, yields a less-impulsive residual than MED ($V = 0.0571$), but a more-impulsive residual than linear prediction ($V = 0.0128$) or the acoustic signal ($V = 0.0065$). For the synthetic case shown in Figure 6-5, we find that the hybrid method, like MED, yields a high varimax norm of 0.0960 compared with linear prediction ($V = 0.0188$) or the original acoustic signal ($V = 0.0036$). In this case, it actually performs better than MED, which has a varimax norm of 0.0673.

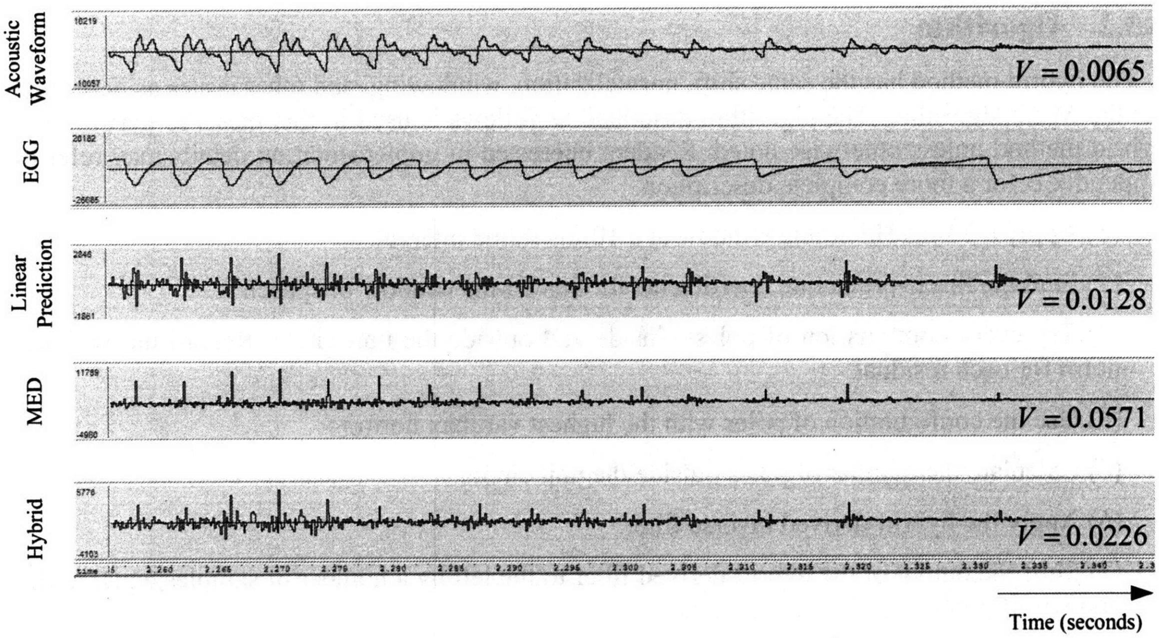


Figure 6-5. Application of order-15 MED, linear prediction, and hybrid method to synthetic phonation with irregular glottal event timings and amplitudes. The bottom three panels show the application of the algorithms to the acoustic waveform and corresponding volume-velocity source in the top two panels. The hybrid residual is more impulse-like than either the linear-prediction residual or the MED residual. The varimax norm V is shown for the acoustic waveform and the results of inverse filtering.

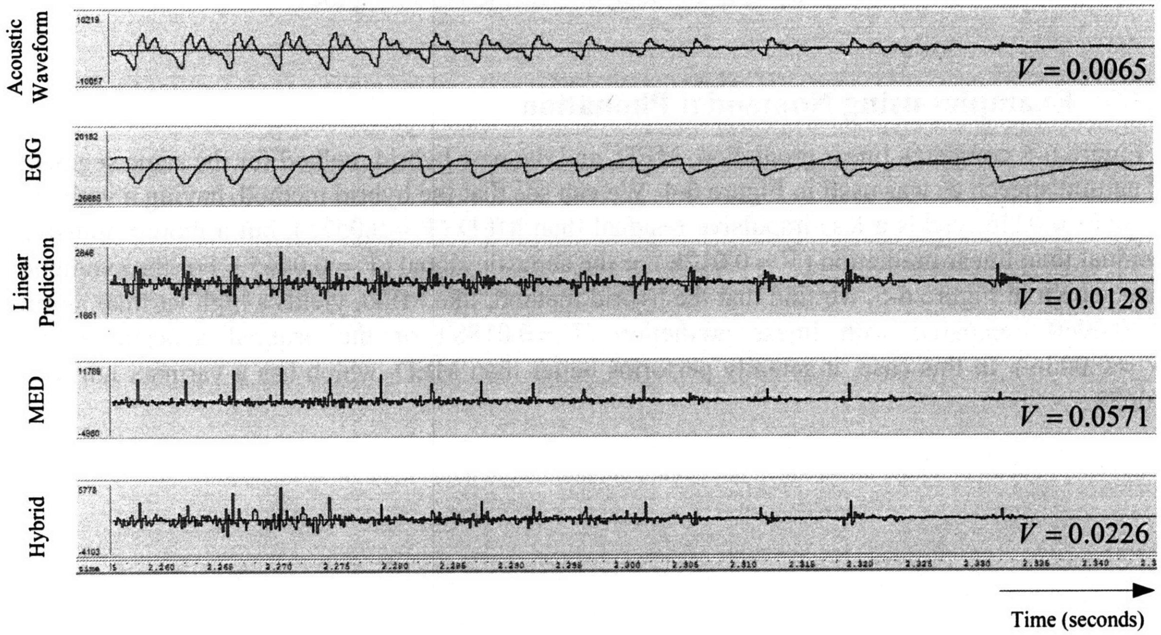


Figure 6-6. Output of order-15 hybrid method compared with order-15 MED compared and order-15 linear prediction for a natural speech signal produced by a female talker. Both the hybrid and MED residuals are more impulse-like than the linear-prediction residual, but MED is the most impulsive. The varimax norm V is shown for the acoustic waveform and the results of inverse filtering.

6.4 Discrete Glottal-Event Extraction

The outputs of linear prediction, MED, and the hybrid method do not yield discrete impulses. Once the output of inverse filtering is obtained, we must further process it with a peak-picking stage in order to extract the locations and amplitudes of the impulses. Two parameters are used in this process, the *dead time* and *threshold.* Figure 6-7 illustrates the glottal-event extraction procedure schematically.

The process begins by applying a threshold to the raw output of the inverse filter, keeping only samples with value greater than this value. The goal of this part of event extraction is to eliminate elements in the inverse-filtered signal due to noise and other deviations from the impulsive source model. The next stage is to determine whether any samples occur closer to each other than the specified dead time. Only the sample having the largest value over the dead time is kept. This pruning procedure is implemented iteratively, by keeping the largest remaining sample and eliminating all smaller than it within plus or minus the dead time. Applying the dead-time constraint allows responses to glottal events in the residual that are not perfect impulses to yield single discrete events. As we have seen, such additional responses close to the times of glottal events are common, especially in real speech.

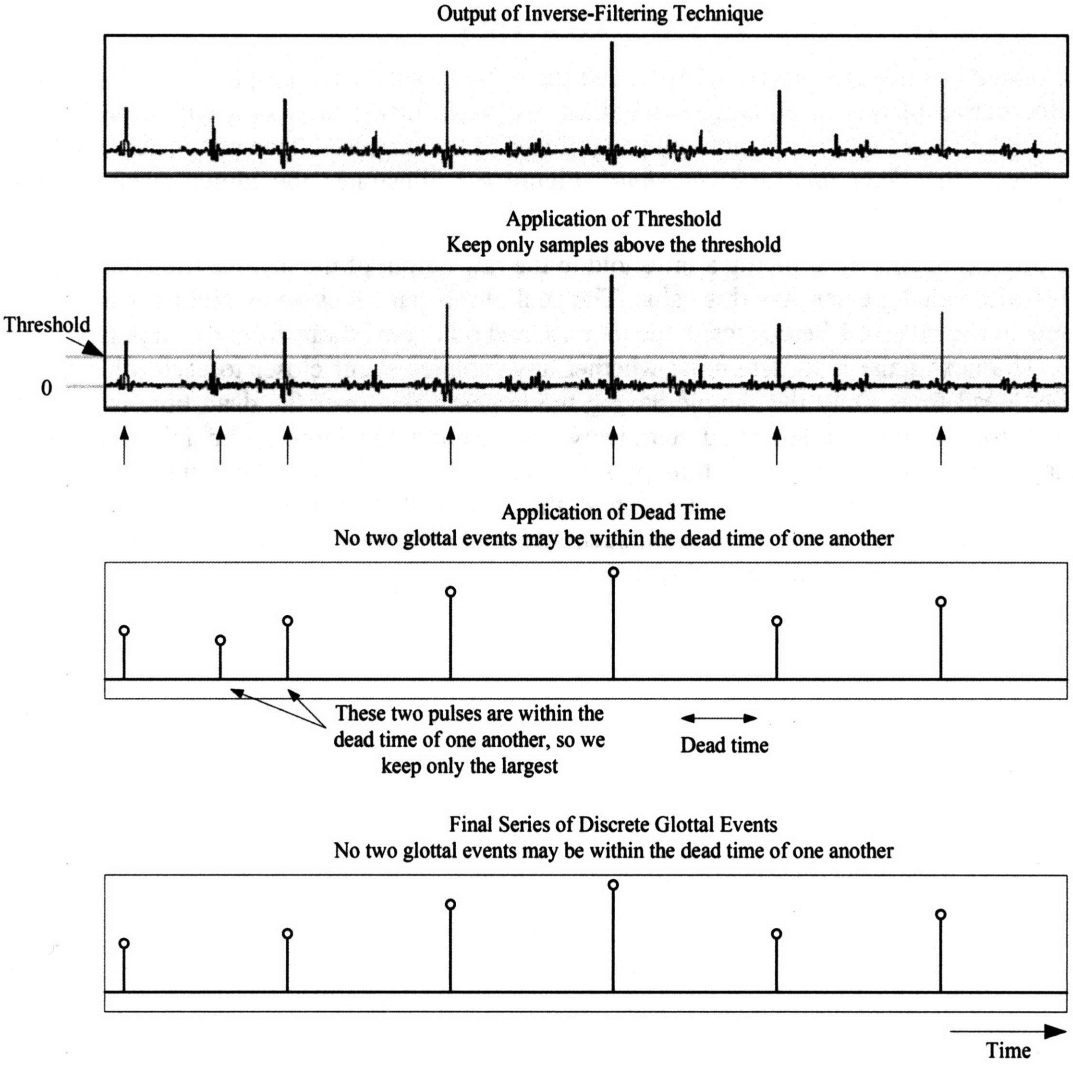

Output of Inverse-Filtering Technique
Application of Threshold
Keep only samples above the threshold
Threshold
0
Application of Dead Time
No two glottal events may be within the dead time of one another
These two pulses are within the dead time of one another, so we keep only the largest
Dead time
Final Series of Discrete Glottal Events
No two glottal events may be within the dead time of one another
Time

Figure 6-7. Schematic of the process of extracting discrete glottal events from the output of linear prediction, MED, or the hybrid approach.

6.5 Evaluation of Composite Impulse Response Estimates

One way to evaluate the performance of the inverse filtering methods is to use synthetic speech and compare the estimated composite impulse responses to the known composite impulse response. In this section, we show the normalized sum-squared error between known and estimated composite impulse responses for a set of synthetic vowels. The normalized sum-squared error is found using

$$SSE_{norm} = \frac{\sqrt{\sum_k (h[k] - a\hat{h}[k-\theta])^2}}{\sqrt{\sum_k h^2[k]}}$$

where a and θ are parameters found using

$$\arg\min_{a,\theta} \sum_k (h[k] - a\hat{h}[k-\theta])^2$$

where $h[k]$ is the known impulse response and $\hat{h}[k]$ is the estimated impulse response.

The results are presented as a set of histograms with the normalized sum-squared error on the x-axis and the percentage of frames obtaining each sum-squared error on the y-axis. As described earlier, analysis is performed using 20-ms Hamming windows with 10-ms frames. The input signals to each analysis method are four five-second synthetic vowels. Each utterance consists of 125-Hz phonation, with a two-pole maximum-phase source response and a three-formant, minimum-phase vocal tract response. The amount of jitter was uniformly distributed over ±2 ms. The four vowels that were synthesized were /ɑ/, /i/, /u/, and /æ/. Results both with and without a first-order high-pass radiation characteristic were obtained.

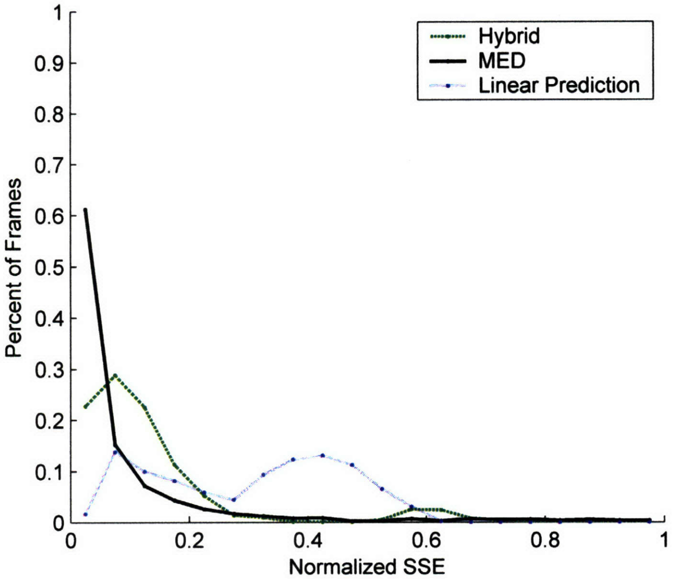


Figure 6-8. Histogram of the percent of frames yielding a certain normalized sum-squared error for each of three different inverse-filtering techniques. Four synthetic vowels, /ɑ/, /i/, /u/, and /æ/ were used with no radiation characteristic. The amount of jitter in the input is uniformly distributed over ±2 ms. The spacing of the histogram bins is 0.05.

In Figure 6-8, the sum-squared error performance of each technique is shown for the case where no radiation characteristic was applied. This allows us to observe how the analysis techniques perform when the input follows an all-pole model. We see that MED outperforms all other methods on average, with over 60-percent of its frames yielding an sum-squared error less than 0.05. The hybrid method has the next-best performance, followed by linear prediction. The poor performance of linear prediction can be attributed to its inability to represent the maximum-phase part of the composite impulse response. This behavior is shown in Figure 6-9 where for a given analysis frame, the hybrid method and linear prediction are both able to estimate the maximum-phase part of the response, but linear prediction converts the maximum-phase portion to minimum-phase. We have used the hybrid method for comparison with linear-prediction since they both use the same estimate of the underlying poles.

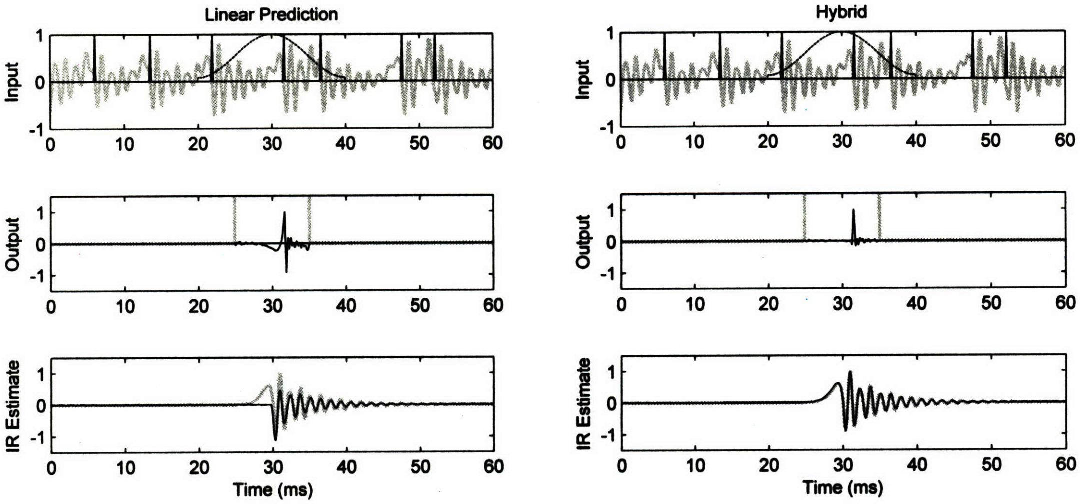


Figure 6-9. Demonstration of improved impulse-response estimation using a mixed-phase system. The synthetic vowel /ɑ/ input signal (top panel) has been created with the impulses illustrated. Processing the input with the illustrated analysis window yields an inverse-filtered output signal for one frame with boundaries denoted by the grey vertical lines (middle panel), The composite impulse response estimated is illustrated using a dark line compared with the known system response drawn with a light line (bottom panel).

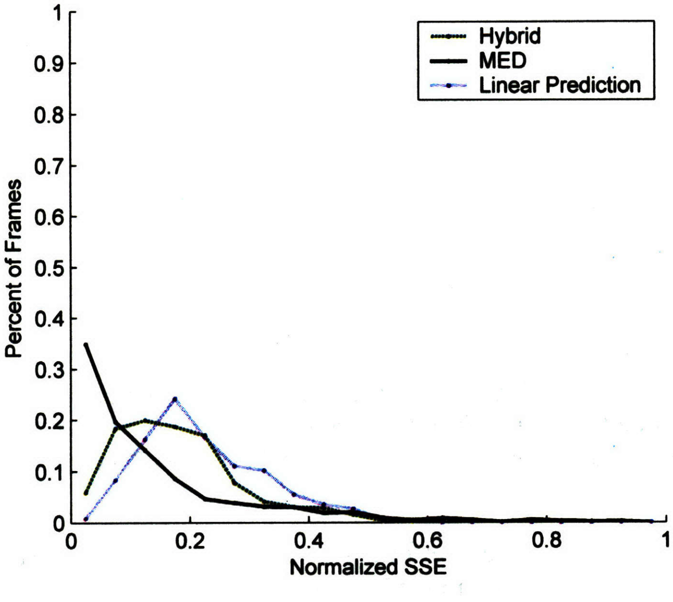


Figure 6-10. Histogram of the percent of frames yielding a certain normalized sum-squared error for each of three different inverse-filtering techniques. Four synthetic vowels, /ɑ/, /i/, /u/, and /æ/ were used with a first-order radiation characteristic, being modeled as a single zero. The amount of jitter in the input is uniformly distributed over ±2 ms. The spacing of the histogram bins is 0.05.

In Figure 6-10, we perform the sum-squared error histogram analysis for the case where the synthetic stimuli include the radiation filter, being modeled as a single zero. This case using the radiation characteristic is comparable to the earlier synthetic examples including Figure 6-5. For MED and hybrid method, we see that the sum-squared error is higher than it was using no radiation characteristic. This is likely because the underlying system no longer fits the all-pole assumption. Despite the degradation in performance, MED still performs the best, with over 30-percent of the frames having an sum-squared error value under 0.05. Rather than undergoing a reduction in performance, linear prediction actually performs *better* with the radiation characteristic than without it. As shown for an example in Figure 6-11, this is because much of the effect of the maximum-phase source response has been eliminated by the radiation characteristic. The resulting near-minimum-phase source response is more amenable to an all-pole minimum-phase fit by linear prediction.

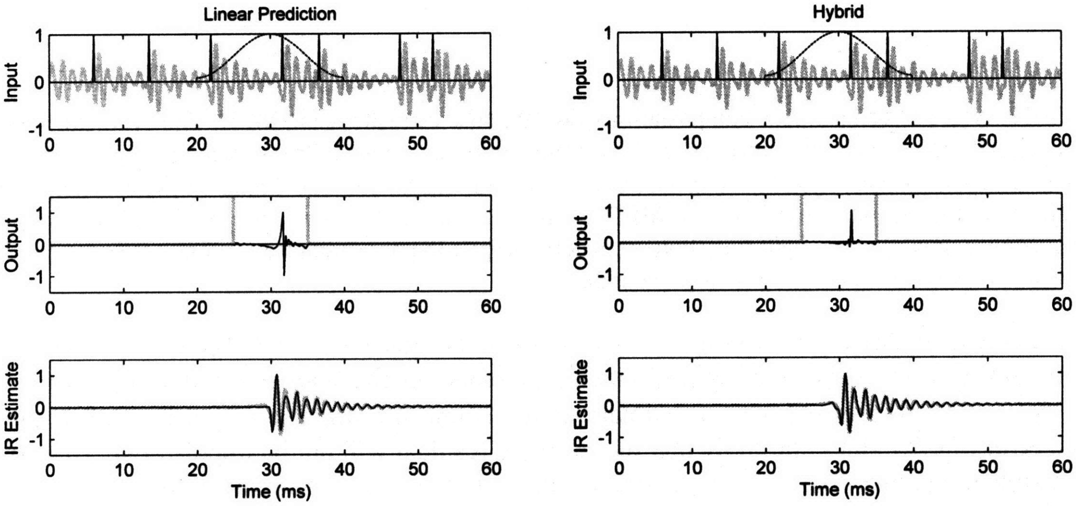


Figure 6-11. Demonstration of improved impulse-response estimation using a mixed-phase system. The synthetic vowel /ɑ/ input signal (top panel) has been created with the impulses shown. Processing the input with the illustrated analysis window yields an inverse-filtered output signal for one frame (middle panel). The composite impulse response estimated is illustrated using a dark line compared with the known system response drawn with a light line (bottom panel).

We have shown that for most synthetic frames, MED produces the best estimates of the true composite impulse response in synthetic cases. Even under these conditions, however, MED occasionally yields unrealistic composite impulse-response shapes. In fact, for each of the sum-squared error experiments shown in Figure 6-8 and Figure 6-10, MED yielded the frames with the maximum single sum-squared error even though it performed with the lowest sum-squared error on average.

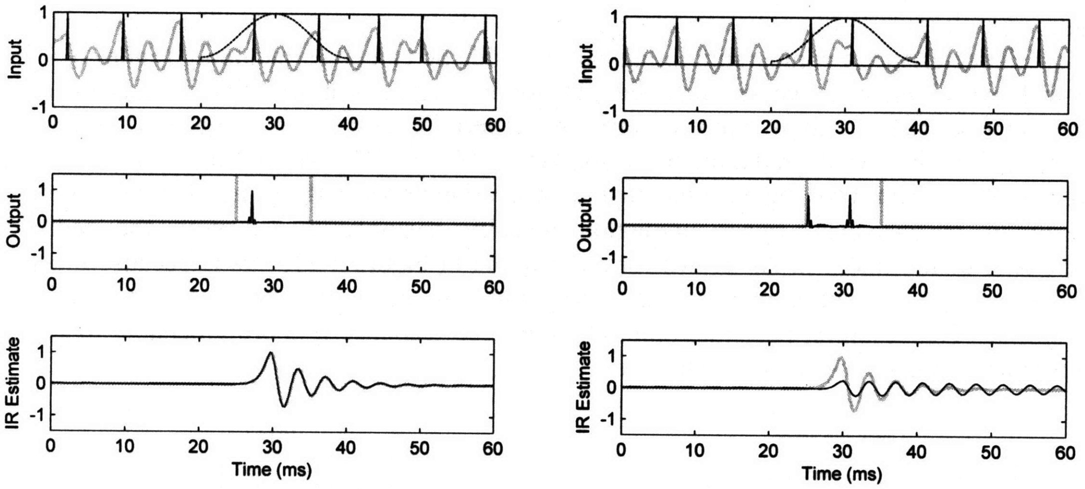
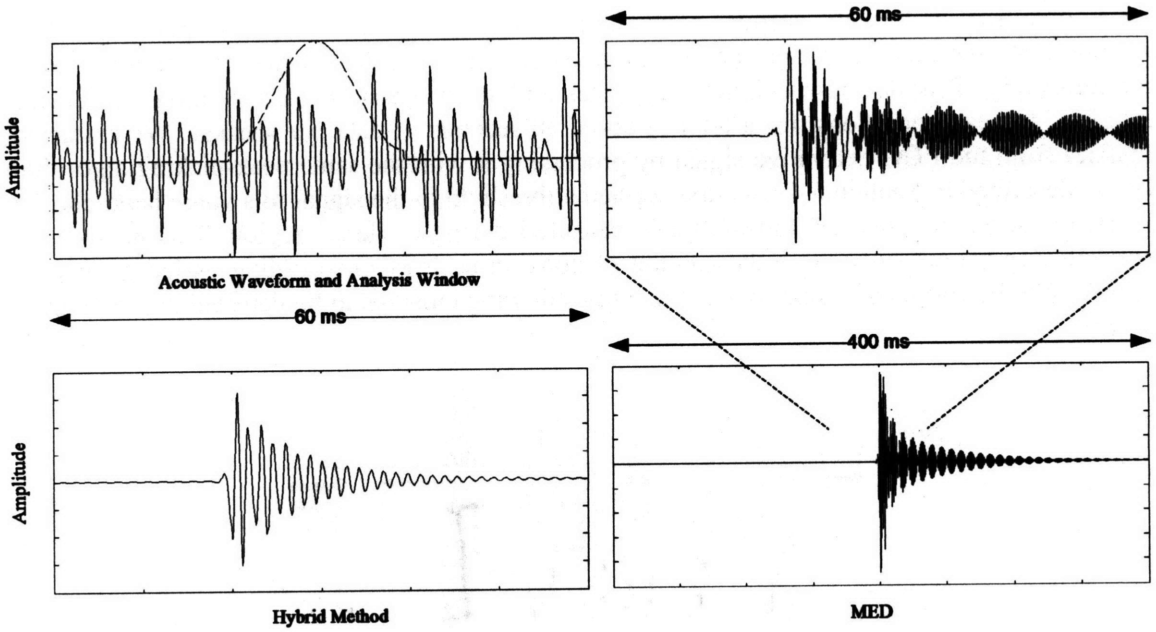


Figure 6-12. A good versus a poor composite-impulse response estimate using MED. The synthetic vowel /i/ input signal (top panel) has been created with the impulses shown. Processing the input with the illustrated analysis window yields an inverse-filtered output signal for one frame (middle panel). The composite impulse response estimated by MED is shown using a dark line compared with the known system response drawn with a light line (bottom panel). Despite the poor impulse-response estimate for the right case, both residuals are impulsive and accurate.

Two examples of MED performing poorly are shown in Figure 6-12 and Figure 6-13. Here, the synthetic vowel /i/ is shown, synthesized without the radiation characteristic. In the right panel of Figure 6-12, we can see that a poor estimate of the composite impulse response, with a sum-squared error of 0.75, does not imply that the resulting residual will not be impulse-like. Not only is the residual impulse-like for this example, these impulses fall near the impulses used for synthesis. This result is compared to a case where the composite impulse response is estimated well, having a sum-squared error of 0.047, at left.

One phenomenon that occurs especially with MED is underestimation of the formant bandwidths. As seen in the example of Figure 6-13, while the hybrid method yields a reasonable impulse response, which decays in about 20 ms, MED estimates a composite impulse response that rings significantly for over 100 ms. As can be seen in the acoustic waveform, the pulses in this example occur close to the edges of the analysis window. This distorts the waveform that MED and linear prediction process. Based on this and other examples, it appears that the hybrid method is not prone to the unrealistic "ringing" behavior that can occur with MED.

The ringing behavior seen in the previous example is likely due to the location of the excitation near the edges of the window. We do not completely understand why MED is sensitive to the position of the signal in the window. These problems, however, illustrate that the hybrid method can produce an improved composite impulse response over MED under certain conditions. By the first example, however, this may not translate to a significantly more accurate estimation of the underlying glottal events.



Acoustic Waveform and Analysis Window

Hybrid Method

MED

Figure 6-13. Application of MED and the hybrid method to a frame of synthetic nonmodal phonation that yields an unrealistic estimated glottal impulse response for MED. The MED impulse response "rings" for nearly 200 ms as can be seen in the bottom-right panel. The input signal is a synthetic vowel /ɑ/ with a radiation characteristic applied.

6.6 Validation of Extracted Pulses for Natural Speech

In the previous section, we have analyzed the performance of our three inverse-filtering techniques in terms of their ability to recover a known composite impulse response from synthetic speech. These results indicate that, on average, MED will recover the most accurate impulse response in a sum-squared error sense, followed by the hybrid method, and linear prediction. In the current section, we analyze the ability of our systems to detect a set of likely glottal events in both real and synthetic speech. This task builds upon the evaluation in the previous section, presenting evidence that our methods can find meaningful glottal events in natural speech.

6.6.1 Experimental Methods

In order to test the glottal-event detection performance of each algorithm, we use a detection-error tradeoff (DET) methodology similar to techniques used in the signal detection community. This process produces a curve that shows how well a given technique finds a series of ground-truth pulses at different threshold levels. For synthetic speech, these ground truth pulses are the impulses used to excite the composite impulse response. For natural speech, one popular measure for ground truth is the electroglottogram (EGG) signal. The EGG signal is derived from a non-acoustic sensor placed on the neck near the larynx which reflects the conductance across the glottis. When the glottis is fully or partially closed, the conductance increases, and when the glottis is open, the conductance decreases. Following convention in the literature, we used the first difference of this signal in order to determine the approximate times of the glottal-closure event. As we will see, the impulses extracted by our techniques represent this glottal-closure event well.

We obtained a speech corpus containing acoustic signals and corresponding EGG measurements and automatic word and phone transcriptions provided by the Air Force Research Laboratory [5]. This database contains six males and seven females, each reading ten TIMIT sentences in clean conditions, for a total of about 30 seconds of data per speaker. We extracted impulses from the EGG-derivative signal by processing it with the same peak-picking algorithm that we described in Section 6.4. We first applied a threshold to the signal to a value hand-chosen for each speaker, roughly 1.5 times a peak measured during a silence region. This hand-tuned approach was necessary because the amount of noise in each EGG recording varied greatly by speaker. We then applied a dead time of ±2.5 ms, allowing closures to be detected at a rate up to 400 Hz.

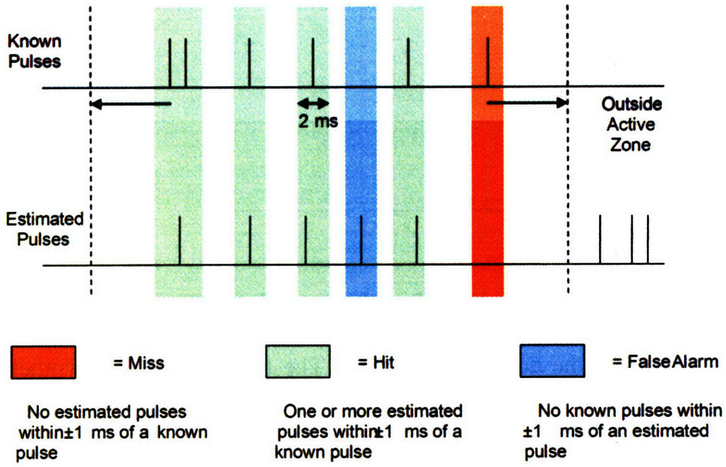

Known Pulses
2 ms
Outside Active Zone
Estimated Pulses
= Miss
= Hit
= FalseAlarm
No estimated pulses within±1 ms of a known pulse
One or more estimated pulses within±1 ms of a known pulse
No known pulses within ±1 ms of an estimated pulse

Figure 6-14. Schematic of the definitions used to calculate the detection accuracy of the objective pulse measures. The "active zone" is defined as a region labeled as a vowel in the transcription—only estimated pulses in this range are considered.

After we estimated the set of glottal-closure events from the derivative of the EGG signal for natural speech, we performed the detection-error evaluation. In order to test the detection ability of the objective pulse measures, we calculated two different variables—misses and false alarms. Figure 6-14 shows schematically how these values were calculated. The process was as follows:

(1) *Setting of the active region*. First, an "active region" was set using the phonetic transcription data provided with the database. Only regions marked as vowels were used. This process was important with running speech as it allowed the analysis to focus on regions with glottal excitation rather than fricatives or other speech sounds.

(2) *Calculation of misses*. Next, each known pulse was compared to the estimated pulses. If there was no estimated pulse within ±1 ms of a known pulse, then a miss was tabulated. This value was divided by the total number of known pulses to arrive at a percentage.

(3) *Calculation of false alarms*. Finally, each estimated pulse not occurring within ±1 ms of a known pulse was counted as a false alarm. This value was divided by the total number of known pulses to arrive at a percentage. This percentage could be greater than one if there were more estimated pulses than known pulses.

The time mentioned above, 1 ms, was used for the detection zone because it was smaller than most pitch periods and yet large enough to capture moderate variation in the measurements.

As will be shown in the next section, we use the measures derived here to plot the percentage of misses versus the number of false alarms for each objective pulse measure. Sweeping the event-detection threshold for each technique traces out a *detection-error tradeoff curve* for each approach, with the percentage of misses at a particular threshold on the y-axis and the number of false alarms on the x-axis.

6.6.2 Detection Experiments

Synthetic Vowels

In order to test the performance of the techniques on synthetic speech, we used the same set of four five-second synthetic vowels as studied in Section 6.5. Each utterance consisted of 125-Hz phonation, with a two-pole maximum-phase source response and a three-formant, minimum-phase vocal tract response. The amount of jitter was uniformly distributed over ±2 ms. The four vowels that were synthesized were /ɑ/, /i/, /u/, and /æ/. Results both with and without a first-order high-pass radiation characteristic were obtained.

Figure 6-15 and Figure 6-16 show the DET plots for the synthetic data using linear prediction, the hybrid method, MED, and a Gaussian random noise signal as input to the event detection stage. Figure 6-15 shows the case for an all-pole system, without the radiation characteristic included. Here, we can see that both MED and the hybrid method exhibit excellent performance, a result that agrees with the accurate composite impulse response estimates in Section 6.5. In comparison to these methods, linear prediction exhibits worse performance, likely due to its poor estimation of the maximum-phase component. This plot supports the assertion that for systems with a significant maximum-phase component, MED and the hybrid method both offer a distinct advantage.

Figure 6-16 presents the DET plot for the three glottal-event extraction methods when the stimuli are processed with a radiation characteristic. As with Figure 6-15, MED and the hybrid method continue to perform nearly perfectly. This plot differs from the previous one, however, in that linear prediction now also performs close to zero-percent miss rate with no false alarms. This result is not surprising, given that the ability of linear prediction to estimate the composite impulse response was much improved when the radiation characteristic was introduced in Figure 6-10. This is likely due to the reduction in the energy of the maximum-phase source component, which is mostly low frequency, by the high-pass response of the radiation-characteristic.

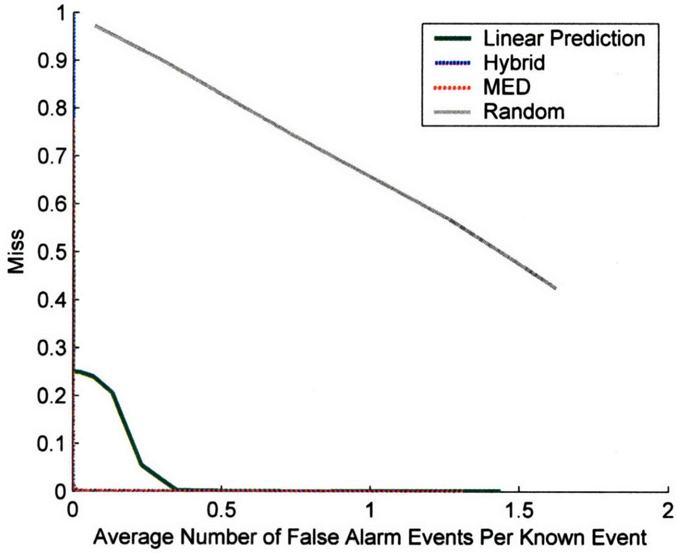


Figure 6-15. Detection-error tradeoff curve for three pulse-finding methods operating on a synthetic jittered signal compared to the known synthetic events. Four synthetic vowels, /ɑ/, /i/, /u/, and /æ/ were used with no radiation characteristic. The amount of jitter in the input is uniformly distributed over ±2 ms. Performance improves as the curves approach the bottom-left corner. Curves for MED and the hybrid method are difficult to see because they are close to the axes. This plot is for a dead time of 2.5 ms.

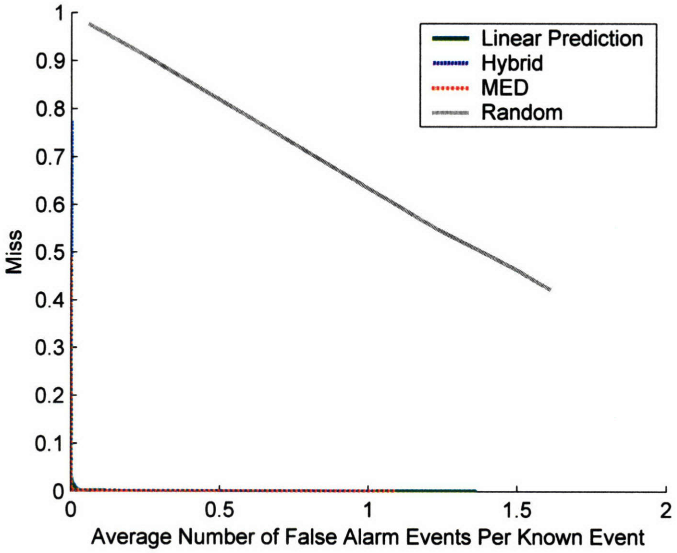


Figure 6-16. Detection-error tradeoff curve for three pulse-finding methods operating on a synthetic jittered signal compared to the known synthetic events. Four synthetic vowels, /ɑ/, /i/, /u/, and /æ/ were used with a first-order radiation characteristic. Here, the amount of jitter in the input is uniformly distributed over ±2 ms. Performance improves as the curves approach the bottom-left corner. Curves for all three methods are difficult to see because they are close to the axes. This plot is for a dead time of 2.5 ms.

Natural Utterances

In order to test the performance of the techniques on natural speech, we used the AFRL TIMIT data produced by both male and female speakers. This set of data allowed us to test the validity of our pulses compared with glottal-closure events over vowel regions. One of the male sentences was omitted from analysis due to a missing transcription.

Figure 6-17 and Figure 6-18 show the DET plots for male speakers using linear prediction, the hybrid method, MED, and a Gaussian random noise signal as input to the event detection stage. We can see that all three techniques perform similarly over both 1-ms and 2.5-ms dead-time conditions. As expected, the number of false alarms increase when the dead time is reduced since this allows the detector to place estimated events closer to one anther. At a threshold yielding one false-alarm per known event, all methods perform at about a 10-percent miss rate.

Figure 6-19 and Figure 6-20 show the DET curves for the set of female speakers. The behavior is different than the males, exhibiting fewer false alarms at a given miss rate. For example, there are fewer than about 0.5 false alarms for every ground truth glottal closure at a miss rate of 10 percent. Additionally, MED provides an advantage over linear prediction and the hybrid method for higher thresholds (towards the left of the curve), allowing under a 0.25 false alarm rate at a miss rate of 10 percent. DET curves for the combined set of male and female data are shown in Figure 6-21 and Figure 6-22.

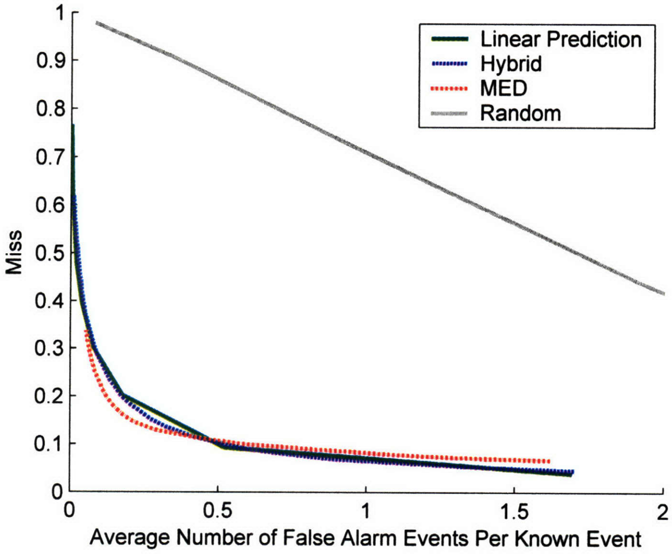


Figure 6-17. Detection-error tradeoff curve for three pulse-finding methods compared to a differentiated-EGG reference signal. Performance improves as the curves approach the bottom-left corner. This is for all males with a dead time of 2.5 ms.

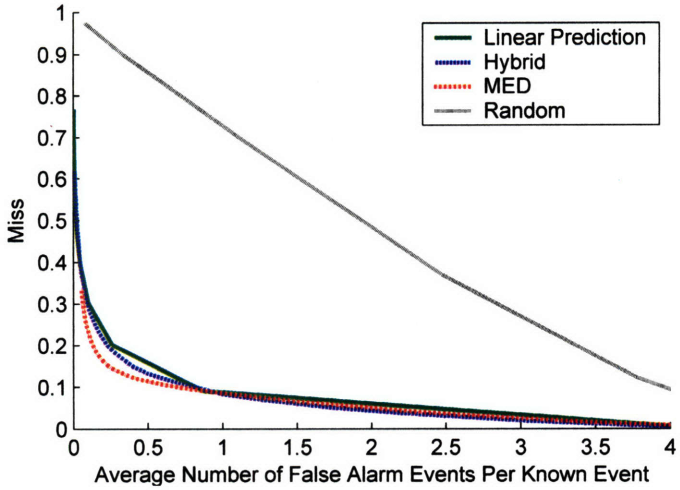


Figure 6-18. Detection-error tradeoff curve for three pulse-finding methods compared to a differentiated-EGG reference signal. Performance improves as the curves approach the bottom-left corner. This is for all males with a dead time of 1 ms.

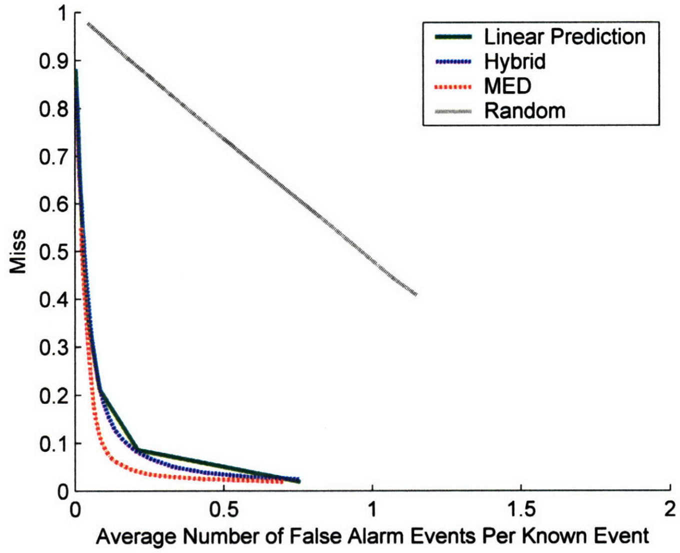


Figure 6-19. Detection-error tradeoff curve for three pulse-finding methods compared to a differentiated-EGG reference signal. Performance improves as the curves approach the bottom-left corner. This is for all females with a dead time of 2.5 ms.

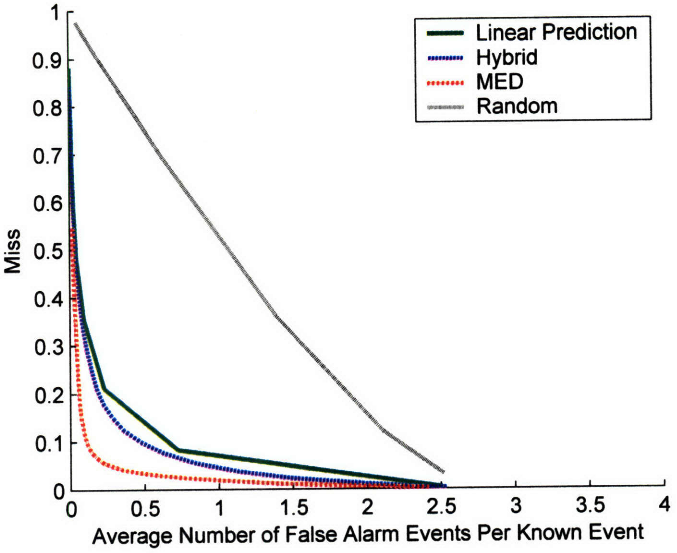


Figure 6-20. Detection-error tradeoff curve for three pulse-finding methods compared to a differentiated-EGG reference signal. Performance improves as the curves approach the bottom-left corner. This is for all females with a dead time of 1 ms.

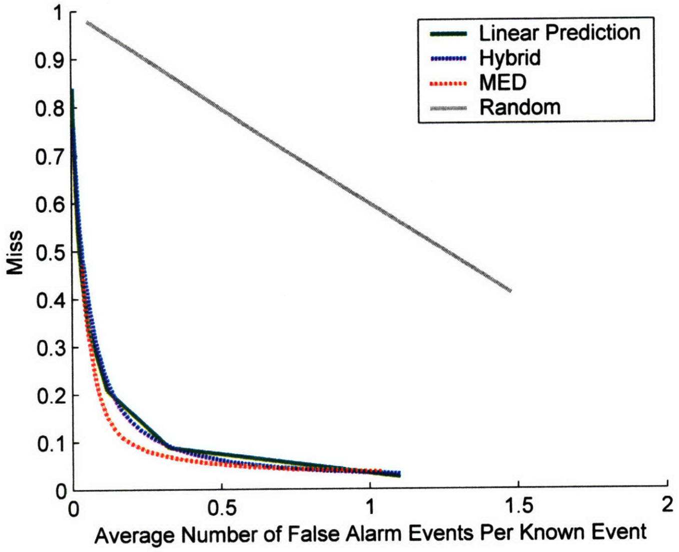


Figure 6-21. Detection-error tradeoff curve for three pulse-finding methods compared to a differentiated-EGG reference signal. Performance improves as the curves approach the bottom-left corner. This is for combined males and females with a dead time of 2.5 ms.

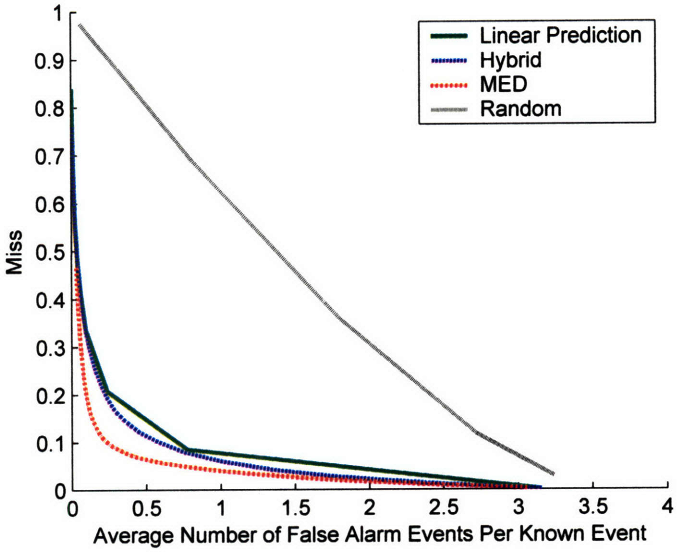


Figure 6-22. Detection-error tradeoff curve for three pulse-finding methods compared to a differentiated-EGG reference signal. Methods do better as the curves approach the bottom-left corner. This is for combined males and females with a dead time of 1 ms.

For the each of the DET curves generated from natural speech, we can see that there are a significant number of extracted events classified as false alarms as the miss rate decreases. For example for males in Figure 6-17 and Figure 6-18, there is about one false-alarm detected for every known event at a miss rate of 10 percent. Figure 6-23 demonstrates that, in many cases, such false alarms actually correspond to additional glottal events not related to closure. The figure shows a section of speech taken from a male speaker of the vowel /i/ from "oily." Here we show the EGG derivative, where positive peaks are thought to correspond to glottal-closure events and negative peaks to glottal-opening events. Our ground truth corresponds only to the positive peaks since these are more reliably extracted from the EGG signal, but we therefore neglect all of the opening events. Using a more accurate ground truth signal to evaluate our event-extraction techniques is a subject for future research.

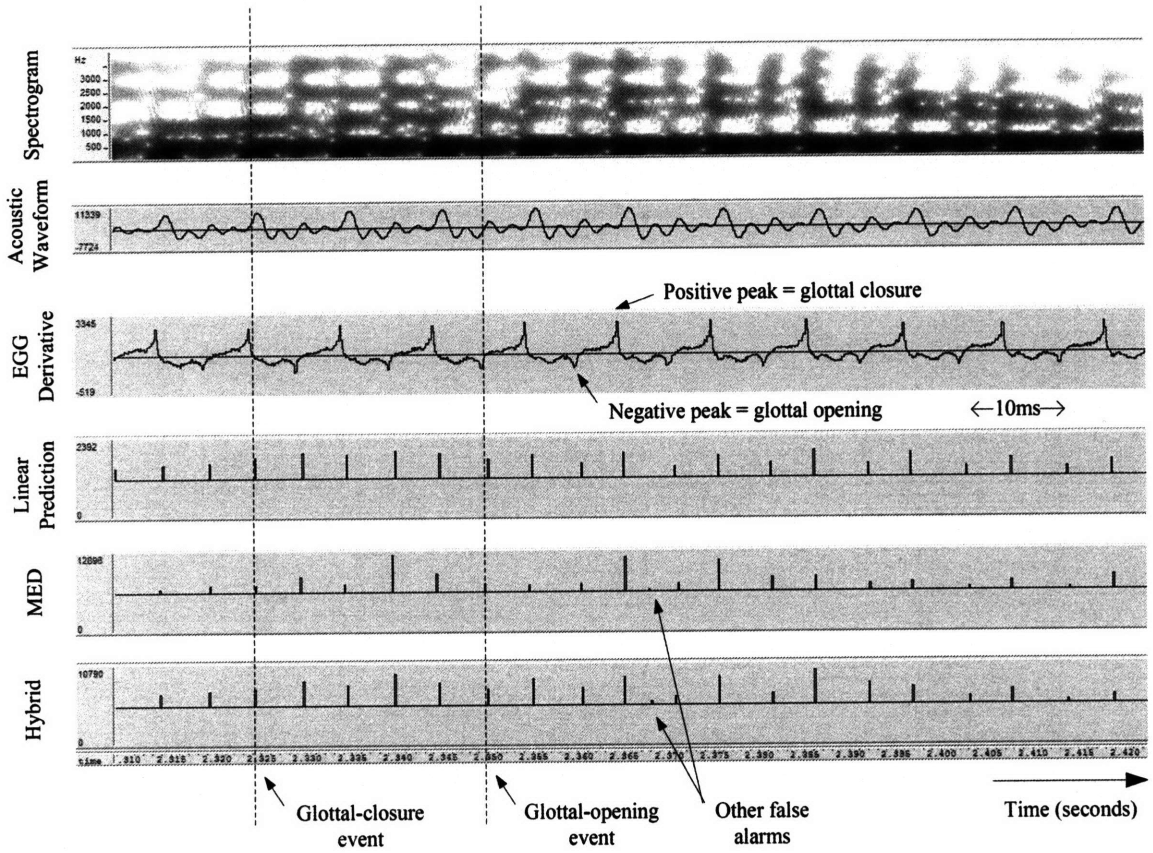


Figure 6-23. Illustration that many false alarms in the DET curves are due to excitation at glottal opening. This example was taken from the vowel /i/ for a male speaker processed with a dead time of 2.5 ms. The spectrogram was created using a 4-ms Hamming window.

6.7 Conclusions

In this chapter, we have shown the result of applying linear prediction, MED, and a hybrid method combining linear prediction and MED to the problem of decomposing nonmodal phonation into a pulse-like residual and a realistic composite impulse response. In particular, we have shown evidence that both MED and a hybrid method can be used to obtain impulse-like residuals and accurate composite impulse responses. Although MED on average produces the best impulse-response fit for synthetic data, we have shown evidence that it is not as robust as the hybrid method. In particular, it occasionally produces composite impulse responses that "ring" excessively.

We have also evaluated the performance of our glottal-event extraction algorithms on both natural and synthetic speech. In this investigation we found that performance on synthetic speech had almost no errors for MED and the hybrid method with and without a high-pass radiation characteristic. Linear prediction was shown to be sensitive to the maximum-phase component, exhibiting much improved performance when the radiation characteristic was applied. Event-extraction performance on natural data using EGG as ground truth was different than with the synthetic speech cases. In particular, many more false alarms were obtained for a given miss rate. We have provided evidence that in many cases, however, such false alarms correspond to real speech events such as glottal openings.

Chapter 7

Speaker Dependence of Nonmodal Glottal Events

In previous chapters, we have discussed the spectral and temporal theory associated with nonmodal glottal events as well as how to extract series of these events from real speech. This chapter addresses the hypothesis that patterns of these glottal events are speaker dependent. We are motivated by investigations in the literature showing a connection between nonmodal phonation and linguistic information, speaker identity, dialect, and the health of the vocal folds.

The results of the experiments reported in this chapter support the assertion that the glottal-event amplitude- and timing-pattern features developed in this thesis provide speaker information on single-session data and consistently across multi-session data. We also conduct experiments that indicate our features provide information complementary to what is carried by conventional mel-cepstral features.

The chapter is organized as follows. In Section 7.1, we describe the experimental methods including the datasets, used in our automatic speaker recognition experiments. Section 7.2 gives previous work. Section 7.3 then presents a series of experiments that illustrate that our event-pattern features carry speaker-dependent information, enough such that a speaker's utterance can be correctly verified over 80 percent of the time using the TIMIT corpus. Section 7.4 compares the speaker-recognition performance of our time-domain features to mel-cepstral features derived from the spectrum of the glottal-event impulse train as presented in Chapter 3. This section reinforces that the time-domain features are able to capture information in the impulse train that is lost in the mel-cepstral representation.

Sections 7.5 and 7.6 address two possibly confounding factors: the leakage of formant information into the glottal events and the dependency of the features on the recording session. In Section 7.5, through a series of experiments using different vowel sounds for training and testing, we show that conventional mel-cepstral features have lower performance than in the matched train/test condition, while the performance of the event-pattern features stays about the same. In Section 7.6, we compare experiments where speakers are trained and tested on the same session with experiments where speakers are trained and tested on different sessions. We find that for MED- and the hybrid-based features, performance remains about the same, and that linear prediction-based features have moderate degradation between sessions compared to within sessions.

Finally, in Section 7.7, we end with a discussion interpreting the speaker-recognition results. To support our analysis, we use visualizations of the feature space as well as our separation method from Chapter 5 for several TIMIT speakers. We present intuition into how the feature spaces for different speakers differ, supporting the promising speaker-recognition results reported earlier in the chapter.

7.1 Automatic Speaker-Recognition Methods

We begin by describing the datasets, features, and pattern recognition components used in this chapter.

7.1.1 Datasets

Two different datasets were used to conduct the experiments in this chapter. The first is the TIMIT database, a database of clean speech with each talker recorded over a single session. This dataset contains 630 speakers, with 438 males and 192 females, subdivided into eight dialect-region groups. Each speaker read ten sentences, each approximately 3 seconds long [1]. Two sentences (designated "SA") are shared by all speakers, 450 sentences (designated "SX") are each shared by seven speakers, and 1890 sentences (designated "SI") are each unique to a certain speaker. Each TIMIT speaker reads two SA sentences, five SX sentences, and three SI sentences. TIMIT data includes hand-corrected labels at the word and phone levels.

The second database used is the Boston University Radio News corpus [52]. While TIMIT data has the advantage of being clean, coming from many speakers, and of being reliably transcribed, it contains only ten sentences for each speaker, all recorded during only one session. In evaluating the speaker-dependency of our features, we would also like to determine the effect of multiple recording sessions. In order to address this issue, we use a database of broadcast radio news stories, recorded over multiple sessions, made over a period of months with "probably not multiple [recordings] per speaker in a day" [51].

This database contains stories from four male and three female speakers. A subset of the data for each speaker has been automatically transcribed at the word and phone levels. All data was read over the air in the speaker's "radio voice," and the broadcast was recorded to tape at the studio. Data regarding the exact recording conditions are not known. In our experiments, we used data from six out of the seven speakers. Talker F3A had only one file per separate news story and was excluded to avoid additional processing that would need to be done to the speech files. Table 7-1 describes the composition of the parts of the BU radio corpus used in this thesis.

Table 7-1. Table describing the composition of the parts of the BU radio corpus used in this thesis

Speaker	F1A	F2B	M1B	M2B	M3B	M4B
Minutes	66	56	55	70	49	142
Stories	43	32	33	33	21	60
Files	276	124	157	213	126	235

Both TIMIT and the BU data were preprocessed by downsampling to 8-kHz and applying a normalization stage based on the energy of the speech regions. The function *activlev* in the Voicebox MATLAB toolbox [11] was used to conduct this normalization using an algorithm based on the ITU-T P.56 standard.

7.1.2 Feature Extraction

Time-Domain Glottal-Event Features

We conducted experiments with the time-domain glottal-event features described in Chapter 5, using the linear prediction, MED, and hybrid deconvolution methods discussed in Chapter 6 to generate event times and amplitudes. In the current chapter, linear prediction and the hybrid method were each implemented using an order-15 inverse filter, 20-ms Hamming window, and a 10-ms frame rate. The MED approach used the same window length and frame rate as the other two methods, but was different in that it was order-25 rather than order-15. This order is larger than the one used for MED in Chapter 6, but was found not to drastically affect behavior. The difference came about because we initially used order-25 MED for the speaker-recognition components of the thesis, but wanted orders to match for direct comparison with other inverse-filtering methods in Chapter 6. Each of these deconvolution techniques was followed by a peak-picking stage, as described in Chapter 6, which extracted discrete impulses from the deconvolved signals using *threshold* and *dead-time* parameters. The parameter values used for each experiment in this chapter vary and are their values are described for each particular configuration.

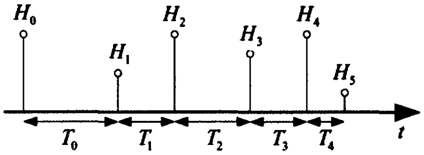


Figure 7-1. Schematic of the heights and times used to compute amplitude- and timing-ratio features.

Recall that the features derived from event sequences, detailed in Chapter 5, are based on ratios between the timings and heights of neighboring glottal events. If Figure 7-1 is the set of events being analyzed, each with the indicated height, H_m and timing, T_m, then the resulting feature vector, F_1, is found as the base-10 logarithm of the ratios between these values:

$$F_1 = \left[\log_{10}(H_1/H_0), \log_{10}(H_2/H_0), \log_{10}(H_3/H_0), \log_{10}(T_1/T_0), \log_{10}(T_2/T_0), \log_{10}(T_3/T_0)\right].$$

This particular example shows a feature vector containing three amplitude and three timing features, for a total length of six. In this chapter, we will refer often to the exact feature vector used with the phrases "number of amplitude features" and "number of timing features." The number of amplitude or timing features refers to how many consecutive amplitude or timing ratios are used. For example the phrase "one amplitude feature and one timing feature" refers to the feature vector:

$$F_1 = \left[\log_{10}(H_1/H_0), \log_{10}(T_1/T_0)\right],$$

and the phrase "two amplitude features and four timing features" refers to the feature vector:

$$F_1 = \left[\log_{10}(H_1/H_0), \log_{10}(H_2/H_0), \log_{10}(T_1/T_0), \log_{10}(T_2/T_0), \log_{10}(T_3/T_0), \log_{10}(T_4/T_0)\right].$$

Mel-Cepstral Features

In addition to the time-domain pulse-pattern features, we additionally evaluated conventional mel-cepstral features performed on the acoustic speech signal and also on the impulse trains resulting from our glottal-event finding methods. For speaker-recognition applications, mel-cepstral features are generally considered state-of-the-art [56]. These features are computed as the inverse discrete cosine transform (DCT) of the log energies from an auditory-like filter bank [56] as follows:

$$C_{mel}[n,m] = \frac{1}{R}\sum_{l=0}^{R-1} \log\{E_{mel}(n,l)\}\cos(\frac{2\pi}{R}lm), \tag{7.1}$$

where the variable m refers to cepstral coefficients [49] calculated at each sample n. $E_{mel}(n,l)$ denotes the energy calculated using a weighted function, $W_l[k]$, of the squared DFT magnitude bins, $X[k]^2$, where l refers to the lth weighting function:

$$E_{mel}(n,l) = \sum_{k=1}^{129}\left(X[k]^2 W_l[k]\right), \tag{7.2}$$

and where a 256-point DFT is invoked. The set of weighting functions used mimics the effect of an auditory-like filter bank using mel scaling which yields linear filter spacing up to 1 kHz and logarithmic spacing above 1 kHz. The shape of each weighting function is triangular with widths based on human-derived critical bands. Davis and Mermelstein [13], who first introduced the method, used a total of 20 bands covering the range from 0 to 4600 Hz. As shown in Figure 7-2, we use in this thesis a set of 20 triangular weighting functions, based on a subset of the 24 filters [56], to weight the squared values of the first 129 bins of a 256-point DFT. As shown, each energy-weighting function has been normalized to sum to one. The set of mel-cepstral features used in this chapter are therefore found as follows, observing that the zeroth cepstral coefficient is dropped to minimize the effect of any DC component in the signal:

$$C_{mel}[n,m] = \frac{1}{20}\sum_{l=0}^{19} \log\{\sum_{k=1}^{129}\left(X[k]^2 W_l[k]\right)\}\cos(\frac{2\pi}{R}lm) \quad \text{for } 1 \le m \le 19. \tag{7.3}$$

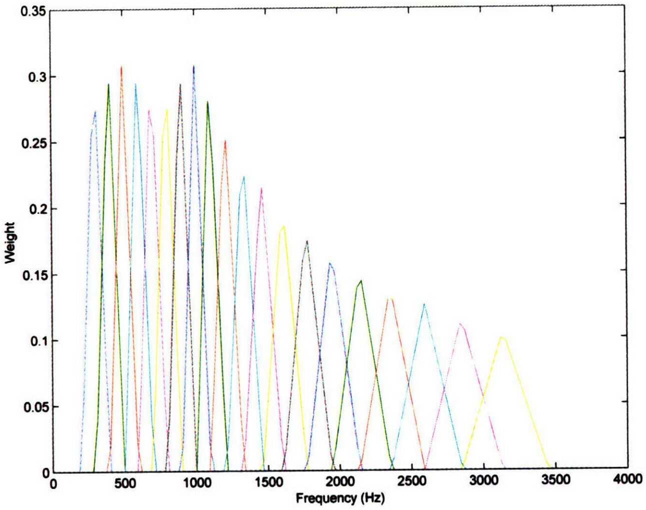


Figure 7-2. Triangular functions used to weight the squared bins of the 256-point DFT magnitude spectrum. Each weighting function has been scaled to sum to a value of one. Nonuniform peaks for the first eight bins, which are uniformly spaced, result from the triangles being undersampled by the DFT bins.

Quatieri [56] notes that "although [mel-cepstral] features are likely to contain some source information, e.g., the spectral tilt of the STFT magnitude influenced by the glottal flow derivative, we have assumed that the primary contribution to the Mel-cepstrum is from the vocal tract system function." This result is thought to be due to factors including the inability of the analysis filters to resolve individual harmonics and the handling of spectral phase information. Pitch information, however, may remain for high-pitched speakers [56]. We shall support this idea that the mel-cepstrum coefficients are primarily sensitive to the vocal-tract system function in later sections.

7.1.3 Speech-Segment Extraction

Both the TIMIT and BU Radionews databases are transcribed at the word and phone levels. As will be discussed for each experimental section, we used these transcriptions in order to isolate segments, namely certain vowels, which were then used in each experiment.

We have written custom software to isolate desired speech segments for each of the three glottal-event-based features as well as for mel-cepstral features. State-of-the-art mel-cepstral systems generally use all speech frames, but we extract vowel speech segments for more direct comparison with the pulse features. For mel-cepstrum performed directly on the speech, the task is relatively easy; we simply keep any frame that partially contains one of the marked phones. This step is performed once the mel-cepstral features have been created. For the pulse features, we must *first* isolate the desired regions of speech before the pulse extraction step. Each of the regions is then independently processed to extract impulses and a series of event-pattern features. These independently-extracted features are then combined in a final step.

7.1.4 Classifier

Each of our speaker-recognition experiments uses a classifier based on Gaussian mixture models (GMMs). This technique currently form the core of state-of-the-art methods of pattern classification for speaker recognition applications [56]. With this method, distributions of features are described as sums of multidimensional Gaussian distributions, allowing arbitrary histograms to be fit [61].

Two different experimental paradigms, *speaker verification* and *speaker identification*, were used. Speaker verification is the process by which we determine whether a presented speaker is or isn't a certain target speaker. The system used as the back end for our experiments is an adapted GMM-based classifier using a universal background model (UBM) [61]. Each GMM uses 512 mixtures. For this process, the TIMIT data is divided into background, training, and testing sets. None of the same utterances are used between the groups of data. For those speakers involved in the training and test stages, eight utterances (the SI and SX sentences) were used for training and two (the SA sentences) were used for testing.

Figure 7-3 depicts the process used to implement the GMM classifier for speaker verification with a UBM. Speakers A and B represent two potential users of the system who have submitted example sentences of their speech. As shown, features are extracted from these speech utterances and are used to create histograms. For simplicity, only one-dimensional features are shown; in general, features can have as many dimensions as desired. These histograms of speech features are then modeled with sums of Gaussians as shown. In addition, a "background" model is created by performing the same process with samples of speech collected from many other speakers, representing the "universe" of speech. As detailed in [61], each of the target-speaker models is adapted from the background model.

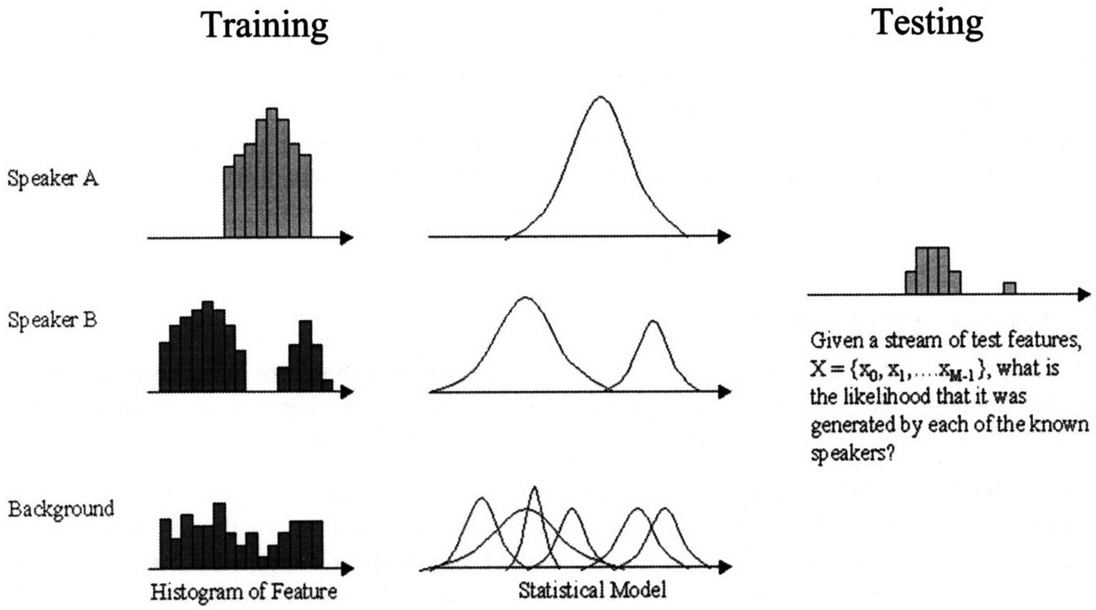

Training
Testing
Speaker A
Speaker B
Given a stream of test features, X = {x₀, x₁, ...x_{M-1}}, what is the likelihood that it was generated by each of the known speakers?
Background
Histogram of Feature
Statistical Model

Figure 7-3. Schematic depiction of the process of modeling a distribution of (here one-dimensional) features with a Gaussian mixture model for the purpose of speaker verification. Cases "A" and "B" represent two speakers who have trained the system to recognize their voices; the "background" case represents all other possible speakers.

Once these models are trained, the situation arises where an unidentified user approaches the system and proposes that he or she is speaker A. The system collects a relatively small number of features, $X=\{x_0,x_1,\ldots,x_{M-1}\}$ and then decides whether or not the speaker comes from the claimed distribution, a hypothesis denoted λ_C. Statistically, the decision consists of determining whether the log-likelihood ratio is above a predetermined threshold [56]:

$$\frac{P(\text{features from claimed speaker})}{P(\text{features from background})} = \frac{P(\lambda_C \mid X)}{P(\lambda_{\bar{C}} \mid X)} = \frac{p(X \mid \lambda_C)P(\lambda_C)/P(X)}{p(X \mid \lambda_{\bar{C}})P(\lambda_{\bar{C}})/P(X)}$$

$$\text{log} - \text{likelihood} - \text{ratio} = \Lambda(X) = \log[p(X \mid \lambda_C)] - \log[p(X \mid \lambda_{\bar{C}})] \tag{7.4}$$

$$\begin{array}{ll} \Lambda(X) \geq \text{threshold}, & \text{accept} \\ \Lambda(X) < \text{threshold}, & \text{reject} \end{array}.$$

In contrast to speaker verification, speaker identification entails selecting which *one* speaker out of a set of known speakers produced a given utterance [60]. As with speaker verification, this process entails training a set of target models that represent each known speaker. Unlike in speaker verification, we do not train a background model, instead relying on a comparison of a set of scores rating the likelihood that a given utterance belongs to each known speaker. The talker with the highest score for each presented utterance is identified as its speaker:

$$\begin{aligned} &\arg\max_{\lambda_C}\left[P(\text{observed features from speaker } \lambda_C)\right] \\ &= \arg\max_{\lambda_C}\left[P(\lambda_C \mid X)\right] = \arg\max_{\lambda_C}\left[p(X \mid \lambda_C)P(\lambda_C)/P(X)\right] = \arg\max_{\lambda_C}\left[p(X \mid \lambda_C)\right] \end{aligned} \tag{7.5}$$

$$\text{speaker score} = \log[p(X \mid \lambda_C)]$$

where λ_C denotes each hypothesized speaker. Here we have assumed that the apriori probabilities of each speaker, $P(\lambda_C)$, are equal.

7.2 Previous Work in Automatic Speaker Recognition using Source Mechanisms

There have been several examples in the literature prior to our experiments of using features that represent the speech source to perform automatic speaker recognition. Plumpe, Quatieri, and Reynolds [54] applied a parametric model of the glottal cycle to the problem. These authors found 69.1% correct identification for males and 73.6% for females in a TIMIT SID experiment using source features. They also obtained 41.1% male and 51.8% female using the mel-cepstrum of their modeled glottal flow-derivative waveform and 95.1% male and 95.5% female using the estimated glottal-flow derivative itself. These experiments used the mel-cepstrum for analysis of their source features, which is different from our approach. Additionally, Quatieri, Jankowski, and Reynolds [57] investigated the use of "energy onset times" for this purpose. The set of features, representing the periods of primary and secondary source pulses, boosted SID results on NTIMIT by 15%, from 55% to 70% over a sine-wave based pitch estimate. Neither of these approaches specifically addressed the general problem of irregular glottal events.

The linear-prediction residual has also been used in automatic speaker recognition by several authors [55, 71]. Prasanna et al report using 5-ms blocks of data from the order-8 LP residual as inputs to a neural network classifier. Only "high voiced" regions are analyzed. They report a performance of 23.8% equal-error rate (EER) for the NIST 2002 primary evaluation condition using their source features only. The equal-error rate (EER) is a performance metric describing the percentage of misses when it equals the percentage of false-alarms. Additionally, they found that performance improved from 8.6% EER to 7.8% EER when combining a Gaussian-mixture-model using linear-prediction cepstral coefficients to represent the vocal-tract contribution with their source features. The experiments use the entire residual, which is different than our technique.

7.3 Speaker Dependency of Glottal-Event Features

The first set of experiments that we report describe a set of speaker-verification experiments operating on the TIMIT corpus. As previously mentioned, TIMIT has the advantage of including many different speakers and being carefully labeled at the phonetic level. Using this database we will show that the glottal-event pattern features that we extract from natural speech are able to represent speaker-dependent properties.

As described in previous chapters, there are many different parameters that can be used in extracting our glottal-event features. Aside from the three different techniques, linear prediction, MED, and the hybrid method, we must also consider the *number of timing and amplitude features* as well as the *threshold* and *dead time* used for event extraction. In addition to supporting the basic result that the glottal-event features are speaker dependent, we will summarize several of the trends that depend on these parameters. A more complete set of the raw data is available in Appendix C for interested readers.

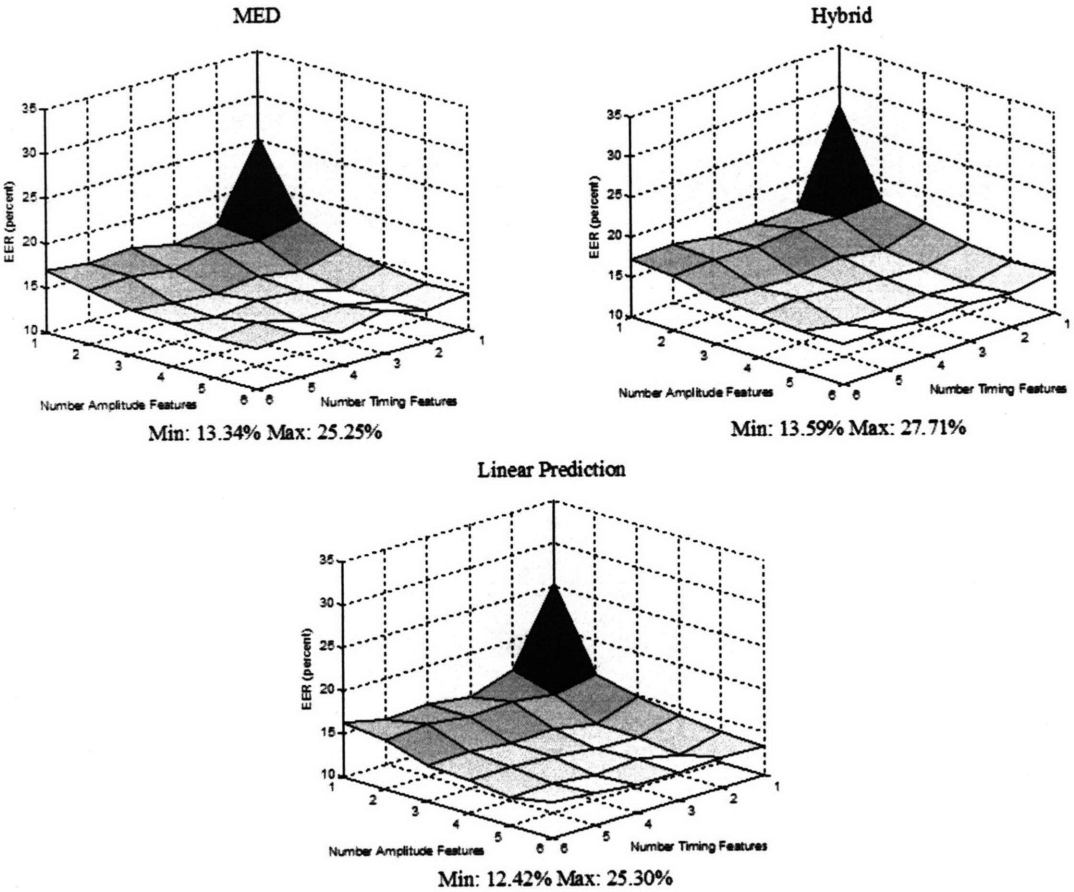
7.3.1 Experimental Setup

To configure the TIMIT speaker-verification experiment, the following subdivisions of the speech files were used. First, all 680 sentences in the "test set" division of TIMIT were used to train a background model. All sentence types (SI, SX, and SA) for both males and females were used for this background training. Next target models were built using eight SA, SI, and SX sentences of 326 male speakers and 136 female speakers from the "train set" division, which were different speakers than used in the background model. Finally, test utterances were created using the two remaining SX sentences of each speaker as used in the target-training procedure.

In order to concentrate on speech produced with the phonation source, only the vowel regions of the speech were used in this experiment. Log-likelihood scores were generated for each test utterance against all of the target models and used to create a detection-error tradeoff (DET) curve which describes the tradeoff between misses and false alarms. The equal-error rate (EER) at which these two values are equal were extracted from this curve and used to describe the performance. Both male and female data were used to create the background model, but experiments using male and female target models and test data were conducted and reported separately. As we will see, male and female results exhibited different properties, justifying this choice.

7.3.2 Results

In the first set of results, we found that our event-pattern features carry speaker-dependent information. In these experiments we fixed the dead time at 2.5 ms and the threshold at zero in order to make a straightforward comparison. For each of the three feature sets, we show the results for males and females and sweep across different numbers of amplitude and timing features. Figure 7-4 shows the results for male speakers and Figure 7-5 shows the results for female speakers. The panels are arranged such that MED is in the top left, hybrid method is in the top right, and linear prediction is at the bottom.


MED

Min: 13.34% Max: 25.25%

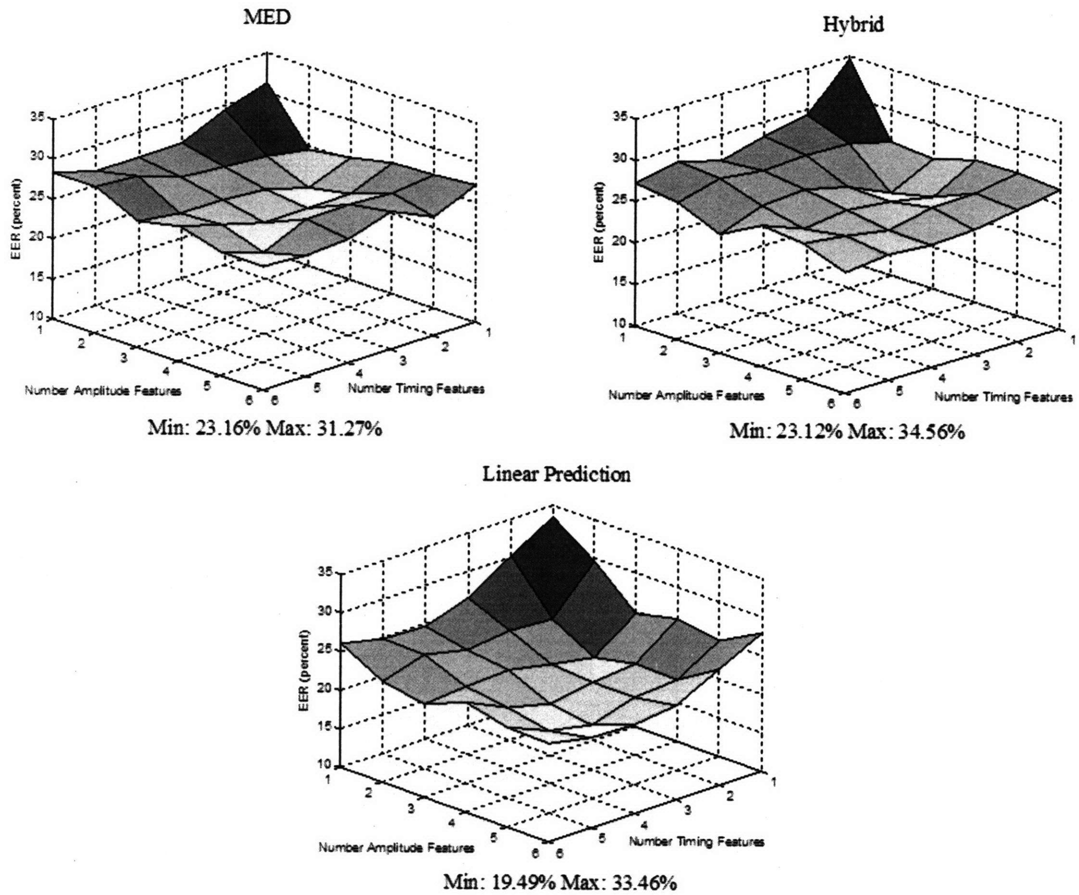
Hybrid

Min: 13.59% Max: 27.71%

Linear Prediction

Min: 12.42% Max: 25.30%

Figure 7-4. Equal-error rates for speaker verification for male TIMIT speakers using MED, hybrid method, and linear prediction.



MED Min: 23.16% Max: 31.27%

Hybrid Min: 23.12% Max: 34.56%

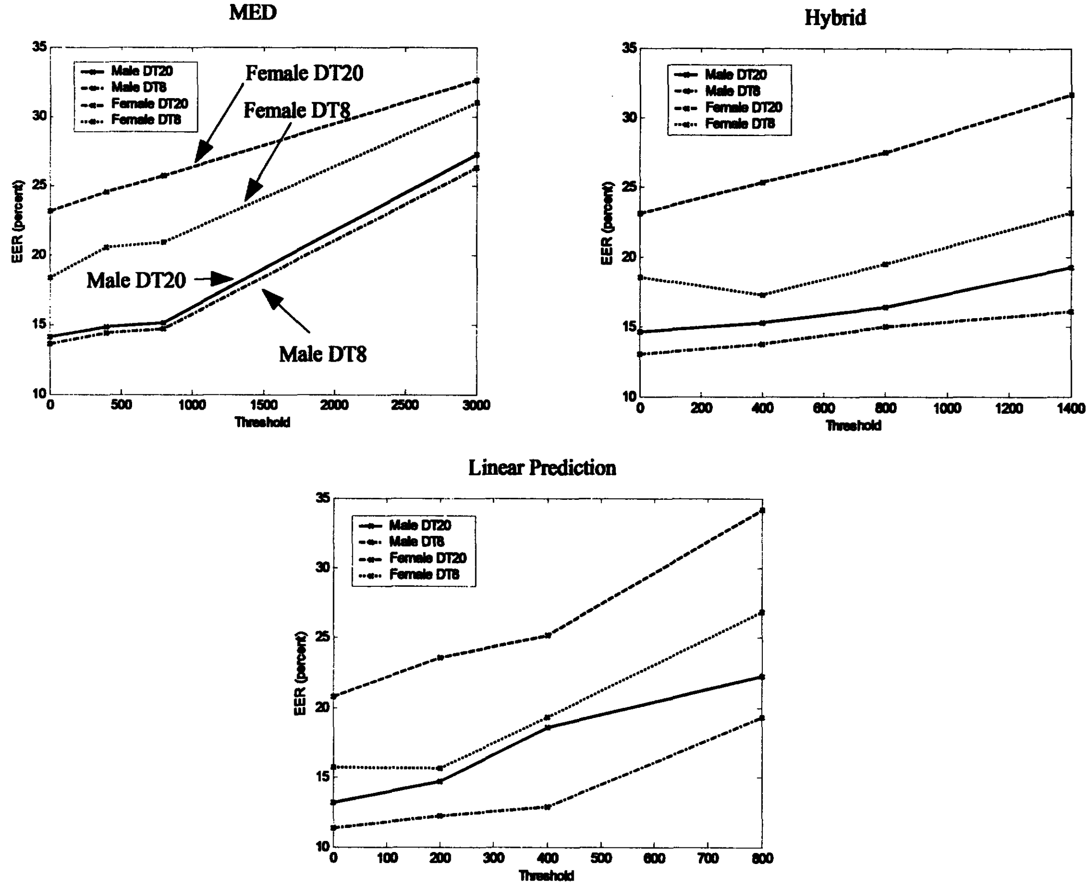
Linear Prediction Min: 19.49% Max: 33.46%

Figure 7-5. Equal-error rates for speaker verification for female TIMIT speakers using MED, hybrid method, and linear prediction.

For comparison, we also performed automatic speaker recognition using mel-cepstral features on the speech itself, using all of the vowel segments. No delta features, cepstral mean subtraction, or RASTA were used. This configuration was chosen to allow the most straightforward comparison with our pulse features, since these methods do not have a direct analog in the event-feature domain. Exploratory experiments also suggested that additional processing actually made mel-cepstrum perform worse in our TIMIT experiment. With 512 mixtures, we obtained 4.02% EER for males and 6.99% EER for females. For a baseline experiment using conventional mel-cepstral features including all speech regions, rather than just the vowels, the results were 2.25% EER for males and 4.01% EER for females.

From the plots in Figure 7-4 and Figure 7-5, we can see a general trend that in each case adding more features decreases the EER. The females show a small counterexample to this trend with increasing EER at high numbers of timing or amplitude features. Additionally, males overall perform better than females, a trend that we will discuss further, through an investigation of the role of reducing the dead time, in the next section. From this data, a good point to investigate is two timing features and three amplitude features. This set of features is extracted from a total of four neighboring glottal events. Although this is not always the best-performing configuration, it is representative of the best EER across a wide range of experimental conditions (see Appendix C for data). Using this configuration and others similar to it will allow us to compare the effect of parameters in a straight-forward manner.

We have shown that the glottal-event-pattern features carry speaker-dependent information. We now investigate how the performance varies with changes to the dead time and threshold parameters. For simplicity, we will focus on the two-timing-ratio, three-amplitude-ratio cases only. Figure 7-6 presents plots for MED (top left), hybrid method (top right), and linear prediction (bottom) depicting the change in EER as a function of the threshold at two dead times, 20 samples (2.5 ms) and 8 samples (1 ms).



MED

Hybrid

Linear Prediction

Figure 7-6. Dependence of the equal-error rate on threshold for male and female TIMIT speakers at two different dead times. (top left) MED, (top right) Hybrid, (bottom) Linear prediction.

From these plots, we observe several trends. The most prominent characteristics of the data are that decreasing dead time and decreasing threshold both lead to improved performance. We see that decreased dead time generally yields the largest performance increase for the female datasets and the smallest for the males. One exception to this result is for the male linear-prediction case, where decreasing the dead time increases performance notably in both the male and female cases. For all three methods, especially linear prediction and the hybrid method, the improvement to female performance with reduced dead time reduces the gap observed between the male and female cases in Figure 7-4 and Figure 7-5.

Although we explored the effects of pulse-extraction parameter settings to some extent in Chapter 6, it is not clear what information is being added at the lower thresholds and dead times that allows for improved performance. One reason that females may be more responsive to lower dead time than males is that females tend to have glottal events that are closer to one another, due to higher pitch. Lowering the dead time allows these events to be resolved. It is curious, however, that males continue to do well at decreased dead time and threshold, though, as we observed many apparent insertion errors with these parameters in Chapter 6. The speaker-recognition experiments suggest that these "errors" actually carry information about the speakers.

7.4 Mel-Cepstral Features for Impulse Patterns

For comparison with the individual experiments described above, we also performed automatic speaker recognition using mel-cepstral features on the impulse trains alone. In Chapter 3, we demonstrated that nonmodality influences the spectrum, showing much sensitivity of the magnitude spectrum to variations in the times and amplitudes of impulses. Because the mel-cepstrum is based on the magnitude spectrum, it is reasonable that it also may carry information about glottal-event sequences used by different speakers. The purpose of this section is to gauge whether using glottal-event spectral information carried by the mel-cepstrum is beneficial, allowing us to resolve subtle differences between speakers, or harmful, acting as noise to make recognition more difficult. We accomplish this by comparing the performance of mel-cepstral features with our time-domain features.

7.4.1 Experimental Setup

In the set of experiments reported in this section, both the mel-cepstral and time-domain pulse-pattern features operate *directly* on the impulse train output by the event-extraction stage, not on the acoustic signal. Unlike the experiments in Section 7.3, each of the time-domain features used three amplitude and three timing features. As described previously, no delta features, cepstral mean subtraction, or RASTA were used with the mel-cepstrum.

The files used to create the background model, train the target models, and act as test utterances were the same as described for the speaker-verification experiments in Section 7.3. As with the previous setup, only the vowel regions of the speech were used.

7.4.2 Results

Figure 7-7 presents the dependence of the equal-error rate on threshold for the time-domain-based features compared with mel-cepstral features. In this figure, the left column shows the results for male speakers and the right column shows the results for female speakers. Each row is for linear prediction, MED, and the hybrid method respectively.

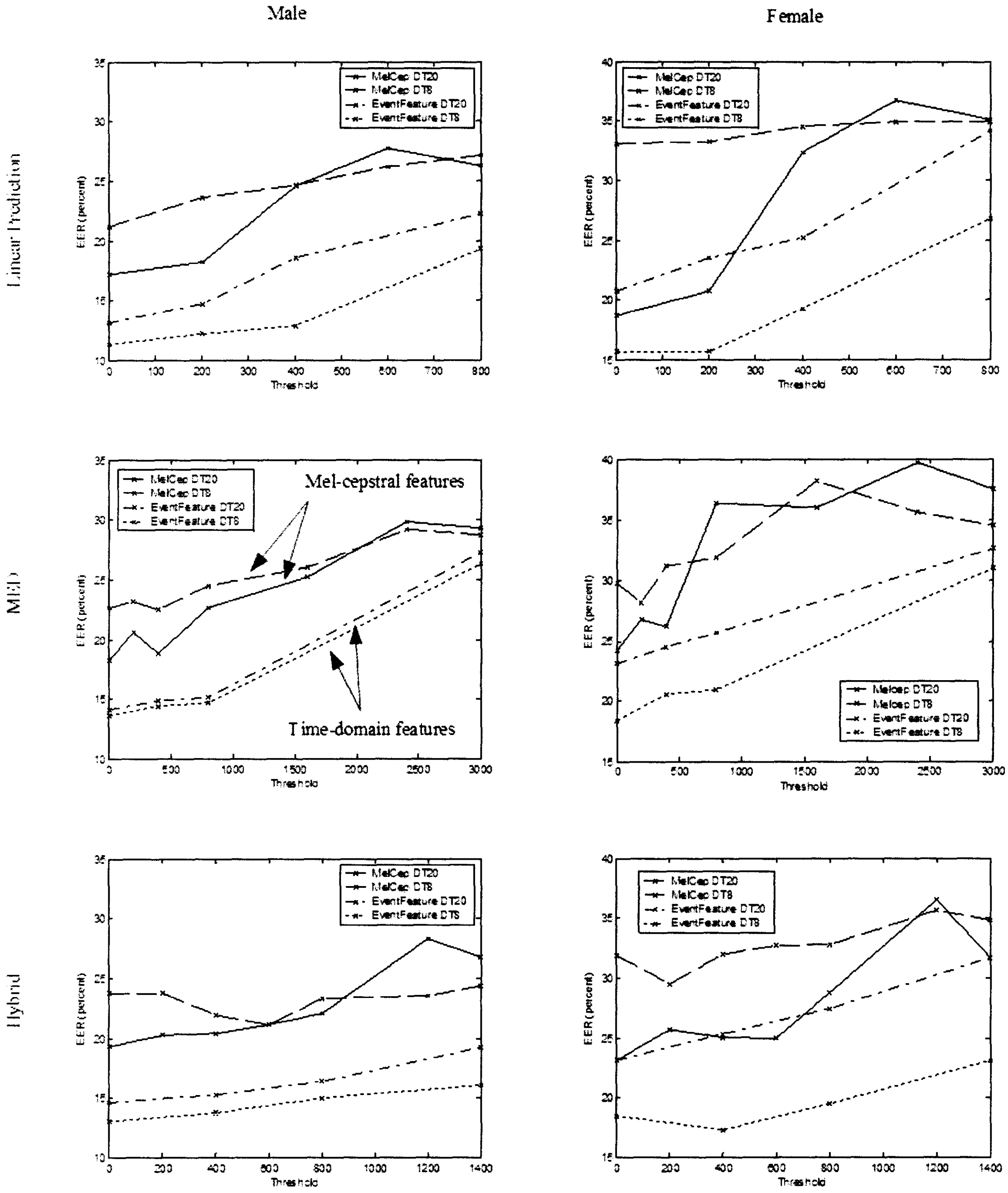


Figure 7-7. Comparison of the performance of time-domain glottal-event features versus mel-cepstral features. Two dead times are used, 2.5 ms (20 samples) and 1 ms (8 samples). For each combination of threshold and dead time, the time-domain features outperform the mel-cepstral features.

The results illustrate that the time-domain glottal-event feature representation of the extracted impulse trains outperform the mel-cepstral representation at each dead time and threshold setting. This supports the assertion that our relative pulse pattern features are able to represent more speaker-dependent information than the spectrum of the impulses. We can interpret these findings as evidence that the sensitivity of the spectrum to nonmodality is detrimental to speaker-recognition performance.

7.5 Mismatched Phone Experiments

One complicating factor in the experiments described above is that they may not entirely represent source phenomena; that is, the formants may leak through the inverse filter. In order to address this factor, we conducted an experiment where one set of vowels was used to train each speaker model while another was used to test. The logic was that if our features truly represent source information and that information is not dependent on the vowel being produced, then a vowel mismatch between training and testing would have a large effect on the mel-cepstral results and a much smaller effect on the pulse-features. In this situation, we hypothesized that the pulse features would outperform the mel-cepstral features alone.

7.5.1 Experimental Setup

In order to test this hypothesis, we used vowel sections of the TIMIT database, splitting these sections into two groups roughly corresponding to open (low) and closed (high) vowels[3]:

Vowel set used for training (open vowels): [ɛ], [æ], [ʌ], [ɔː], [ɑ], [ə]

Vowel set used for *matched*-vowel testing (open vowels): [ɛ], [æ], [ʌ], [ɔː], [ɑ], [ə]

Vowel set used for *mismatched*-vowel testing (closed vowels): [iː], [ʊ], [uː], [ɪ], [oʊ]

This strategy allows us to train and test on a large number of vowels from across the vowel space without training and testing on the same phones. Some source information may depend on the vocal-tract configuration, so this strategy allowed us to cover a wide range of vocal-tract configurations without training and testing on the same vowel.

To configure the TIMIT speaker-verification experiment, the same target training, and testing divisions of the TIMIT dataset were used as described in the Section 7.3.1. First, all 680 sentences in the "test set" division of TIMIT were used to train a background model. All sentence types (SI, SX, and SA) for both males and females were used for this background training. Only the open vowel-sections of the speech were used to construct this background model. Next target models were built using 8 SA, SI, and SX sentences of 326 male speakers and 136 female speakers from the "train set" division, which were different speakers than used in the background model. As with the background model, only the open-vowel sections of the speech were used to train these models.

Finally, test utterances were created using the two remaining sentences of each speaker used in the target-training procedure. In the matched-vowel experiments, only the open-vowel segments of the speech were used, while only the closed-vowel segments were used for the mismatched-vowels experiments. It is important to note that the only difference between the two kinds of experiments was the test sets—exactly the same background and target models were used.

7.5.2 Results

The results of this experiment, shown in Figure 7-8 and Figure 7-9, demonstrate different behaviors of the mel-cepstral features applied to the acoustic signal compared with the time-domain glottal-event features. Three amplitude and two timing features were used for linear prediction and MED, both with a threshold of 0 and dead time of 2.5 ms. As described

[3] The specific TIMIT symbols used were [eh], [ae], [ah], [ao], [aa], [ax], and [ax-h] for the open vowels and [iy], [uh], [uw], [ux], [ih], [ix], and [ow] for the closed vowels.

previously, no delta features, cepstral mean subtraction, or RASTA were used with the mel-cepstrum.

The results show that, while each of the time-domain features (middle bars in each set) perform comparably between the matched and mismatched cases, performance with the mel-cepstral features (left bars in each set) degrades significantly. We also have shown the effect of an equal linear fusion of the log-likelihood scores from the mel-cepstral and time-domain glottal event features. This fusion significantly improves the performance over mel-cepstral features alone in the mismatched cases for males and the matched cases for females.

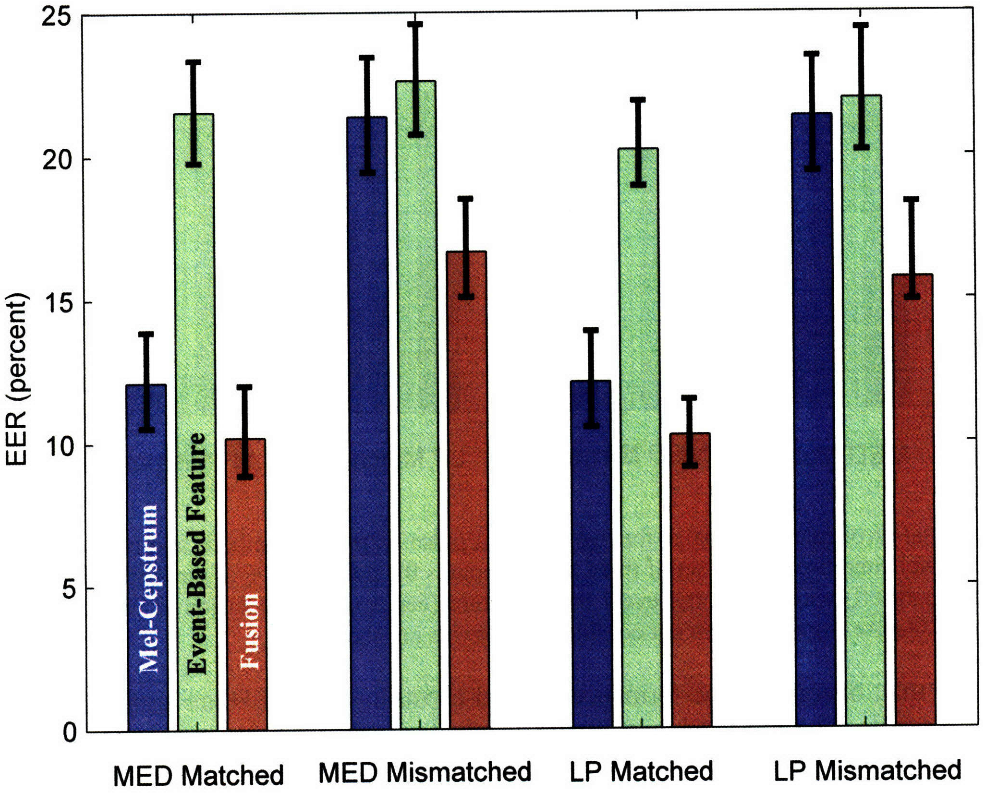


Figure 7-8. Equal-error rate for males for two different pulse-extraction methods and both matched and mismatched vowel conditions. Each set of three bars compares the performance of mel-cepstrum applied to the acoustic signal (left) with the glottal-event based feature (center), and the fusion of mel-cepstrum and the event-based feature (right). 95-percent confidence intervals are shown.

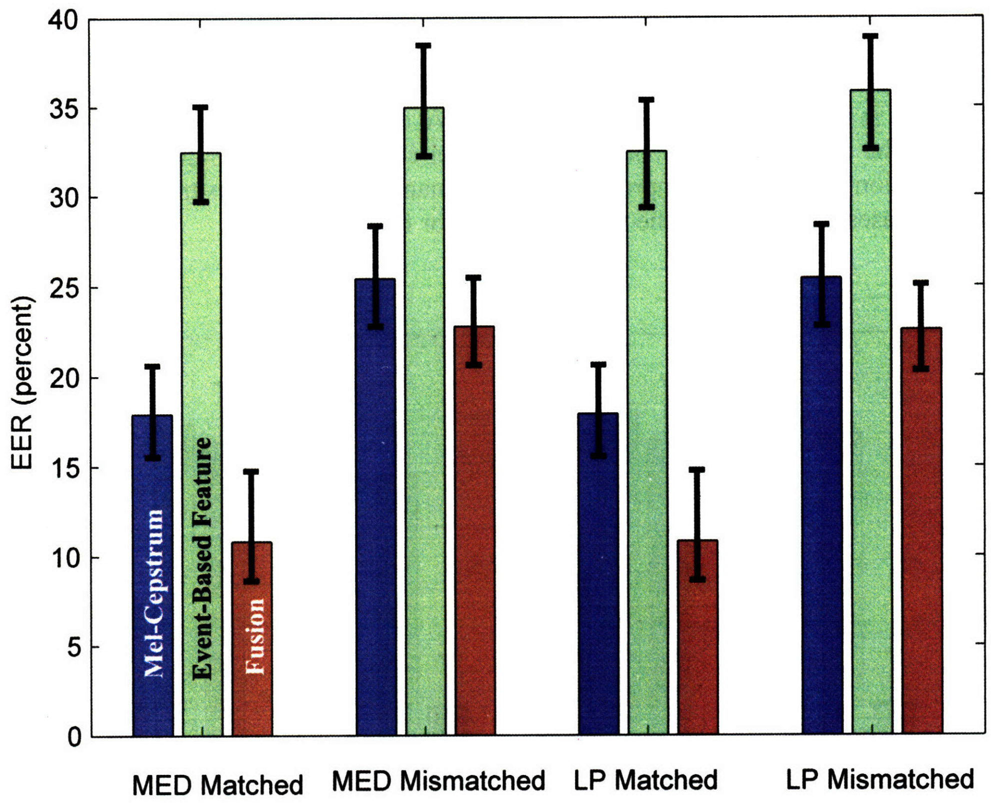


Figure 7-9. Equal-error rate for females for two different pulse-extraction methods and both matched and mismatched vowel conditions. Each set of three bars compares the performance of mel-cepstrum applied to the acoustic signal(left) with the glottal-event based feature (center), and the fusion of mel-cepstrum and the event-based feature (right). 95-percent confidence intervals are shown.

We can see that, overall, the performance for mel-cepstrum and MED are significantly worse than when the whole dataset is used, even in the matched case. This observation is not unexpected as reducing the number of vowels eliminates large amounts of information about the vowel space for each speaker and also decreases the total amount of speech available for training background and target models. For the mismatched case, the decrease in performance is *expected* by design since the target models are built using a different vowel space than contained in the test set. Although they contain different vowels, the formant spaces between the mismatched training and testing sets are not independent. The two different sets, open and closed vowels, may share similar values for the second formant and above in a given talker. This is one possible explanation for the significantly-better-than-chance performance of the mel-cepstral features in the mismatched-vowel configuration.

7.6 Multiple-Session Experiments

We have thus far presented evidence that patterns of the timings and amplitudes of glottal events convey information about the identity of a speaker. The experiments using TIMIT, however, can only address this question for within a single session. *Between* sessions, it is less clear that a speaker retains properties of the glottal source that are captured in patterns of glottal events. Speakers, for example, may have different levels of hydration, different levels of fatigue, and varying levels of overall health between sessions. Emotional factors such as psychological stress, mood, and state of mind may also vary between sessions and affect speech, a fact that is exploited in the field of automatic affect recognition [18]. In the following experiment, we conduct a set of intersession closed-set speaker-identification experiments that support the claim that our features represent speaker-dependent source information both within and between sessions.

7.6.1 Experimental Setup

We set up the multiple-session experiments as closed-set speaker-identification tasks using the BU radio database described in Section 7.1.1. Two different overall configurations were used in these experiments as detailed in Figure 7-10 and Figure 7-11. The number of session and number of minutes used from each speaker as well as the target-training and test-set divisions are detailed in these figures. As can be observed, the strategy of the first configuration was to use a single training model with two different test sets, either matched or mismatched to the pooled session used in the training model. In contrast, the second configuration used a single test set with either a matched or mismatched set of pooled sessions used to train the target model. As shown, files used in the test sets were always different than those used to train the target models.

The logic behind using two different configurations is that they complement one another. As we are comparing the performance of two different experiments, one with matched sessions and one with mismatched sessions, we would like to minimize the sensitivity to the particular sets of data used. The configuration with the single target model assumes that relationships observed between the performance of a speaker-identification experiment is more sensitive to the particular set of data used to *train* the target models. In contrast, the configuration with the single test set but different target models assumes that the results of an experiment are most sensitive to the particular set of sessions used in the *test* set. Performing tests with both configurations allows us to observe the results under both hypotheses.

The closed-set speaker-identification experiments were run by modeling each of the six target speakers using a mixture of 512 Gaussians. In the testing phase, each test utterance was scored against each of the target speakers and the model yielding the largest score was deemed to be the talker who produced the utterance. The *closed-set classification rate* was obtained for each speaker by finding the percentage of cases for which an utterance produced from that talker was correctly identified by the system. The rate reported for each experiment in the next section is the *average* of the speakers' closed-set classification rates.

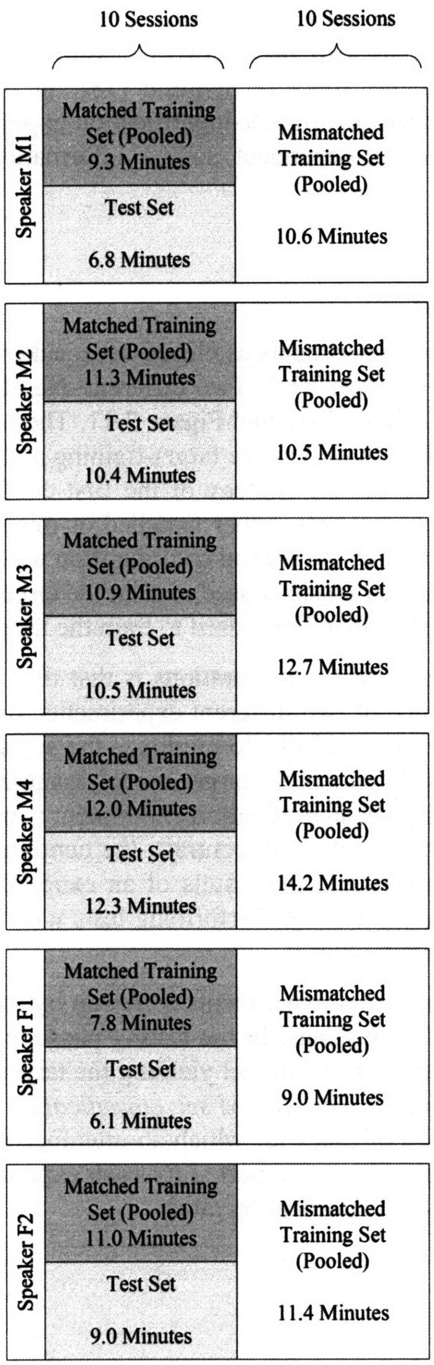

10 Sessions
10 Sessions
Speaker M1
Matched Training Set (Pooled) 9.3 Minutes
Test Set
6.8 Minutes
Mismatched Training Set (Pooled)
10.6 Minutes
Speaker M2
Matched Training Set (Pooled) 11.3 Minutes
Test Set
10.4 Minutes
Mismatched Training Set (Pooled)
10.5 Minutes
Speaker M3
Matched Training Set (Pooled) 10.9 Minutes
Test Set
10.5 Minutes
Mismatched Training Set (Pooled)
12.7 Minutes
Speaker M4
Matched Training Set (Pooled) 12.0 Minutes
Test Set
12.3 Minutes
Mismatched Training Set (Pooled)
14.2 Minutes
Speaker F1
Matched Training Set (Pooled) 7.8 Minutes
Test Set
6.1 Minutes
Mismatched Training Set (Pooled)
9.0 Minutes
Speaker F2
Matched Training Set (Pooled) 11.0 Minutes
Test Set
9.0 Minutes
Mismatched Training Set (Pooled)
11.4 Minutes

Figure 7-10. Configuration of the multiple-session closed-set speaker-identification experiment with a single test set. Each target model is trained using a pooled set of ten sessions which uses different files than the test set.

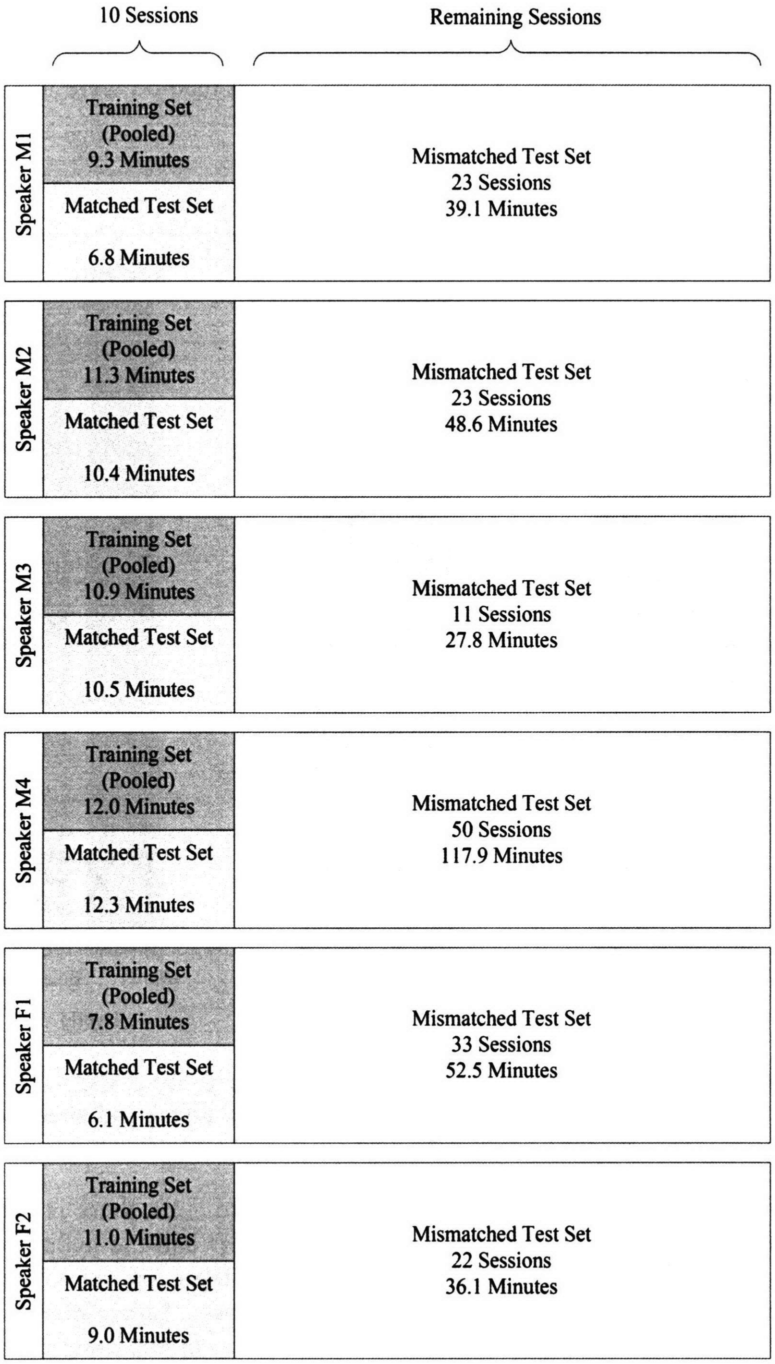

10 Sessions
Remaining Sessions
Speaker M1
Training Set (Pooled) 9.3 Minutes
Matched Test Set 6.8 Minutes
Mismatched Test Set 23 Sessions 39.1 Minutes
Speaker M2
Training Set (Pooled) 11.3 Minutes
Matched Test Set 10.4 Minutes
Mismatched Test Set 23 Sessions 48.6 Minutes
Speaker M3
Training Set (Pooled) 10.9 Minutes
Matched Test Set 10.5 Minutes
Mismatched Test Set 11 Sessions 27.8 Minutes
Speaker M4
Training Set (Pooled) 12.0 Minutes
Matched Test Set 12.3 Minutes
Mismatched Test Set 50 Sessions 117.9 Minutes
Speaker F1
Training Set (Pooled) 7.8 Minutes
Matched Test Set 6.1 Minutes
Mismatched Test Set 33 Sessions 52.5 Minutes
Speaker F2
Training Set (Pooled) 11.0 Minutes
Matched Test Set 9.0 Minutes
Mismatched Test Set 22 Sessions 36.1 Minutes

Figure 7-11. Configuration of the multiple-session closed-set speaker-identification experiment with a single target-training set. Each target model is trained using a pooled set of ten sessions which uses different files than either of the test sets. Observe that in this configuration, each speaker has a different amount of test data.

7.6.2 Results

Figure 7-12 shows the performance of the closed-set speaker-identification experiments using linear prediction and MED glottal-event features for the fixed target training set configuration. The interested reader may refer to Appendix C for the raw data. The results using linear-prediction features show differences between the within-session and between-session experiments, dropping from an average of 83 percent correctly identified to 66 percent correctly identified. In comparison, the results for MED and the hybrid method were much more uniform between sessions, both close to 90 percent.

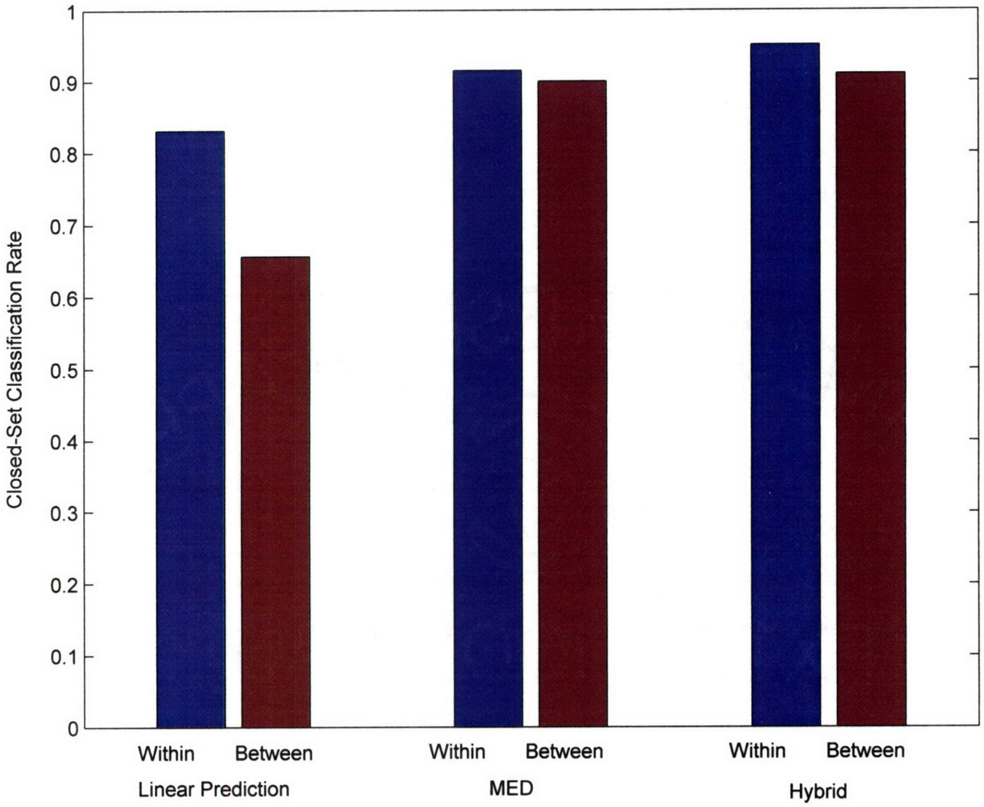


Figure 7-12. Comparison of within- and between- session experiments using MED and linear-prediction-based features. For this experiment, the training set was fixed. Three timing and three amplitude features, order 512, threshold 200, and dead time 2.5 ms were used.

For the configuration using a fixed test set, shown in Figure 7-13, there is a greater decrease between the within- and between-session experiments for MED and the hybrid method than for the fixed target-model configuration, dropping an average of 10 percent for MED and 12 percent for the hybrid method between the matched and mismatched cases. Linear prediction, in contrast, drops only 4 percent between the matched and mismatched cases in this experiment.

The minor 1- to 12-percent performance decrease that we have observed in all but the fixed-target-model linear prediction results supports the assertion that we are finding properties of the speaker source that are persistent in a given speaker across multiple sessions, and that are not particularly sensitive to the state of the speaker at the time of the recording. For a feature based on the speech source, which is known to vary due to many factors, at least a small drop in recognition performance is logical. Indeed, such sensitivity to these factors may be desirable in certain fields such as the analysis of fatigue and clinical voice disorders.

The finding that the linear-prediction based method breaks down between sessions in one of the experimental configurations could be due to multiple factors. In Chapter 6, we found differences between linear prediction, MED, and the hybrid method in terms of their abilities to represent a set of ground-truth pulses, the accuracy of their estimated composite impulse responses, and the impulsiveness of the residuals they produce. It is likely that these differences lead to sensitivity to the recording session. Both this break down of linear prediction and the overall differences between the fixed-target-model and fixed-test-set experiments require future work, likely using a multiple-session experiment with a greater number of speakers.

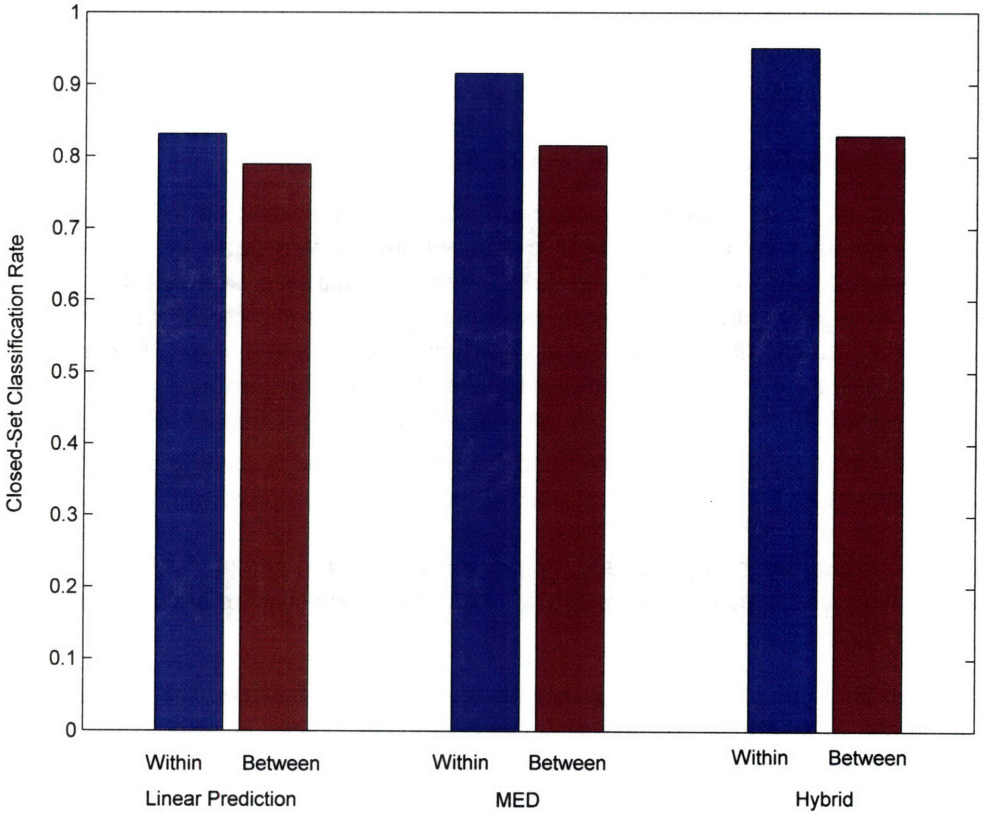


Figure 7-13. Comparison of within- and between- session experiments using MED and linear-prediction-based features. For this experiment, the test set was fixed. Three timing and three amplitude features, order 512, threshold 200, and dead time 2.5 ms were used.

7.7 Pattern Analysis

The experiments in this chapter indicate that different speakers differ in patterns of their glottal events. We have seen that information over several neighboring glottal events are important to achieving good speaker-recognition performance. For example, using a time-domain glottal-event feature vector with one amplitude and one timing feature, the performance is much lower than with three amplitude and two timing features. This observation agrees with the examples of real speech analyzed in Chapter 6, where we saw many patterns of glottal events which involved three or more impulses. We have shown in this chapter that this richness of patterns is speaker-dependent.

The question remains how exactly patterns in the events differ from one speaker to another. In this section we compare the pulse-pattern timing distributions for three different male TIMIT speakers. We will see that these speakers tend to differ in two key ways: (1) What percentage of the features represent 1-long, 2-long, and 3-long patterns and (2) where the patterns are located in feature space.

Figure 7-14, Figure 7-15, and Figure 7-16 show the event feature space for timing patterns of three male TIMIT speakers. Here, the top panels of each figure show the total feature space and a histogram of the amounts of features identified as 1-long (modal), 2-long, 3-long, and other sequences. These different kinds were separated using the algorithm described in Chapter 5. A prediction error of 0.1 was allowed between each predicted and observed next feature, with six minimum repeats required for a particular sequence of features to be identified as having a certain pattern. The rest of the panels show the feature space for features identified to belong to each type of pattern. We have concentrated on only the first two timing features in order to simplify our discussion. The actual space used throughout this chapter is richer, involving a combination of multiple timing and amplitude features, but is more difficult to visualize.

As we saw when we investigated the meaning of the impulses extracted in Chapter 6, 2-long patterns representing the sequence of glottal events within a given pulse period are common. These patterns are related to both within-cycle behaviors, primarily glottal opening and closing events, as well as 2-long patterns between neighboring glottal cycles. In each of the three examples we have shown, these phenomena contribute significantly; over 20 percent of the patterns in each speaker are 2-long. For the first speaker, they are the dominant pattern and can be seen to have a $\log_{10}(T_1/T_0)$ value mostly in the range $[-0.2 \quad 0.2]$, meaning that each period of the pattern is between a scale factor of 0.63 and 1.6 relative to the preceding pattern. It appears that there are peaks at -0.1 and 0.1, corresponding the repeating relative timing interval pattern $[1, 0.8]$. This pattern was different than those for the second and third speakers who exhibited dominant relative timing interval patterns of approximately $[1, 0.6]$ and $[1, 0.9]$. There were also differences in how much the patterns were spread, with the second speaker having elements out to about $[-0.4 \quad 0.4]$ or the pattern $[1, 0.4]$.

The percentage of 1-long patterns can also be seen to differ between the speakers. This is especially apparent between the second and third speakers. There is also evidence that the residual spaces are different, reflecting all of the phenomena that cannot be put into one of the pattern categories. In particular, for some speakers, patterns beyond 3-long may be significant. We can see this in the residual signal as clusters rather than the smoothly varying residual as for the first speaker. The third speaker has distinct clusters for example that are not represented well in the lower-order patterns. These plots indicate that speakers differ both in the overall percentage of events in each of the patterns categories as well as the specific distribution of each of those categories.

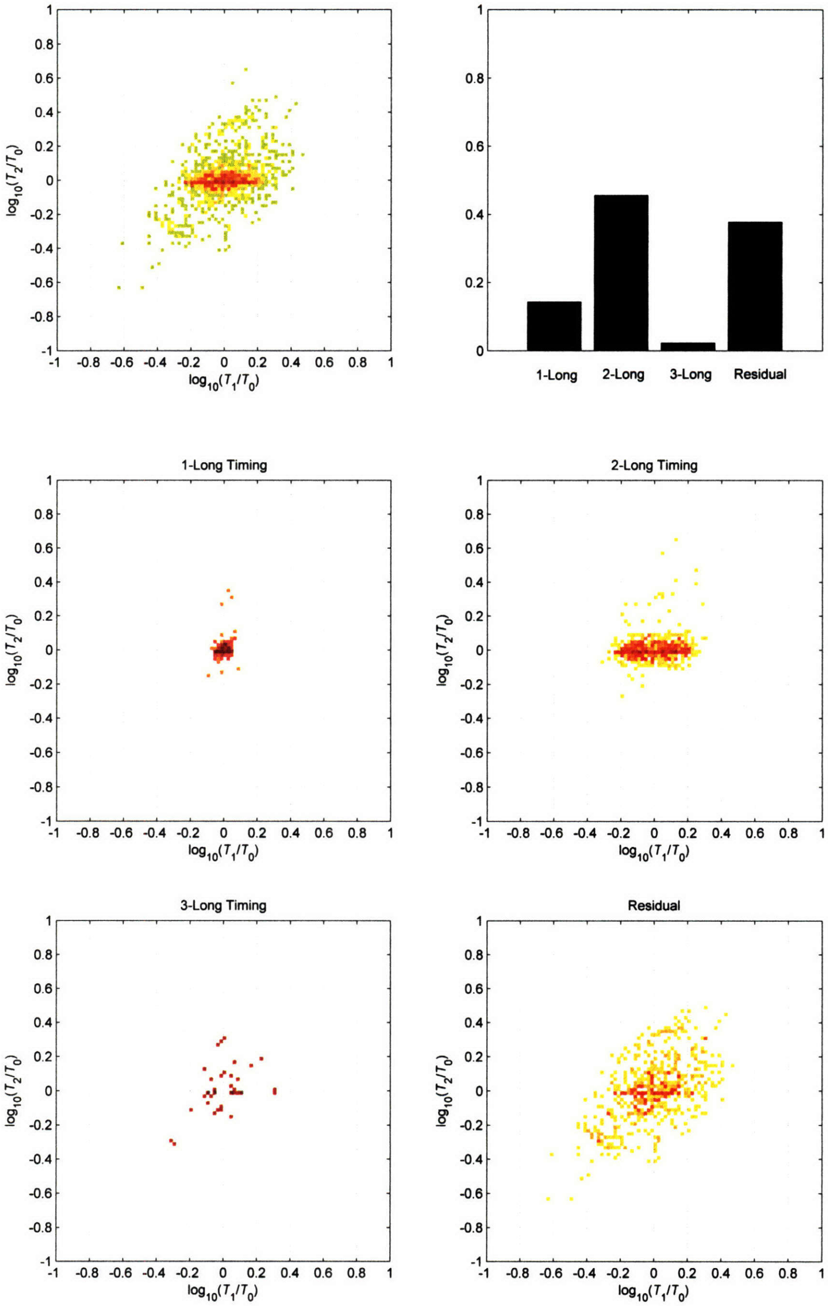


Figure 7-14. Automatic separation of the timing-feature space into features representing its modal, diplophonic, triplophonic, and irregular components for male DR1 TIMIT speaker MEDR0.

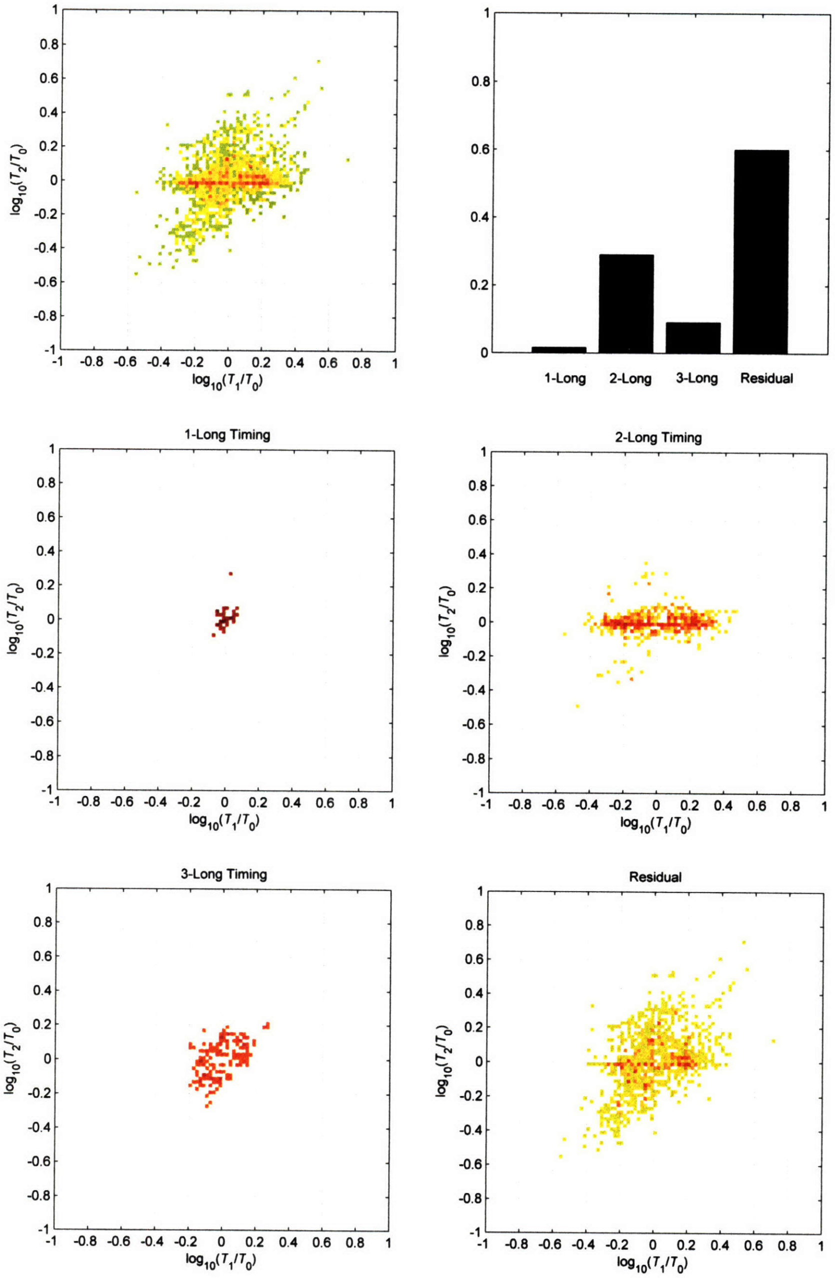
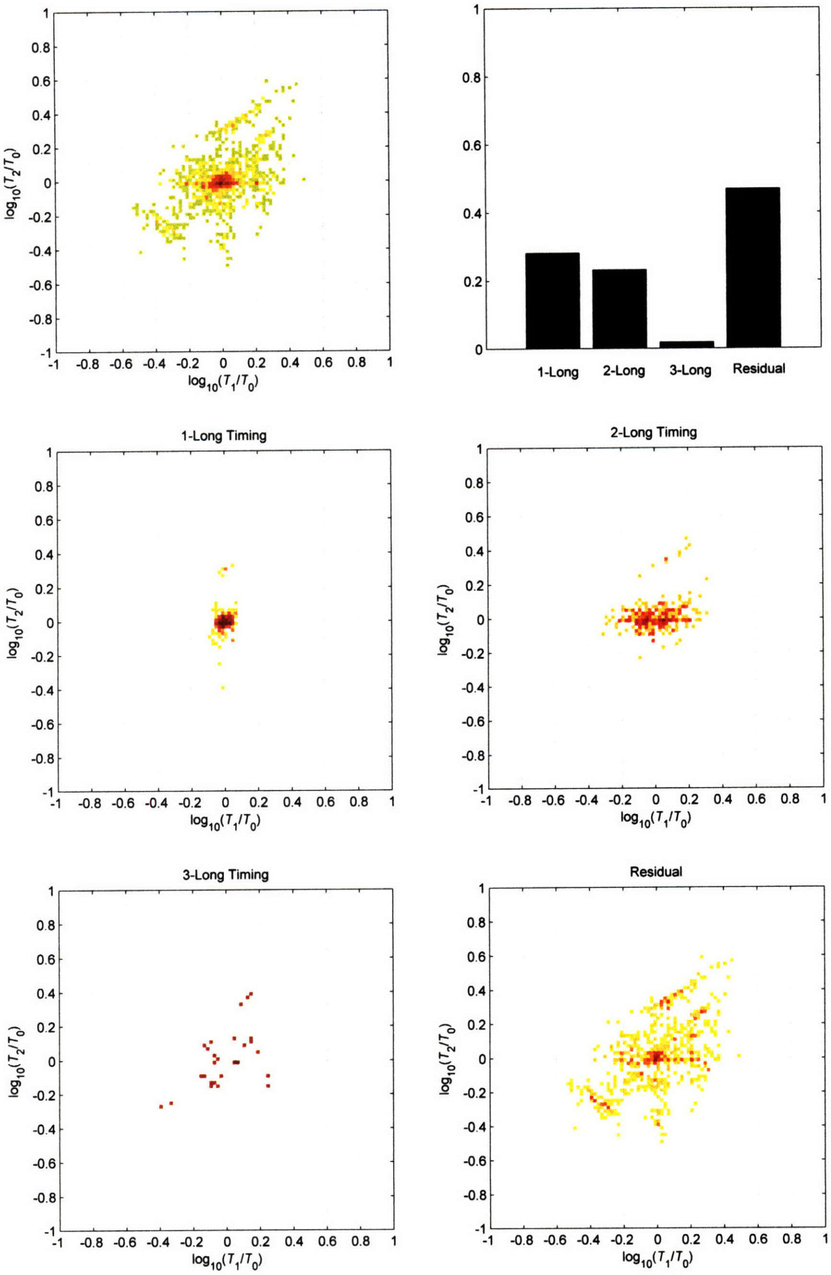


Figure 7-15. Automatic separation of the timing-feature space into features representing its modal, diplophonic, triplophonic, and irregular components for male DR1 TIMIT speaker MPGR0.



1-Long Timing

2-Long Timing

3-Long Timing

Residual

Figure 7-16. Automatic separation of the timing-feature space into features representing its modal, diplophonic, triplophonic, and irregular components for male DR1 TIMIT speaker MRWS0.

To complement the discussion of the timing space of the speakers in Figure 7-14, Figure 7-15, and Figure 7-16, we have also plotted the event feature spaces for these speakers' amplitude patterns in Figure 7-17, Figure 7-18, and Figure 7-19. The top panels of each figure show the total feature space and a histogram of the amounts of features identified as 1-long (modal), 2-long, 3-long, and other sequences. Since the amplitude features covered a wider area of the feature space than the timing features, a prediction error of 0.4 was allowed between each predicted and observed next feature, with six minimum repeats required for a particular sequence of features to be identified as having a certain pattern. The rest of the panels show the feature space for features identified to belong to each type of pattern.

As with the timing features, we can see that the talkers differ both in terms of the overall percentage of events in each of the amplitude-pattern categories as well as the specific distribution within each of those categories. In addition, we observe that the amplitude features overall represent a larger range of ratios between neighboring events. For example, the 2-long patterns centered in the vicinity of 1 and -1 represent ratios of about 10 between neighboring event amplitudes, much greater than the scale factors under 2 that we observed in the 2-long timing patterns. From these amplitude examples, we can also see that patterns in the event timings are not necessarily mirrored in the event amplitudes. In Figure 7-17 for example, we see no 1-long amplitude patterns even though there are many 1-long timing patterns for the same speaker in Figure 7-14. This complementary information between timing and amplitude features helps explain the speaker-verification results in Section 7.3, where we saw the best performance for a combination of both timing and amplitude features.

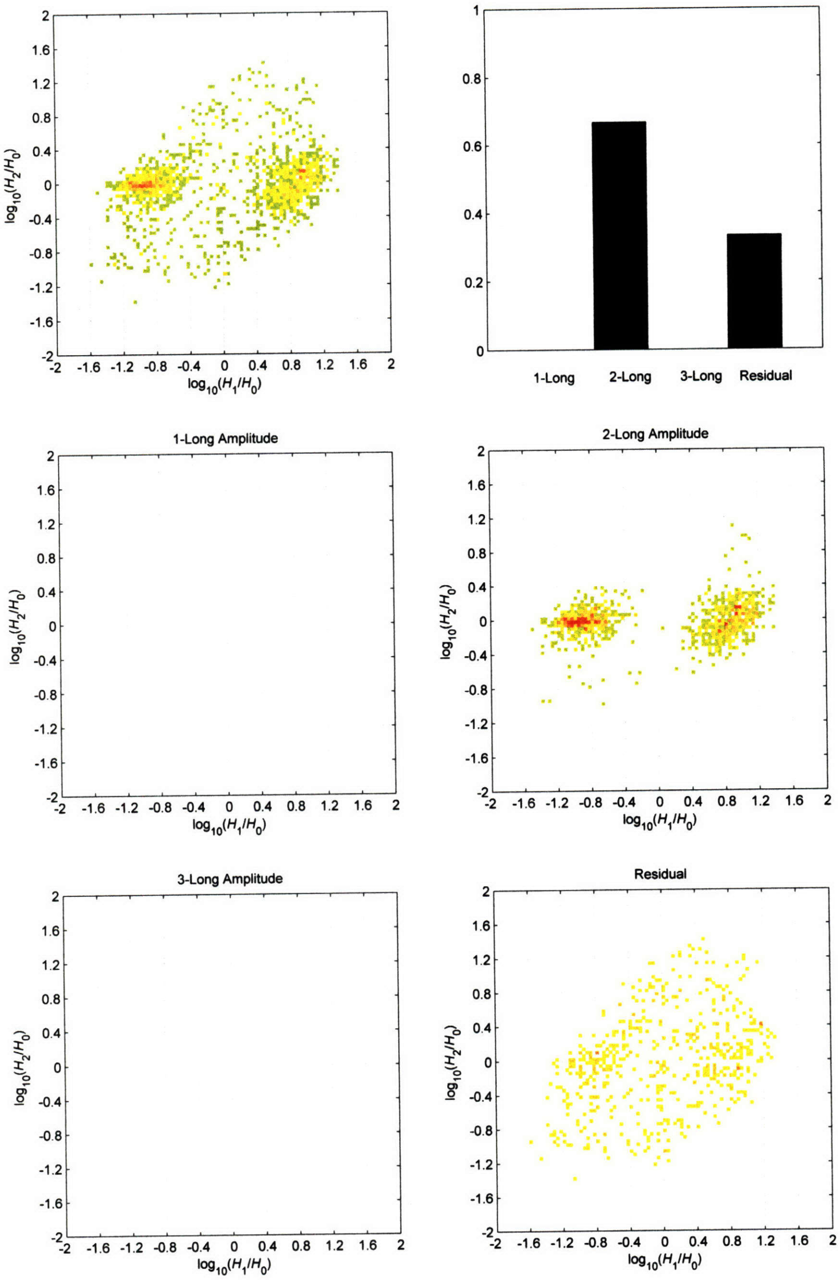


Figure 7-17. Automatic separation of the amplitude-feature space into features representing its modal, diplophonic, triplophonic, and irregular components for male DR1 TIMIT speaker MEDR0.

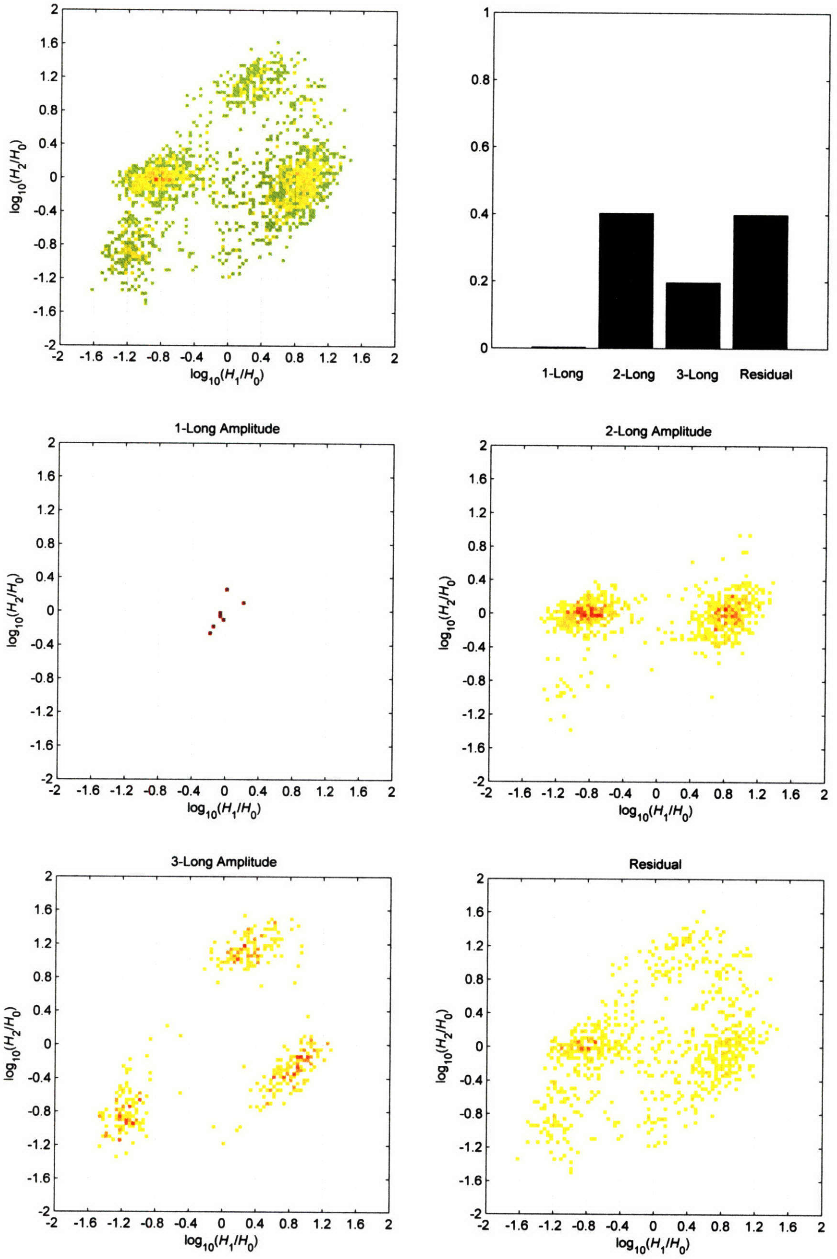
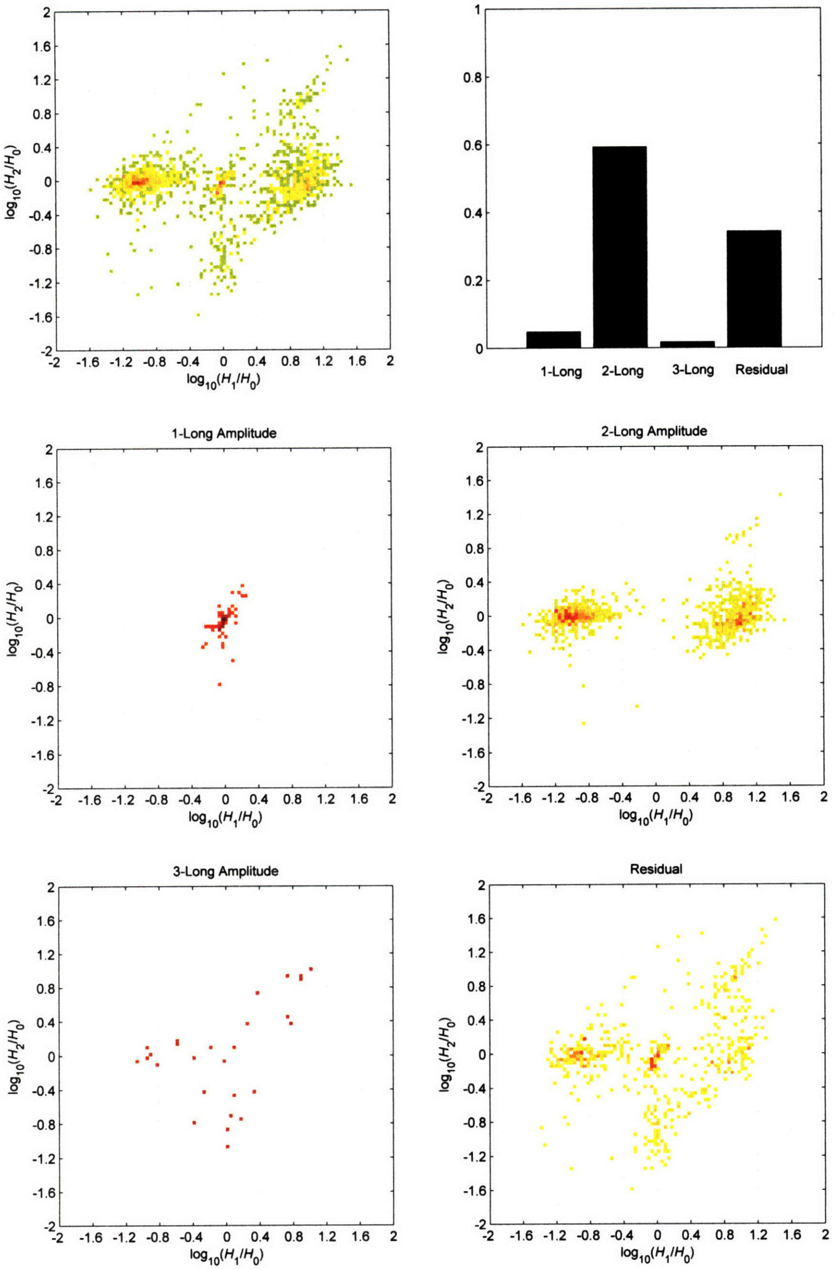


Figure 7-18. Automatic separation of the amplitude-feature space into features representing its modal, diplophonic, triplophonic, and irregular components for male DR1 TIMIT speaker MPGR0.


1-Long Amplitude

2-Long Amplitude

3-Long Amplitude

Residual

Figure 7-19. Automatic separation of the amplitude-feature space into features representing its modal, diplophonic, triplophonic, and irregular components for male DR1 TIMIT speaker MRWS0.

7.8 Conclusion

In this chapter, we have conducted experiments to gauge the usefulness of features capturing glottal events for speaker recognition applications. We have shown promising results, indicating that features derived from sequences of glottal events do carry speaker-dependent source information, and that they are able to perform close to 10 percent equal-error rate in the best cases when using the TIMIT database. We also showed how this performance varies as the number of timing and amplitude features, dead time, and threshold parameters are swept. The general trends are better performance with larger number of features, smaller dead time, and smaller threshold. We showed that the performance of the MED and hybrid-method features hold both within and between data-collection sessions. Also, the event-pattern separation method used supported the speaker-recognition experiments, showing explicit differences across speakers in their nonmodal pattern behavior.

In addition, we compared the relative pulse amplitude and timing features against the mel-cepstrum of the impulses alone, finding that the pulse-features provide an advantage over mel-cepstrum. We also presented evidence that formant information has a much greater affect on mel-cepstrum than on our glottal-event features, supporting that our features primarily represent the speech source mechanisms.

Chapter 8

Conclusions and Future Work

8.1 Summary

In this thesis, we have described the theory and practical elements of working with a representation of speech based on discrete glottal excitation events. We began in Chapter 2 by describing a representation of nonmodal speech as the convolution of a series of glottal events, each modeled as an impulse with a set of corresponding composite impulse responses. Although the underlying glottal source signal is continuous, we argued through both synthetic and natural speech examples that the filtered impulsive-excitation model was a reasonable abstraction. This was especially true away from DC, at the frequencies most relevant to exciting the vocal-tract transfer function.

The next three chapters discussed the theoretical properties of the impulse-excited representation. In Chapter 3, we addressed the effect of both deterministically- and stochastically-perturbed impulse trains on the spectrum. We found that the spectrum was sensitive to even small perturbations of impulses. For deterministic patterns, this sensitivity manifested as a changes in the spectral shaping function applied to the flat source line spectrum. For stochastic impulse perturbations, the effect was an additive high-pass noise floor and a low-pass filtering of the harmonic spectrum. These effects indicated the complexity in trying to spectrally represent nonmodality. This chapter was a prelude to the use of mel-cepstral features in the speaker-dependence experiment of Chapter 7.

In Chapter 4, we then investigated an alternative spectral representation to the short-time Fourier transform viewpoint, using harmonically-related sinusoids. The goal was to determine the meaning of pitch for nonuniform glottal events, as a possible way to represent patterns. The chapter finds that for a given set of excitation impulses, the instantaneous fundamental frequency is ambiguous, with many possible pitch tracks yielding the same observed signal. In a related way, we also found that discrete measurements of the true fundamental frequency do not represent a unique impulse train. Our findings are relevant to the thesis in that they motivated us to find a *unique* time-domain representation for a sequence of impulses. They are also relevant to speech-technology in general, and we presented one application, sinewave analysis-synthesis, as an example.

Chapter 5 described a time-domain representation of sequences of glottal events. This representation was shown to have the property that similar patterns formed clusters in the feature space. We found that different event patterns mapped to regions of the feature space based on their overall period of repetition (1-long, 2-long, etc) as well as the specific pattern embodied in this repetition. We also introduced an algorithm to identify regions of speech with different patterns by taking advantage of the order in which feature vectors occurred in time. The time-domain features introduced in this chapter were used in the automatic speaker-recognition experiments in Chapter 7 and explored in Appendix A with application to measuring voice quality.

Having presented aspects of the theoretical analysis of glottal-event sequences, in Chapter 6 we turned to the practical task of extracting nonmodal glottal events from natural speech. The extraction methods we introduced were based on three different inverse-filtering techniques, linear prediction, minimum-entropy deconvolution (MED), and a hybrid method combining the two other approaches. Evaluations of these algorithms showed that MED had advantages over the others in terms of its ability to estimate the true composite impulse response and to estimate the locations of a set of known glottal events. MED, however, was shown to occasionally yield highly inaccurate composite-impulse response estimates, which the hybrid method and linear prediction did not do. Because of these tradeoffs, we decided to experiment with all three methods in the automatic speaker-recognition experiments of Chapter 7.

Using the time-domain features from Chapter 5 and the automatic event extraction techniques developed in Chapter 6, Chapter 7 dealt with using patterns of glottal events for automatic speaker recognition. The results of our experiments indicated that glottal-event features were able to capture speaker-dependent characteristics including aspects that were complementary to conventional mel-cepstral features. This was shown to be true both with the large single-session TIMIT database and the smaller multi-session BU RadioNews database. In this chapter we also presented evidence that the spectra of glottal-event patterns were less beneficial than the temporal representation for automatic speaker recognition.

8.2 Contribution of Thesis

The general contribution of this thesis is a focus on the *nonmodal* glottal-source versus near-periodic phonation. We have investigated the effect of nonmodality from its definition, to its analysis using short-time spectra, pitch, and time-domain features, to its application to natural speech. Although aspects of nonmodality have been studied before in the signal-processing literature, this thesis is the first work to describe a set of theory and technologies specifically designed to address these traditionally difficult signals.

Our theoretical investigation of nonmodal-source spectra contributes the first comprehensive mathematical derivations of the effects of deterministic and stochastic source perturbation on the short-time spectrum. Before this point, researchers had commented on the addition of subharmonic spectral line components when a glottal-pulse train was perturbed, but had not developed a theoretical foundation and explanation using the source envelope function that we derived. Likewise, random jitter and shimmer were known to appear noise-like in the spectrum, but this thesis is the first work to include the derivation of a formula describing the theoretical origin of the noise component. The spectral investigation is important to the study of nonmodality because short-time spectra are used commonly as the basis for voice analysis.

The theoretical examination of a pitch representation of nonmodality gives an alternative viewpoint on the frequency content that does not rely on Fourier bases. Many sources in the literature imply that there is an underlying fundamental frequency that must be found in nonmodal regions of phonation, but our work shows that such a unique fundamental-frequency contour does not exist. As we will discuss in the future-work section, this fact yields many interesting questions regarding the pitch that human listeners *perceive* for nonmodal sequences of impulses. It also yields insights into why conventional sinewave analysis-synthesis tends to break down in nonmodal regions.

Finally, we have devised a novel set of time-domain features for the representation of glottal-event patterns. These features cluster similar patterns close to each other, a property that can be used to characterize the glottal source properties of different speakers. These features have led to the contribution of an algorithm to extract phonation types explicitly based on their glottal event sequences. Other authors have devised detectors for "irregular phonation" or "glotttalization," but rely on hand labels of regions to train the detectors. As discussed in Appendix A, our methods are unique in that they are able to find and separate different classes of behavior based on properties of the glottal events themselves, without using such labels.

8.3 Future Work

This section describes areas for future work related to topics discussed in the thesis. It is subdivided into seven areas: refined definition of glottal events, improvements to glottal-event extraction, application to automatic recognition problems, separation of glottal event patterns, voice-quality measurement, sinewave represention, and perceptual implications. The discussion does not follow the chapter outline as many of the topics span the concepts presented in multiple chapters.

8.3.1 Defining Glottal Events

In Chapter 1, we presented a definition of glottal events motivated by work in the literature and also real speech and synthetic examples. In particular, we presented a model for speech as a set of impulses each with an associated composite impulse response. This definition still requires much refinement, including the formalization of a relationship between events and the concept of instantaneous excitation energy in the production system, as explored by Teager and others [57, 70].

A more rigorous definition is important to speech research where terms like "epoch," "glottal closure instant" and others are commonly used. This is especially important when considering source phenomena where the movements of the vocal folds differ from the traditional opening and closing cycle and include complicated vibration patterns, lack of synchrony between the folds, and lack of complete closure [8, 45, 66]. Aeroacoustic events such as ripples in the airflow and vortices may also create excitation sources. By creating a better model of the occurrence of impulsive excitation, we can create better methods to detect these events in natural speech. It is likely, also, that an impulsive-excitation model is only viable for a subset of complex speech phenomena. Therefore, an important area for future research is studying the characteristics of this subset. As part of such an investigation, synthesis of natural speech using an impulsive model may provide an important way to judge the appropriateness of the glottal-event representation.

8.3.2 Spectral Representations

Using the formulations for the effect of nonmodal glottal events on the spectrum derived in Chapter 3, we hope to answer questions related to three fundamental techniques in speech signal processing [56]. The approaches we are interested in are linear-prediction analysis, sinusoidal analysis-synthesis, and spectrally-derived features such as the mel-cepstrum, as well as in the analysis of disordered voices. For linear-prediction analysis, we are primarily concerned with how the spectral envelope created by the repeating impulse pattern affects the linear-prediction coefficients. This exploration is important to understanding the effect of nonmodality on, for example, the inverse-filtering of speech.

Sinusoidal analysis/synthesis may also benefit from our study. Sinusoidal-based coders and speech modifiers, for example, analyze the spectrum for prominent peaks to assign sinusoids. As we have shown, simple perturbations can alter the prominences of different line components and, in the stochastic case, can even make pulse-like signals appear like noise in the spectrum. Understanding how the spectrum evolves in time with different nonmodal signals will be important to understanding the meaning of sinusoids extracted by the sinewave system.

The effect of nonmodality on spectrally-derived features such as the mel-cepstrum is also an important area. We are interested in how such measures are affected by perturbations, whether fluctuations in the source can obfuscate formant frequency and bandwidth information, and whether useful information about perturbations in the source is present in spectrally-derived features or if such information is not recoverable. The results of the speaker-verification experiments in Chapter 7 using mel-cepstrum directly on the pulse train provide evidence that the sensitivity of the spectrum to event perturbations is detrimental when compared with the time-domain representation.

Finally, we are interested in applying our findings to the clinical voice-analysis domain where objective measurements of nonmodality are important (see [72] and [47] pp.481-519). In the clinic, it is standard to analyze stationary sounds that exhibit different kinds of timing and amplitude perturbations. Our methods may help in extending the current measurement techniques to be effective on running speech in contrast to sustained vowel utterances or with severely pathological cases.

In addition to the above areas, the effect of natural pitch variations on our derivations must be addressed in future work. Natural pitch intonation occurs in running speech and is known to cause, among other spectral effects, blurring in the high-frequency harmonics (see [56] p.548). For the time-domain representation of Chapter 5, we have found that linear changes in the pitch yield a simple constant offset, while the effect on the spectrum remains to be derived. A full comprehension of this phenomenon will be critical to understanding how our findings relate to the frequency-domain analysis of natural speech.

8.3.3 Improvements to Glottal-Event Extraction

Although we have contributed a new perspective on glottal-event extraction with our analysis of MED and the hybrid method, this area still requires further work. Five particular areas for future efforts are (1) synthesis experiments (2) a more realistic ground-truth metric for glottal events, (3) a method for specifically testing performance in highly irregular regions, (4) combining multiple event-extraction techniques, and (5) a more advanced glottal impulse-detection algorithm.

The first area, synthesis, would provide an additional method by which to evaluate glottal-event extraction algorithms. The strategy of such experiments would be to analyze speech using linear prediction, MED, and the hybrid method and then synthesize an estimate of the speech using the composite-impulse response and impulse train extracted for every frame. By measuring the error between the acoustic signal and its estimate, the model could be validated.

A more detailed ground truth signal is very important for understanding the performance of glottal-event extraction and the meaning of the impulses we obtain. One alternative to the EGG technology used in this thesis is high speech imaging of the vocal folds, which is becoming more readily available [14]. Additionally, other nonacoustic sensors such as GEMS and accelerometer may eventually be useful as a ground truth. A better ground truth is necessary if we are to evaluate how well our measures extract different events in the physical system such as glottal openings and closures, as well as other behaviors in between the two such as closure of only part of the vocal folds.

We would also like to specifically evaluate the ability of our algorithms to deal with highly-irregular regions of natural speech. In this thesis, we evaluated our measures for significant synthetic perturbation, but did not restrict our analysis to such regions in our study of natural speech. A multi-pass method, using the pattern-separation technology in Chapter 5 to find likely nonmodal regions would be one way to specifically evaluate highly-nonmodal regions.

Another subject for future research is combining multiple event-extraction techniques in order to exploit their different advantages. A combination of MED, linear prediction, and the hybrid method, for example, could take advantage of each of their respective strengths. As we have seen, MED tends to produce an impulse-like residual at the expense of unrealistic composite impulse responses, linear prediction produces spectrally well-behaved estimates of the impulse response but is limited to minimum phase, and the hybrid method is between the two, allowing mixed-phase responses while remaining spectrally well-behaved. Although not discussed in the body of the thesis, we have also observed anecdotally that MED can behave strangely for non-vowel voiced regions, such as nasals, whereas linear prediction performs comparably between the two kinds of regions. This property further motivates combining deconvolution approaches.

Finally, we must also make improvements to the glottal-impulse extraction algorithm. Currently, our approach is a simple peak-picking algorithm using only two parameters, dead time and threshold, held constant over all utterances and across each utterance. An area for future work is the improvement of these algorithms. In particular, the threshold could be adaptively adjusted based on the local level of the signal and the dead time modified based on estimates of a speaker's pitch. As we have seen in our experiments, female speech responds differently to dead time than male speech, hypothetically because of the higher fundamental frequency for females. More investigation needs to be done to understand how to tune our detection algorithm to take this into account.

8.3.4 Application to Speaker, Dialect, and Language Recognition

The potential of our features for speaker recognition is promising. Our work supports the claim that the mel-cepstrum and other similar spectral approaches do not represent the subtleties of the source to the same degree as our event-pattern features. One might then ask why modern state-of-the-art systems do not use more source information.

One issue that must be better studied and addressed is the robustness of such features to channel distortion and noise. There is a large dynamic range of the energy of glottal events, meaning that noise may make it difficult to detect small but important events, while only mildly affecting others. Future work in this area should revolve around how source-based information changes when exposed to additive and convolutive distortion. Future work must also evaluate how to best implement glottal-event extraction techniques in these situations.

In addition to automatic speaker recognition, language and dialect recognition may additionally provide a good place to look for advantages of source-event patterns. Glottal-event features may extend to dialect and language recognition in the same way that other feature sets such as mel-cepstra have extended to these problems from speaker recognition in the past. Nonmodality is also text- and structure- dependent in many languages and dialects, occurring with more frequency in certain word contexts and prosodic locations. It is also a phonologically-distinctive feature, used in some languages to discriminate between different phonemes. By combining higher-level knowledge of prosody and phonetics with low-level source information, future efforts may be able to improve on current recognition approaches.

8.3.5 Separation of Glottal Event Patterns

The algorithm that we have described for separating different patterns of glottal events relies on the notion that, during a region with certain glottal-event characteristics, we can predict future feature vectors based on previous ones. Our approach requires two parameters—the minimum length of a repeating sequence and the amount that the prediction can differ from the observed feature.

An alternative approach for future work may be a statistical model, such as a Markov model, that allows for the evolution between a set of pattern *states*. This type of model would allow us to handle temporary deviations from the expected pattern, such as due to missed glottal events or inserted glottal events. It may also allow us to learn a set of patterns related to a given speaker and patterns of transitions between these patterns. Such an approach may allow us to learn characteristic sequences of patterns such as modal-diplophonia-irregularity that are characteristic of a given speaker.

The separation algorithms that we have developed may also be useful for automatic speaker recognition. The separation technique would allow each type of phonation for each speaker to be modeled separately from the others. Separating different source behaviors would also allow modeling of the dynamics of changes between each phonation type for different speakers.

8.3.6 Voice-Quality Measurement

Appendix A contributes a preliminary discussion of how glottal-event-based analysis of the acoustic source may improve voice-quality measurement for many applications in speech science, clinical practice, and engineering. The application of our approaches to these areas and refinement of the algorithms to be practical is thus an important area of future work.

One promising application of our tools is to quantitatively measure aspects of voice quality in regions of speech labeled by humans. For example, we can select all the word-initial vowels or all of the regions perceptually labeled as "rough" by a panel of human subjects. We may then ask how regions differ from one another other in a quantitative way. Another application is to use our automatic methods to select *regions of interest* that can be further analyzed by humans or machines. This is especially important when there is too much data for humans to label efficiently. Quantitative measures can be used to augment linguistic rules that currently help labelers select sections of speech to analyze [64].

8.3.7 Sinewave Representation

Chapter 4 outlined several of the implications of our work with nonmodal sequences on a sinewave representation. We found that measurement of the underlying fundamental frequency does not lead to a unique impulse train. There is much work left to be done in this area, particularly with showing that the theory accounts for missing, moved, and additional glottal pulses observed in current sinewave speech systems. Developing a formal argument for the modification of underlying pulses in actual speech will require a generalization to the relations presented in this chapter and specifically will require addressing two properties of real speech: First, standard sinusoidal analysis is not constrained to yield frequency harmonicity or continuity such as we have described in our model [44]. In nonmodal speech regions, sinewave analysis via peak-picking in the frequency domain can yield what appear to be frequencies that are erratic spectrally and temporally, as we saw in Chapter 3. An important consequence is that, although impulses are generated approximately in synthesis, the resulting frequency tracks are broken and sporadic and appear to have little meaning physiologically.

The second property to address is that natural speech is filtered by bandlimiting and vocal-tract filtering, as we outlined in Section 4.3. We must address how such manipulations can be represented while keeping a sinusoidal representation. Although we have argued in Chapter 4 that real speech will suffer from the same sinusoidal-model uniqueness problems of impulse trains, we do not yet fully understand how associated sinusoidal tracks change with the introduction of vocal-tract linear filtering and bandlimiting. A promising area for future work is an extension of the discussion begun in Section 4.3 to connect the work of Bovik [10] on the filtering of single sinusoids to understanding how filtering affects the harmonically-related sinusoidal model.

8.3.8 Perceptual Implications

The theory developed in this thesis to address the effect of nonmodal phonation on the spectrum and fundamental frequency leads to many important questions for human perception. From the spectral point of view, we have seen that the harmonic structure of speech becomes distorted when the glottal cycle becomes nonmodal. How humans perceive sections of deterministic and stochastic glottal-cycle perturbations is thus another subject for future research. Sun [68] has begun to address the question of pitch for certain 2-long deterministic patterns.

The issue of fundamental frequency provides another avenue for perceptual testing. Since the underlying fundamental-frequency track of nonmodal regions is ambiguous, a subject for future research is which of the possible theoretical fundamental-frequency contours, if any, human listeners perceive. By studying this issue, we may begin to better understand how listeners encode time-varying pitch and how it relates to the spectrum and the time-domain occurrences of glottal pulses.

Appendix A

Measuring Voice Quality

This thesis has presented a methodology for capturing and representing a variety of phonation behaviors in terms of their underlying impulsive excitation events. We have focused on the speaker-dependent aspects of these phenomena, presenting evidence that our features contain speaker-dependent information that is complimentary to conventional spectral-based features.

In addition to automatic speaker recognition, there are many applications to which our findings can be applied. This appendix describes how the theory and techniques described in this thesis can be applied to research problems which require the measurement of voice quality. Our primary goal is to make our findings accessible to nonspecialists, showing them the immediate value of our ideas to their work.

The appendix is organized as follows. First, we motivate the measurement of voice quality, describing the different factors that contribute to it, and highlighting the importance of the acoustic source. We then discuss the acoustic-source contribution to voice quality in two parts, *what* acoustic source phenomena occur and *where* these phenomena occur. Finally, we describe approaches to measuring these elements of the source contribution using the techniques developed in this thesis. We show examples of glottal-event analysis on natural speech examples including cases where it complements conventional analysis techniques.

A.1 Motivation

Speech researchers, including linguists, engineers, and clinicians, are confronted by a wealth of voice qualities in their practices. In a rigorous sense voice quality is difficult to quantify and define, but is clear from the literature that researchers *do* try to arrive at definitions. There are strong disagreements, though, especially with regard to the terminology and perceptual, physiological, and acoustic correlates that make up a certain voice quality [20]. Different fields have worked out their own taxonomies of voice quality, which define relationships between terms for different voice qualities. In the clinical voice therapy domain, for example, a basic taxonomy of voice quality might look like Figure A-1, where arrows denote voice qualities that are subsets of other voice qualities. See [40] for more detailed information regarding this particular taxonomy and how it was derived from the literature.

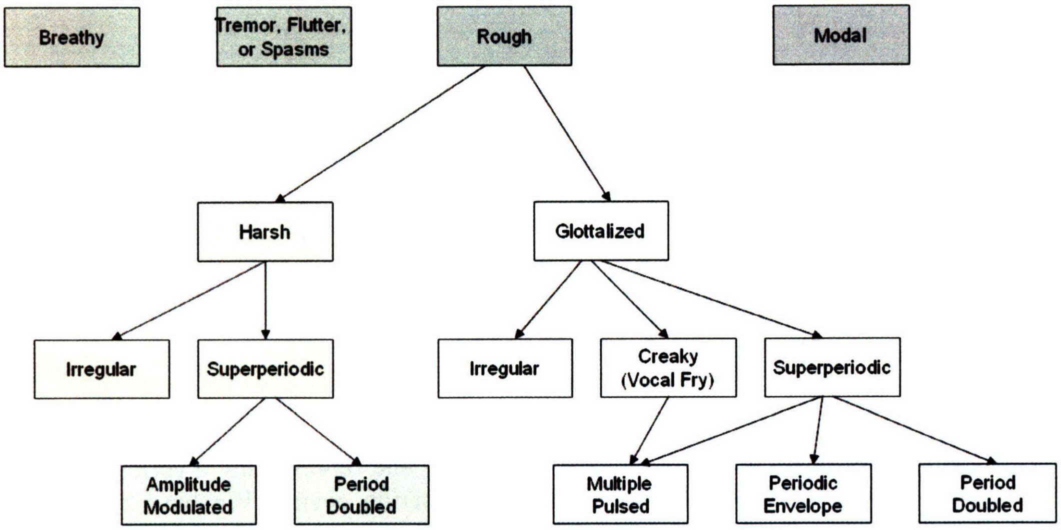


Figure A-1. An example taxonomy for voice quality, derived from the clinical literature.

Despite differences in the systems used for each field, many aspects of taxonomies for voice quality stem from activity taking place at the *glottal source*, as depicted in Figure A-2, where the glottis is the space between the vocal folds. The focus is on how the vocal tract is excited by this source rather than the vocal tract itself.

Development of a taxonomy requires many inputs including aspects of the vocal-fold physiology, perceptual judgments, acoustic measurements, and linguistic and prosodic context. These inputs are depicted schematically in Figure A-3. Depending on the field, these influences may be given different weights; for example voice-quality definitions in the clinical community depend heavily on the physiological health of the vocal folds [75] and the perceptual quality of the voice [38], but not to linguistic and prosodic context. Given these challenging issues, we are not trying to create taxonomies or assign labels in this appendix. Instead, our objective is to describe the kinds of *acoustic source properties* that are important to voice quality and methods for their measurement.

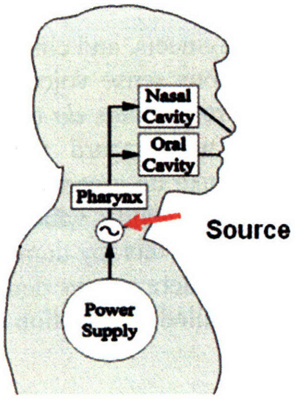


Figure A-2. Location of the glottal source, the origin of many acoustic phenomena related to voice quality.

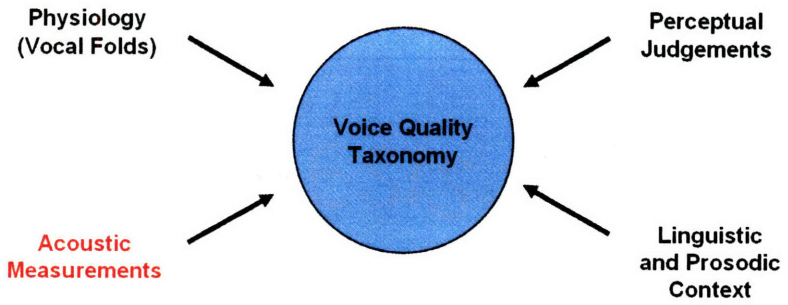


Figure A-3. Schematic of the different types of information that contribute to the development of voice-quality taxonomies. The focus in this appendix is on acoustic measurements.

A.2 Acoustic Source Contribution to Voice Quality

As we have discussed, the glottal source is a major contributor to voice quality. It consists of two primary components, *what* the acoustic characteristics of the source are and *where* they occur in speech. The acoustic characteristics are generated by behaviors in the physiology and aeroacoustics at the glottis. In this thesis, the voice production model is a glottal source which excites the vocal tract. As we have shown throughout the thesis, there are many different variations that may occur in the source including regular cycles, irregular cycles, and differences in the pulse shape, illustrated in Figure A-4. The implication of our focus on the acoustic characteristics of the source is that changes in the source will change the perceptual quality of the output.

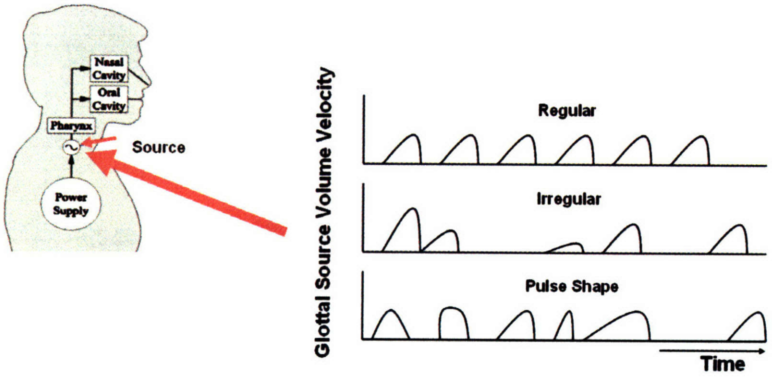
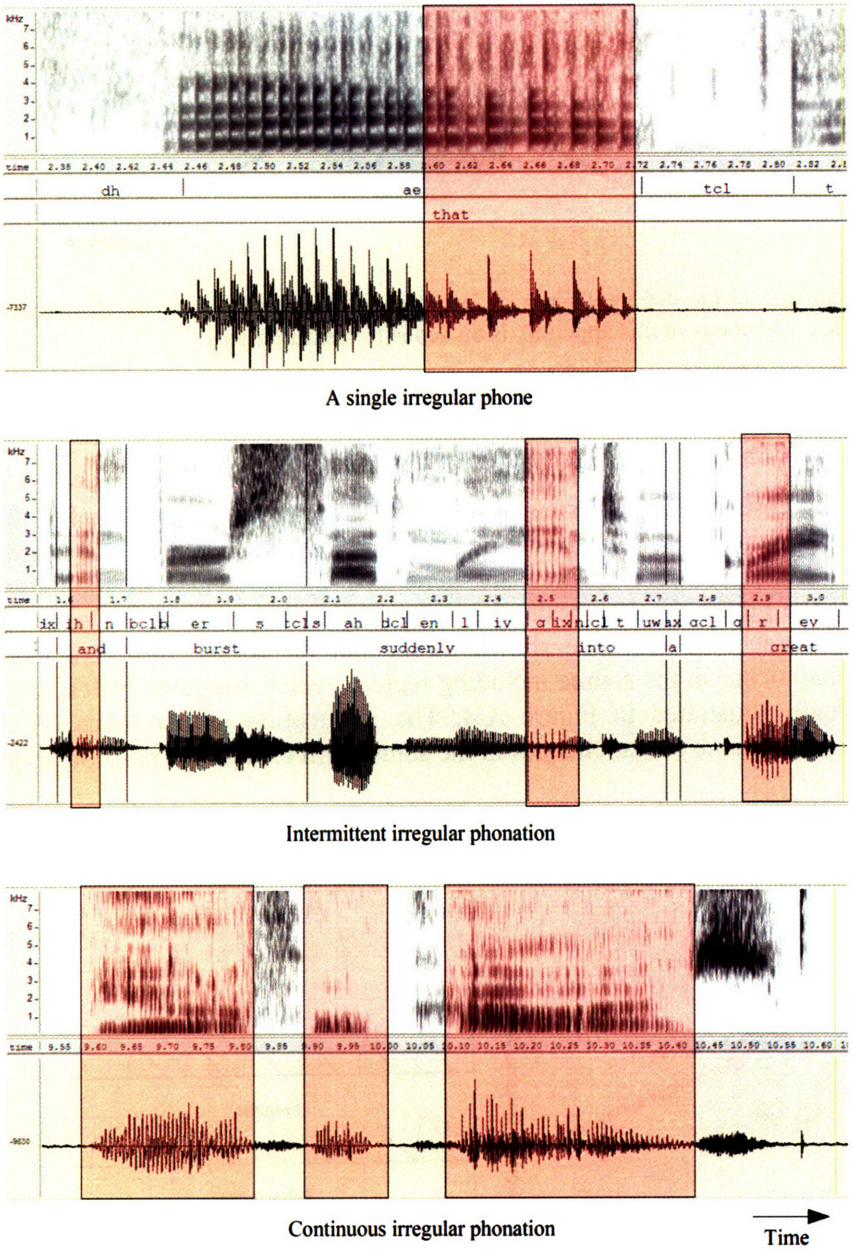


Figure A-4. Examples of variations of the source acoustics that affect voice quality.

The other important aspect of the source contribution is where different acoustic characteristics occur, including their location and duration. The duration of voice qualities span a range of behaviors including short acoustic phenomena ("that phone has a creaky quality"), the effect of the relative positioning of many intermittent short acoustic phenomena ("our current subject is a persistent glottalizer"), and acoustic properties occurring over an extended period of time ("she is hoarse today"). Figure A-5 depicts the contribution of acoustic phenomena at different scales contributing to voice quality. As we will discuss further, the specific location of the acoustic phenomena depends on many factors including linguistics, phonetics, prosodic structure, and the state of the vocal folds.



A single irregular phone

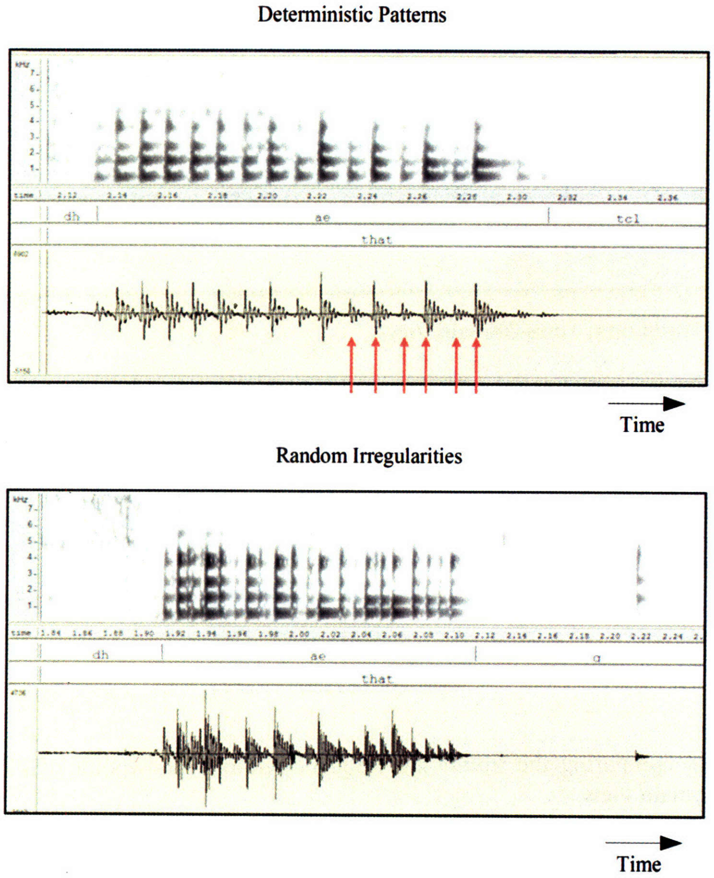
Intermittent irregular phonation

Continuous irregular phonation

Figure A-5. The contribution of acoustic phenomena of various durations to voice quality. Highlighted regions indicate three different scales over which acoustic source behaviors take place. Phone and word transcriptions are included in the top two panels for reference.

A.2.1 Influences on Acoustic Glottal-Source Behaviors

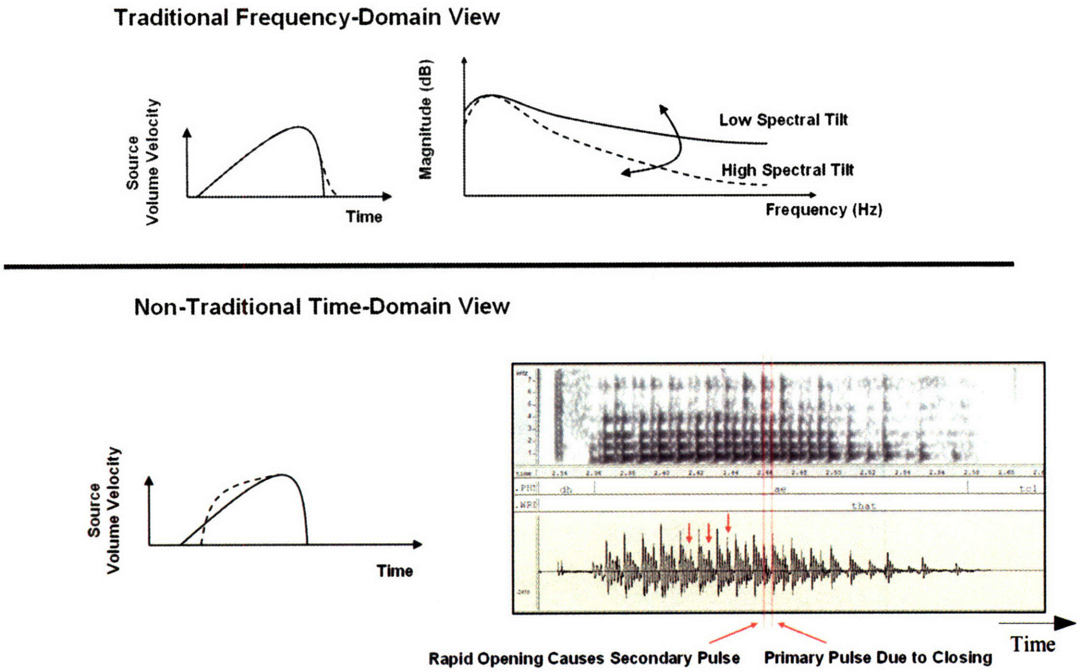
We can break the acoustic source characteristics themselves into two categories: relationships between the amplitudes and timings of neighboring glottal cycles and changes to the glottal-pulse shape. The first aspect is related to the between-cycle glottal events discussed in Chapter 2. As we have shown in the thesis, relationships between glottal cycles have varying degrees of both deterministic and stochastic aspects, forming a continuous range of behaviors. Examples of deterministic and random properties of neighboring glottal cycles is depicted in Figure A-6.


Deterministic Patterns

Random Irregularities

Figure A-6. Illustration of deterministic and random aspects of relationships between the amplitudes and timings of glottal cycles.

The second aspect of acoustic characteristics, glottal shape, is traditionally analyzed in the literature in the spectral domain [23]. As depicted in Figure A-7, typically this contribution is measured using the *tilt* of the spectrum, proportional to the rate at which the glottal source rolls off in frequency. Based on the observations about intra-cycle glottal events in this thesis, though, we can also see that there is a non-traditional *time-domain* view of the source contribution that complements the spectra view. Here, steep glottal onsets lead to additional glottal events within a single glottal cycle. The figure depicts this contribution with a steep glottal onset leading to secondary impulsive excitation.



Traditional Frequency-Domain View

Non-Traditional Time-Domain View

Figure A-7. Schematic comparing the traditional frequency-domain viewpoint of glottal shape with the non-traditional time-domain view.

A.2.2 Influences on Location of Acoustic Glottal-Source Behaviors

Language and dialect are two of the major influences on the locations, durations, and overall frequency of regions of nonmodal phonation. These factors help govern the relationship between aspects of phonetic and prosodic structure and the occurrence of nonmodality. In English, for example, certain features of prosodic structure have been linked to an increased percentage of nonmodal glottal cycles. In particular, word-initial vowels that are at the beginnings of intonational phrases are linked to nonmodality, especially when they are found in a pitch-accented word [15]. Intonational-phrase-final nonmodality is also common, where it is often known as "creak" [26] and is related to a lowering of the fundamental frequency [39]. The overall degree of nonmodality also has been shown to depend on dialect factors, as has been shown for British English [26].

In addition to language and dialect, other aspects can also affect the location of nonmodal phonation. Hagen [22] surveys the effects of many factors on the occurrence of nonmodal phonation. Speakers themselves also seem to exhibit specific styles in addition to their broader sociophonetic background. "Habitual creakers" [26] in the Modified Northern dialect of British English are one extreme example, exhibiting nearly constant nonmodal phonation. Speakers of the same language have been shown in the literature to differ in terms of how often they produce nonmodal phonation [15, 59]. Speaker-dependence of glottal events, the application explored in Chapter 7 of this thesis, uses information about these factors obtained from the source in order to distinguish different talkers. We show that glottal-event sequences differ between speakers allowing us to tell them apart.

A.3 Approaches to Measurement of Source Phenomena

We have thus far discussed the importance of *what* the acoustics of the glottal source are and *where* nonmodal phenomena occur. We now explore the contribution of this thesis to measuring these two elements. Through several examples, we will motivate the use of glottal-event extraction as the basis for quantifying source information.

A.3.1 Acoustic-Source Characteristics

Our approach to measuring acoustic-source characteristics is based on explicitly extracting a set of impulses from speech that represent the timing and amplitudes of a set of underlying impulsive glottal events. As shown in Figure A-8 for near-modal phonation, highly-irregular phonation, and a 2-long deterministic timing pattern, the event representation captures meaningful impulsive excitation in the glottal cycle. As with other analysis approaches in signal-processing, such as the spectrogram, there is a set of parameters that tunes the behavior of the analysis. The two parameters used in our glottal-event extraction algorithm are the *threshold* and *dead time*, discussed in Chapter 6. Guidelines for setting these parameters in speech-science applications have not yet been developed, although we have explored some of the parameters' effects in Chapters 6 and 7.

In Chapter 5, we developed a feature space with which to visualize sequences of extracted glottal events. These are shown in the right panels of Figure A-8 for each of the acoustic phenomena. We see that in this space, near-modal phonation clusters around the origin, random timing variation causes dispersion of the features away from the origin, and deterministic timing patterns form distinct clusters away from the origin. As nonmodality is a continuum, so is the feature space, able to represent sequences that are nonmodal to different degrees and with different characteristics.

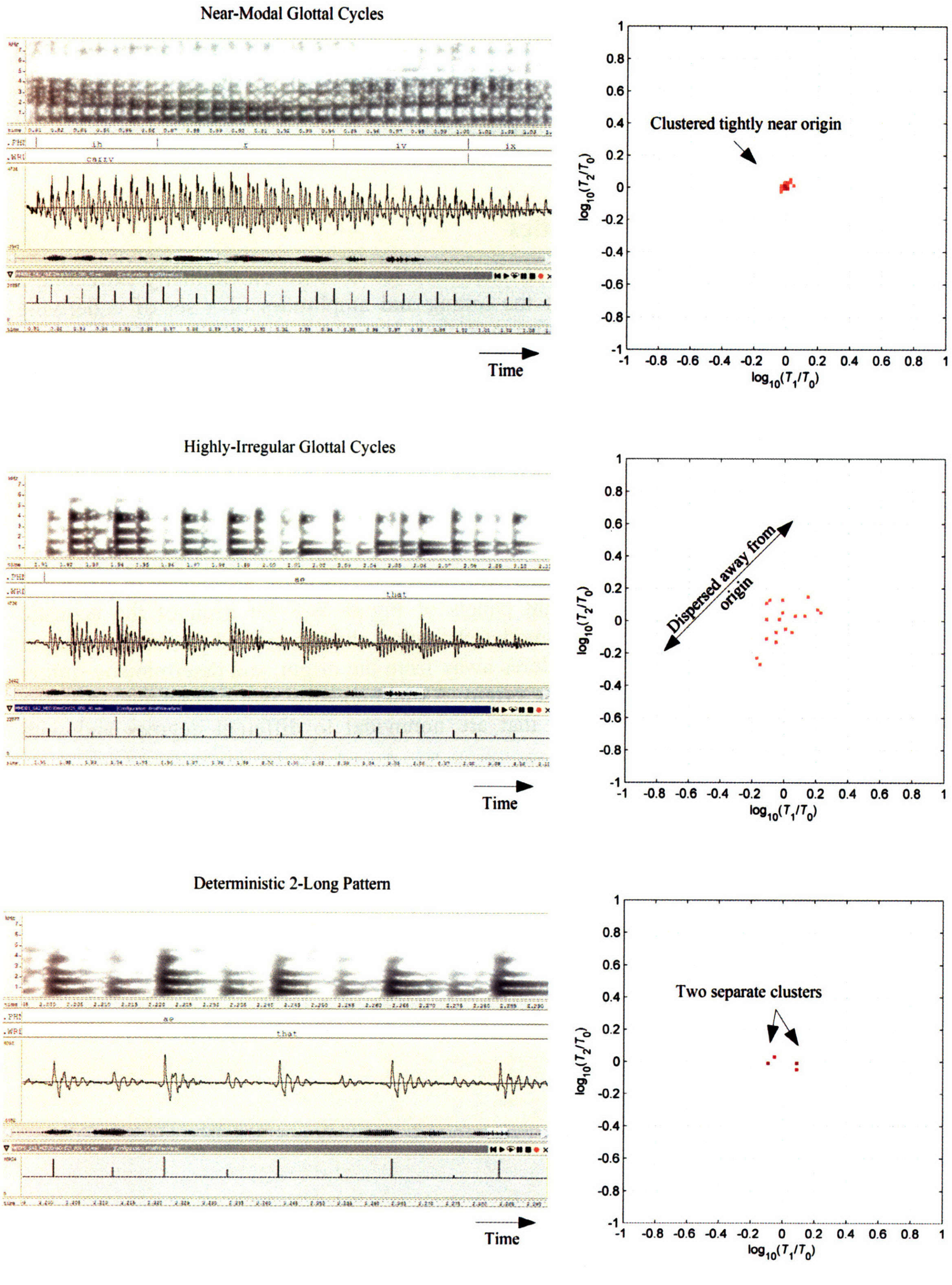


Figure A-8. Quantitative measurement of the acoustic phenomena comparing three speech segments from male speakers. One region contains mild jitter, another exhibits highly-irregular phonation, and the last is a short segment exhibiting a 2-long timing and amplitude pattern. The timing-feature space for the mild jitter is tight around the origin whereas it is dispersed for the highly-nonmodal case. For the 2-long pattern, two clusters are formed symmetric about the origin. MED was used for glottal-event extraction with a dead time of 5 ms for the top two panels and 8 ms for the bottom panel.

Glottal-event features may also be used to characterize differences in the source shape. Figure A-9 presents an example exhibiting the effect of strong glottal openings. The average fundamental frequency of this male speaker is about 105 Hz, measured across the entire utterance, and we can see the openings manifesting as extra excitation points about 4 to 5 ms before each closure. In the segment shown, a pitch estimate using the conventional ESPS tool in WaveSurfer [65] is nearly constant at about 100 Hz. We can see that the excitations at the glottal openings in the latter half of the utterance, though, are large. Although we leave a discussion of perception to future research, it is interesting to note that the extra events in the second half of this example yield a higher-frequency percept than the first half. *This perceptual change in the pitch is not reflected in the fundamental-frequency estimate.*

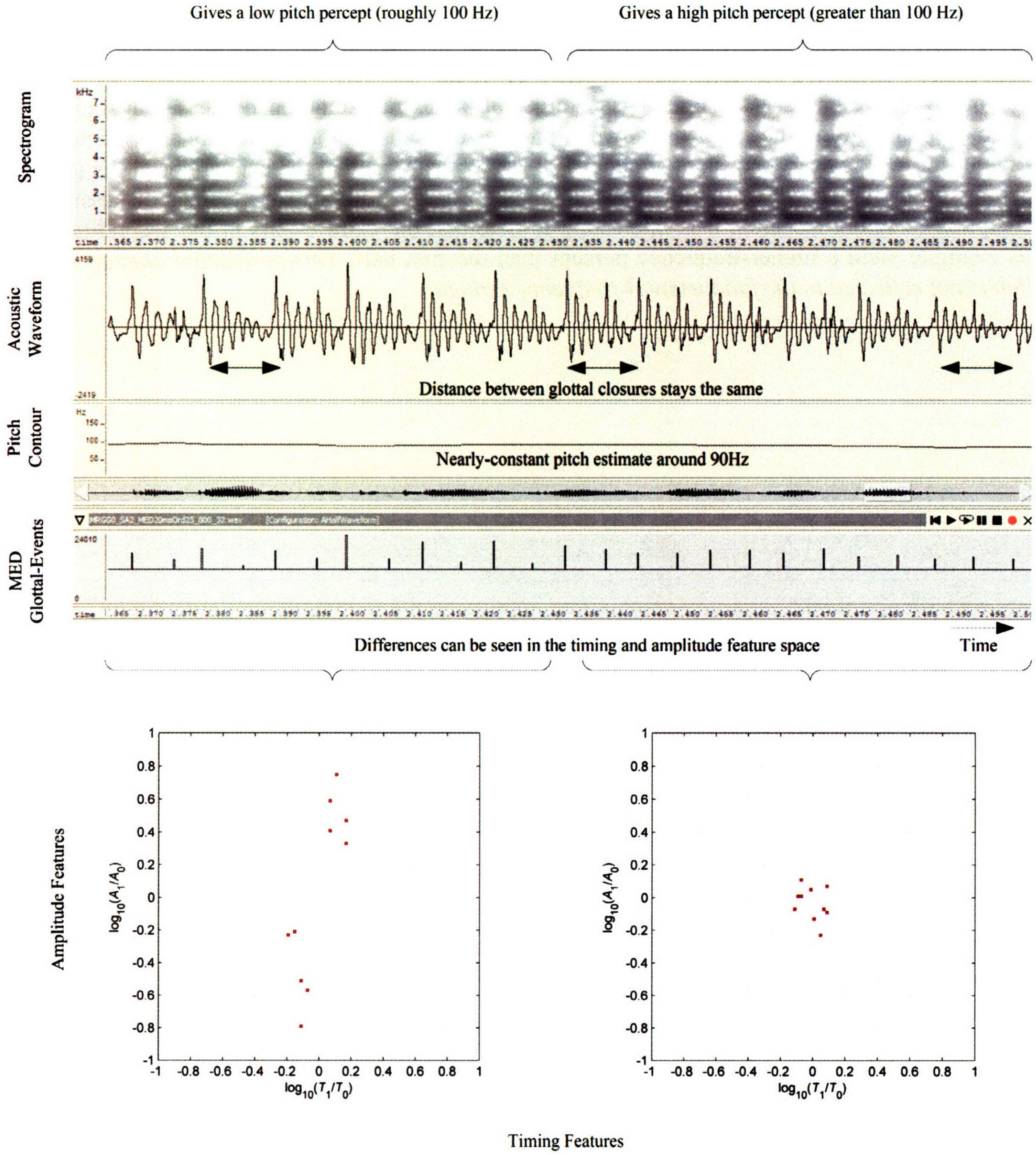


Figure A-9. Glottal-event features augmenting a conventional pitch estimate with information about the source shape. MED was used for glottal-event extraction with a dead time of 4 ms.

We can quantify the difference between the first and second half of the segment using our glottal-event features by plotting the first amplitude feature on the *y*-axis and the first period feature on the *x*-axis as shown in the lower two panels of Figure A-9. The two sections are significantly different in this space, with the left plot indicating two clusters, reflecting a 2-long amplitude pattern and the right plot clustered loosely about the origin. This example illustrates the promise of using our new features to augment conventional measurements such as pitch in order to learn more about the source.

A.3.2 Location of Acoustic Source Characteristics

We can also apply the glottal-event features to measuring where different acoustic source characteristics occur. In Figure A-10, we use a version of the glottal-event pattern separation algorithm presented in Chapter 5 to automatically separate the voiced regions of a speech segment from the phrase "The clouds bulged downward and burst suddenly into a great black funnel" into different classes of behavior. Here, we have classified two sections as having near-modal timing, one section as having a 2-long timing pattern, and a third as having irregular timing. The identified regions appear reasonable except for the /n/ in "into" which is labeled as irregular because of errors in the MED event-extraction process resulting in a missed closure and a slight shift in the last two events.

The procedure for finding the regions in time consists of first building the glottal-event features and separating the 1-long, 2-long, 3-long, and all other sequences that do not fit these categories (given the term "irregular" here) using our separation algorithm. The starting and ending times of the impulse sequences used to create each feature are also recorded. We then combine all the samples covered between the beginning and end times of each of the features in a particular class of sequence to generate the output signal detecting that sequence. The algorithm in its current form only separates glottal event sequences based on whether it is a 1-long, 2-long, or 3-long pattern. However, further criteria could be used to achieve more detailed separations depending on the application. For example, one could further classify contiguous segments exhibiting a certain pattern based on their durations. Another direction would also be to add further information to each separated section, such as the degree of deviation from modal.

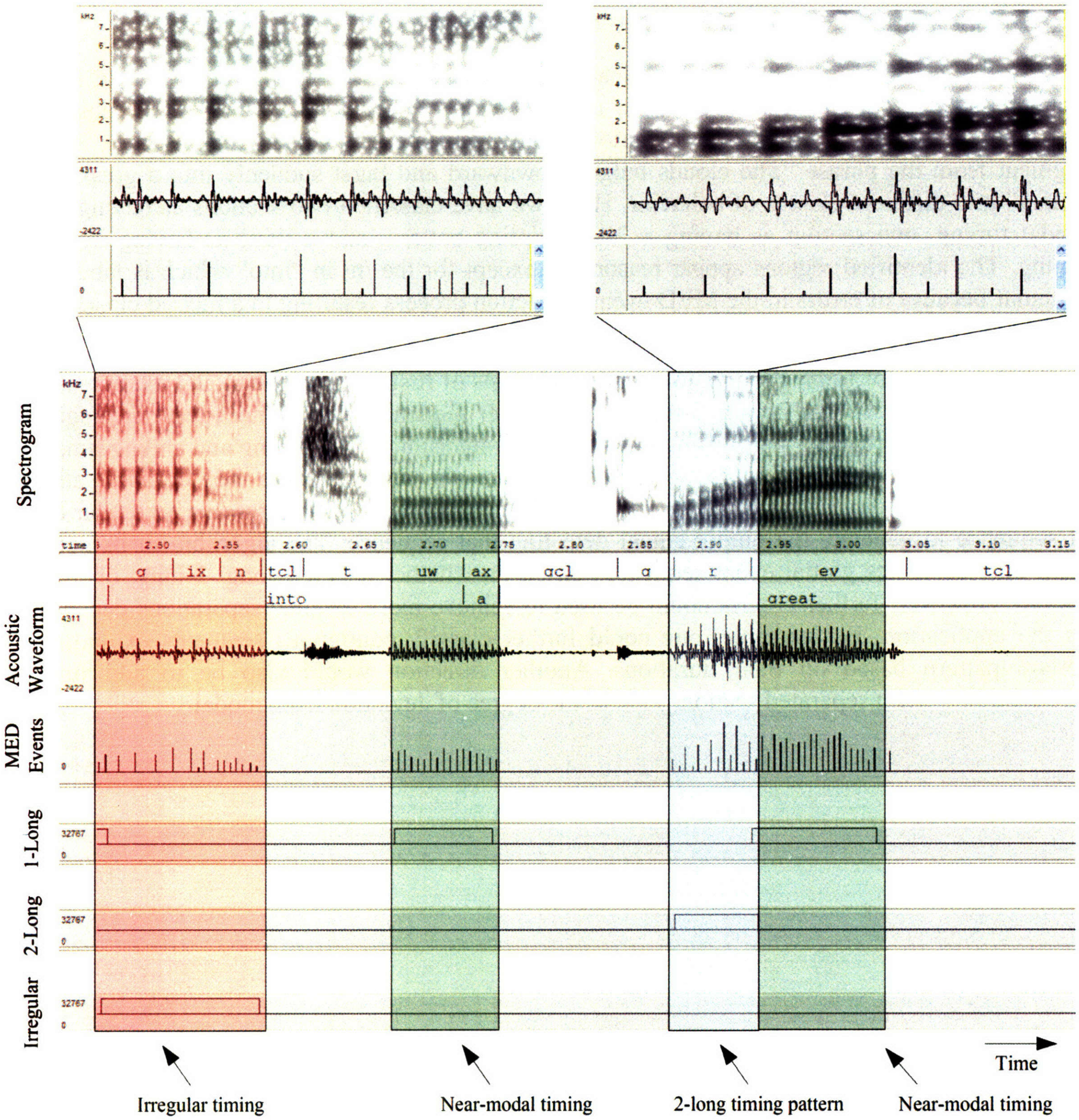


Figure A-10. Automatic tracking of different regions of three different types of glottal-event patterns for speech from a female talker. The algorithm has separated regions containing near-modal timing, 2-long timing patterns, and irregularly-timed phonations. The output of each of these detectors is plotted and the corresponding parts of the spectrogram, acoustic waveform, and MED events are highlighted. MED was used for glottal-event extraction with a dead time of 3 ms.

A.3.3 Application to Speech-Science Experiments

The examples we have shown highlight the usefulness of the approaches in this thesis for determining the acoustic source properties and their locations in speech. Speech-science experimentation can benefit from quantitative measurements. For example, we might select all the word-initial vowels or all of the regions perceptually labeled as "rough" by a panel of human subjects. We may then ask how these regions differ from one another in a quantitative way, based on the source characteristics they exhibit. Such quantitative measurements also avoid unnatural hard categories such as specifying that a region is "irregular" versus "regular" as is common in the literature [22, 32, 69, 73]. Our glottal-event approach allows for a continuum of different source behaviors, as exist in natural speech, to be quantified and compared.

Another experimental application is to use the automatic methods to select the locations of regions of interest that can be further analyzed by humans or machines. An example speech-science application would be in an experiment to determine how 2-long patterns of glottal events affect the percept of pitch. The pattern-separation algorithm could be used to automatically extract all such regions from a large speech database for pitch rating by a human listener.

When regions of interest are analyzed automatically, they may reveal differences between populations that are difficult to find by hand. For example, they may find a series of regions with subtle 3-long timing patterns in two different speakers and show that these 3-long patterns differ significantly in their feature space. This kind of machine analysis of speech could be used to automatically discover important phonation phenomena that discriminate speakers, dialects, and voice disorders. Humans can then use this information to inform the creation of voice-quality taxonomies. In this way, our work is important in fields such as sociolinguistics.

A.4 Conclusions

This appendix has discussed the acoustic source contribution to voice quality measurement. We have seen that the source consists of two distinct components, its local acoustic characteristics and the locations of these characteristics in a larger context. The glottal-event-based analysis approaches in this thesis provide a way to measure the acoustic source contribution, important to voice quality based explicitly on the timing and amplitude of impulse source-excitation events. Through several examples, we have shown how this perspective can be used to characterize the source and track changes in its contribution across time.

Appendix B

Event-Extraction Algorithms

This appendix presents details of the event-extraction algorithms used in this thesis. Each algorithm is presented in schematic form with the important components described in accompanying text. For the MED and hybrid methods, sections of the code are included for researchers wanting to replicate our findings.

B.1 Linear Prediction

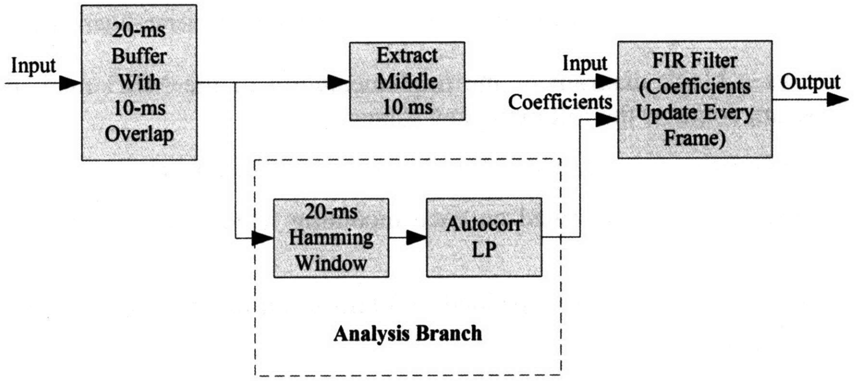


Figure B-1. Block diagram of the implementation of the linear-prediction method of inverse-filtering.

Figure B-1 shows a block-diagram of the implementation of the linear-prediction (LP) method of inverse-filtering. This diagram contains the following key stages:

Buffer: Block processing is used, with 20-ms buffers and 10-ms overlap

Analysis Branch: The standard autocorrelation linear prediction block in MATLAB Simulink [2] DSP blockset is used to extract coefficients from the 20-ms-long Hamming-windowed input.

FIR Filter: The standard digital filter block in MATLAB Simulink DSP blockset is used. This block performs filtering in the time domain with the coefficients updated for each incoming 10-ms block.

B.2 MED

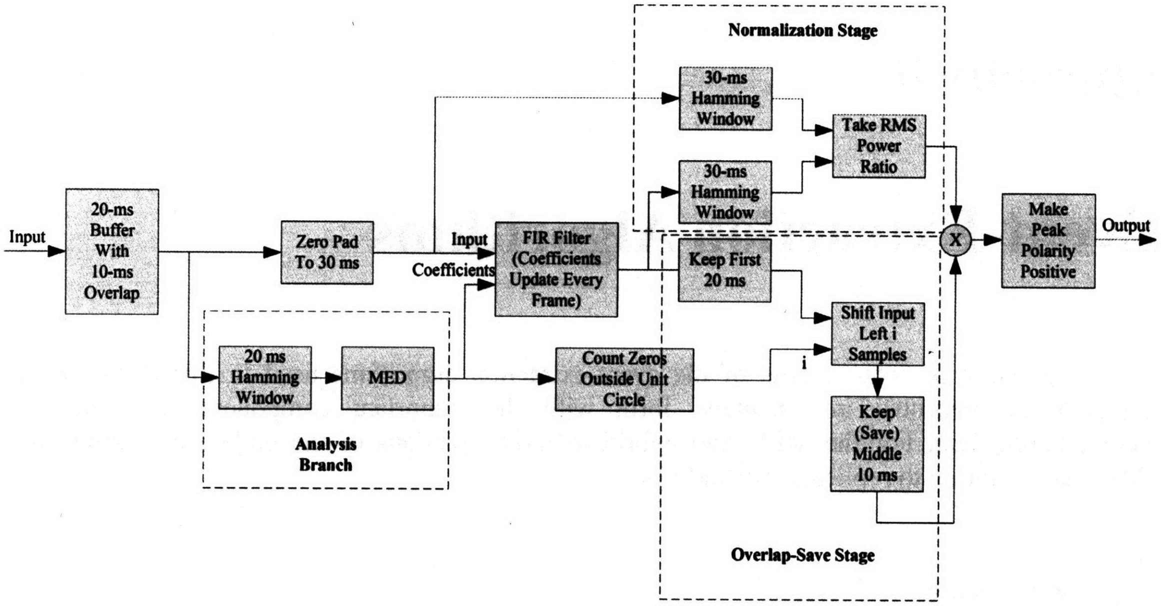


Figure B-2. Block diagram of the implementation of the MED method of inverse-filtering.

Figure B-2 shows a block diagram of the implementation of the MED method of inverse-filtering. This diagram contains the following key stages:

Buffer: Block processing is used, with 20-ms buffers and 10-ms overlap

Analysis Branch: MED is performed on each incoming window to calculate the FIR filter coefficients.

Zero Pad: Each input buffer is zero padded to 30 ms to allow the ringing of each processed frame to be captured.

FIR Filter: The standard digital-filter block in MATLAB Simulink DSP blockset is used. This block performs filtering in the time domain with the coefficients updated for each incoming 10-ms block.

Normalization Stage: The RMS power of the Hamming-windowed zero-padded input frame and the Hamming-windowed FIR-filtered frame are extracted. The ratio of these powers determines the scale factor applied to the output signal. The window is applied to a zero-padded signal and was intended to smooth only the beginning of the signal, whereas the end of the signal is allowed to ring out.

Overlap-Save Stage: As described in Chapter 6, an overlap-save scheme is used to apply the filter to each block.

Polarity Correction: The polarity of each normalized output block is corrected if necessary by making the peak polarity positive. This step is necessary since the polarity of the residual for a frame generated by MED is ambiguous.

The primary segment of MATLAB code used to compute MED is given below. For simplicity, the entire file is not shown, and the segment is part of a larger function. The variable `best_f` is the best set of filter coefficients found by the MED algorithm. The important parameters are:

`x`: the current 20-ms Hamming windowed input frame

`filterLength`: the length of the FIR filter being devised; equal to the order plus 1

`num_monte_carlo_trials`: number of randomly-generated initial filter values used

`num_iterations`: number of iterations of the MED algorithm to find coefficients

```
% initial best filter guess
best_f = zeros(filterLength,1); % impulse in the center
best_f(ceil(filterLength/2)) = 1;

max_V = 0; %initialize best V

for monteCarloIteration = 1:num_monte_carlo_trials
    % randomly initialize the filter coefficients
    f = 2.*rand(filterLength,1) - 1;

    for index1 = 1:num_iterations

        % calculate inverse-filtered waveform
        y = conv(f,x);

        % calculate varimax norm of the inverse-filtered waveform
        V = sum(y.^4)/((sum(y.^2)^2)+epsilon);

        % calculate simple variance
        u = sum(y.^2) + epsilon;

        % find R matrix which is a symmetric Toeplitz matrix
        autocorr_x = xcorr(x);
        autocorr_x =
autocorr_x(ceil(length(autocorr_x)/2):length(autocorr_x));
        autocorr_x = autocorr_x(1:filterLength);
        R = V.*toeplitz(autocorr_x)./(u);

        % find g vector
        crosscorr_y3_x = xcorr(y.^3,x);
        crosscorr_y3_x =
crosscorr_y3_x(ceil(length(crosscorr_y3_x)/2):length(crosscorr_y3_x));
        crosscorr_y3_x = crosscorr_y3_x(1:filterLength);
        g = (crosscorr_y3_x.')./(u^2);

        f = inv(R)*g;

    end

    % calculate inverse-filtered waveform
    y = conv(f,x);
    % calculate varimax norm of the inverse-filtered waveform
    V = sum(y.^4)/((sum(y.^2)^2)+epsilon);
    % See if we have found a better estimate according to varimax norm
criteria
    if (V > max_V)
        max_V = V;
        best_f = f;
    end
end
```

B.3 Hybrid Method

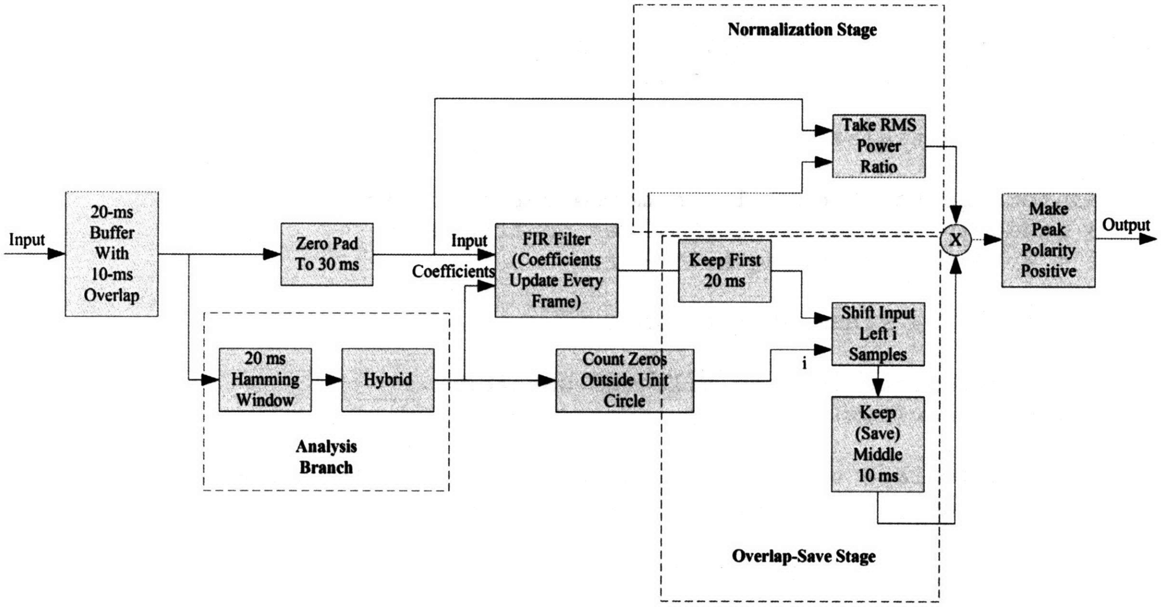


Figure B-3. Block diagram of the implementation of the hybrid method of inverse-filtering.

Figure B-3 shows a block-diagram of the implementation of the hybrid method of inverse-filtering. This diagram contains the following key stages:

Buffer: Block processing is used, with 20-ms buffers and 10-ms overlap

Analysis Branch: The hybrid method is performed on each incoming window to calculate the FIR filter coefficients. A segment of code describing this algorithm is included below

Zero Pad: Each input buffer is zero padded to 30 ms in order to allow the ringing of each processed frame to be captured.

FIR Filter: The standard digital filter block in MATLAB Simulink DSP blockset is used. This block performs filtering in the time domain with the coefficients updated for each incoming 10-ms block.

Normalization Stage: The RMS power of the zero-padded input frame and the FIR-filtered frame are extracted. The ratio of these powers determines the scale factor applied to the output signal. The process is different than for MED since no windowing is applied before the powers are calculated.

Overlap-Save Stage: As described in Chapter 6, an overlap-save scheme is used to apply the filter to each block.

Polarity Correction: The polarity of each normalized output block is corrected by making the peak polarity positive. This step is necessary since the polarity of the residual for a frame generated by the hybrid method is ambiguous.

The primary segment of MATLAB code used to compute the Hybrid filter is given below. For simplicity, the entire file is not shown, and the segment is part of a larger function. The variable `numBest` is the best set of filter coefficients found by the Hybrid algorithm. The important inputs are:

`inputWaveform`: the current 20-ms Hamming windowed input frame

`order`: the order of the FIR filter being found

```
% estimate locations of the zeros of the inverse-filter
minPhaseTF = real(lpc(inputWaveform, order));
[z,p,k] = tf2zpk(minPhaseTF,1);

% seperate conjugate poles from individual poles
z_conj = [];
z_real = [];
for index1 = 1:length(z)
    if isreal(z(index1))
        z_real = [z_real;z(index1)];
    else
        z_conj = [z_conj;z(index1)];
    end
end

% sort conjugate poles from low to high frequency
[tmpy,tmpi] = sort(abs(angle(z_conj)));
z_conj = z_conj(tmpi);

% sort real poles from most negative to most positive
[tmpy,tmpi] = sort(z_real);
z_real = z_real(tmpi);

% find impulse response corresponding to each combination
% of individual poles and pairs of conjugate poles
num_combinations = 2^(length(z_conj)/2 + length(z_real));
tmp1 = dec2bin([0:num_combinations-1]);
phase_matrix = bin2dec(tmp1(:,1));
for index1 = 2:(length(z_conj)/2 + length(z_real))
    phase_matrix = [phase_matrix,bin2dec(tmp1(:,index1))];
end
phase_matrix_conj = phase_matrix(:,1:(length(z_conj)/2));
phase_matrix_real = phase_matrix(:,(length(z_conj)/2 +
1):size(phase_matrix,2));


VBest = 0;
numBest = minPhaseTF;
for index1 = 1:num_combinations % iterate through all combinations of
zeros inside and outside unit circle
    tmp_z_conj = z_conj;
    tmp_z_real = z_real;
    tmp_k = k;
    for index2 = 1:size(phase_matrix_conj,2)
        if phase_matrix_conj(index1,index2)
            tmp_k = tmp_k*real((tmp_z_conj(2*(index2-1)+1))*
(tmp_z_conj(2*(index2-1)+2)));
            tmp_z_conj(2*(index2-1) + 1) =
1/tmp_z_conj(2*(index2-1) + 1);
            tmp_z_conj(2*(index2-1) + 2) =
1/tmp_z_conj(2*(index2-1) + 2);

        end
    end
    for index2 = 1:size(phase_matrix_real,2)
        if phase_matrix_real(index1,index2)
            tmp_k = tmp_k*(-1*tmp_z_real(index2));
            tmp_z_real(index2) = 1/tmp_z_real(index2);
        end
    end
    tmp_z = [tmp_z_conj;tmp_z_real];
    [num, den] = zp2tf(tmp_z,[],tmp_k);

    % calculate inverse-filtered waveform
    y = conv(num,inputWaveform);
    % calculate varimax norm of the inverse-filtered waveform
    V = sum(y.^4)/((sum(y.^2)^2)+epsilon);


    if (V>VBest)
        VBest = V;
        numBest = num;
    end
end

% make sure that filter has correct number of coefficients
if length(numBest)<(order+1)
    numBest = [numBest,zeros(1,order+1-length(numBest))];
end
```

Appendix C

Data from Speaker-Recognition Experiments

This appendix chapter presents the data from the primary speaker-recognition experiments. Only the raw results are presented, with interpretation provided in Chapter 7.

C.1 Speaker-Verification Experiment

Table C-1. Male speaker-verification EER results for linear prediction at dead-time 2.5 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	25.3	16.56	14.88	15.58	15.18	16.24
2 Amp	16.04	15.18	15.84	15.44	16.1	15.66
3 Amp	14.88	13.19	13.96	13.8	14.28	14.13
4 Amp	14.11	13.03	13.14	13.63	13.96	13.84
5 Amp	13.65	12.42	13.38	12.87	13.46	13.18
6 Amp	13.34	13.46	13.16	13.17	13.65	14.06
Thresh 200						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	28.37	16.26	16.57	17.64	17.48	17.48
2 Amp	17.48	16.56	17.35	17.39	18.08	18.25
3 Amp	16.23	14.72	14.58	15.47	16.26	17.18
4 Amp	16.26	14.81	15.39	14.72	15.8	16.26
5 Amp	15.7	15.15	15.35	14.59	15.18	15.03
6 Amp	16.45	15.55	15.91	15.47	15.34	15.98
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	33.9	21	19.63	19.68	20.89	22.07
2 Amp	20.08	18.16	19.79	19.88	20.09	21.93
3 Amp	18.87	18.56	18.87	18.25	19.38	19.33
4 Amp	19.08	18.87	18.6	19.17	18.38	20.33
5 Amp	19.05	18.2	19.33	17.66	19.48	20.16
6 Amp	19.33	19.02	18.87	18.4	18.72	20.1
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	39.65	28.37	26.69	25.46	26.29	26.99
2 Amp	26.05	23.31	24.54	23.42	23.97	26.07
3 Amp	23.53	22.25	25	24.69	24.56	26.84
4 Amp	25	22.49	25	24.85	24.38	26.54
5 Amp	25	24.03	25	23.77	25.31	25.68
6 Amp	26.8	26.78	26.38	24.1	25.15	26.52

Table C-2. Male speaker-verification EER results for linear prediction at dead-time 1 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	26.22	21.51	17.35	15.49	14.72	13.31
2 Amp	16.26	13.54	12.84	12.42	11.5	11.35
3 Amp	14.12	11.36	10.58	10.61	10.46	10.17
4 Amp	12.12	10.78	10.43	10.43	9.97	9.88
5 Amp	12.04	10.46	10	10.04	9.99	9.78
6 Amp	11.2	10.5	9.82	10.06	9.83	9.51
Thresh 200						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.14	14.47	12.12	12.58	13.04	12.88
2 Amp	15.64	12.72	11.81	11.78	11.47	11.79
3 Amp	15.11	12.24	11.64	11.2	11.01	11.04
4 Amp	14.35	12.12	11.5	11.35	11.26	11.12
5 Amp	14.19	12.27	11.78	10.89	11.31	11.35
6 Amp	13.83	12.66	11.77	11.35	10.69	11.28
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	28.22	14.89	14.42	14.42	13.96	14.96
2 Amp	16.72	13.96	14.11	14.13	14.11	14.07
3 Amp	15.9	12.88	14.02	13.93	13.34	13.5
4 Amp	15.96	14.11	14.42	13.83	13.38	14.36
5 Amp	16.2	14.36	14.17	13.65	13.65	14.16
6 Amp	16.53	14.98	14.26	14	13.95	14.11
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.13	21.78	20.71	21.63	22.08	23.31
2 Amp	22.18	19.49	19.47	20.5	21.31	22.09
3 Amp	21.85	19.35	19.99	21.44	21.34	22.01
4 Amp	21.73	20.86	20.03	20.09	21.63	21.67
5 Amp	23.9	21.17	20.09	20.67	20.45	22.55
6 Amp	25.46	21.78	21.43	21.32	21.17	23.42

Table C-3. Female speaker-verification EER results for linear prediction at dead-time 2.5 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	33.46	30.15	26.47	24.63	24.88	26.1
2 Amp	29.56	23.9	24.15	23.53	24.63	23.05
3 Amp	24.63	20.75	21.7	22.79	22.43	22.06
4 Amp	25.89	21.91	21.32	20.38	21.76	24.18
5 Amp	25	21.69	21.32	19.49	20.59	22.92
6 Amp	27.94	25	22.19	21.47	21.69	22.71
Thresh 200						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.41	29.42	24.52	25.32	24.16	26.42
2 Amp	27.42	23.84	22.07	23.98	26.47	25.37
3 Amp	28.31	23.55	24.6	25.33	23.16	25.74
4 Amp	30.88	25.4	25.37	21.69	26.1	24.8
5 Amp	30.25	27.91	26.02	22.86	23.9	25.74
6 Amp	31.99	27.94	26.84	24.71	24.33	24.95
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	35.43	26.78	27.57	27.21	28.21	27.79
2 Amp	33.01	28.57	25.37	24.33	26.26	27.21
3 Amp	31.99	25.19	24.43	26.1	25.75	26.8
4 Amp	32.66	30.88	26.86	28.31	26.02	27.94
5 Amp	34.56	31.78	27.57	24.67	26.47	28.44
6 Amp	33.82	29.58	27.25	27.57	27.21	29.53
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	43.38	39.71	34.93	33.82	35.29	33.09
2 Amp	37.83	33.2	36.61	35.74	34.19	34.99
3 Amp	39.17	34.19	35.29	33.82	34.19	32.35
4 Amp	39.34	36.4	35.66	34.56	31.99	33.59
5 Amp	40.44	33.85	35.96	31.62	31.62	33.43
6 Amp	41.18	37.5	35.56	33.82	34.68	33.46

Table C-4. Female speaker-verification EER results for linear prediction at dead-time 1 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	30.51	24.1	16.92	15.81	17.17	18.75
2 Amp	20.6	16.91	16.62	16.54	15.48	15.44
3 Amp	16.91	15.72	15.56	16.05	15.5	15.14
4 Amp	18.01	16.01	15.44	16.18	15.07	15.51
5 Amp	18.01	15.89	15.92	15.44	15.58	14.63
6 Amp	17.28	16.89	15.73	15.75	15.44	15.38
Thresh 200						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	32.42	18.27	15.96	16.18	16.31	16.91
2 Amp	18.75	16.18	16.18	16.54	15.81	16.87
3 Amp	18.89	15.7	16.35	15.72	15.47	16.18
4 Amp	17.65	15.44	16.97	16.54	15.94	15.24
5 Amp	16.54	16.54	17.28	16.18	16.08	15.81
6 Amp	17.09	16.91	17.17	15.81	15.48	15.62
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	36.76	22.97	21.82	20.34	19.49	20.96
2 Amp	21.97	20.08	23.87	18.79	19.49	19.49
3 Amp	21.69	19.3	23.53	19.12	18.9	17.8
4 Amp	21.91	19.22	21.46	19.34	19.85	18.28
5 Amp	21.4	20.22	20.83	18.65	19.12	19.12
6 Amp	22.45	20.96	22.06	19.49	19.25	17.3
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	36.76	35.74	29.32	28.54	29.41	27.48
2 Amp	27.97	31.56	29.04	26.1	26.1	27.57
3 Amp	26.62	26.84	25.39	25.92	26.47	25.37
4 Amp	26.84	26.1	26.1	24.4	25.57	25.76
5 Amp	27.57	26.25	26.1	26.1	26.1	27.31
6 Amp	30.51	26.16	26.04	24.63	26.18	25.02

Table C-5. Male speaker-verification EER results for MED at dead-time 2.5 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	25.25	16.87	15.95	16.87	16.41	16.94
2 Amp	17.33	16.23	16.68	15.64	16.26	15.95
3 Amp	15.38	14.11	14.88	14.26	14.57	15.03
4 Amp	14.57	13.35	14.36	14.72	14.12	14.72
5 Amp	13.9	13.34	14.11	13.7	14.93	14.26
6 Amp	14.11	13.5	14.7	13.65	14.84	14.51
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	26.4	16.41	16.1	17.34	17.21	17.98
2 Amp	17.41	16.86	17.76	17.48	17.79	18.01
3 Amp	16.45	14.83	15.64	15.64	16.19	16.14
4 Amp	15.05	14.11	15.34	15.49	15.84	16.24
5 Amp	14.72	14.11	14.84	15.06	15.72	16.03
6 Amp	14.78	14.43	14.93	14.9	15.34	15.68
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	30.34	18.71	18.84	18.25	19.25	19.48
2 Amp	18.87	17.64	19.03	19.42	19.08	18.84
3 Amp	17.48	15.14	16.92	17.94	17.33	18.39
4 Amp	16.65	16.19	16.72	16.98	17.85	18.22
5 Amp	16.72	15.87	17.65	16.36	17.94	18.58
6 Amp	16.87	16.91	17.46	16.62	17.79	17.33
Thresh 3000						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	38.34	33.55	31.02	31.06	31.13	32.21
2 Amp	30.53	28.28	28.14	28.07	30.67	30.18
3 Amp	28.37	27.29	27.66	28.62	30.54	30.33
4 Amp	29.45	25.77	27.84	28.07	30.52	29.45
5 Amp	29.91	26.97	27.91	27.76	28.23	30.09
6 Amp	30.67	29.68	26.71	28.68	27.76	30.21

Table C-6. Male speaker-verification EER results for MED at dead-time 1 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	32.7	24.72	19.48	19.01	19.63	18.81
2 Amp	16.26	15.36	15.18	14.26	14.06	13.8
3 Amp	15.18	13.61	13.22	12.01	12.75	12.66
4 Amp	14.77	13.34	12.27	12.44	12.24	12.27
5 Amp	15.06	12.78	12.34	11.96	11.96	12.12
6 Amp	13.8	13.14	12.42	11.91	11.5	11.37
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.02	16.82	15.54	15.49	15.54	15.34
2 Amp	16.87	14.24	13.96	13.7	13.8	13.97
3 Amp	16.42	14.44	13.34	13.14	13.28	13.11
4 Amp	16.89	14.49	13.82	13.64	12.76	13.04
5 Amp	16.67	14.56	13.9	13.34	12.29	12.78
6 Amp	16.2	14.56	13.96	13.19	12.89	12.83
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.67	16.36	16.47	16.1	16.26	17.33
2 Amp	17.02	15.03	14.94	14.39	14.57	16.1
3 Amp	16.94	14.72	14.19	14.26	13.96	14.84
4 Amp	16.87	14.57	13.94	13.96	14.11	14.26
5 Amp	16.91	16.07	15.03	14.21	14.72	14.84
6 Amp	17.64	17.06	15.35	15.21	15.03	15.52
Thresh 3000						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	40.15	31.07	28.68	29.58	29.75	30.14
2 Amp	28.99	26.16	27.45	26.53	28.07	29.6
3 Amp	27.41	26.35	25.25	27.45	27.45	29.3
4 Amp	27.24	24.54	26.53	26.53	26.99	28.53
5 Amp	28.25	27.45	25.8	26.23	27.15	28.68
6 Amp	30.37	27.76	27.65	26.78	27.08	28.68

Table C-7. Female speaker-verification EER results for MED at dead-time 2.5 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.27	29.54	26.99	26.84	26.84	28.25
2 Amp	24.59	25.45	25.82	26.17	27.94	28.31
3 Amp	25.37	23.16	24.63	25	25	25.74
4 Amp	26.47	24.26	24.07	23.9	25.37	26.84
5 Amp	26.65	26.1	26.25	26.1	23.61	25.37
6 Amp	27.15	24.86	27.21	25.4	25	25.4
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	30.63	26.1	26.47	25.83	26.1	28.68
2 Amp	26.63	25.25	25.38	25.74	25.74	26.23
3 Amp	25.37	24.56	23.9	25	26.26	26.06
4 Amp	26.55	24.31	24.26	25.37	27.41	29.59
5 Amp	27.37	26.58	26.5	26.43	26.42	25.41
6 Amp	29.41	28.16	27.94	25.37	26.67	25.95
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.41	26.75	27.34	26.47	28.38	29.41
2 Amp	30.51	27.21	27.44	27.57	28.31	27.17
3 Amp	28.31	25.74	25.97	25.74	27.21	25.66
4 Amp	28.07	25.95	26.47	25.37	27.57	28.95
5 Amp	30.15	26.39	25.74	25.52	25.47	26.01
6 Amp	30.77	28.38	27.99	29.02	24.63	26.95
Thresh 3000						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	40.07	35.6	34.34	32.11	34.19	34.17
2 Amp	38.6	34.93	32.28	35.49	34.93	37.5
3 Amp	36.2	32.62	32.43	32.35	34.89	33.46
4 Amp	36.4	32.94	31.85	31.25	33.46	33.46
5 Amp	34.56	33.09	31.62	30.57	32.72	31.25
6 Amp	36.4	34.19	32.81	32.64	29.41	31.06

Table C-8. Female speaker-verification EER results for MED at dead-time 1 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	30.81	25.87	23.53	24.6	24.84	24.69
2 Amp	21.69	20.66	19.5	20.15	19.49	20.68
3 Amp	21.45	18.38	18.06	17.65	17.78	17.65
4 Amp	19.49	18.32	18.01	18.35	17.65	16.54
5 Amp	20.59	19.12	18.44	17.88	16.54	18
6 Amp	19.49	19.12	17.65	17.15	15.81	16.54
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	35.99	23.68	21.69	20.89	20.22	21.32
2 Amp	23.9	21.02	19.67	18.38	18.4	19.74
3 Amp	20.96	20.58	18.38	18.38	18.75	19.67
4 Amp	20.68	19.43	19.89	20.22	19.27	18.86
5 Amp	20.59	20.45	21.02	19.71	19.68	19.66
6 Amp	21.32	20.79	21.68	20.41	19.85	19.75
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.5	23.53	21.69	23.31	22.93	25.37
2 Amp	22.94	21.69	22.79	22.16	21.68	22.27
3 Amp	22.38	20.96	21.32	21.32	21.36	21.5
4 Amp	20.99	22.43	22.06	21.32	21.32	20.96
5 Amp	22.79	23.03	23.16	22.43	21.69	21.88
6 Amp	24.11	24.68	23.59	23.36	23.53	22.56
Thresh 3000						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	41.98	35.06	29.78	30.88	33.46	31.32
2 Amp	36.03	33.09	32.33	33.1	33.46	30.79
3 Amp	34.77	31.03	30.26	30.88	31.99	30.53
4 Amp	33.45	31.25	30.15	30.14	29.58	31.66
5 Amp	33.82	30.88	29.68	29.72	30.51	31.58
6 Amp	34.56	33.35	32.53	30.15	30.15	32.27

Table C-9. Male speaker-verification EER results for hybrid at dead-time 2.5 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	27.71	16.54	16.26	16.47	17.33	17.18
2 Amp	17.22	17.06	17.42	16.87	17.02	16.83
3 Amp	16.26	14.67	15.58	14.88	15.18	15.71
4 Amp	15.29	14.12	14.49	15.03	15.64	15.34
5 Amp	14.72	13.59	14.11	13.8	13.97	14.88
6 Amp	15.03	14.26	14.7	14.47	15.02	14.97
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	28.09	17.79	17.17	17.33	16.72	17.18
2 Amp	19.07	16.89	18.25	17.31	17.63	16.87
3 Amp	17.18	15.31	15.49	15.23	15.64	16.35
4 Amp	16.15	15.05	15.49	15.54	16.31	16.72
5 Amp	15.95	15.29	15.49	15.06	16.03	15.64
6 Amp	16.16	15.49	16.26	15.49	15.15	14.94
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.96	18.3	18.87	19.57	19.39	20.6
2 Amp	19.66	17.85	19.33	19.63	19.94	20.4
3 Amp	19.11	16.41	17.79	17.33	18.05	18.79
4 Amp	17.54	16.87	17.64	17.69	18.62	18.29
5 Amp	17.94	17.14	18.2	17.67	17.94	17.48
6 Amp	17.48	17.57	18.1	17.67	17.33	17.71
Thresh 1400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	36.48	21.32	21.32	21.57	22.92	24.08
2 Amp	21.25	19.48	20.97	20.88	22.62	23.16
3 Amp	20.35	19.26	19.48	19.63	20.9	22.55
4 Amp	19.79	19.4	19.63	19.96	21.01	21.63
5 Amp	21.17	19.52	20.28	19.44	21.17	21.01
6 Amp	21.26	21.17	20.55	20.49	21.77	21.89

Table C-10. Male speaker-verification EER results for hybrid at dead-time 1 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.34	22.39	18.4	17.94	17.02	16.72
2 Amp	17.03	15.18	14.89	14.18	13.51	13.65
3 Amp	15.49	13.04	12.4	11.89	11.66	11.75
4 Amp	14.51	12.51	11.5	11.5	11.2	10.89
5 Amp	13.84	11.97	11.35	10.79	10.7	10.58
6 Amp	13.59	12.09	11.26	10.78	10.28	10.24
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	30.52	17.23	15.64	15.03	15.18	14.82
2 Amp	17.48	14.26	14.07	12.73	12.42	12.88
3 Amp	17.33	13.8	12.76	12.16	11.84	12.12
4 Amp	16.43	13.36	12.27	11.81	11.94	11.66
5 Amp	15.51	13.04	12.27	11.54	11.47	11.25
6 Amp	15.18	13.32	12.42	11.93	11.66	11.04
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.6	16.39	15.49	14.98	16.1	16.21
2 Amp	17.64	14.57	14.51	14.42	14.72	14.57
3 Amp	18.27	15	14.57	14.34	14.26	14.42
4 Amp	18.71	15.32	13.88	14.42	13.71	13.63
5 Amp	17.75	15.64	14.1	13.45	13.5	13.67
6 Amp	17.64	15.87	15.03	13.73	13.85	14.44
Thresh 1400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.14	18.32	17.02	18.1	19.45	19.79
2 Amp	19.66	16.23	16.01	16.48	17.32	17.79
3 Amp	19.08	16.08	15.69	16.3	16.41	17.75
4 Amp	18.4	16.1	15.21	15.42	16.37	17.48
5 Amp	19.09	17.03	16.77	16.56	16.41	16.66
6 Amp	20.25	19.17	17.79	17.29	17.22	16.88

Table C-11. Female speaker-verification EER results for hybrid at dead-time 2.5 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	34.56	29.41	28.32	26.88	28.31	27.24
2 Amp	26.22	27.44	27.6	27.3	27.89	26.47
3 Amp	25.74	23.12	25.37	26.61	26.18	24.26
4 Amp	27.21	24.26	25.19	25.74	25	26.88
5 Amp	27.21	26.33	26.02	24.93	25.74	26.33
6 Amp	26.84	25.85	25.01	24.7	25	24.46
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	31.99	30.15	26.5	28.5	29.78	29.18
2 Amp	28.31	26.84	27.7	27.58	27.21	28.31
3 Amp	26.84	25.37	27.35	25.47	27.21	27.54
4 Amp	29.23	26.47	27.57	26.84	27.21	27.57
5 Amp	30.15	27.94	27.21	27.21	26.84	27.57
6 Amp	30.88	29.45	29.2	26.62	27.88	28.27
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	29.04	28.26	29.78	29.78	30.88	30.51
2 Amp	29.78	30.88	29.78	31.25	28.31	29.72
3 Amp	28.4	27.49	27.94	28.31	28.7	29.84
4 Amp	30.15	27.21	27.94	27.57	28.54	28.75
5 Amp	30.97	27.58	28.08	27.4	27.21	29.04
6 Amp	31.47	29.55	29.21	26.79	27.74	26.84
Thresh 1400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	38.97	32.65	33.77	31.99	32.04	30.88
2 Amp	34.93	33.82	31.62	31.62	30.88	31.47
3 Amp	35.66	31.69	30.59	30.15	30.19	31.59
4 Amp	34.32	31.14	29.78	30.15	31.29	30.39
5 Amp	35.29	30.51	30.72	27.94	29.78	30.51
6 Amp	33	32.24	30.94	29.04	30.88	29.41

Table C-12. Female speaker-verification EER results for hybrid at dead-time 1 ms and four thresholds

Thresh 0						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	36.03	27.21	19.49	20.96	20.96	20.96
2 Amp	23.27	19.21	17.87	18.64	18.38	18.01
3 Amp	20.96	18.5	18.1	16.85	17.28	17.65
4 Amp	20.96	18.58	17.66	17.28	16.91	16.91
5 Amp	20.96	18.5	17.65	16.91	16.54	15.68
6 Amp	19.85	18.03	17.26	16.54	15.96	15.2
Thresh 400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	34.62	24.52	18.53	19.49	18.38	19.78
2 Amp	22.54	19.12	18.3	17.71	17.28	17.74
3 Amp	19.49	17.28	17.94	16.95	16.1	16.22
4 Amp	19.16	18.14	17.65	16.92	16.95	17.28
5 Amp	19.98	19.06	18.55	17.19	16.93	16.54
6 Amp	19.65	19.19	17.7	17.39	17.28	16.47
Thresh 800						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	34.56	21.96	21.73	20.96	22.84	21.69
2 Amp	24.26	19.85	20.22	20.22	20	20.68
3 Amp	21.96	19.49	21.31	20.19	19.03	18.65
4 Amp	21.85	20.01	21.69	20.76	19.43	20.03
5 Amp	21.15	20.59	20.96	19.96	19.09	18.97
6 Amp	22.32	21.56	21.69	20.23	19.77	19.77
Thresh 1400						
	1 Timing	2 Timing	3 Timing	4 Timing	5 Timing	6 Timing
1 Amp	34.19	26.1	25.44	24.26	24.63	25.3
2 Amp	25	22.79	23.04	21.69	22.79	23.95
3 Amp	24.94	23.16	22.43	22.22	23.46	22.79
4 Amp	25.5	23.83	22.65	22.78	23.65	23.74
5 Amp	27.21	25.23	23.61	24.28	22.06	21.32
6 Amp	26.1	25.04	22.06	24.63	22.25	22.02

C.2 Mel-Cepstrum on Extracted Event Impulse Trains

Table C-13. Male speaker-verification EER results using mel-cepstral features on the impulse train extracted using MED

Threshold/Dead time	1 ms	1.5 ms	2 ms	2.5 ms	3 ms
0	18.31	18.71	20.4	22.7	22.7
200	20.64	18.93	21.17	23.21	24.23
400	18.87	19.94	20.43	22.57	21.97
800	22.7	20.21	21.01	24.54	22.39
1600	25.22	25.61	24.39	26.07	26.53
2400	29.86	26.07	26.53	29.16	29.75
3000	29.36	32.1	27.29	28.78	29.65

Table C-14. Female speaker-verification EER results using mel-cepstral features on the impulse train extracted using MED

Threshold/Dead time	1 ms	1.5 ms	2 ms	2.5 ms	3 ms
0	24.26	27.25	30.15	29.78	36.05
200	26.84	24.08	26.47	28.23	37.5
400	26.19	31.62	26.1	31.22	34.37
800	36.4	28.08	34.93	31.89	34.56
1600	36.03	38.97	33.86	38.24	33
2400	39.71	34.82	32.48	35.66	36.76
3000	37.53	41.18	35.49	34.56	33.83

Table C-15. Male speaker-verification EER results using mel-cepstral features on the impulse train extracted using linear prediction

Threshold/Dead time	1 ms	1.5 ms	2 ms	2.5 ms	3 ms
0	17.18	17.74	19.17	21.24	24.39
200	18.23	18.1	18.33	23.6	22.55
400	24.58	22.85	21.47	24.69	25.61
600	27.76	25.15	23.62	26.17	28.23
800	26.23	30.98	25.15	27.15	29.61

Table C-16. Female speaker-verification EER results using mel-cepstral features on the impulse train extracted using linear prediction

Threshold/Dead time	1 ms	1.5 ms	2 ms	2.5 ms	3 ms
0	18.7	24.63	29.3	33.09	33.82
200	20.72	28.31	29.8	33.23	36.84
400	32.35	32.19	28.62	34.56	38.65
600	36.76	35.64	28.49	34.93	34.93
800	35.11	38.89	31.25	34.93	30.59

Table C-17. Male speaker-verification EER results using mel-cepstral features on the impulse train extracted using hybrid

Threshold/Dead time	1 ms	1.5 ms	2 ms	2.5 ms	3 ms
0	19.3	19.02	20.2	23.76	24.08
200	20.25	20.55	19.83	23.77	24.7
400	20.4	19.48	21.01	21.93	23.94
600	21.17	20.82	20.37	21.17	22.81
800	22.12	20.1	21.36	23.31	22.39
1200	28.28	24.52	22.55	23.54	25.79
1400	26.77	26.38	22.59	24.39	25

Table C-18. Female speaker-verification EER results using mel-cepstral features on the impulse train extracted using hybrid

Threshold/Dead time	1 ms	1.5 ms	2 ms	2.5 ms	3 ms
0	23.16	27.34	29.41	31.91	33.79
200	25.72	26.47	28.68	29.49	32.97
400	25.07	27.31	28.76	31.99	35.11
600	25	29.89	28.68	32.72	34.56
800	28.75	30.51	29.34	32.81	34.59
1200	36.5	33.19	30.38	35.65	32.72
1400	31.69	33.82	32.98	34.81	36.94

Table C-19. Confusion matrix for the within-session speaker-identification experiment

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	19	0	0	1	0	0	95
Male 2	2	18	11	0	0	0	58
Male 3	0	0	30	0	0	0	100
Male 4	0	0	0	20	0	0	100
Female 1	0	0	0	0	23	1	96
Female 2	0	0	0	0	0	20	100

MED with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 91 percent

Table C-20. Confusion matrix for the between-session speaker-identification experiment

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	109	0	0	0	0	0	100
Male 2	5	104	33	1	0	0	73
Male 3	0	0	60	0	0	0	100
Male 4	2	0	0	188	0	0	99
Female 1	0	21	12	0	157	32	71
Female 2	0	1	1	0	0	80	98

MED with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 90 percent.

C.3 Mismatched Vowel Speaker Verification

Table C-21. Equal-error rates for male TIMIT speakers in the mismatched-vowel experiment

	Matched Condition	Mismatched Condition
Mel-Cepstrum	12.11%	21.36%
MED	21.54%	22.60%
Mel-Cepstrum/MED Fusion	10.19%	16.65%
Linear Prediction	20.19%	21.97%
Mel-Cepstrum/LP Fusion	10.25%	16.56%

Table C-22. Equal-error rates for female TIMIT speakers in the mismatched-vowel experiment

	Matched Condition	Mismatched Condition
Mel-Cepstrum	17.91%	25.41%
MED	32.46%	34.95%
Mel-Cepstrum/MED Fusion	10.82%	22.76%
Linear Prediction	32.46%	35.82%
Mel-Cepstrum/LP Fusion	10.82%	22.56%

C.4 Intersession Speaker Identification Experiments

Table C-23. Confusion matrix for the within-session speaker-identification experiment using linear prediction and the fixed target model configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	15	0	3	2	0	0	75
Male 2	1	20	4	0	0	6	65
Male 3	0	1	29	0	0	0	97
Male 4	1	0	0	19	0	0	95
Female 1	0	0	1	0	21	2	88
Female 2	0	0	0	0	4	16	80

Linear prediction with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 83 percent.

Table C-24. Confusion matrix for the between-session speaker-identification experiment using linear prediction and the fixed target model configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	72	17	5	15	0	0	66
Male 2	5	71	38	1	5	23	50
Male 3	4	3	35	2	0	16	58
Male 4	14	0	0	176	0	0	93
Female 1	0	7	6	0	144	65	65
Female 2	0	1	1	0	29	51	62

Linear prediction with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 66 percent.

Table C-25. Confusion matrix for the within-session speaker-identification experiment using linear prediction and the fixed test set configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	15	0	3	2	0	0	75
Male 2	1	20	4	0	0	6	65
Male 3	0	1	29	0	0	0	97
Male 4	1	0	0	19	0	0	95
Female 1	0	0	1	0	21	2	88
Female 2	0	0	0	0	4	16	80

Linear prediction with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 83 percent.

Table C-26. Confusion matrix for the between-session speaker-identification experiment using linear prediction and the fixed test set configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	16	0	3	1	0	0	80
Male 2	1	16	14	0	0	0	52
Male 3	0	2	28	0	0	0	93
Male 4	0	0	0	20	0	0	100
Female 1	0	0	1	0	20	3	83
Female 2	0	0	3	0	4	13	65

Linear prediction with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 79 percent.

Table C-27. Confusion matrix for the within-session speaker-identification experiment using MED and the fixed target model configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	19	0	0	1	0	0	95
Male 2	2	18	11	0	0	0	58
Male 3	0	0	30	0	0	0	100
Male 4	0	0	0	20	0	0	100
Female 1	0	0	0	0	23	1	96
Female 2	0	0	0	0	0	20	100

MED with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 91 percent.

Table C-28. Confusion matrix for the between-session speaker-identification experiment using MED and the fixed target model configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	109	0	0	0	0	0	100
Male 2	5	104	33	1	0	0	73
Male 3	0	0	60	0	0	0	100
Male 4	2	0	0	188	0	0	99
Female 1	0	21	12	0	157	32	70
Female 2	0	1	1	0	0	80	98

MED with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 90 percent.

Table C-29. Confusion matrix for the within-session speaker-identification experiment using MED and the fixed test set configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	19	0	0	1	0	0	95
Male 2	2	18	11	0	0	0	58
Male 3	0	0	30	0	0	0	100
Male 4	0	0	0	20	0	0	100
Female 1	0	0	0	0	23	1	96
Female 2	0	0	0	0	0	20	100

MED with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 91 percent.

Table C-30. Confusion matrix for the between-session speaker-identification experiment using MED and the fixed test set configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	20	0	0	0	0	0	100
Male 2	2	17	11	0	1	0	55
Male 3	0	7	22	0	1	0	73
Male 4	5	0	0	15	0	0	75
Female 1	0	0	0	0	23	1	96
Female 2	0	0	1	0	1	18	90

MED with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 82 percent.

Table C-31. Confusion matrix for the within-session speaker-identification experiment using the hybrid method and the fixed target model configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	20	0	0	0	0	0	100
Male 2	1	28	2	0	0	0	90
Male 3	0	1	29	0	0	0	97
Male 4	0	0	0	20	0	0	100
Female 1	0	0	0	0	20	4	83
Female 2	0	0	0	0	0	20	100

The hybrid method with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 95 percent.

Table C-32. Confusion matrix for the between-session speaker-identification experiment using the hybrid method and the fixed target model configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	108	0	0	1	0	0	99
Male 2	12	124	7	0	0	0	87
Male 3	0	0	60	0	0	0	100
Male 4	2	0	0	188	0	0	99
Female 1	0	22	16	0	147	37	66
Female 2	0	3	1	0	0	78	95

The hybrid method with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 91 percent.

Table C-33. Confusion matrix for the within-session speaker-identification experiment using the hybrid method and the fixed test set configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	20	0	0	0	0	0	100
Male 2	1	28	2	0	0	0	90
Male 3	0	1	29	0	0	0	97
Male 4	0	0	0	20	0	0	100
Female 1	0	0	0	0	20	4	83
Female 2	0	0	0	0	0	20	100

The hybrid method with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 95 percent.

Table C-34. Confusion matrix for the between-session speaker-identification experiment using the hybrid method and the fixed test set configuration

Presented/Detected	Male 1	Male 2	Male 3	Male 4	Female 1	Female 2	Row Percent Correct
Male 1	20	0	0	0	0	0	100
Male 2	4	11	15	0	1	0	35
Male 3	1	8	21	0	0	0	70
Male 4	0	0	0	20	0	0	100
Female 1	0	0	0	0	22	2	92
Female 2	0	0	0	0	0	20	100

The hybrid method with three timing and three amplitude features, threshold 200 and dead time 2.5 ms were used. Each entry is the number of utterances for the presented talker (row) that were identified as produced by the detected talker (column). The "row percent correct" is the percentage of a presented subject's utterances that were correctly detected as produced by that talker. The average of the rows' correct percentages is 83 percent.

Appendix D

Synthetic Event Pattern Specifications

This appendix presents the specifications for the synthetic event patterns used in Chapter 5. Each signal described includes 5 seconds of one particular pattern in the form of a train of impulses with equal amplitudes having a sampling rate of 8 kHz.

1-long timing patterns were generated using a 5-ms-long signal using an 8-ms period and including ±0.125 ms (±1 sample) of random timing perturbation. Irregular signals were generated by randomly perturbing the times of a periodic impulse train having period 8 ms from ±1 sample to ±28 samples with a discrete uniform distribution. Each of the 28 perturbation amounts was included in a separate 5-second-long signal.

The specifications for each of the 2-long deterministic timing patterns are shown in Table D-1 with the times of each interval between two events rounded to the nearest 0.1 ms. The two intervals used in a particular 2-long pattern are described with the set [T_1, T_2]. Likewise, the specifications for each of the 3-long deterministic timing patterns are shown in Table D-2 with the times of each interval between two events rounded to the nearest 0.1 ms. The three intervals used in a particular 3-long pattern are described with the set [T_1, T_2, T_3]. Each pattern described was included in a separate 5-second-long signal.

Table D-1. 2-long timing patterns used to study the feature space

[	0.5 ms	15.5 ms	]	[	4.8 ms	11.3 ms	]	[	9.0 ms	7.0 ms	]	[	13.3 ms	2.8 ms	]
[	0.6	15.4	]	[	4.9	11.1	]	[	9.1	6.9	]	[	13.4	2.6	]
[	0.8	15.3	]	[	5.0	11.0	]	[	9.3	6.8	]	[	13.5	2.5	]
[	0.9	15.1	]	[	5.1	10.9	]	[	9.4	6.6	]	[	13.6	2.4	]
[	1.0	15.0	]	[	5.3	10.8	]	[	9.5	6.5	]	[	13.8	2.3	]
[	1.1	14.9	]	[	5.4	10.6	]	[	9.6	6.4	]	[	13.9	2.1	]
[	1.3	14.8	]	[	5.5	10.5	]	[	9.8	6.3	]	[	14.0	2.0	]
[	1.4	14.6	]	[	5.6	10.4	]	[	9.9	6.1	]	[	14.1	1.9	]
[	1.5	14.5	]	[	5.8	10.3	]	[	10.0	6.0	]	[	14.3	1.8	]
[	1.6	14.4	]	[	5.9	10.1	]	[	10.1	5.9	]	[	14.4	1.6	]
[	1.8	14.3	]	[	6.0	10.0	]	[	10.3	5.8	]	[	14.5	1.5	]
[	1.9	14.1	]	[	6.1	9.9	]	[	10.4	5.6	]	[	14.6	1.4	]
[	2.0	14.0	]	[	6.3	9.8	]	[	10.5	5.5	]	[	14.8	1.3	]
[	2.1	13.9	]	[	6.4	9.6	]	[	10.6	5.4	]	[	14.9	1.1	]
[	2.3	13.8	]	[	6.5	9.5	]	[	10.8	5.3	]	[	15.0	1.0	]
[	2.4	13.6	]	[	6.6	9.4	]	[	10.9	5.1	]	[	15.1	0.9	]
[	2.5	13.5	]	[	6.8	9.3	]	[	11.0	5.0	]	[	15.3	0.8	]
[	2.6	13.4	]	[	6.9	9.1	]	[	11.1	4.9	]	[	15.4	0.6	]
[	2.8	13.3	]	[	7.0	9.0	]	[	11.3	4.8	]	[	15.5	0.5	]
[	2.9	13.1	]	[	7.1	8.9	]	[	11.4	4.6	]	[			]
[	3.0	13.0	]	[	7.3	8.8	]	[	11.5	4.5	]	[			]
[	3.1	12.9	]	[	7.4	8.6	]	[	11.6	4.4	]	[			]
[	3.3	12.8	]	[	7.5	8.5	]	[	11.8	4.3	]	[			]
[	3.4	12.6	]	[	7.6	8.4	]	[	11.9	4.1	]	[			]
[	3.5	12.5	]	[	7.8	8.3	]	[	12.0	4.0	]	[			]
[	3.6	12.4	]	[	7.9	8.1	]	[	12.1	3.9	]	[			]
[	3.8	12.3	]	[	8.0	8.0	]	[	12.3	3.8	]	[			]
[	3.9	12.1	]	[	8.1	7.9	]	[	12.4	3.6	]	[			]
[	4.0	12.0	]	[	8.3	7.8	]	[	12.5	3.5	]	[			]
[	4.1	11.9	]	[	8.4	7.6	]	[	12.6	3.4	]	[			]
[	4.3	11.8	]	[	8.5	7.5	]	[	12.8	3.3	]	[			]
[	4.4	11.6	]	[	8.6	7.4	]	[	12.9	3.1	]	[			]
[	4.5	11.5	]	[	8.8	7.3	]	[	13.0	3.0	]	[			]
[	4.6	11.4	]	[	8.9	7.1	]	[	13.1	2.9	]	[			]

Each set of two values refers to the first and second intervals (in milliseconds) between glottal events in the repeating pattern. A different signal was created for each pattern, with five continuous seconds of impulses exhibiting the particular pattern. ±0.125 ms of random timing perturbation was applied.

Table D-2. 3-long timing patterns used to study the feature space

[	2.0 ms	11.5 ms	3.5 ms	]	[	5.5 ms	13.3 ms	5.3 ms	]	[	10.3 ms	2.9 ms	12.9 ms	]
[	2.0	12.5	4.1	]	[	5.5	13.6	4.8	]	[	10.3	5.8	7.3	]
[	2.3	4.4	4.3	]	[	5.6	2.5	11.1	]	[	10.5	7.4	4.1	]
[	2.3	10.0	13.5	]	[	5.6	9.4	12.9	]	[	10.6	9.1	4.5	]
[	2.3	12.0	2.5	]	[	5.6	11.1	7.8	]	[	10.6	9.9	3.9	]
[	2.4	6.5	2.3	]	[	5.8	8.6	5.9	]	[	10.8	12.8	8.1	]
[	2.5	8.3	9.3	]	[	5.9	4.9	5.5	]	[	10.9	8.0	12.6	]
[	2.6	4.1	8.9	]	[	5.9	8.6	11.1	]	[	11.1	8.1	8.6	]
[	2.6	7.3	6.0	]	[	6.0	5.9	9.9	]	[	11.1	13.8	10.3	]
[	2.6	8.8	12.9	]	[	6.1	13.8	11.4	]	[	11.3	12.0	5.9	]
[	2.9	6.1	5.3	]	[	6.3	2.6	4.3	]	[	11.5	8.6	4.1	]
[	3.0	2.6	12.6	]	[	6.4	3.6	1.3	]	[	11.8	8.5	6.1	]
[	3.1	11.1	12.5	]	[	6.4	3.9	7.1	]	[	11.9	4.6	11.8	]
[	3.5	2.6	12.1	]	[	6.6	11.1	7.8	]	[	12.0	6.8	2.8	]
[	3.5	7.1	7.9	]	[	6.8	12.1	8.9	]	[	12.0	7.8	10.8	]
[	3.5	10.1	12.8	]	[	7.1	10.0	2.1	]	[	12.6	6.8	5.9	]
[	3.5	11.3	9.3	]	[	7.4	5.0	5.0	]	[	12.9	7.5	6.0	]
[	3.6	5.6	3.3	]	[	7.5	2.8	7.0	]	[	12.9	12.9	4.8	]
[	3.6	8.0	3.1	]	[	8.1	3.0	8.9	]	[	13.0	2.3	8.6	]
[	3.8	6.1	2.0	]	[	8.1	11.6	10.0	]	[	13.1	9.1	6.4	]
[	3.8	8.9	9.4	]	[	8.1	13.5	9.3	]	[	13.1	10.1	8.0	]
[	3.8	12.4	10.1	]	[	8.3	7.6	3.1	]	[	13.3	5.8	12.4	]
[	3.9	7.3	8.0	]	[	8.4	5.3	9.3	]	[	13.3	8.1	5.9	]
[	4.0	2.1	9.6	]	[	8.8	12.3	3.5	]	[	13.4	2.3	7.5	]
[	4.0	9.1	2.1	]	[	8.9	3.1	9.3	]	[	13.4	3.5	9.4	]
[	4.0	10.3	10.6	]	[	8.9	8.1	10.9	]	[	13.4	5.9	13.0	]
[	4.4	9.5	8.0	]	[	8.9	12.9	5.1	]	[	13.5	2.0	13.0	]
[	4.6	5.4	5.5	]	[	9.4	2.8	2.4	]	[	13.5	8.3	2.9	]
[	4.8	5.8	6.0	]	[	9.4	4.4	9.1	]	[	13.6	2.4	11.9	]
[	4.9	8.9	3.1	]	[	9.4	9.4	9.1	]	[	13.8	7.8	11.9	]
[	5.0	6.5	10.9	]	[	9.5	12.0	10.0	]	[	13.9	2.6	12.6	]
[	5.0	9.6	3.9	]	[	9.8	11.3	11.8	]	[	13.9	6.1	7.4	]
[	5.3	12.6	13.9	]	[	9.8	11.5	8.6	]	[				]
[	5.4	11.9	2.6	]	[	10.0	2.3	5.5	]	[				]

Each set of three values refers to the first, second, and third intervals (in milliseconds) between glottal events in the repeating pattern. A different signal was created for each pattern, with five continuous seconds of impulses exhibiting the particular pattern. ±0.125 ms of random timing perturbation was applied.

References

[1] "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus: NIST Speech Disc CD1-1.1," National Institute of Standards and Technology, 1990.

[2] "MATLAB," 6.5 ed: The MathWorks, Inc., 2002.

[3] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 23, pp. 562-570, 1975.

[4] T. 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*, vol. 27, pp. 309-319, 1979.

[5] T. R. Anderson, "Unpublished EGG Database," Air Force Research Laboratory, 2001.

[6] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," presented at ICASSP 1982, Paris, France, 1982.

[7] M. S. Bartlett, "The spectral analysis of point processes," *Journal of the Royal Statistical Society, Series B*, vol. 29, pp. 264-296, 1963.

[8] D. A. Berry, "Mechanisms of modal and nonmodal phonation," *Journal of Phonetics*, vol. 29, pp. 431-450, 2001.

[9] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech and Hearing Research*, vol. 39, pp. 126-134, 1996.

[10] A. C. Bovik, J. P. Havlicek, M. D. Desai, and D. S. Harding, "Limits on discrete modulated signals," *IEEE Transactions on Signal Processing*, vol. 45, pp. 867-879, 1997.

[11] D. M. Brookes, "VoiceBox: Speech Processing Toolbox for MATLAB," 1.3 ed, 2005.

[12] J. Chowning and D. Bristow, "FM Theory and Applications By Musicians for Musicians," 1986.

[13] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, pp. 357-366, 1980.

[14] D. Deliyski, "High-speed videoendoscopy: recent progress and clinical prospects," presented at Advances in Quantitative Laryngology, Voice and Speech Research, Groningen, the Netherlands, 2006.

[15] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, pp. 423-444, 1996.

[16] G. Fant, *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. s'Gravenhage: Mouton, 1960.

[17] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow.," presented at the French-Swedish Symposium, Grenoble, France, 1985.

[18] R. Fernandez, "A computational model fro the automatic recognition of affect in speech," PhD Thesis in Media Arts and Sciences, Cambridge, MA: MIT, 2004.

[19] A. Fourcin, J. McGlashan, and B. Blowes, "Measuring voice in the clinic - Laryngograph® speech studio analyses," presented at the 6th Voice Symposium of Australia, Adelaide, Australia, 2002.

[20] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, pp. 365-381, 2001.

[21] G. Gonzalez, R. E. Badra, R. Medina, and J. Regidor, "Period estimation using minimum entropy deconvolution (MED)," *Signal Processing*, vol. 41, pp. 91-100, 1995.

[22] A. Hagen, "The Linguistic Functions of Glottalizations and their Language Specific Use in English and German," MS Thesis: Erlangen University/MIT Speech Group, 1997.

[23] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *Journal of the Acoustical Society of America*, vol. 106, pp. 1064-1077, 1999.

[24] H. M. Hanson, K. N. Stevens, H. K. J. Kuo, M. Y. Chen, and J. Slifka, "Towards models of phonation," *Journal of Phonetics*, vol. 29, pp. 451-480, 2001.

[25] P. Hedelin and D. Huber, "Pitch period determination of aperiodic speech signals," in *ICASSP 90, Vols 1-5*, 1990, pp. 361-364.

[26] C. Henton and A. Bladon, "Creak as a sociophonetic marker," in *Language, Speech, and Mind: Studies in Honour of Victoria Fromkin*, L. M. Hyman and C. N. Li, Eds. London; New York: Routledge: published in the USA in association with Routledge, Chapman and Hall, 1988, pp. 3-29.

[27] H. Herzel, D. Berry, I. R. Titze, and M. Saleh, "Analysis of vocal disorders with methods from nonlinear dynamics," *Journal of Speech and Hearing Research*, vol. 37, pp. 1008-1019, 1994.

[28] J. Hillenbrand, "A methodological study of perturbation and additive noise in synthetically generated voice signals," *Journal of Speech and Hearing Research*, vol. 30, pp. 448-461, 1987.

[29] A. Hirson and M. Duckworth, "Glottal fry and voice disguise - a case-study in forensic phonetics," *Journal of Biomedical Engineering*, vol. 15, pp. 193-200, 1993.

[30] Y. Horii, "Fundamental-frequency perturbation observed in sustained phonation," *Journal of Speech and Hearing Research*, vol. 22, pp. 5-19, 1979.

[31] Y. Horii, "Vocal shimmer in sustained phonation," *Journal of Speech and Hearing Research*, vol. 23, pp. 202-209, 1980.

[32] C. T. Ishi, "Analysis of autocorrelation-based parameters for creaky voice detection," presented at The 2nd International Conference on Speech Prosody, Nara, Japan, 2004.

[33] J. J. Jiang, Y. Zhang, and C. N. Ford, "Nonlinear dynamics of phonations in excised larynx experiments," *Journal of the Acoustical Society of America*, vol. 114, pp. 2198-2205, 2003.

[34] P. Jinachitra, "Glottal closure and opening detection for flexible parametric voice coding," presented at Interspeech 2006, Pittsburgh, PA, 2006.

[35] D. H. Klatt, "Software for a cascade-parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, pp. 971-995, 1980.

[36] D. H. Klatt, "Description of the cascade/parallel formant synthesizer," 1990.

[37] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, pp. 820-857, 1990.

[38] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality - Review, tutorial, and a framework for future research," *Journal of Speech and Hearing Research*, vol. 36, pp. 21-40, 1993.

[39] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge [Eng.]; New York: Cambridge University Press, 1980.

[40] N. Malyska, "Automatic voice disorder recognition using acoustic amplitude modulation features," MS Thesis in EECS, Cambridge, MA: MIT, 2004.

[41] N. Malyska and T. F. Quatieri, "Analysis of nonmodal phonation using minimum entropy deconvolution," Interspeech 2006, Pittsburgh, PA, 2006.

[42] N. Malyska and T. F. Quatieri, "Spectral representations of nonmodal phonation," *IEEE Transactions on Audio, Speech and Language Processing*, to appear January, 2008.

[43] M. R. Matausek and V. S. Batalov, "A new approach to the determination of the glottal waveform," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 28, pp. 616-622, 1980.

[44] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 34, pp. 744-754, 1986.

[45] P. Mergell, H. Herzel, and I. R. Titze, "Irregular vocal-fold vibration - High-speed observation and modeling," *Journal of the Acoustical Society of America*, vol. 108, pp. 2996-3002, 2000.

[46] P. Milenkovic, "Least mean-square measures of voice perturbation," *Journal of Speech and Hearing Research*, vol. 30, pp. 529-538, 1987.

[47] F. D. Minifie, "Introduction to Communication Sciences and Disorders." San Diego, CA: Singular Publishing Group, 1994.

[48] P. J. Murphy, "Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals," *Journal of the Acoustical Society of America*, vol. 107, pp. 978-988, 2000.

[49] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1999.

[50] A. V. Oppenheim, A. S. Willsky, and I. T. Young, *Signals and Systems*: Prentice-Hall, 1983.

[51] M. Ostendorf, 2007, pp. personal communication.

[52] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Linguistic Data Consortium, 1995.

[53] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech," *Journal of Speech Language and Hearing Research*, vol. 44, pp. 327-339, 2001.

[54] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 569-586, 1999.

[55] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243-1261, 2006.

[56] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[57] T. F. Quatieri, C. R. Jankowski, Jr., and D. A. Reynolds, "Energy onset times for speaker identification," *IEEE Signal Processing Letters*, vol. 1, pp. 160-162, 1994.

[58] C. R. Rabinov, J. Kreiman, B. R. Gerratt, and S. Bielamowicz, "Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter," *Journal of Speech and Hearing Research*, vol. 38, pp. 26-32, 1995.

[59] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, pp. 407-429, 2001.

[60] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *The Lincoln Laboratory Journal*, vol. 8, pp. 173-192, 1995.

[61] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[62] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," *Journal of the Acoustical Society of America*, vol. 49, pp. 583-590, 1971.

[63] J. Schoentgen, "Spectral models of additive and modulation noise in speech and phonatory excitation signals," *Journal of the Acoustical Society of America*, vol. 113, pp. 553-562, 2003.

[64] W. Shen, S. Strassel, and C. Cieri, "Software Tools for Transcription and Annotation of Interview Recordings," presented at NWAV 2007, Philadelphia, PA, 2007.

[65] K. Sjölander and J. Beskow, "WaveSurfer," 1.8.5 ed, 2005.

[66] J. Slifka, "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of Voice*, vol. 20, pp. 171-186, 2006.

[67] K. N. Stevens, *Acoustic Phonetics*. Cambridge, Mass.: MIT Press, 1998.

[68] X. Sun and Y. Xu, "Perceived pitch of synthesized voice with alternate cycles," *Journal of Voice*, vol. 16, pp. 443, 2002.

[69] K. Surana and J. Slifka, "Acoustic cues for the classification of regular and irregular phonation," presented at Interspeech 2006, Pittsburgh, PA, 2006.

[70] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modeling*, vol. 55, *NATO Adv. Study Inst. Series D*, H. W.J. and A. Marchal, Eds. Bonas, France: Kluwer Academic Publishers, 1990, pp. 241-262.

[71] P. Thévenaz and H. Hügli, "Usefulness of the lpc-residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145-157, 1995.

[72] I. R. Titze, "Workshop on acoustic voice analysis. Summary statement," National Center for Voice and Speech, Denver, CO 1995.

[73] S. Vishnubhotla and C. Y. Espy-Wilson, "Automatic detection of irregular phonation in continuous speech," presented at Interspeech 2006, Pittsburgh, PA, 2006.

[74] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol. 16, pp. 21-35, 1978.

[75] W. R. Wilson, J. B. Nadol, Jr., and G. W. Randolph, *Clinical Handbook of Ear, Nose, and Throat disorders*. New York, NY: The Parthenon Publishing Group, 2002.

[76] C. Wu-Ton and C. Chong-Yung, "Deconvolution and vocal-tract parameter estimation of speech signals by higher-order statistics based inverse filters," presented at Higher-Order Statistics, 1993., IEEE Signal Processing Workshop on, 1993.