Investigating Speaker Features From Very Short Speech Records

Brian L. Berg

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

A. A. (Louis) Beex, Chair
Joe Ball
Ira Jacobs
Jeffrey H. Reed
Hugh F. VanLandingham

July 23, 2001
Blacksburg, Virginia

Keywords: Digital Signal Processing, Speech Processing, Speech Analysis

Investigating Speaker Features
From Very Short Speech Records

Brian L. Berg

(ABSTRACT)

A procedure is presented that is capable of extracting various speaker features, and is of particular value for analyzing records containing single words and shorter segments of speech. By taking advantage of the fast convergence properties of adaptive filtering, the approach is capable of modeling the nonstationarities due to both the vocal tract and vocal cord dynamics.

Specifically, the procedure extracts the vocal tract estimate from within the closed glottis interval and uses it to obtain a time-domain glottal signal. This procedure is quite simple, requires minimal manual intervention (in cases of inadequate pitch detection), and is particularly unique because it derives both the vocal tract and glottal signal estimates directly from the time-varying filter coefficients rather than from the prediction error signal. Using this procedure, several glottal signals are derived from human and synthesized speech and are analyzed to demonstrate the glottal waveform modeling performance and kind of glottal characteristics obtained therewith. Finally, the procedure is evaluated using automatic speaker identity verification.

Acknowledgments

This work was largely inspired by summer internships with Bellcore and Bell Atlantic, under the mentorship of Dr. Timothy Feustel and Mr. Alex McAllister who provided an exciting introduction into the professional world of speech processing. Financial support was also provided through a fellowship from the Bradley Endowment, established by the late Mrs. Marion Bradley Via and continued by Mr. Edward Via. Their generosity enabled the pursuit of this very interesting research topic. Finally, employment with Agilent Technologies/Hewlett Packard allowed this work to continue until completion.

Perhaps the biggest contributers, in terms of their valuable time spent reviewing this document and providing exceptional coaching thoughout the duration of this research, would be the Ph.D. committee members Dr. Ball, Dr. Jacobs, Dr. Reed, Dr. VanLandingham and, most of all, from committee chairman Dr. Beex. The suggestions and meticulous attention to detail provided by Dr. Beex have contributed immeasurably to the quality of this work. But even more valuable is the exceptional education provided by these professors and others at Virginia Tech. A special thanks to Dean (formerly Department Chair) and Mrs. Stephenson for their hospitality and for making the author's family feel at home at Virginia Tech. The author would like to thank his parents, Gerald and Linda, for their endless support, for instilling the importance of education from day one, and the confidence required to pursue it and "never give up". The author would like to thank Hector and Bernie for their constant encouragement, Blaire, Kayla, Morgan, and "Baby Berg" for being wonderful children and providing so much joy in life, and most of all, Gretchen, for being such a devoted mother and wife (and forgiving her husband for forgetting their anniversary two weeks prior to his dissertation defense). Words cannot express the gratitude for the endless support and encouragement received from these and so many other people throughout this work.

Table of Abbreviations

AFRIF	Adaptive Forced Response Inverse Filtering
AR	Autoregressive
ASR	Automatic Speech Recognition
CELP	Code Excited Linear Prediction
CGI	Closed Glottis Interval
DFT	Discrete Fourier Transform
DPCM	Differential Pulse Code Modulation
DSP	Digital Signal Processor
DTW	Dynamic Time Warping
EER	Equal Error Rate
FFT	Fast Fourier Transform
GIF	Glottal Inverse Filtering
HMM	Hidden Markov Modeling
LF	Liljencrants–Fant
LPC	Linear Predictive Coding
LPF	Lowpass Filter
MSE	Mean-Squared Error
OGI	Open Glottis Interval
OQ	Open Quotient
PARCOR	Partial Correlation
PCM	Pulse Code Modulation
PDF	Probability Density Function
RLS	Recursive Least Squares
SIV	Speaker Identity Verification
VQ	Vector Quantization
VSELP	Vector Sum Excited Linear Prediction

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Features found within short speech records are particularly important for modeling the characteristics of a speaker's voice, in contrast to longer term vocabulary features. Such speaker features are valuable in applications such as speech coding and synthesis. In general, these speaker features are both behavioral and inherent. Inherent (or anatomical) features depend upon the anatomy of the vocal cord and vocal tract (i.e., air passages above the vocal cords). The vocal tract anatomy refers to the size and shape of the vocal tract and is determined by the word being spoken (which is a function of mouth and tongue position) and to some degree the sex, size, and age of the speaker. Formants (i.e. speech resonances) are common features used to define the vocal tract. The vocal cord anatomy determines the pitch, breathiness, and vocal register (e.g. creakiness) of the speaker. Examples of behavioral features are dialect and voice expressiveness and these relate to the longer-term dynamics of the speaker's vocal tract (particularly, the movement of the tongue and jaw) and vocal cords.

Various approaches have been used to extract these speaker features. The most popular speech analysis algorithms use linear predictive analysis (such as the Autocorrelation or Covariance techniques [10]) to extract formants as they vary over time. The vocal cord features remain within the residual signal, which is essentially the prediction error of the analysis algorithm. Speech coding schemes have parameterized this information in various ways; popular approaches include regular pulse excited linear prediction and code excited linear prediction [2].

In pursuit of more accurate speech models, researchers in speech synthesis have developed an analysis approach called glottal inverse filtering [22, 6]. This approach uses shorter linear prediction analysis frames, centered over a specific region within each pitch period to minimize inaccuracies due to speech nonstationarities. Based on this accurate linear prediction model, a residual signal is computed that represents what is known as the glottal waveform. This glottal waveform estimate has been compared to results obtained from a physical analysis of the vocal cords [6] and with older inverse filtering techniques [26] and has been found consistent therewith. It has also been adopted as the source signal for popular speech synthesis systems known as formant synthesizers [31, 24].

Speaker features existing in short speech records, of word length and less, are investigated and an analysis algorithm is presented that accurately extracts and efficiently models the features without requiring manual intervention. The proposed procedure is based on standard adaptive (i.e. recursive) filtering, thus computing a formant estimate at every sample, rather than once per frame [20]. Hence the operations required for this analysis procedure consist of an adaptive filter, inverse filter, pitch detector, voiced/unvoiced (v/uv) detector, and endpoint detector. These processing blocks are quite common today. Furthermore, simple detection routines (v/uv, endpoint, and pitch) are adequate for successful operation.

Speech records from several speakers are analyzed demonstrating the performance as well as convenience of this method for analyzing fine speech detail. Such speech detail includes the vocal cord dynamics that produce each pitch period, as the analysis method is shown to be able to model the classical differential glottal waveform. An example of how this procedure can be used successfully for speaker identity verification is also shown.

Chapter 2

Speech Processing Concepts

This chapter discusses concepts upon which many of today's popular speech processing systems are based. Section 2.1 provides a short introduction to basic speech physiology. Next, in Section 2.2, some speech processing applications that are popular today are presented. Then Section 2.3 discusses the standard LPC analysis approach as used in most current speech analysis systems for extracting vocal tract speech characteristics. Glottal inverse filtering is then introduced in Section 2.4 to demonstrate an approach for extending basic LPC analysis to obtain a model of the vocal cord operation. Finally, in Section 2.5, common approaches for mathematically evaluating speech features, as computed by these speech analysis algorithms, are discussed.

2.1 Speech Production

The physiology of the human vocal apparatus is well understood and so it is often convenient to relate the features extracted from the speech signal analysis algorithm with the known characteristics of the vocal mechanism. This section gives only a simple overview of the speech mechanics, and so a textbook that specializes in speech production should be referred to for more detail [33, 13, 41, 46].

The physical systems that are responsible for producing the speech waveform are the vocal cords and vocal tract. Voiced sounds are produced by the vibrating vocal cords which flap open and shut (corresponding to the open and closed glottis cycles, respectively) as the air is forced out of the lungs. The sound produced during the open glottis interval produces a waveform with a triangular shaped pulse, whereas the closed glottis generally corresponds to an interval of approximately zero excitation. The main vocal cord excitation occurs when the glottis suddenly flaps shut at the end of the open glottis interval. The resulting waveform is known as the glottal volume-velocity waveform $g(t)$ (or simply, glottal waveform) [45]. Figure 2.1 shows one period of a typical glottal waveform pattern. The pitch period is denoted as T_0. The opening time T_p corresponds to the interval of the pulse with a positive slope, and the closing time T_n is the interval of the pulse with a negative slope. Since it is quite common to specify the relative glottal waveform intervals, the following variables are introduced

$$OQ = \frac{T_p + T_n}{T_0} \tag{2.1}$$

and

$$SQ = \frac{T_p}{T_n} \tag{2.2}$$

referred to as the open quotient and glottal pulse skew, respectively. Also note another common term referred to as the fundamental frequency which is

$$F_0 = \frac{1}{T_0}. \tag{2.3}$$

The value A is commonly referred to as the amplitude of voicing.

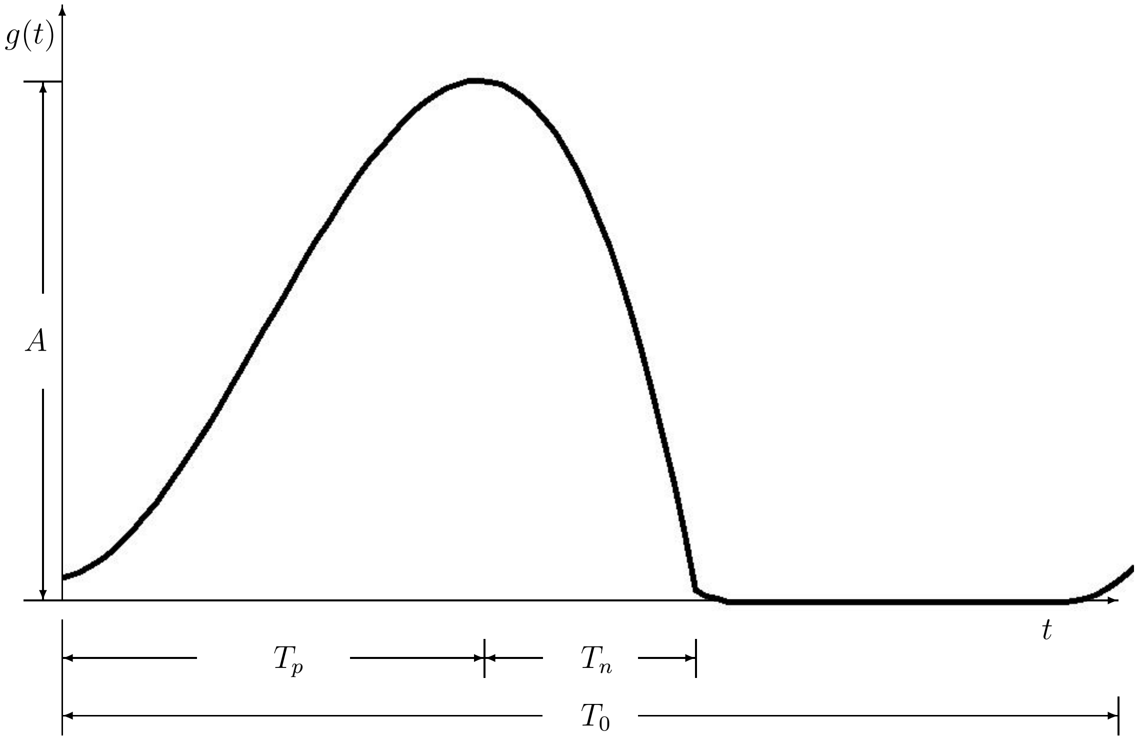

Figure 2.1: A traditional glottal waveform.

The glottal source signal (which is rich in harmonics due to the approximate triangular shape) then excites the system above the vocal cords known as the vocal tract. The vocal tract, acting as an acoustic tube characterized by its natural resonances called formants, filters the glottal excitation so as to produce the desired speech.

The study of human speech communication falls into the field known as linguistics. According to linguistics, the fundamental speech unit that makes up our vocabulary of speech sounds is referred to as a phoneme. Another popular speech unit in speech processing applications is the diphone which is the segment from the stationary portion of one phoneme to the stationary portion of the next phoneme. As implied above, the vocal tract is responsible for producing the desired speech phoneme and is determined by the vocal tract shape. Speakers control the vocal tract shape by positioning the tongue and jaw. Besides its vocabulary dependence, the vocal tract is speaker-dependent as well. For example, differences in dialect may be captured by comparing the way two speakers pronounce the same vowel, or change the shape of the vocal tract when pronouncing a diphthong as in the word “I”. More inherent speaker features incorporated in the vocal tract would be the overall size, static shape, unique vocal obstructions, and even dental work.

In order to add speech expression and character, speakers may alter their vocal cord vibration patterns (i.e. glottal waveform) by changing pitch or intensity of vibration. The latter are the two primary characteristics of the vocal cord operation which the speaker is able to control, and are accomplished by varying the length of the vocal cords via certain laryngeal muscles, and by changing the flow rate of air from the lungs, respectively. Unlike the vocal tract, the vocal cords are essentially unaffected by factors that determine vocabulary (assuming voiced speech). However, some pitch variations have been shown to occur with specific vowels [56]. Despite these sources of variation to the vocal cord characteristics, studies based on speech during normal discourse have observed inter-speaker variations to be more significant than intra-speaker variations [26, 36, 48]. Hence, physical characteristics such as vocal cord size, are the primary factors that influence the shape of the glottal signal. In fact, the common vocal registers–the modal (or normal) register, falsetto register (often associated with high pitched speech), breathy register, harshness register, and vocal fry (or creak)–have been associated with different vocal cord lengths and thicknesses [58, 8, 23]. There are different voice qualities within these main registers as well. So, although a speaker may be able to alter his or her voice register, in normal discourse it is possible (and quite common) to generalize the overall voice quality of a speaker [6] to a particular vocal register.

A common assumption for the speech mechanism is that the vocal tract and the excitation are independent of each other so that the operation of the vocal cords does not affect the acoustical properties of the vocal tract. Many synthesizers now used operate under this assumption. However, it is now known that during glottal opening, the vocal tract properties become coupled with the properties of the system below the glottis resulting in a slight variation of the resonances, particularly at low frequencies. This source-tract interaction is intelligible and many speech synthesizers now attempt to model it via the excitation source or vocal tract [1, 47].

2.2 Speech Processing Applications

Digital speech processing is used in several popular applications today including speech coding, speech synthesis, and speech/speaker recognition. Here an overview of these popular applications is given and it is shown that, although the objectives of each of these applications may seem very different, the speech analysis approach used is quite universal. This universal approach is commonly based on parametric signal modeling techniques.

Parametric signal modeling is perhaps the most important aspect of speech analysis in current speech processing applications. One of the main objectives of signal modeling is to reduce the amount of information needed to represent a signal with a certain degree of accuracy. Hence in speech analysis, a good signal modeling routine removes all information that does not improve the accuracy of the particular speech processing application. The most common information that gets discarded in speech processing systems relates to the redundancy of the speech signal. Another example of information to discard would be that which characterizes a particular speaker for applications using speaker-independent speech recognition systems. This section discusses some of these basic operating principles of popular speech processing applications, such as the signal modeling goals and how the accuracy of the model is measured for the particular application.

2.2.1 Speech Synthesis

The speech processing application that converts text to a speech signal is known as speech synthesis. A common implementation is illustrated in the block diagram in Figure 2.2. Basically the system first analyzes the text within the linguistic processor and determines the speech units, from the pronunciation dictionary, that provide an appropriate representation of the input speech. The speech unit coder then loads the respective speech parameters from the speech unit dictionary. The speech units are then generated with a natural sounding prosody (speed) and pitch and are concatenated to produce the entire speech representation of the input text. The remainder of this section will focus mainly on the bottom three blocks of the diagram; specifically, how to obtain the parameters for the speech unit dictionary and how to use them to generate speech.

The contents of the speech unit dictionary mentioned above varies between speech synthesis systems, and depends upon the assumed speech model. One very popular system uses an excitation sequence which drives a time-varying filter to model the various speech sounds. Such a speech synthesis system is shown in Figure 2.3. The voiced/unvoiced input represents a decision concerning whether to produce a voiced sound, by making $\{p_n\}$ a train of scaled unit pulses with an

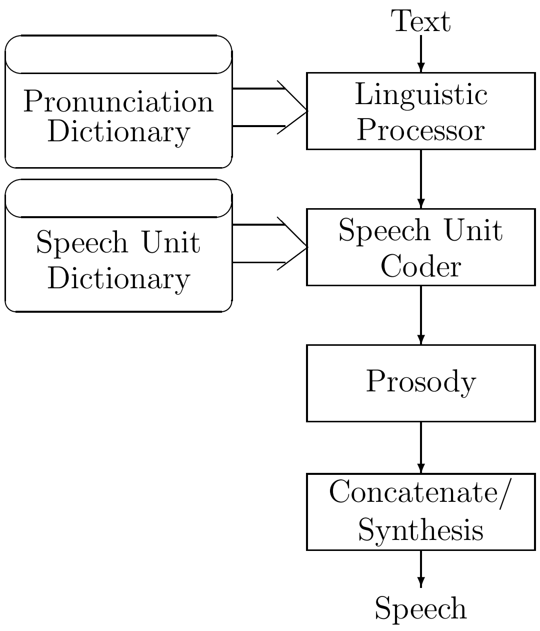

Figure 2.2: Block diagram of a speech synthesis system.

appropriate pitch period, or to produce an unvoiced sound, such as “sh,” by using discrete white noise. The coefficients of the synthesis filter, $A(z)$, represent the physical effects of the vocal tract. These parameters are typically updated every 10 to 25 ms corresponding to the duration of a typical phoneme.

One important benefit of this signal modeling approach for speech processing, demonstrated in this application, is that the number of parameters needed to implement the above speech synthesis system is typically much less than the number of data points in the block. The excitation sequence may not even need to be stored, such as when an “equivalent” sequence is generated as a random uncorrelated sequence $e_n = w_n$, for an unvoiced block and a pulse train for the

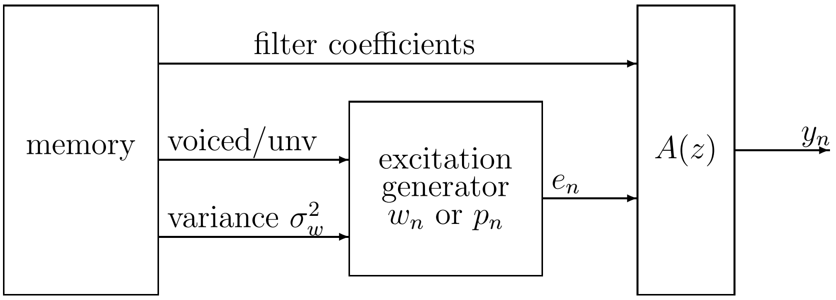

Figure 2.3: A typical speech synthesis system.

voiced block, $e_n = p_n$. In this case, storage is only required for the few coefficients of $A(z)$, which is usually in the range of 8 to 12 values. However, compressing the excitation sequence to this extent affects the speech accuracy as this model is known to generate mechanical sounding speech [44]. Hence, for a more accurate reproduction, the parameters, as well as the entire error sequence, may need to be stored. Even so, compression is still accomplished since the uncorrelated sequences $\{w_n\}$ and $\{p_n\}$ have a smaller dynamic range than the data sequence $\{y_n\}$, thus requiring fewer bits for representation.

2.2.2 Controlling Speech Rate, Pitch, and Spectrum

Some speech synthesis systems provide the ability to adjust the speech and speaker characteristics of the synthesized speech–to change the speech prosody (speed), pitch, or even choose between a male voice and a female voice. The latter could be implemented by storing speech templates for several different types of speakers. However some speech systems have devised rules on how to change given parameters to produce certain speech characteristics providing the following additional advantages:

- It is possible to transform speech recorded by a reference speaker into a more desirable target speech, eliminating the need to record and store additional dictionaries. For example, one can convert speech recorded from a male speaker to female or child-like speech (and vice versa).
- For people with hearing loss at a certain frequency range, it is adventagous to be able to shift the pitch of the synthesized speech.
- Being able to control the speech rate without altering the pitch and spectrum can be useful when one has a lot of information to listen to, and wishes to speed it up during the less important speech and slow it down for that of most interest.

Early synthesis systems have allowed pitch-scale and time-scale modifications for the synthesis of more natural sounding speech. These modifications have been performed in the time and frequency domains, as well as with parametric filters. Figure 2.4 shows how these particular synthesizers evolved and their references.

One of the first of these synthesis systems was the phase vocoder. The phase vocoder is an analysis-synthesis system which uses parameters that describe the short-time Fourier transforms of a signal. This system became feasible when the

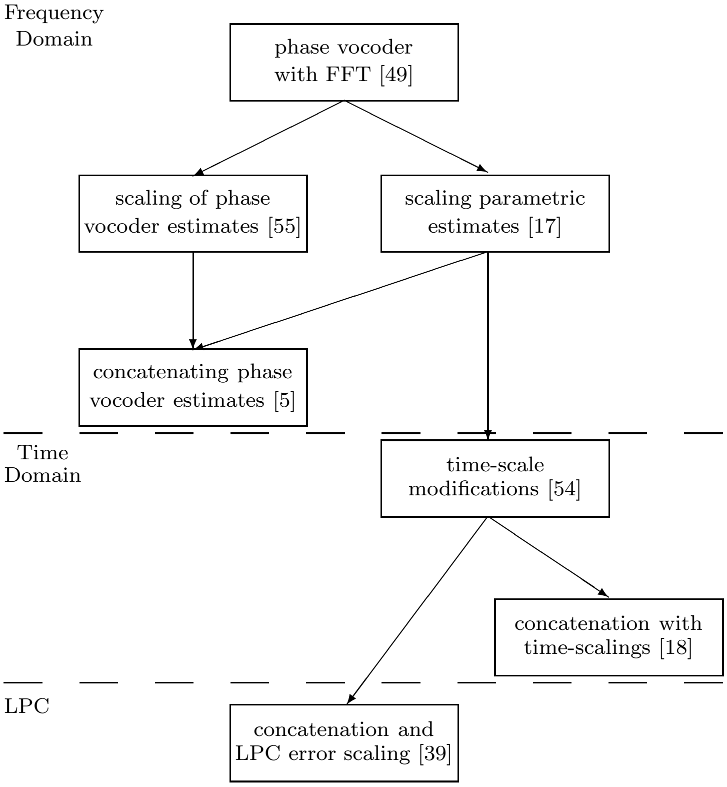
Frequency Domain
phase vocoder with FFT [49]
scaling of phase vocoder estimates [55]
scaling parametric estimates [17]
concatenating phase vocoder estimates [5]
Time Domain
time-scale modifications [54]
concatenation with time-scalings [18]
LPC
concatenation and LPC error scaling [39]

Figure 2.4: Evolution of synthesis systems that allow prosody and pitch scaling.

fast Fourier transform was applied to the problem [49]. This short-time Fourier transform approach was eventually enhanced, enabling modifications to be made to the excitation and/or the spectrum of a speech signal [55, 17]. Phase vocoder algorithms were later extended to handle memory efficient diphone concatenation [5] using an approach referred to as frequency domain, pitch-synchronous overlap add (FD-PSOLA) and requires storage of the raw speech waveforms, hence imposing the need for a diphone dictionary of approximately 7 Mbytes when sampled at 16 kHz.

To demonstrate the basic idea behind these types of modifications, first short-time speech signals $x_m(n) = h_m(t_m - n)x(n)$ are obtained, where $x(n)$ is the speech amplitude at time n, $h_m(\cdot)$ is some window, and t_m mark each pitch period making the algorithm "pitch-synchronous" (set at random points for unvoiced speech). The length of h_m is typically chosen to be long enough to cover at least three times the longest pitch period over which the short-time Fourier transforms are then computed. Finally, frequency domain modifications (pitch) are made on the harmonics (i.e. compression–expansion, etc.) and time (duration) modifications are made back in the time domain to obtain the desired synthesis modification such as rate or pitch. This approach allows a great deal of flexibility in pitch-scaling, but faces the problem that natural sounding artificial harmonics must be produced when the pitch (frequency) is decreased.

Other systems have been presented that bypass the frequency domain manipulations and perform time modifications directly on the waveform to save computations [54, 18]. In this approach, the short-time speech signals are again obtained within each pitch period. To modify the pitch, the pitch period is altered depending upon whether the speech signal is wide-band or narrow-band. For slowing down the speech signal some short-time signals are repeated, whereas the speech signal is accelerated by selectively eliminating some short-time signals. This approach has also been used to control other speech qualities such as vocal register [57]. Since the raw diphone waveforms are stored as in FD-PSOLA, a large database is required with this time-domain approach. Furthermore, attenuation of certain frequency bands for some narrow-band signals often results in a reverberant sound.

Another approach that has been proposed referred to as linear-prediction, pitch-synchronous overlap add (LP-PSOLA) also utilizes time-scale and pitch-scale modifications, but instead of performing these modifications directly on the time-domain or frequency-domain waveform, they are performed on the signal that will excite the parametric filter [39]. The computational burden of the latter technique is smaller than that of the frequency domain technique and it also requires less memory than the previous two approaches. However, as the LP-PSOLA procedure currently stands, operation solely on the residual provides limited flexibility compared to FD-PSOLA.

Another class of well-known synthesizers that provide control of the speech characteristics are known as formant synthesizers [30, 31, 6, 24]. These systems do not store templates or speech waveforms that represent segments of speech, but a dictionary of rules on how to choose and modify synthesis parameters. The parameters provided in these formant synthesizers perform adjustments to the formant amplitudes and bandwidths, as well as the shape of the excitation or glottal waveform, as it is specifically referred to. Rules have been devised for not only controlling the prosody and pitch, but also the vocal register of the voice; even whether the voice should sound like a male or female speaker. More details on the glottal waveform and its significance in speaker characterization are discussed in Section 2.4.

2.2.3 Speech Coding

The transmission of digitized speech for communications applications, as shown in the block diagram in Figure 2.5, involves what is referred to as speech coding. The speech first passes through a coder where it is analyzed and compressed. The encoded speech is then transmitted through a channel like a wire or a wireless propagation channel. Finally the encoded speech parameters are received in the decoder and synthesized to a speech signal.

Similar signal processing operations as described in Section 2.2.1 for speech

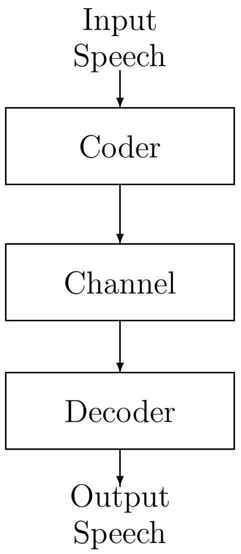

Figure 2.5: Block diagram of a speech coding system.

synthesis are also commonly used for speech coding, however, the challenges are often different. For speech coding it is necessary to transmit an audio signal as accurately as possible, and yet to compress the bandwidth as much as possible, to allow more signals to be transmitted over the same channel. Immediately, the tradeoff between accuracy and compression emerges, particularly for low bit-rate speech coding systems. Furthermore, the telephone is a real-time system (i.e., it is not possible to store the signal, process it, and then transmit it); hence, the signal modeling algorithm must be sufficiently fast and must rely on assumptions about the signal characteristics for modeling. These are the challenging issues faced in speech coding systems.

As an example consider a modeling approach similar to that used for speech synthesis in Section 2.2.1. In low bit-rate speech coding systems where $\{e_n\}$ is synthesized at the receiver instead of being transmitted, it is necessary to transmit the voiced/unvoiced decision, gain parameters and, in the case of a voiced decision, a pitch period. Hence for each N length block, a set of, say, p filter coefficients for $A(z)$ are transmitted as well as 1 bit for the voiced/unvoiced decision, about 6 bits for the pitch period and about 5 bits for the gain parameter. Typical rates with such schemes are $72F_s$ bits per second, where F_s is the number of blocks per second. For $F_s = 100$, 67, and 33, respectively, this results in bit rates of 7200, 4800, and 2400 bits per second [51]. A popular variation of this type of system is known as LPC-10 [10]. In transmission, it should be noted that synchronization is a potential problem since it is necessary to distinguish which part of the transmission stream contains the coefficients for $A(z)$, and which part corresponds to the excitation parameters. Also, separate consideration must be made to the coding schemes used for $\{e_n\}$ and $A(z)$ since they each have different sensitivities to transmission errors [16].

Many encoding schemes have been devised that attempt to achieve high compression of the filter coefficients and excitation information while maintaining quality that is acceptable for a given application. In terms of the modeling filter coefficients, particular care must be taken in computing and quantizing the coefficients for $A(z)$ since roundoff, truncation, and quantization errors can result in an unstable modeling filter. Commonly the coefficients are converted to another representation, for example partial correlation (PARCOR) or line spectrum pair (LSP) coefficients, to allow straightforward stability verification [10].

Vector quantization (VQ) is another data compression technique that is often used for quantizing the speech coefficients that represent $A(z)$. VQ involves mapping vectors that are capable of taking on an infinite number of values onto a finite set of vectors which are referred to as code words (or centroids). Greater compression is achieved by using a small collection of code words (i.e. a small codebook). For example, if a codebook is designed with 16 code words to represent 10 filter coefficients in a block of speech, then a binary index of only $log_2 16 = 4$ bits is needed to address each respective code word as opposed to transmitting the 10 floating point values. Hence VQ is of interest whenever compression of speech parameters is desired, not only for reducing bit rate in speech coding, but also for conserving memory and computations in speech synthesis, recognition or speaker identity verification (SIV), as will be shown in Section 2.5.2.

As for speech excitation compression, the simple speech coding system presented at the beginning of the section gave an example of how the excitation information can be encoded to achieve low bit rates using simply a voiced/unvoiced flag, the pitch period and a gain parameter. However, as discussed for speech synthesis, the problem with this model is the simplicity with which the excitation is represented as well as the difficulty in estimating the pitch period. These problems have contributed to mechanical sounding speech [44]. To achieve an acceptable level of speech quality, more accurate models of the excitation signal are used [44]. A popular alternative to artificially generating the excitation signal is to directly encode $\{e_n\}$ into a transmission stream using waveform encoding such as differential pulse-code modulation (DPCM) [51]. Since the pitch information is retained in $\{e_n\}$, pitch detection is not necessary. Furthermore, although the modeling filter is unable to uniformly reduce dynamic range over all analysis blocks, a low bit rate is achieved by adaptively quantizing $\{e_n\}$. A more elaborate quantization scheme known as multipulse LPC has been used to even further improve speech quality. Following the progression for encoding the modeling coefficients, VQ is a logical candidate for encoding the excitation sequence when even lower bit rates are desired. Such coding systems include code-excited linear prediction (CELP) and vector sum excited linear prediction (VSELP) which are popular in wireless communications [2, 32].

2.2.4 Speech and Speaker Recognition

There are other speech processing applications where the interest is not in reproducing speech, as in synthesis and coding, but in obtaining information from the speech. Such is the case for the two applications discussed in this section – speech recognition and speaker identity verification. These applications both utilize a matching routine for comparing speech information.

The technology known as speech recognition has become very popular today. Considered the inverse function of speech synthesis, speech recognition takes speech as the input and outputs the text equivalent. A block diagram of a typical system is shown in Figure 2.6. In this system, the extracted speech features for a given word are compared to stored features that represent a dictionary of words. When the input word is similar to more than one word in the dictionary, this system uses linguistic rules to help determine which word fits the sentence the best. Perhaps the biggest attraction of speech recognition is as a friendly human interface for computer control.

In the area of voice security applications, speaker recognition systems, such as for speaker identity verification (SIV), are popular. SIV may be used for secure voice access to a facility such as a home or computer, or to services like money in a bank account, credit or phone card (validation), or personal information (e.g. medical records). SIV may also be a useful legal or correctional tool, for forensics or home incarceration, to verify that the convict is indeed at home. A diagram for an SIV system is given in Figure 2.7. Note that the task of speaker identity verification (SIV) only requires a binary decision as to whether the claimed speaker

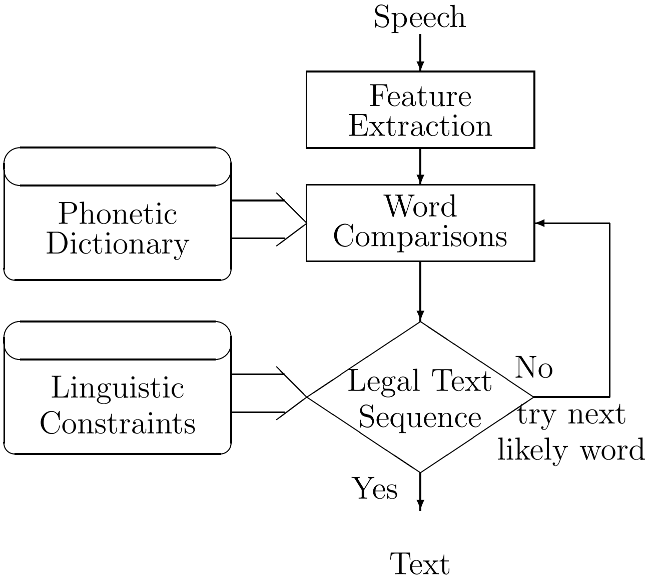

Figure 2.6: Block diagram of a speech recognition system.

is who he or she claims to be. In order to “recognize” a particular person’s voice, the system is trained for each speaker, each of which are referred to as subscribers of the system. These trained feature sets, also called templates, are compared to the feature sets that are analyzed during the actual SIV testing process. During the SIV test phase, the user must provide some kind of non-speech identification as to which subscriber he or she claims to be. The templates for the particular subscriber are then loaded from the dictionary and compared to the test features extracted from the current user. Users who are not the subscriber that they claim to be are referred to as imposters.

So the task of automatic speech recognition (ASR) systems is to identify the

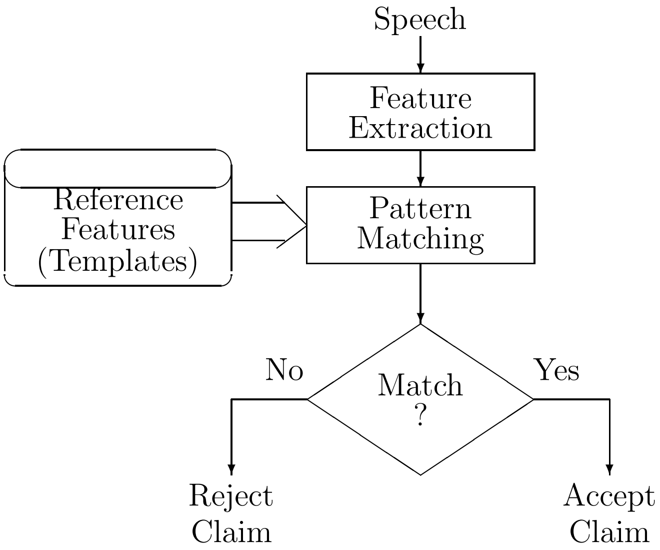

Figure 2.7: Block diagram of an SIV system.

word or phrase that makes up a given input speech record. On the other hand, the process of distinguishing a particular speaker by his or her voice from a set of known speakers is referred to as speaker recognition.

As a result, ASR is concerned with extracting the vocabulary-dependent information, while SIV is mainly interested in the speaker-dependent features. Consequently, the objectives of SIV and ASR are quite different. In fact, speaker independent ASR attempts to remove all speaker-dependent information, in contrast with vocabulary-independent SIV which attempts to remove all vocabulary-dependent features.

Despite these differences, the most common SIV and ASR systems utilize virtually identical front-end feature extraction algorithms. The reason for this is that the formants are the main source of phonetic information needed in ASR, while their speaker-dependent aspects have made them the most effective SIV feature as well. Hence, algorithms that accurately extract the formants are of interest for both the SIV and ASR areas. Other features that are vocabulary- and speaker-dependent include the energy contour (i.e., the variation over time) and the zero-crossing rate of the speech signal.

Another similarity between ASR and SIV is the manner in which the specimen speech record is compared to a given model. In ASR, the model represents a word or phrase, or speaker-dependent features in SIV. A popular approach for ASR and vocabulary-dependent SIV involves dynamic time-warping in which the time-varying features of two utterances are nonlinearly aligned according to a certain distance measure [45]. Another comparison approach utilizes Hidden Markov Modeling (HMM) which statistically models a particular speech utterance based on some training data. The specimen utterance is compared to the HMM using a maximum likelihood computation. Comparison routines like these will be discussed in more detail in Section 2.5.

Hence, ASR and SIV share many of the same algorithms for analysis and matching. As mentioned earlier, the primary operational differences of the two systems occur as an ASR system moves toward speaker-independent operation whereas the SIV system relies on the speaker-dependent information. Hence the signal modeling problem in these cases requires that the ASR system de-emphasizes and the SIV system emphasizes the speaker-dependent information.

Transformations or rotations of the feature space can be made as well to emphasize the feature's vocabulary-dependence for ASR, or speaker-dependence for SIV [45]. Such an operation typically involves a training procedure in order to accurately represent the feature distributions for each class (i.e. speakers in SIV or words in ASR) in multi-dimensional space. For speaker-independent ASR, the trained class should represent several different speakers saying the same word corresponding to the class. Once the distributions are obtained, a transformation is devised in an attempt to maximize the distance between class distributions.

One other practice to make an ASR system less speaker-dependent, not involving any feature manipulation as in the above techniques, is to train the model (e.g. HMM, or VQ) that corresponds to the word with sufficient data from a representative population of different speakers. It is advised, however, to first obtain features (perhaps using a transformation) that are less sensitive to speaker variations so that the training data features will be more stable between speakers, and the model criteria will be tight enough to accurately discriminate the words. This approach can also be applied to achieve a more vocabulary-independent SIV system. Most vocabulary-independent verifiers use similar analysis procedures, for obtaining the pitch, energy and formant information, as the vocabulary-dependent verifiers. Instead of storing each computed value, however, they are typically averaged over large speech segments to obtain a good average of the feature values [40]. For example, if large enough speech records from both an adult male and a child were observed, one would expect the formant frequencies to be higher for the child's speech than for the adult male's speech. In order to achieve acceptable error rates for this case, often more than 30 seconds of speech is required for training. In many situations such a requirement may be unacceptable due to limited speech data, limited computation time, or simply the human factor cost related to the training time required for each user.

Another popular approach to make an ASR system less speaker-dependent is to normalize the formant frequencies so that the slight formant variations between speakers for a given phoneme do not affect the pattern-matching phase [45]. A similar approach is to perform a frequency-domain transformation in hopes that the phonetic content will be emphasized over the speaker-dependent information. Such transformations are often referred to as perceptual spectral scaling procedures as they attempt to model human auditory perception. Many speech scientists feel that the human auditory system concentrates on the phonetic content of the speech, which is the reason that mel and Bark scales found early application in automatic speech recognition (ASR) [9, 21]. Thus speech scientists believe that the speaker dependent information is extraneous, which is somewhat justifiable in that people have relied mainly on sight for the task of speaker identification.

The mel-frequency scale is based on the pitch unit *mel* and can be approximated by a function which, when plotted versus frequency, is linear below 1000 Hz and logarithmic above. The mel-frequency scale is believed to describe the ear's perception of pitch [45]. The mel-frequency cepstral coefficients are commonly obtained [9, 35] by spanning the mel-frequency axis with N triangular bandpass filters to derive the log energy X_k for each critical band $0 \leq k < N$ from which the M mel-frequency cepstral coefficients are computed

$$MFCC_i = \sum_{k=1}^{N} X_k cos[i(k - 1/2)\pi/N], \text{ for } i = 1, 2, \cdots, M. \tag{2.4}$$

The Bark scale is another form of perceptual spectral scaling which is believed to more directly model the ear's perception of vowels [37, 19, 12], as it has been observed that the auditory system processes nonspeech sounds differently than it does speech sounds [45]. The function

$$f_{Bark} = \frac{26.81f}{1960 + f} - 0.53 \tag{2.5}$$

provides an approximate mapping of the frequency axis f in Hz to the auditory Bark frequency scale f_{Bark} [42].

2.3 Standard LPC Analysis

One approach to modeling some of the speech features presented in the previous section, which has been used successfully in a wide range of speech processing applications for several years, is called LPC analysis. Standard LPC analysis is used to model speech formants as they vary with the vocal tract dynamics. This section describes how this is accomplished.

Linear Prediction algorithms attempt to represent each sample of the digital speech signal s_n in the form

$$s_n = \hat{a}_1 s_{n-1} + \hat{a}_2 s_{n-2} + \cdots + \hat{a}_p s_{n-p} + e_n \tag{2.6}$$

where s_n is a speech sample (e.g. pulse code modulation or PCM value) at discrete time n, p is known as the prediction order and is set at a value slightly higher than twice the expected number of formants (for reasons discussed later), and $\{\hat{a}_i\}_{i=1}^{p}$ and $\{e_n\}$ are results computed by the algorithm. The relationship between the extracted LP parameters and the physical speech features can be more readily observed in the frequency domain. So converting to the z-domain,

$$\begin{aligned} \mathscr{Z}(s_n) &= S(z) && (2.7) \\ &= a_1 z^{-1} S(z) + a_2 z^{-2} S(z) + \cdots + a_p z^{-p} S(z) + E(z), && (2.8) \end{aligned}$$

and gathering the $S(z)$ terms,

$$S(z)(1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_p z^{-p}) = E(z), \tag{2.9}$$

gives the z-domain version of $\{s_n\}$,

$$S(z) = \frac{E(z)}{1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_p z^{-p}}$$

$$= \frac{z^p}{(z-\rho_1 e^{j2\pi f_1})(z-\rho_2 e^{j2\pi f_2})\cdots(z-\rho_p e^{j2\pi f_p})} E(z), \qquad (2.10)$$

where $\rho_i e^{j2\pi f_i}$ for $i = 1, 2, \cdots, p$ are the complex roots of the pth order polynomial in the denominator with ρ_i representing the root radius and f_i the root angle. Finally, to express this in terms of frequency, the substitution $z = e^{j2\pi f}$ is made where $j = \sqrt{-1}$, and f represents the fractional (or normalized) frequency variable (so that $-0.5 <= f = f_{Hz}T <= 0.5$ where T is the sampling period). Hence,

$$\begin{aligned} S(f) &\equiv S(e^{j2\pi f}) \\ &= \frac{e^{j2\pi f p}}{(e^{j2\pi f}-\rho_1 e^{j2\pi f_1})(e^{j2\pi f}-\rho_2 e^{j2\pi f_2})\cdots(e^{j2\pi f}-\rho_p e^{j2\pi f_p})} E(f). \quad (2.11) \end{aligned}$$

Note that when $f = f_i$ and ρ_i is close to one, the denominator becomes small so that $S(f_i)$ increases. The LP algorithms attempt to identify the "pole" parameters f_i and ρ_i representing these frequency domain peaks in terms of the coefficients $\{a_i\}$ in (2.10).

Since the speech resonances (i.e., formants) show up as peaks in the frequency domain as well, LP is a tool for estimating them. Usually there are five main formants (or fewer) in a speech signal, so conceptually they could be represented by ten factor terms ($p = 10$) in the denominator of $S(z)$. (The order p must be double the expected number of formants since the poles of the speech signal form complex conjugate pairs. Hence for every positive frequency pole there is a corresponding negative frequency pole.) In practice, however, the estimates obtained from LP analysis algorithms tend to accentuate stronger formants while masking lower level ones, requiring additional poles to detect these weaker formants [29].

If the coefficients $\{a_i\}_{i=1}^{p}$ are used to represent the vocal tract characteristics (i.e. formants), then the remaining speech information must reside in $\{e_n\}$. As observed from the expression for $S(f)$ in (2.11), the frequency domain representation of the vocal tract multiplied by the frequency domain representation of $\{e_n\}$ (i.e. $E(f)$) gives the speech spectrum.

Up to this point, no mention has been made of how the parameters $\{a_i\}_{i=1}^{p}$ and $\{e_n\}$ are extracted from $\{s_n\}$. This is done by computing the prediction error signal so that it minimizes the sum of squared errors cost functional

$$\sum_{n=n_1}^{n_2} |e_n|^2 = \sum_{n=n_1}^{n_2} |s_n - \sum_{k=1}^{p-1} a_k s_{n-k}|^2. \tag{2.12}$$

The Autocorrelation and Covariance LP algorithms differ in the values n_1 and n_2 used in the cost functional and the assumptions made on the speech data. Particularly, for the Autocorrelation method, $n_1 = 1$ and $n_2 = N + p - 1$ where N is referred to as the frame length. With this range of n it is observed from the cost functional that values for $s_{2-p}, \cdots, s_{-1}, s_0, s_{N+1}, s_{N+2}, \cdots, s_{N+p-1}$ are required. Since not all of these are available, those that are not are assumed to be zero. The Autocorrelation method turns out to be the most computationally efficient LP algorithm, but the assumptions that the speech samples are zero outside the frame make this technique inaccurate for short speech records. Under this situation, the

Covariance method is preferred. In this case, $n_1 = p$ and $n_2 = N$ so that all of the necessary information is available to minimize the cost functional. The Covariance method has been shown to perform well for short speech frames, but at the expense of additional computation. The Autocorrelation and Covariance LP algorithms, as well as the various efficient algorithms available for obtaining their LP parameters, are discussed in detail in various books [20, 29, 51]. It should be noted that these "frame adaptive" LP approaches characterize an entire frame of speech samples with a single set of formants. Hence the frame length must be chosen large enough so that reliable parameter estimates can be obtained from the LP analysis routine, but small enough so that the formant variation within the frame is small (i.e., the speech is considered stationary). Then the time-varying nature of the formants is extracted by analyzing several formant estimates per speech record.

Before proceeding, another routine that will come up later and should be mentioned involves preemphasis. As eluded to previously, LP algorithms may use more than one pole to model a strong peak, and if the order p is not chosen large enough, other formants may be ignored. This is particularly of concern when analyzing speech which tends to have strong low frequency components. To avoid missing the lower energy high frequency formants, preemphasis is often performed on the speech signal prior to LP analysis. Pre-emphasis is often achieved using the simple transformation $x_n = s_n - \alpha s_{n-1}$ $(0 \leq \alpha < 1)$, which acts as a gradual high pass filter amplifying high frequency components and attenuating those at lower frequencies.

2.4 Glottal Inverse Filtering

Standard LPC analysis, introduced in the previous section, has become very popular in speech processing because of its ability to model the slowly time-varying speech formants, while still being a very simple procedure. This section discusses an approach that extends the capabilities of LPC analysis in order to model the finer speech characteristics that result from the vocal cord behavior.

Standard LPC analysis involves analyzing a speech record using several consecutive analysis frames. The frequency resolution of the LPC analysis approach is known to improve with increasing frame size, hence 30 ms frames are quite common for speech sampled at 8 kHz. For smoother tracking of speech formants, these frames are usually overlapped by a third to a half of the frame length. Hence the number of analysis frames for a given duration of speech will be the same for two different records. This is not the case for the approach described in this section, which attempts to set the frame location over a specific portion of a pitch period. Such analysis approaches are often called "pitch-synchronous."

Hence the focus of linear predictive, glottal inverse filtering (GIF) is to overcome the problems of modeling nonstationarities suffered by conventional frame adaptive algorithms, such as LPC analysis, in order to obtain a model of the vocal cord behavior [6]. GIF models this behavior using the glottal waveform.

A flow diagram illustrating glottal inverse filtering is given in Figure 2.8. The procedure avoids the open glottis region and analyzes over the closed glottis interval (CGI) only using Covariance analysis. Since this region is basically free from source-tract interaction, the formant estimates via $\{a_i\}_{i=1}^{p}$ better represent the true static vocal tract characteristics over that pitch period. These coefficients are then used to obtain the differential glottal waveform $u'_{g,n}$ for that pitch period, which is obtained by running the speech samples $\{s_n\}$ within the entire pitch period through the inverse filter:

$$u'_{g,n} = e_n = s_n - a_1 s_{n-1} - a_2 s_{n-2} - \cdots - a_p s_{n-p}, \tag{2.13}$$

The nonstationarities associated with the open glottis region are hence imbedded in the glottal waveform estimate $u_{g,n}$. The reason that $\{e_n\}$ is considered the differential glottal waveform is that it is assumed to contain the radiation effects of the lips as well. These effects are approximately modeled by the crude high pass

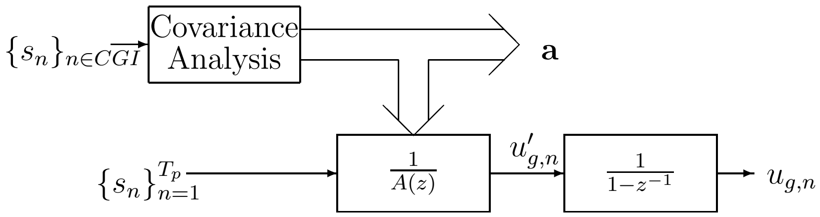

Figure 2.8: Flow diagram illustrating the GIF analysis procedure.

filter $1/(1-z^{-1})$ so that

$$U_g(z) = E(z)/(1-z^{-1}) \tag{2.14}$$

or

$$U_g(z) - z^{-1}U_g(z) = E(z). \tag{2.15}$$

The time-domain expression is

$$e_n = u_{g,n} - u_{g,n-1} \equiv u'_{g,n}. \tag{2.16}$$

Then to find the glottal waveform, the approximate integral, sometimes referred to as the digital integral, of $\{e_n\}$ is computed by simply arranging this equation so

$$u_{g,n} = e_n + u_{g,n-1}. \tag{2.17}$$

This glottal waveform estimate has been compared to results obtained from a physical analysis of the vocal cords by Childers and Lee [6] and with older inverse filtering techniques by Hunt, et al [26] and has been found reasonably consistent therewith.

There are several attractive implications of this special LP analysis procedure. First of all, more consistent vocal tract estimates for a given pitch period are expected since the analysis frame positions are always restricted to the stationary closed glottis interval. Since the frames are typically short so as to cover only the closed glottal interval, tracking of the vocal tract variations can be achieved with high temporal resolution. Lastly, the glottal estimate obtained via the residual can be used as an additional feature.

Unfortunately, some very serious practical problems emerge using this analysis technique. One involves precise location of the closed glottis interval. A popular approach, proposed by Wong et al [60], involves computing a different set of LPC parameters via the Covariance method at every speech sample by sliding a short window of speech sample-by-sample while performing the analysis in an attempt to obtain the precise location of glottal closure (i.e., the interval of approximate zero excitation). A similar method [43] computes the Discrete Fourier Transform (DFT) at every sample. Both approaches require heavy computation which hinders real-time implementation. Another weakness is the issue of frame length. The short duration of the closed glottis interval precludes the use of the efficient Autocorrelation analysis because of poor estimation performance. For high pitched voices, the closed glottis interval can be too short for even the Covariance technique to perform reliably, often restricting this procedure to male voices.

The last problem entails the computation of $\{u_{g,n}\}$. Since the digital integral operation is very sensitive to bias, the glottal waveform becomes badly distorted unless high quality recording equipment is used [60]. It has been shown that $\{u_{g,n}\}$ computed using GIF is severely distorted by low-frequency phase distortion as well [25]. Because of these problems, glottal inverse filtering is limited to applications where it is cost effective to use high quality recording equipment and where it can be done off-line (such as for speech synthesis, as opposed to real-time speaker identity verification or speech translation). An alternative GIF strategy has been presented that alleviates the first two problems by using a recursive least-squares algorithm with a variable forgetting factor [7], rather than the frame-based Covariance estimation approach. Like standard GIF, this sequential adaptive filtering approach obtains the glottal waveform from the analysis residual signal, and is thus vulnerable to the same distortions such as bias. Nonetheless, both have been valuable tools in off-line analysis for speech synthesis.

Before leaving glottal inverse filtering, a comment should be made concerning the vocabulary-dependence of the glottal waveform obtained using this technique. Since the vocal tract estimate is taken over the closed glottis interval which is quite stationary, inverse filtering over this region will yield a glottal estimate that is essentially free of the vocabulary-dependent vocal tract information. However in the open glottis interval where the formants fluctuate and deviate from the estimate within the closed glottis interval, the inverse filtering operation will not totally remove the nonstationary formants. Consequently, the resulting glottal waveform will then be vocabulary-dependent.

2.5 Feature Parameterization and Evaluation

The accuracy of speech analysis algorithms such as those just discussed in

Sections 2.3 and 2.4 is no doubt crucial since it determines the amount of information that will be available to characterize the speech. It also determines the "quality" of the extracted information; that is, the degree to which the assumed information is free from the corruption with irrelevant information. Methods have been proposed that attempt to filter out the irrelevant information that might creep into the analysis results, while often resulting in some degradation of the desired information. Another approach is to transform the feature vectors in an attempt to attenuate the irrelevant information. Often a transformation of the feature vector can improve the quality of the feature, even if it is not corrupted by irrelevant information, by providing a weighting of the feature elements that is closer to optimal.

The next section describes common methods for evaluating the quality of the features extracted by speech analysis systems. Hence, Section 2.5.1 describes methods for parameterizing the vocal tract and vocal cord features; specifically, how these parameter vectors are transformed and finally compared.

2.5.1 Linear Transformations

For a certain utterance or segment spoken by a particular speaker i, let $\mathbf{f}_i = [f_{i1} f_{i2} \cdots f_{im}]^T$ denote the set of extracted features. If different feature estimates can be computed by analyzing a sufficiently large amount of data (i.e., training), then the intraclass (i.e. within-class) covariance matrix can be computed as

$$W_i = E((\mathbf{f}_i - \bar{\mathbf{f}}_i)(\mathbf{f}_i - \bar{\mathbf{f}}_i)^T) \tag{2.18}$$

where $E(\mathbf{f}_i) = \bar{\mathbf{f}}_i$ denotes the statistical mean of $\mathbf{f}_i$. Commonly it is assumed that $\mathbf{f}_i$ is normally distributed, in which case the pooled intraclass covariance matrix is simply the average of the covariance matrices over all speakers $W = E(W_i)$.

Given the above notation and using the maximum likelihood criterion, it can be shown [11] that speaker i minimizes the functional

$$D_{ML,i}(\mathbf{f}) = (\mathbf{f} - \bar{\mathbf{f}}_i)^T W_i^{-1}(\mathbf{f} - \bar{\mathbf{f}}_i) + ln|W_i|. \tag{2.19}$$

Many simplifications of this functional have been devised and used successfully in pattern matching applications. The first simplification is to assume that W_i is the same for all speakers so that the covariance matrix for each speaker is the same as the pooled covariance matrix. This simplification reduces storage requirements as well as computation since $ln|W_i|$ is constant for all speakers and may be dropped from the functional, resulting in

$$D_{M,i}(\mathbf{f}) = (\mathbf{f} - \bar{\mathbf{f}}_i)^T W^{-1}(\mathbf{f} - \bar{\mathbf{f}}_i). \tag{2.20}$$

This functional is known as the Mahalanobis distance. A further simplification is to assume that W is a diagonal matrix, $W = diag(\sigma_1^2, \sigma_2^2, \cdots \sigma_m^2)$, in which case the functional becomes a simple weighted Euclidean distance

$$D_{E,i}(\mathbf{f}) = \sum_{j=1}^{m} \frac{1}{\sigma_j^2}(f_j - f_{ij})^2. \tag{2.21}$$

When the features are virtually uncorrelated so that W is diagonally dominant, then commonly W is made into a diagonal matrix by simply dropping the nondiagonal terms. Yet sometimes the nondiagonal terms are not negligible, and it has been suggested that the (false) assumption does have an impact on performance [27]. So another alternative is to perform a simple diagonalization of W to remove correlations.

The above inverse variance weighted Euclidean distance is a very simple calculation once the feature variances are known. The motivation of applying a smaller weighting on features which have higher variances (i.e., are noisy) is to help prevent a large feature distance, $f_{j'} - f_{ij'}$, resulting from a particular feature j' deviating significantly from its mean, from dominating the calculation of the overall distance $D_{E,i}$. The only drawback with this assumption is that it does not take into account the interclass (i.e., between class) feature variations or, specifically, the separability of the features for a given population of speakers. For example, a feature having a large variance for each speaker may still be an effective discriminator if it happens to be very distinctive between speakers. An approach that incorporates intraspeaker variation in the form of another simple linear transformation is referred to as discriminant analysis and requires the computation of the interclass covariance matrix

$$B = E((\bar{\mathbf{f}}_i - \bar{\mathbf{f}})(\bar{\mathbf{f}}_i - \bar{\mathbf{f}})^T). \tag{2.22}$$

Due to the heavy training cost (as mentioned in Section 2.2.4), this method tends to be impractical and is rarely used for SIV.

2.5.2 Vocal Tract Feature Evaluation

Cepstral coefficients have been used successfully in SIV and ASR and have become the representation of choice when comparing vocal tract features. Since an AR vector $\{a_n\}$ can be transformed to the cepstral vector $\{\hat{c}_n\}$ using the recursion [10]

$$\hat{c}_n = a_n + \sum_{k=1}^{n-1} \frac{k}{n} \hat{c}_k a_{n-k} \text{ for } n \geq 1, \tag{2.23}$$

the efficient and stable linear prediction algorithms can be used to derive the cepstral coefficients. AR derived cepstral coefficients have been shown to yield similar SIV performance as those obtained using direct nonparametric cepstral analysis [14]. The mel-frequency cepstral coefficient representation introduced in Section 2.2.4 is a variation of this standard cepstral vocal tract representation, as is the popular delta cepstrum [10].

Cepstral coefficients are typically compared by using the simple Euclidean metric. The most complicated task in vocal tract feature matching, however, lies in time-scaling the sequence of cepstral vectors when an utterance is spoken at different rates between training and testing. The most direct approach to comparing a specimen utterance with utterances obtained in training involves template matching. A template contains the features that were extracted from a single utterance (e.g. the sequence of cepstral vectors) during the training session. Then, in verification, the feature set from the specimen utterance is directly matched against the template(s). Time-alignment of specimen and template is accomplished by compressing or expanding the time scale. Specifically, to align the features from utterances spoken at different rates, vectors obtained from the faster utterance may be repeated an appropriate number of times to match up with events occurring within the slower utterance. Due to time-variations within even a word, along with inconsistent endpoint detection, simple linear compression or expansion of the time scale does not provide accurate alignment. A popular approach for template matching, especially in systems that compare isolated words, involves dynamic time-warping (DTW). In DTW, the time-varying features of two utterances are nonlinearly aligned by computing various local distances (i.e., between particular vectors from within the specimen and template feature sequences), and matching the time-varying features that utilize the smallest local distances given some simple constraints.

Figure 2.9 shows a particular example of the DTW process for aligning and comparing the specimen feature set $\{a_i\}$ against the template $\{b_j\}$ (where $1 \leq i \leq I$ and $1 \leq j \leq J$). The matrix of dots represents the various distances which result when matching each of the specimen vectors with each of the template vectors. The size of the dot corresponds to the relative value of the distance between two vectors.

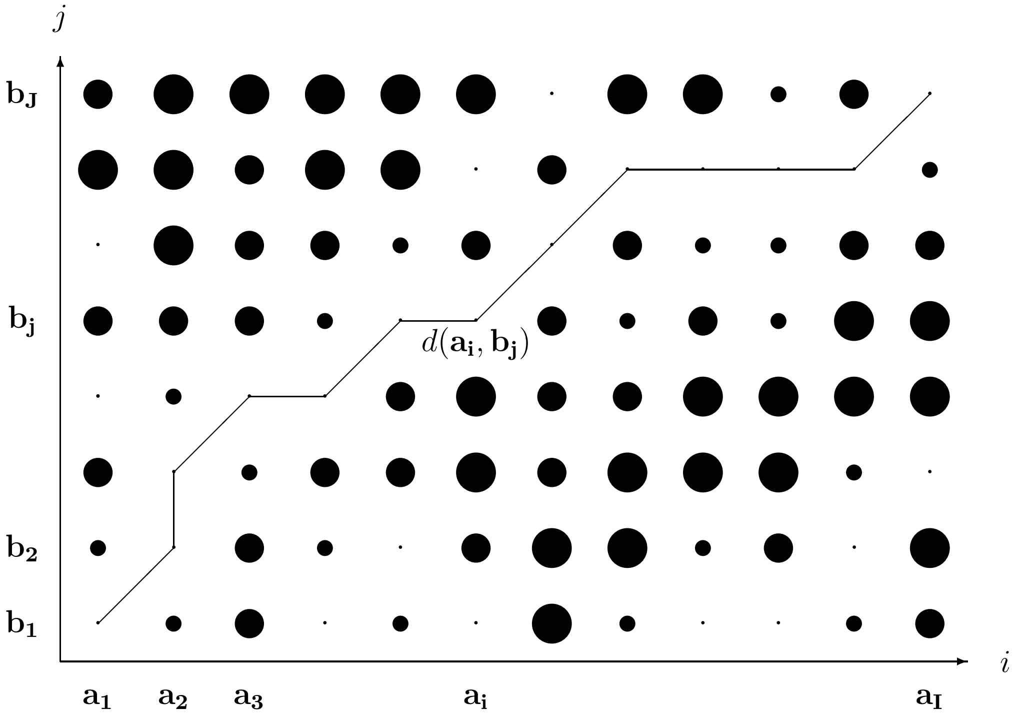

Figure 2.9: DTW process to align the vector sequences of the specimen utterance with the template utterance.

Starting with the first feature vectors from both sets, a_1 and b_1, the local time axis of the specimen utterance is, in effect, expanded by choosing the match (a_1, b_2). The same goes for the template utterance if (a_2, b_1) is chosen. No local time-scale modification occurs when (a_2, b_2) is chosen. The best local alignment is assumed to occur by choosing the match which yields the smallest distance, $d(a_1, b_2)$, $d(a_2, b_1)$, or $d(a_2, b_2)$. The same decision process takes place at each point in the alignment path (i, j) (where (a_i, b_j), $1 \leq i \leq I$ and $1 \leq j \leq J$, was the match that was last chosen) until the last vector of one of the sequences a_I or b_J is reached. The overall match of the two vector sequences can be represented by the normalized sum of the local distances computed along the chosen alignment path.

A less direct technique than template matching for comparing the specimen and training utterances is to incorporate a statistical model of the vector sequences corresponding to the training utterances. One such statistical model is referred to as the hidden Markov model (HMM) [28, 50]. In HMM, the sequence is modeled by a finite number of states (which is generally fewer in number than the length of the feature sequence), along with a state transition description which provides the time-alignment capabilities to the model. A specimen utterance is compared to a particular model by presenting the specimen vector sequence to the respective HMM. From this, a maximum likelihood computation is performed and the probability that the specimen sequence matches the model is provided. The most popular HMM application to SIV, for example, is to have an HMM trained for each speaker and each word in the vocabulary. Then a given specimen word is presented to the respective HMM of the claimed subscriber and the likelihood of a correct match is computed. If the likelihood is smaller than a predefined threshold, the identification claim is rejected.

VQ may be used in speech and speaker recognition systems in various ways. Perhaps the most obvious way is to substitute for the spectral parameters (i.e., the AR coefficients or equivalent representations such as cepstral coefficients or reflection coefficients) with the code words (via their addresses) simply to reduce memory requirements associated with storing templates. In the context of pattern matching, recall that VQ maps a vector to the "closest" (preferably in terms of perception) code word according to a particular distortion measure (which is similar to a distance measure but only requires the property that $dist(\mathbf{x}, \mathbf{y}) \geq \mathbf{0}$ with equality when $\mathbf{x} = \mathbf{y}$). Hence another approach for matching utterances spoken at different rates is to present the test speech utterance to the codebook which has been designed for a particular subscriber and determine if the test utterance came from the claimed subscriber by computing the overall average distortion. The idea is that a codebook trained for a single speaker will be sufficiently speaker dependent so that the same speaker will produce a lower average distortion than someone else using the same codebook. This approach has been applied in text-independent situations since it is relatively insensitive to temporal variations as well as to the sequence in which particular coefficients are input [53, 4, 61]. Consequently, no templates need to be stored and no time-alignment procedure is required.

There are various tradeoffs between using DTW, HMM, or VQ. HMM does a better job of retaining temporal features (which may provide speaker-dependent information) and, like VQ, requires fewer computations than DTW and less memory for storing the templates. However, the amount of training data required for HMM and VQ is generally a lot more than for DTW. Furthermore, updating the HMM is more difficult than updating a template (i.e. a simple replacement with the feature set obtained in a verification session that resulted in an accepted verification) which is important given that a speaker's voice is known to change over time [15]. Finally, it isn't as easy to interpret the HMM and VQ parameters for analysis and testing purposes. So if the time-alignment computations and large memory requirement aren't a problem for a particular application, the template approach may be desired. However, if training data is adequate and it is possible to periodically retrain or update the model, then HMM and VQ have an edge.

2.5.3 Likelihood of Errors

Now that the common techniques for comparing speech features have been introduced, probabilistic detection theory concepts will be utilized for evaluating the likelihood of the two types of speech/speaker recognition errors that can occur–false identifications (fid) (the system erroneously decides that the specimen feature set was produced by subscriber T_k when it was actually produced by subscriber T_j ($1 \leq k, j \leq N$, $j \neq k$)) and missed identifications (mid) (the system rejects the correct speaker). The results obtained here could also be tailored for several common speech recognition systems as well. For the purposes of this dissertation, however, the analysis will focus on speaker recognition and the tradeoffs between false and missed identification errors according to the decision threshold.

In order to analyze the speaker recognition errors formally given a particular criterion, it is necessary to introduce some notation. First, $H_{i,j}$ will represent the hypothesis that the comparison is between utterances from subscriber T_i and subscriber T_j ($1 \leq i, j \leq N$). For the case $i = j$ let $H_{i,j} = H_i$. The dissimilarity score resulting from the comparison (e.g., via Euclidean or other distance functional as presented in Section 2.5.1) between the specimen utterance and the template for subscriber T_i is x_i. The probability density function (PDF) with respect to the continuum of dissimilarity scores X associated with a particular hypothesis $H_{i,j}$ is denoted as $f_{X|H_{i,j}}(x|H_{i,j})$ and can be computed after extensive training.

The most common decision procedure in speaker identification is to compare the specimen feature set to each template feature set provided by the N subscribers in training and choose the subscriber whose template matched the specimen best (i.e., gave the lowest dissimilarity score). The probability of a false identification for such a system is equivalent to the probability that the specimen was actually produced by another subscriber, T_j, rather than the speaker T_k which yielded the lowest dissimilarity score x_k with the specimen. The region which represents the probability of these occurrences is shown in Figure 2.10 and the probability of a false identification is computed as

$$P(fid) = \sum_{j \neq k} P(H_j) \int_0^{x_k} f_{X|H_{j,k}}(x|H_{j,k})dx. \tag{2.24}$$

The lowest probability of a false identification results when x_k is small so that

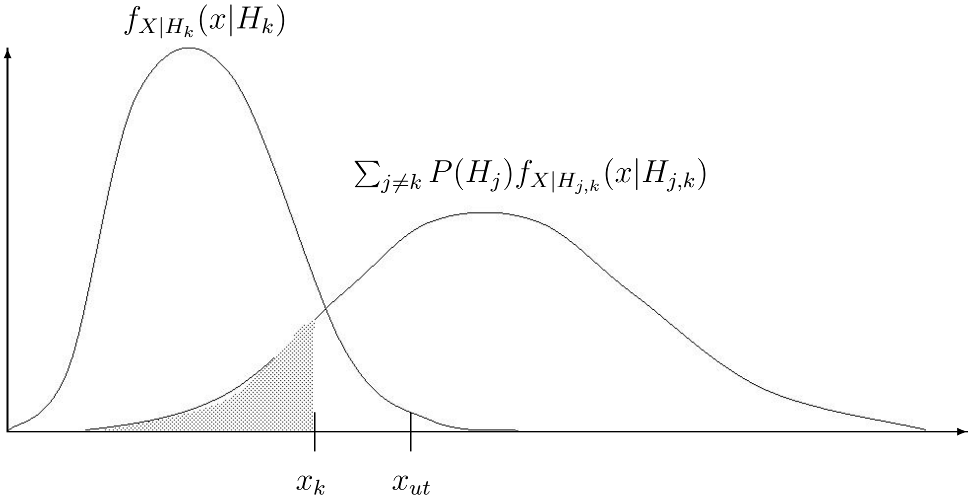

Figure 2.10: Speaker identification criteria.

the score is less likely to be associated with a hypothesis $H_{j,k}$ $(j \neq k)$ (where the template from T_k is hypothesized as being compared to a specimen utterance from another talker T_j ($j \neq k$)). In contrast, the highest probability of a false identification occurs when x_k is high so that it is closer to the mean of $f_{X|H_{j,k}}(x|H_{j,k})$. The range shown in Figure 2.10 can be better understood by thinking of the value of x_k as a fixed threshold $x_t = x_k$ so that, assuming the other dissimilarity scores x_j ($j \neq k$) are fixed, hypothesis H_k will be chosen whenever $x_k \leq x_t$. It is noted that if in this system speaker T_k is falsely identified, then naturally the correct subscriber has been missed. Hence $P(fid) = P(mid)$.

If the cost of a false identification is much greater than the cost of a missed identification, a "don't know" threshold may be added to the system to limit the probability of a false identification. Thus the "don't know" state is commonly achieved when the lowest score x_k exceeds a certain upper threshold x_{ut}. Consequently, the upper limit on the probability of false identification is

$$P(fid)_{max} = \sum_{j \neq k} P(H_j) \int_0^{x_{ut}} f_{X|H_{j,k}}(x|H_{j,k})dx. \tag{2.25}$$

As long as $x_k \leq x_{ut}$, $P(fid) = P(mid)$ as above. When $x_k > x_{ut}$, the "don't know" state results so that $P(fid) = 0$ and $P(mid) = 1$. The probability that this will happen for subscriber T_k is

$$P(x_k > x_{ut}) = \int_{x_{ut}}^{\infty} f_{X|H_k}(x|H_k)dx. \tag{2.26}$$

Given the PDF's for the N subscribers, the average probability of a "don't know" state occurring may be computed as

$$P(x_1 > x_{ut} \cup x_2 > x_{ut} \cup \cdots x_N > x_{ut}) = \sum_{j=1}^{N} P(H_j) \int_{x_{ut}}^{\infty} f_{X|H_j}(x|H_j)dx. \tag{2.27}$$

Normally a missed identification, particularly one resulting from the “don't know” condition, has some cost associated with it, perhaps related to the human factors aspect. Commonly the identification subject will be asked to repeat the utterance in order to resolve the “don't know” condition. Consequently the condition $x_k < x_{ut}$ is forced to hold so that the maximum probability of false identification is also $P(mid) = P(fid)$. Note that on the average, as long as $P(x_k < x_{ut}) > 0.5$, an identification decision will be made on the first utterance with no retest. In general, the probability that it will take $n - 1$ retests before resolving a “don't know” is $P(x_k \leq x_{ut})P(x_k > x_{ut})^{n-1}$. If there are a limited number of retests, M, that are allowed to be performed for each identification session, a “don't know” state must again be a possible outcome. So again, a “don't know” will result in a miss rather than risking a false identification. Hence the condition that $P(fid) = 0$ and $P(mid) = 1$ can again occur. However, now the probability of such an occurrence would be equal to

$$1-\sum_{n=1}^{M} P(x_1 > x_{ut} \cup x_2 > x_{ut} \cup \cdots x_N > x_{ut})^{n-1}(1-P(x_1 > x_{ut} \cup x_2 > x_{ut} \cup \cdots x_N > x_{ut})). \tag{2.28}$$

The other popular application of speaker recognition, SIV, simply gives a binary decision to either reject the identity of the speaker, or accept it as corresponding to the claimed subscriber. In this case, the missed identification and false identification errors are considered instead as false rejection and false acceptance errors (where accepting the SIV claim made by some speaker T_j instead of subscriber T_k is analogous to choosing the wrong subscriber in speaker identification). Using this analogy, the concepts above can also be applied to SIV. A common performance measurement of SIV systems is known as the equal-error rate (EER), which is the percentage of errors that occurs when the SIV decision threshold is set so that the probability of false acceptance equals the probability of false rejection, or

$$\int_0^{x_{ut}} f_{X|H_{j,k}}(x|H_{j,k})dx = \int_{x_{ut}}^{\infty} f_{X|H_k}(x|H_k)dx. \tag{2.29}$$

2.5.4 Vocal Cord Feature Representation

The glottal waveform is perhaps the most popular vocal cord representation, and has been used successfully to replicate voices with various vocal registers [1, 6, 22, 30]. Parameters that have been used to define the glottal waveform include pitch (i.e., fundamental frequency), glottal opening time, glottal closing time, and amplitude, as defined in Section 2.1. Other effects, such as source-tract interaction, have also been incorporated by making slight changes to the wave [1, 47]. Certain perceptually significant characteristics are more apparent in the frequency domain, prompting new glottal excitation generators that directly control the spectrum. Common spectral characteristics include spectral tilt (i.e., magnitude roll-off with increasing frequency), relative harmonic intensities, and interharmonic noise [6].

Extracting glottal features from the time-domain signal is attractive, as the features are easy to relate to physical speech characteristics. Furthermore, since the vocal cords vibrate relatively consistently over all voiced speech, it is possible to model typical vocal cord operation over a short voiced segment (containing at least one pitch period). Doing so in the time-domain will generally require less data than short-time analysis techniques (such as spectral and statistical methods) require, resulting in features that can be easily normalized to pitch. This also results in shorter delays for real-time applications.

On the other hand, variations of the glottal signal over all speakers makes automatic time-domain analysis a difficult task. For example, the opening pulse for breathy voices may be weak, and there may not be a closed interval as the glottis may never close completely. Vocal fry (or "creaky") voices, on the other hand, may not have a detectable opening pulse and may contain more than one excitation per pitch period. The above properties of the glottal signal for breathy and vocal fry voices may help distinguish a particular speaker, but also make pitch detection and consistent estimation of glottal events very difficult.

Other sources of inconsistency of the glottal signal may arise as a result of changes in airflow and obstructions, such as mucous. For example, the glottal signal may contain an occasional large pulse that may be mistaken for an excitation event.

Or if the airflow from the lungs is not strong enough, the glottal signal may alternate between breathy excitation and unvoiced excitation (i.e., no glottal vibration). The glottal signal has also been observed to be inconsistent for some vocal fry voices and for speakers with certain speech pathologies. In these cases, attempting time-domain analysis over a couple of pitch periods may be quite difficult and provide a poor generalization of the glottal signal.

The reason for posing the above difficulties (rare as they may be) before even describing the characteristics of the glottal signal is to serve as a forewarning when designing a time-domain extraction routine that is based on overly optimistic assumptions. For example the problem of pitch detection has long been acknowledged as a very difficult task [22], hence designing an extraction routine that is based heavily on the assumption of accurate pitch detection could result in a set of computed features that bear no resemblance to the actual ones when the assumption is violated.

Heeding the above warnings, only the primary time-domain features of the glottal signal are discussed here as well as their variation for vocal registers at various extremes. As illustrated in Figure 2.1, the most significant glottal features are the glottal opening and glottal closing phases. In the glottal waveform, the glottal opening phase appears as a pulse (i.e., makes up about a tenth to a third of the entire pitch period, depending upon the speaker) and always precedes the glottal closing phase, which shows up as a sharp drop. As a result, except for larger males and speakers with vocal fry voices that require little vocal effort to open the glottis, the glottal opening interval is usually very distinguishable. The glottal closing interval is also quite distinguishable, particularly for lower pitched speakers. The glottal opening and glottal closing phase make up the open glottis interval (OGI). The closed glottis interval (CGI) makes up the remainder of the period and is often assumed to be the stationary region within the pitch period. The CGI is typically much shorter for female speakers than for male speakers and may be difficult to distinguish for breathy speakers as the glottis may not close completely.

Using the above generalizations, one might do quite well identifying main events within glottal signals for the vast majority of speakers. The glottal opening pulse is typically quite large and easily identifiable. Although such a pulse may not occur for some low-pitched male speakers, their closing phase tends to be quite sharp and easily identifiable. If LPC analysis is being performed on the speech, the prediction error signal can also be used to assist in locating the glottal closure interval, since the largest prediction errors occur during main excitation. Several other parameters have also been used to locate specific glottal events, particularly in the context of pitch prediction [22]. Therefore, it is possible to locate a main excitation within a certain interval quite consistently, even for problematic voices such as those that are breathy or have multiple excitations, simply by finding the largest peak and/or the sharpest drop within an interval that contains at least one pitch period.

Chapter 3

Adaptive Forced Response Inverse Filter (AFRIF) Model

A key feature of the speech analysis procedure proposed in this chapter is the adaptive filtering operation that is performed to obtain the vocal tract and glottal waveform estimates. To demonstrate the operation of recursive filtering on speech, Figure 3.1 shows time-varying spectral estimates from the standard recursive, least-squares algorithm over a single speech period. These 41 speech samples were obtained by digitally recording a male child (corresponding to CMBSB in the following chapter) making an /uw/ vowel sound, sampled at 8 kHz. These samples correspond to samples 6610 to 6650 of the waveform in Figure 3.2. As with standard LPC analysis, the adaptive filter uses a 10th order autoregressive model.

The high temporal resolution provided by these time-varying spectral estimates allows continuous tracking of the formant behavior as they are affected by the vocal cord behavior. Although the first 10 estimates are quite stationary, in the next

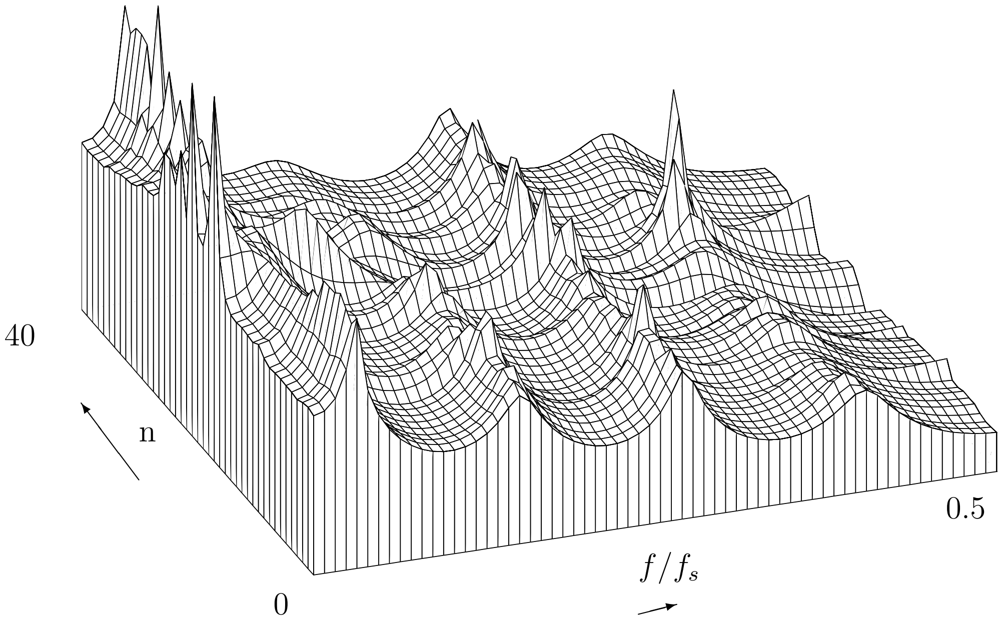

Figure 3.1: Instantaneous spectra from RLS.

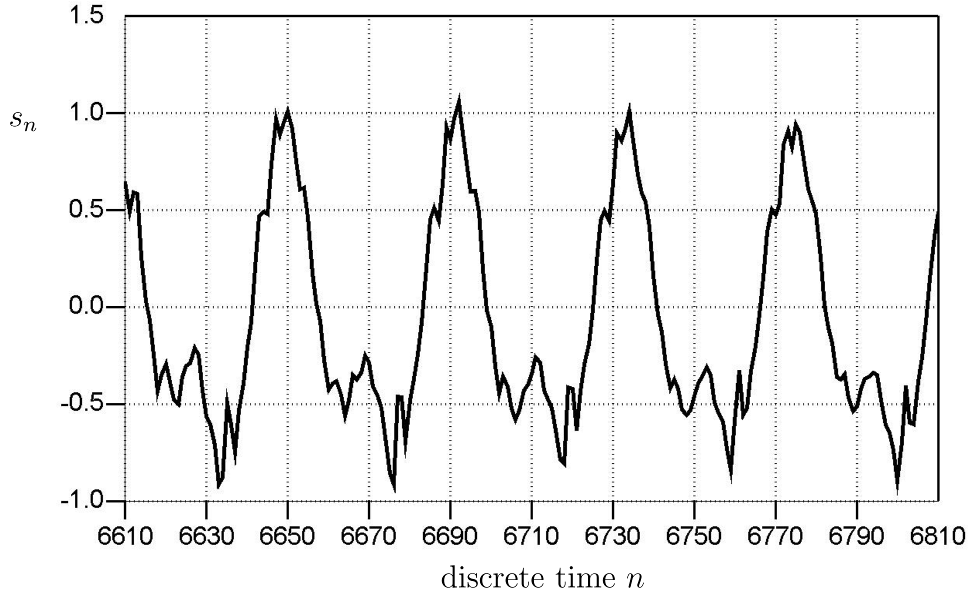

Figure 3.2: Speech record used for illustration.

5 to 10 samples the formants dampen and broaden. At the 25th estimate the formants are excited, and the bandwidths suddenly decrease. For the following 5 to 10 samples, the energy in the high frequency formants increases, while the energy in the lowest formants decreases.

These events correspond with the results observed in other studies, which have gone a step further by relating these events to the specific vocal cord mechanics. The 25th sample at which the formants are excited is the key event, because it results in the generation of the next period as shown in Figure 3.2 corresponding to $n = 6635$. This is known as the main excitation and occurs at the beginning of the glottal closure phase [22]. A well-known consequence of this is the strong excitation of the higher harmonics that was observed for the following 5 to 10 samples [31]. So accordingly, the 10 to 15 sample interval prior to the 25th sample corresponds to when the glottis is open. The observed time-varying frequency response certainly confirms this assumption, as other studies have also found that glottal opening indeed tends to cause a damping of the lower frequency harmonic [60].

In summary, the above observations of the time-varying frequency response complement those observed in other studies. Due to the popularity of the glottal waveform, made possible by glottal inverse filtering, over expensive physical measurements, these events are quite well understood today. However, the significance of the analysis made here is that, unlike in the standard glottal inverse filtering technique, the above observations of glottal events were made using the prediction coefficients rather than the prediction error signal. This was only possible because of the continuous tracking provided by the recursive analysis algorithm, allowing a new set of coefficients to be computed for every sample within the pitch period.

In this chapter, Sections 3.1 and 3.2 will describe how the proposed AFRIF procedure obtains the vocal tract and vocal cord information from time-varying spectral estimates of the input speech signal. Section 3.3 provides an analysis of the glottal waveform modeling capability of the AFRIF procedure using artificial speech, generated from the classical glottal waveform model, as the AFRIF input.

3.1 Adaptive Vocal Tract Modeling

As observed in Figure 3.1, the recursive filter analysis algorithm provides several linear predictive formant estimates over each pitch period. It was also observed that, although the linear prediction parameters modeled the vocal tract (via the formants), many of the parameter vectors were affected by the vocal cord operation as well. To accurately model the vocal tract without excitation effects, a linear prediction parameter vector should be chosen from within the closed glottis interval (CGI), as is done in glottal inverse filtering.

Since the vocal tract features are relatively stationary over 20 to 30 ms intervals, it is not necessary to detect the closed glottis intervals and store the vocal tract parameters for every pitch period. Rather, the proposed analysis algorithm attempts to identify each closed glottis interval but chooses only one vector, say $\mathbf{a}_{n_c}$, within a given speech frame from which to extract the vocal tract estimate. Note that the amount of memory required to store the parameters, extracted with this fixed frame pitch-synchronous analysis, depends only upon the length of the utterance and not on the pitch.

The recursive least-squares (RLS) algorithm has been selected to perform the adaptive filtering operation for computing the time-varying filter coefficients $A_n(z)$. The RLS procedure has been shown to converge after about $2p$ data samples, where p is the filter order, regardless of the properties of the data (i.e. the conditioning, as described in Haykin [20]). This convergence property makes the RLS procedures attractive for accurate glottal signal tracking and formant extraction from short speech intervals or from high pitched speech. Hence, RLS is capable of extracting accurate vocal tract estimates from much shorter closed glottis regions than is possible with the Covariance method used for glottal inverse filtering.

3.2 Forced Response Inverse Filtering for Vocal Cord Modeling

Figure 3.1 demonstrated how the prediction coefficients obtained from the recursive analysis algorithm contain a description of the glottal behavior. A procedure is described here that aims to model this behavior using a time-domain waveform, rather than by analyzing the variations of the p prediction coefficients. Section 3.3 will show that the resulting signal tracks the classical glottal waveform, thus providing an alternative for extracting the classical glottal waveform.

The first step of the procedure is to obtain the step response r_n of the time-varying filter defined by the prediction coefficients:

$$r_n = u_n + \sum_{i=1}^{p} a_{n,i} r_{n-i}. \tag{3.1}$$

Finally, to model how the statistics of this signal change over a pitch period as a result of the vocal cord behavior, the step response is injected into the time-invariant inverse filter defined by the prediction coefficient vector $\mathbf{a}_{n_c}$, obtained in Section 3.1:

$$g_n = r_n - \sum_{i=1}^{p} a_{n_c,i} r_{n-i}. \tag{3.2}$$

The overall flow diagram of this proposed procedure is given in Figure 3.3.

Note that since the time-invariant filter, $A_{nc}(z)$, represents the inverse of the vocal tract model for the corresponding period (which is assumed to be stationary), its output $\{g_n\}$ is expected to be virtually free from vocal tract information. If the speech signal itself is stationary over the particular pitch period (which it never is), the output of the inverse filter, $A_{nc}(z)$, will be unity. Results different from unity, within the pitch period, thus reflect the nonstationarities associated primarily with the vocal cord operation. Hence the inverse filter output $\{g_n\}$ is referred to as the

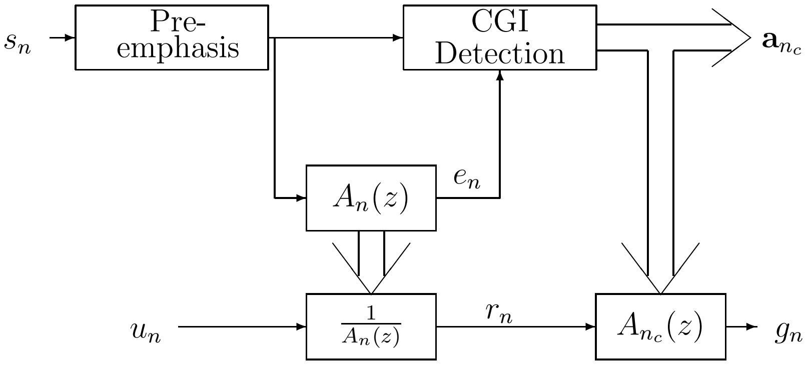

Figure 3.3: Flow diagram illustrating the proposed analysis procedure.

glottal signal. Figure 3.4 shows the glottal signal estimate for first half of the speech period in Figure 3.2. Section 3.3 will show how the resulting glottal signal estimates correspond to the popular differential glottal waveform introduced in Section 2.4.

The proposed approach is noted to have many similarities to classical glottal inverse filtering. The main differences of the two techniques lie primarily in how the vocal tract estimates are obtained, and in the signal used for inverse filtering. Recall that in conventional glottal inverse filtering [60], preprocessing initially takes place to find the closed glottis interval from which the vocal tract estimate is obtained. The proposed approach, on the other hand, recursively performs linear predictive analysis at each speech sample and then selects a linear prediction vector at a particular sample within the closed glottis interval as the vocal tract estimate. The remaining instantaneous prediction coefficients do not get discarded (as do those

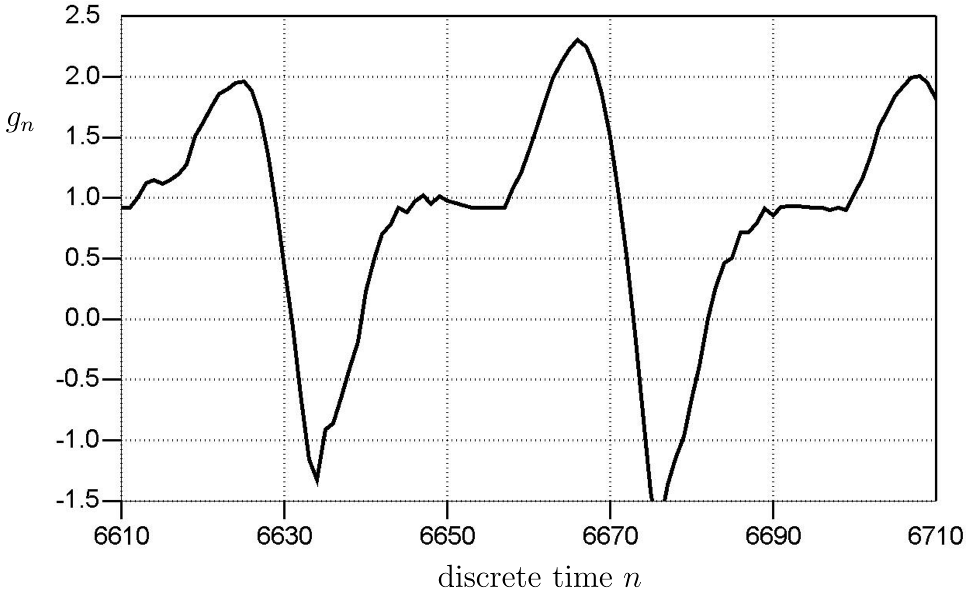

Figure 3.4: Glottal signal estimate.

obtained in the preprocessing procedure of glottal inverse filtering), instead they are characterized by their time-varying unit step response to derive the glottal signal. This leads to the other main difference between glottal inverse filtering and the proposed procedure; specifically the signal to be inverse filtered. Recall that glottal inverse filtering techniques use the speech waveform. The proposed approach, however, uses the step response which is basically the speech signal after much of the random noise component and excitation harmonics have been removed and replaced by the unit step signal.

3.3 AFRIF Vocal Cord Modeling Analysis

This section analyzes glottal behavior, as was observed in the time-varying frequency response in Figure 3.1, which will be shown to result in an estimate of the classical glottal waveform estimate $\{g_n\}$. In order to provide a more controlled analysis of the glottal signal extraction operation of the AFRIF procedure, various phonemes of synthesized speech will be generated using an artificial glottal waveform source. The RLS algorithm will be used for the adaptive filtering operation of the AFRIF procedure and will demonstrate the effects of convergence and filter order on glottal waveform modeling performance.

3.3.1 Experiment Setup

The speech synthesis system used in this section is based on the Klatt formant synthesizer [30] and uses the Liljencrants–Fant (LF) glottal waveform model as in the revised Klatt synthesizer [31]. The LF model represents the glottal pulse with the function

$$g(t) = at^2 - bt^3 \tag{3.3}$$

where a and b determine the open quotient, glottal pulse skew, and voicing amplitude. This section makes use of the differential glottal waveform representation

$$g'(t) = 2at - 3bt^2. \tag{3.4}$$

These representations can be interchanged using simple first-order filters to convert between the glottal and the differential glottal waveform models. Unless otherwise stated, the default values given in [31] were used throughout the analysis. This includes the default sampling rate of 10 kHz. The advantage of this analysis is that the glottal waveform of the (synthesized) speech is completely known, so the glottal estimates from AFRIF can be directly evaluated. Furthermore, since the voicing, pitch, location of all of the glottal events, and formants are known, automatic pitch, endpoint, and voiced/unvoiced detection routines are not necessary in this analysis, and are manually specified according to the speech synthesis model parameters. Hence, the location within each pitch period at which the AFRIF procedure extracts the vocal tract estimate is manually specified as the tenth sample prior to each glottal opening pulse that occurs in the LF glottal waveform model.

3.3.2 Experiment on Typical Vowels

The AFRIF procedure, using 10th order RLS analysis, is first applied to synthesized speech of the long vowel sound /uw/ as in the word *do* shown in Figure 3.5. Figure 3.6 shows a period of the glottal signal estimates as well as the actual differential glottal waveform used to generate the synthesized speech denoted by the dashed-dotted line. As will be the case for the remaining figures in this section, the analysis was performed using exponential forgetting factors of $\lambda = 0.8$ and $\lambda = 0.9$ denoted by a solid line and a dashed line, respectively. The

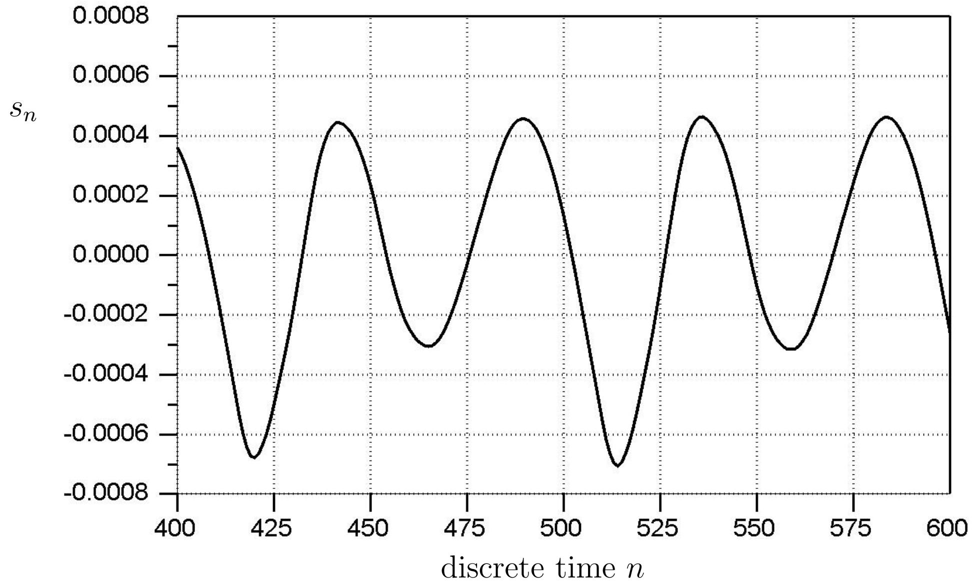

Figure 3.5: Synthesized speech of /uw/.

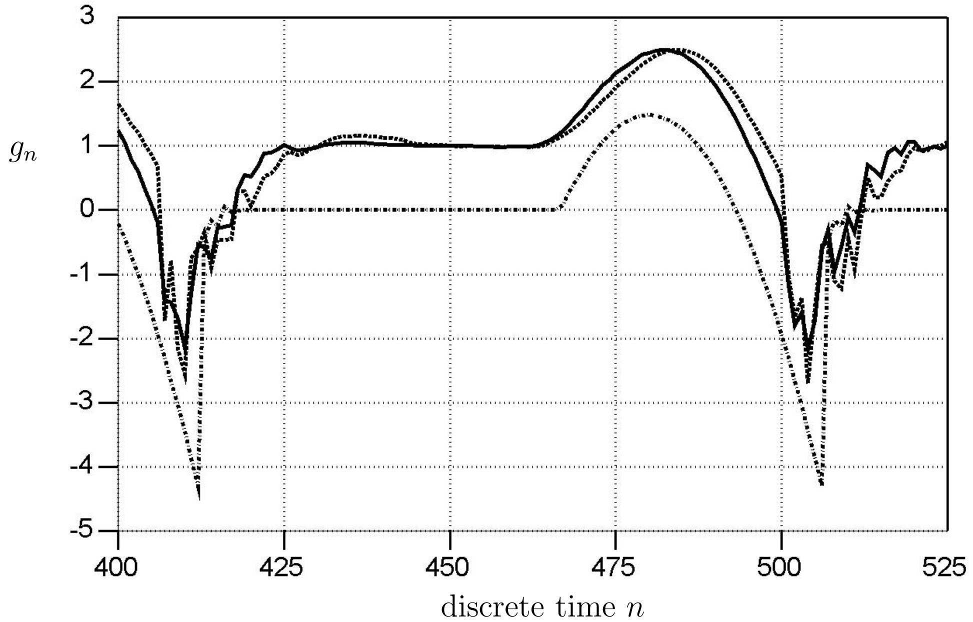

Figure 3.6: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /uw/ from a synthesized waveform (dashed-dotted).

forgetting factor is a variable applied to the RLS cost functional as follows

$$\sum_{i=1}^{n} \lambda^{n-i} |e_i|^2, \tag{3.5}$$

and provides a way to reduce the influence of past data and thus improve estimation while tracking nonstationary data.

Using the same artificial glottal signal source as for the synthesized /uw/ vowel, the long vowel /iy/ as in the word *see* was also synthesized as shown in Figure 3.7. Figure 3.8 reveals similar results as were observed for the analysis of the synthesized vowel /uw/. Open quotient (OQ) is another glottal waveform parameter that affects speech characteristics, and has been associated with breathy phonation [31]. To determine how AFRIF is able to track smaller glottal opening pulses resulting

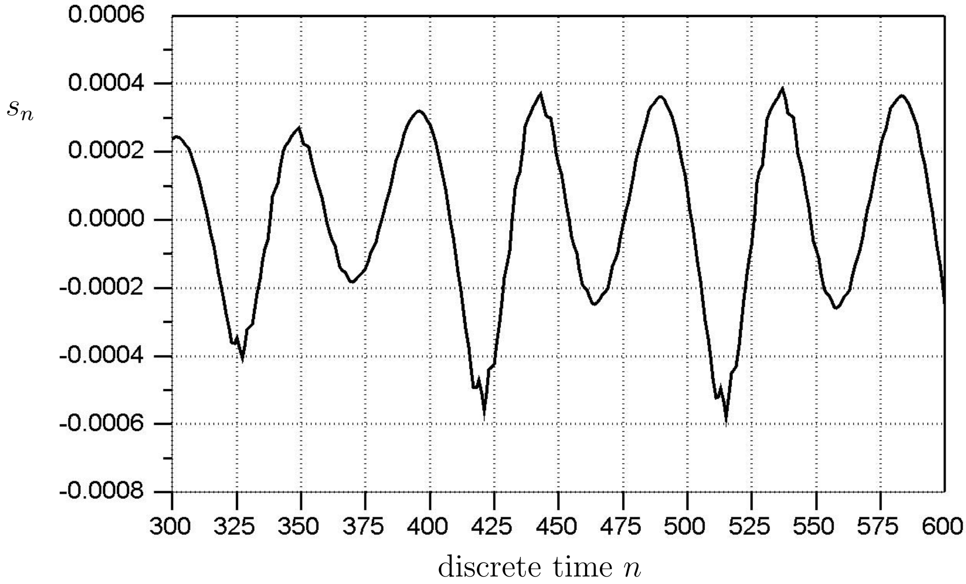

Figure 3.7: Synthesized speech of /iy/.

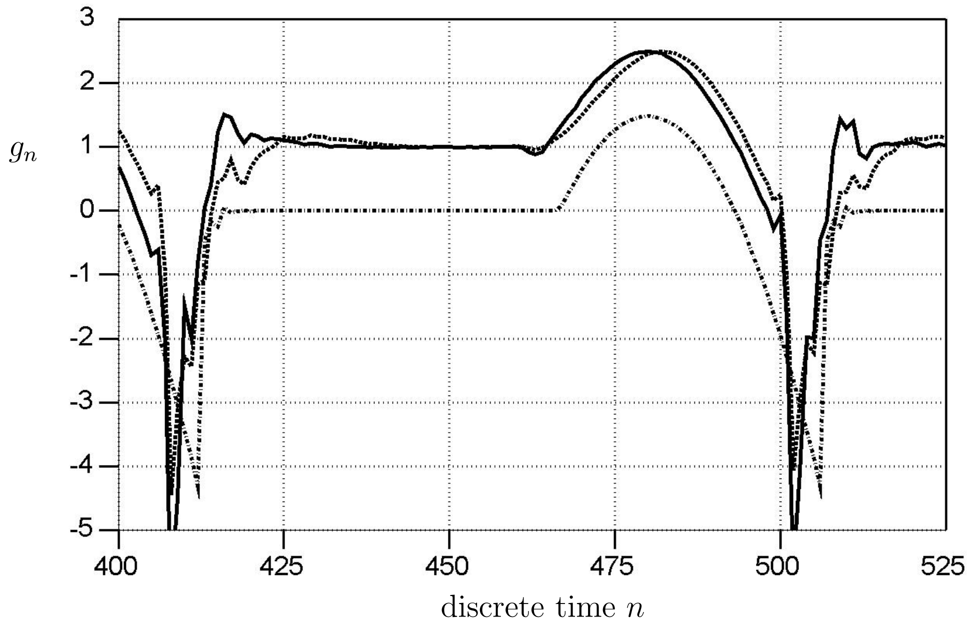

Figure 3.8: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ from a synthesized waveform (dashed-dotted).

from a small OQ, the open quotient was decreased to less than half of its default value of 43%, as in Figure 3.7, to 21% given in Figure 3.9. As observed from the estimated glottal signals in Figure 3.10, the AFRIF procedure models the smaller glottal opening pulse associated with a decreased OQ.

The above analysis was also performed for different filter orders. Figures 3.11 and 3.12 show the results when analyzing the synthesized vowel /uw/ in Figure 3.5 with 14th and 18th order adaptive filters, respectively. Figures 3.13 and 3.14 correspond to 14th and 18th order analyses of the vowel /iy/. Finally, Figures 3.15 and 3.16 give the results of the 14th and 18th order analysis of the speech with a smaller open quotient corresponding to Figure 3.9.

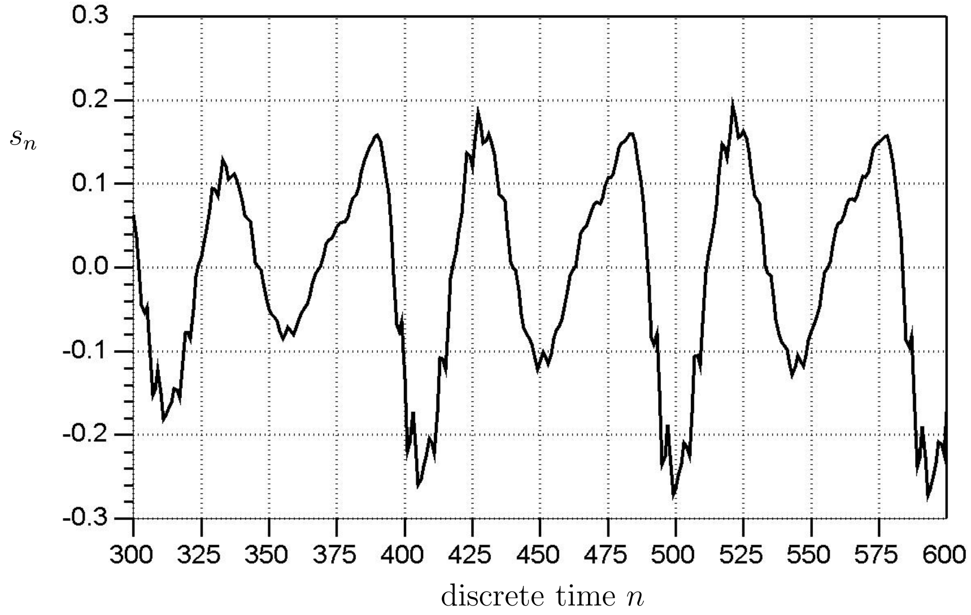

Figure 3.9: Synthesized speech of /iy/ with 21% OQ.

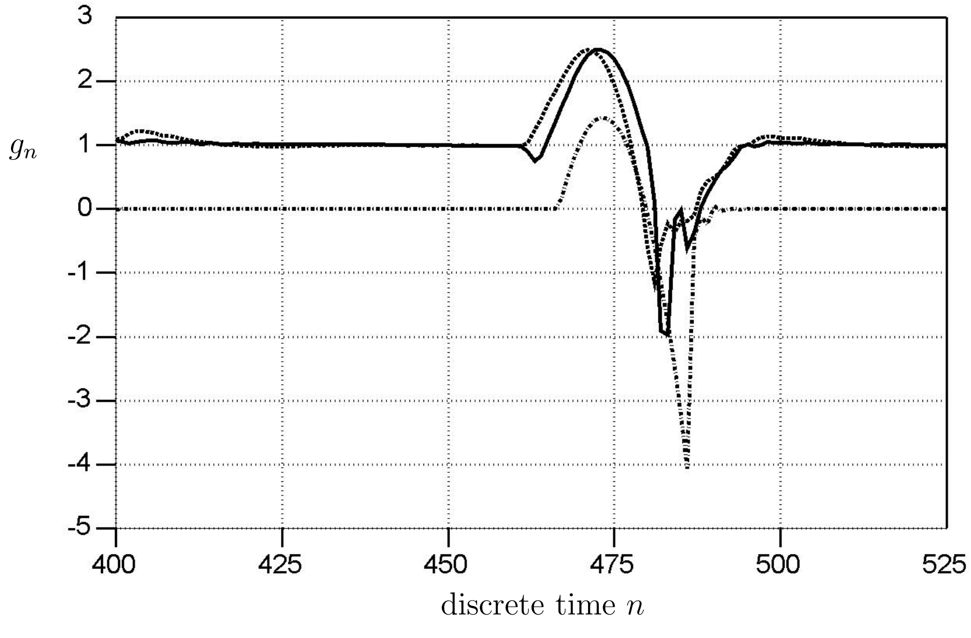

Figure 3.10: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ from a synthesized waveform (dashed-dotted) with 21% OQ.

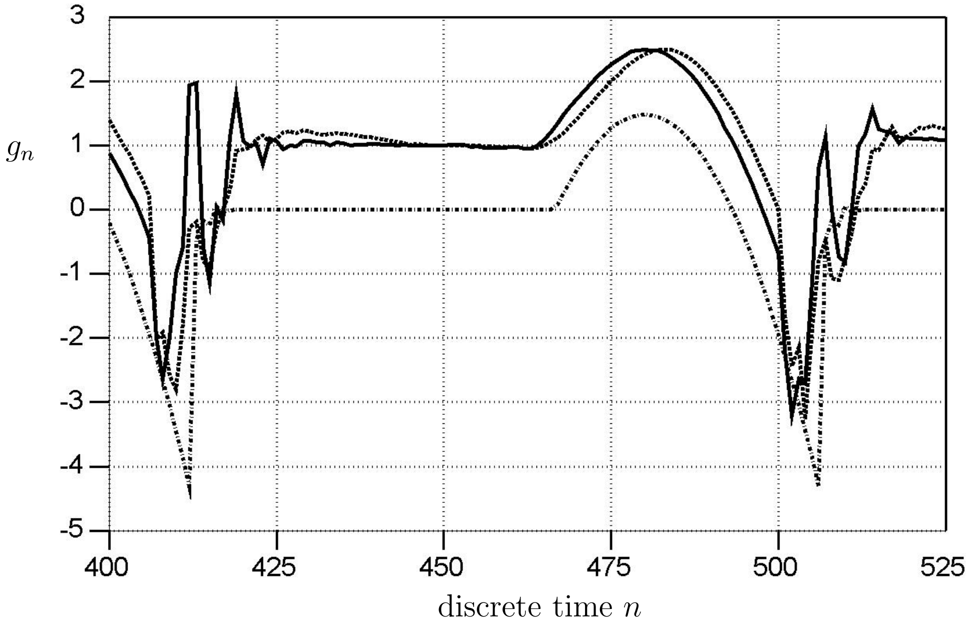

Figure 3.11: 14th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /uw/ from a synthesized waveform (dashed-dotted).

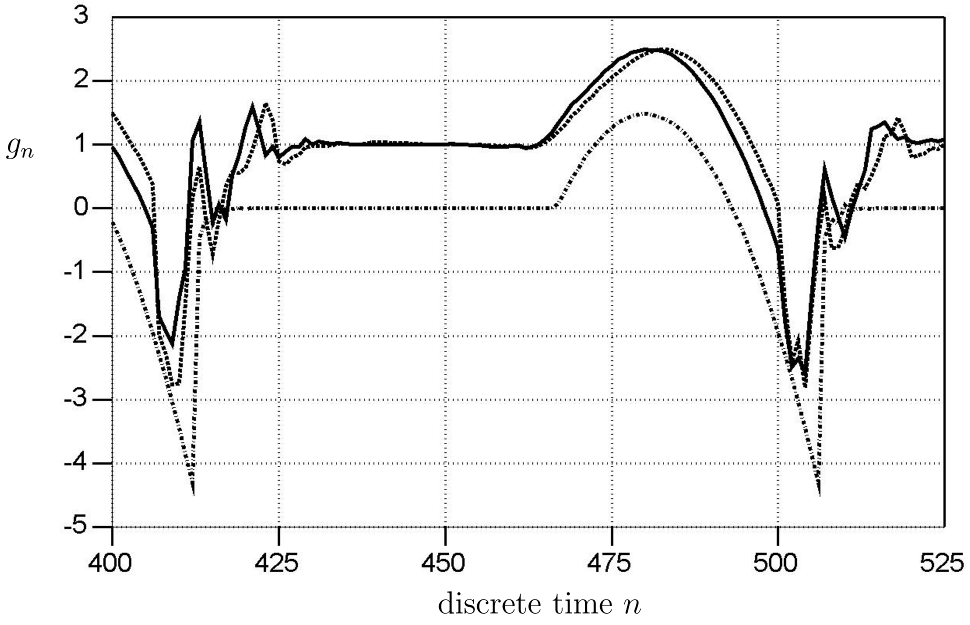

Figure 3.12: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /uw/ from a synthesized waveform (dashed-dotted).

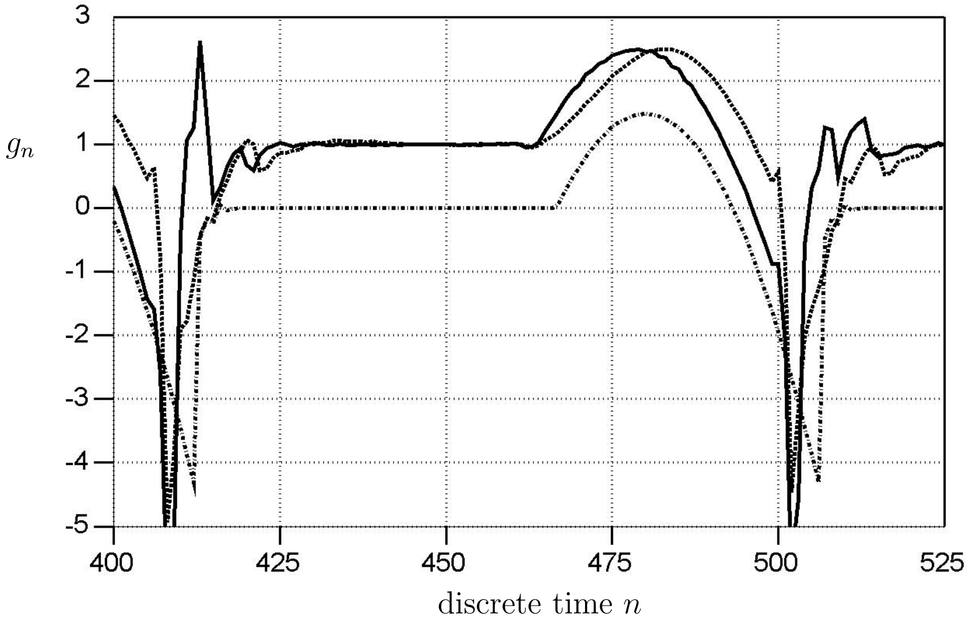

Figure 3.13: 14th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ from a synthesized waveform (dashed-dotted).

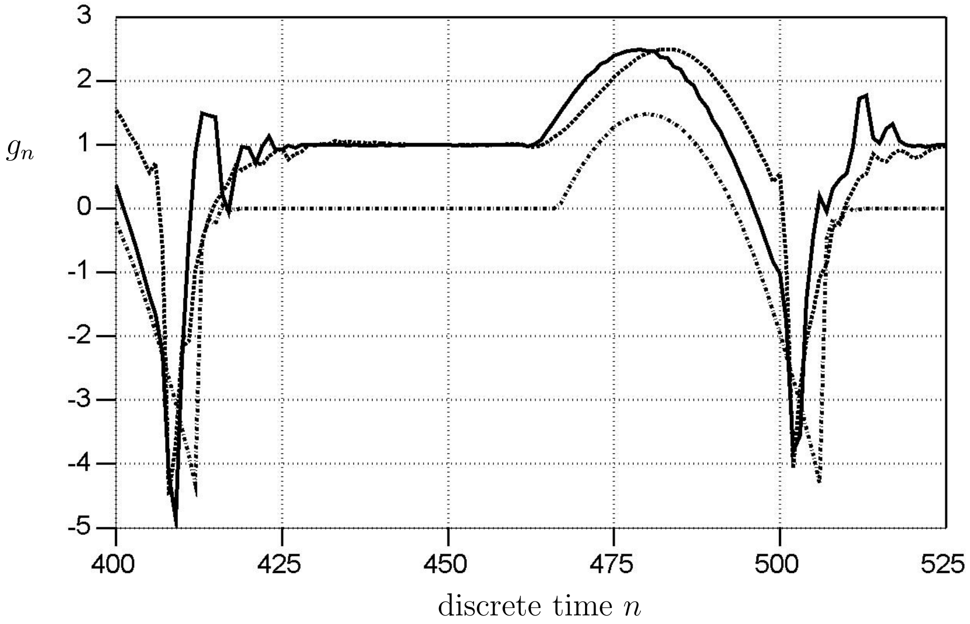

Figure 3.14: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ from a synthesized waveform (dashed-dotted).

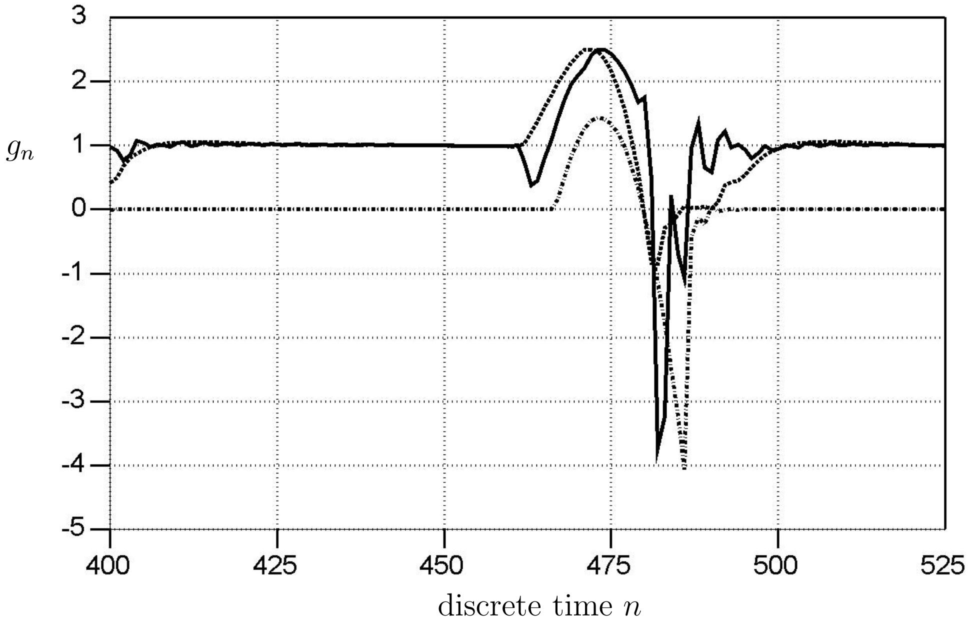

Figure 3.15: 14th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ from a synthesized waveform (dashed-dotted) with 21% OQ.

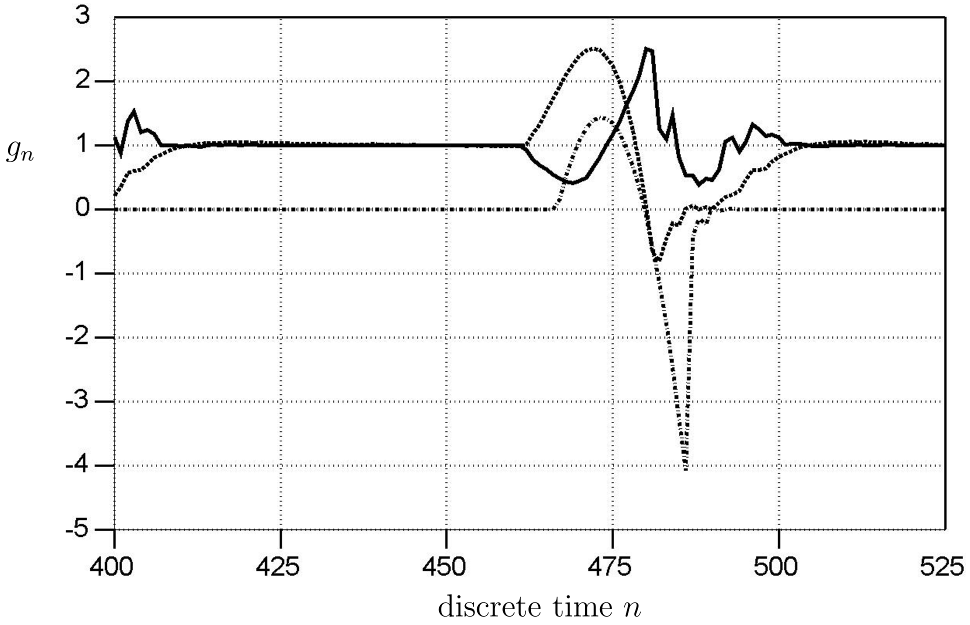

Figure 3.16: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ from a synthesized waveform (dashed-dotted) with 21% OQ.

3.3.3 Experiment Observations

For the most part, the results in the above experiment were quite insensitive to changes to the data and analysis parameters. Comparing Figures 3.6, 3.11, and 3.12 obtained from the synthesized /uw/ vowel, and Figures 3.8, 3.13, and 3.14 from the synthesized /iy/ vowel, shows that the glottal waveform modeling performance tends to be quite consistent for the different analysis filter orders.

The portion of the waveform that was the most difficult for the AFRIF procedure to track was during the glottal closure interval. This is not completely surprising since main excitation occurs in this interval. For example, consider Figure 3.8. In this plot, the glottal closure interval occurred from sample 478 to sample 505. The AFRIF procedure was able to track the waveform until the 499th sample where the estimated glottal signals sharply decrease. Furthermore, the estimated glottal signals were not able to instantaneously reach the constant closed-glottis interval which occurred at sample 506 but tended to oscillate (for the small λ case) or gradually climb (for the large λ case) to the constant value of one after approximately 10 to 20 samples.

The results in the above experiment confirmed that larger forgetting factors translated to slower waveform tracking, particularly at the transition from the closed-glottis interval to glottal opening. For example in Figure 3.14, the waveform modeling at the glottal opening transition at $n = 465$ was not as sharp for the $\lambda = 0.9$ case and lagged the estimate obtained using $\lambda = 0.8$. On the other hand, smaller forgetting factors occassionally yield a dip in the estimate within the glottal opening interval, as observed over the interval from $n = 460$ to $n = 480$ in Figure 3.16 for the $\lambda = 0.8$ case. Similar results will be observed in Section 4.1.1 when inaccurate RLS estimates, that have not adequately converged after excitation, are used as vocal tract estimates. For problems associated with small forgetting factors, inaccurate vocal tract estimates can also result due to another RLS estimation error referred to as misadjustment [20].

3.3.4 Experiment on Breathy Vowels

A common speech quality relates to the breathiness of a speaker. For example, female speakers are often characterized according to their breathy phonation [31, 47]. The revised Klatt synthesizer generates breathy speech using its aspiration noise parameter (AH) which simply adds random noise to the differential glottal waveform [31]. Figures 3.17 and 3.18 show the synthesized vowels /iy/ and /uw/ as synthesized for Figures 3.7 and 3.5, respectively, but with the aspiration noise parameter set to 52 dB (instead of the default where no aspiration noise is assumed; i.e., 0 dB). The 10th and 18th order AFRIF glottal waveform analyses of the breathy vowel /iy/, and the 10th and 18th order AFRIF analyses of the breathy vowel /uw/, are shown in Figures 3.19 through 3.22.

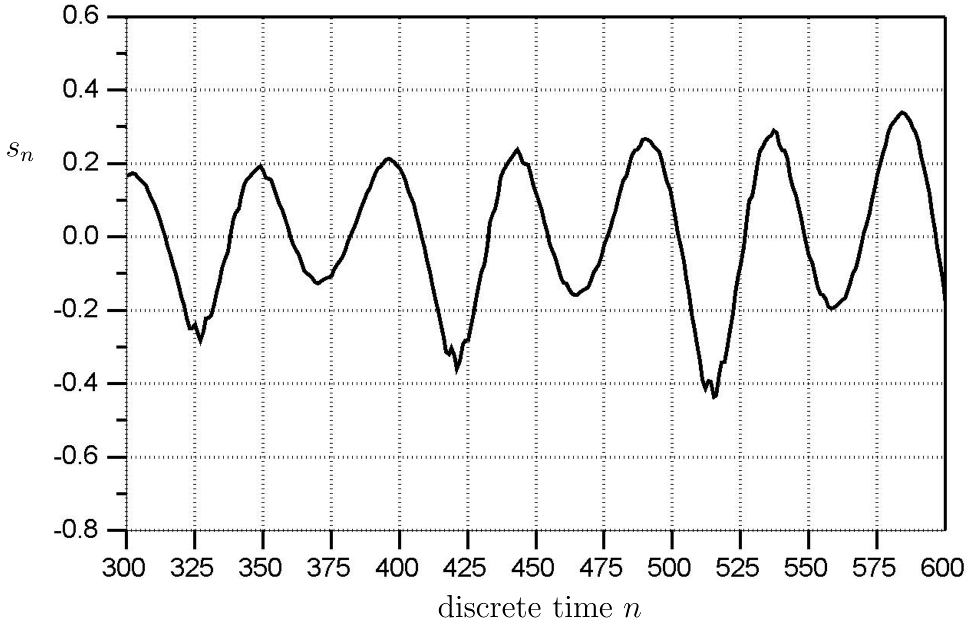

Figure 3.17: Synthesized breathy speech of /iy/.

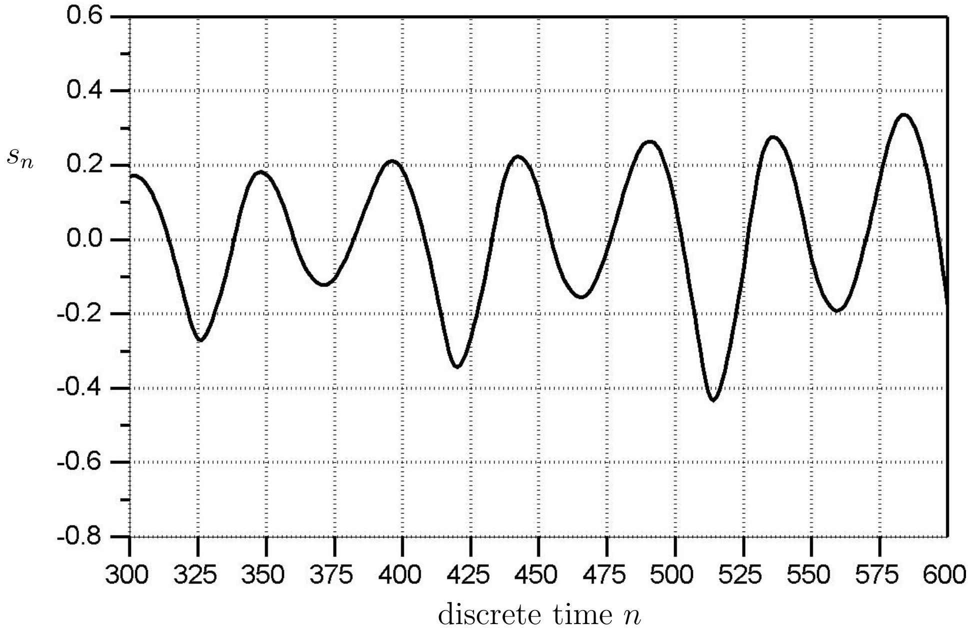

Figure 3.18: Synthesized breathy speech of /uw/.

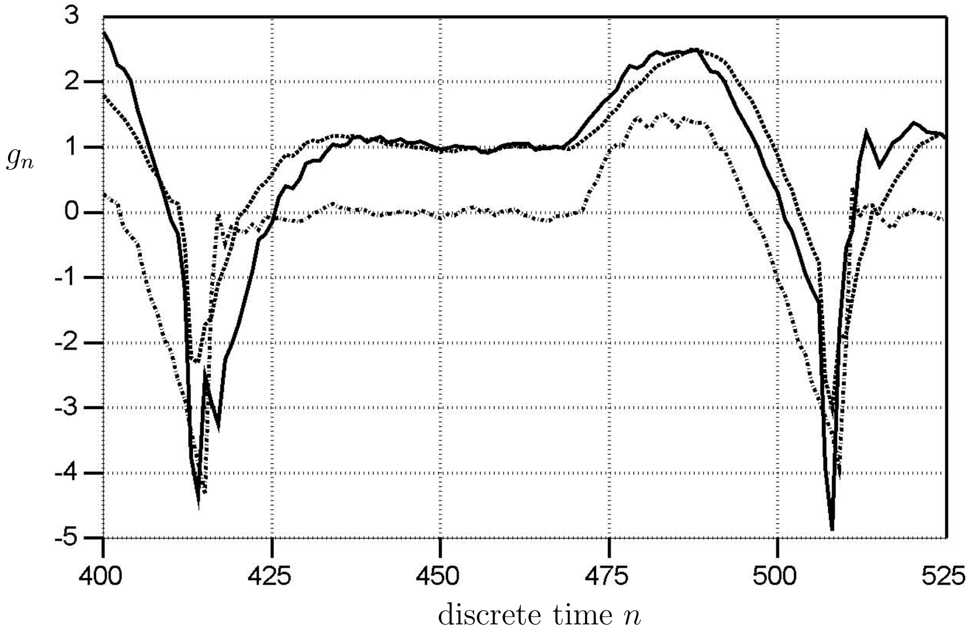

Figure 3.19: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of breathy /iy/ (Figure 3.17) from a synthesized waveform (dashed-dotted).

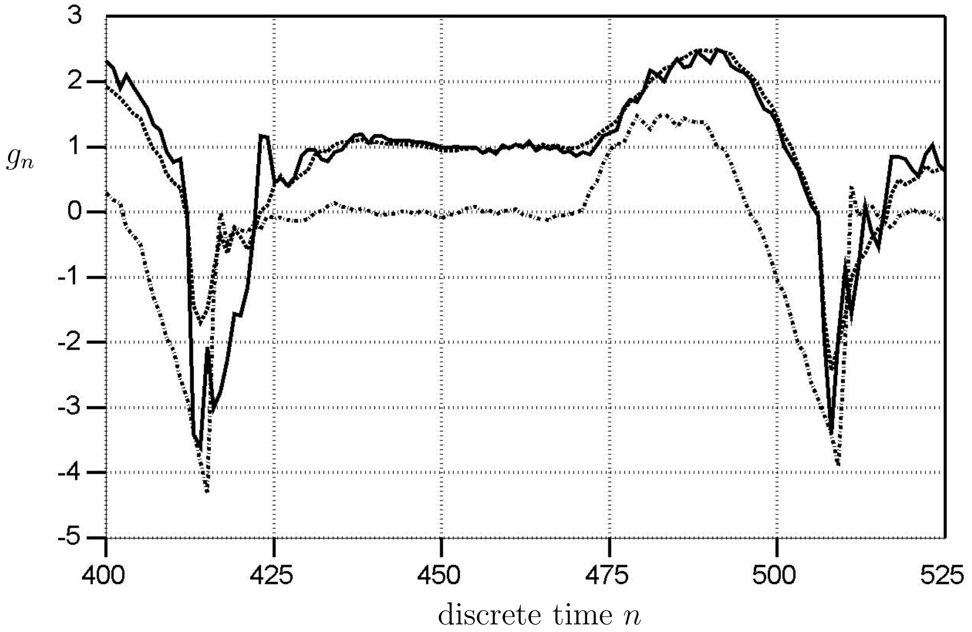

Figure 3.20: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of breathy /iy/ (Figure 3.17) from a synthesized waveform (dashed-dotted).

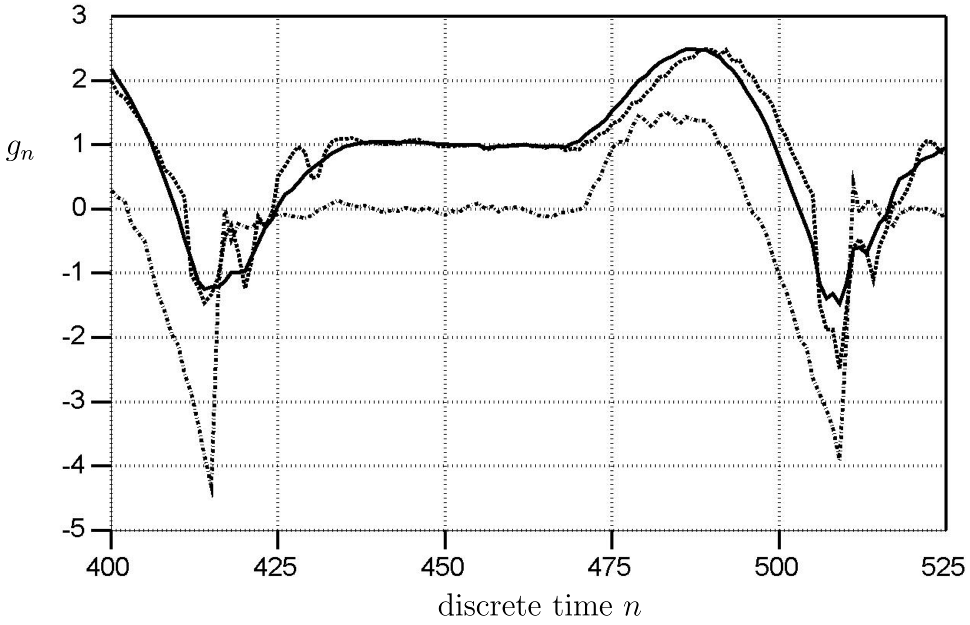

Figure 3.21: 10th order Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of breathy /uw/ (Figure 3.18) from a synthesized waveform (dashed-dotted).

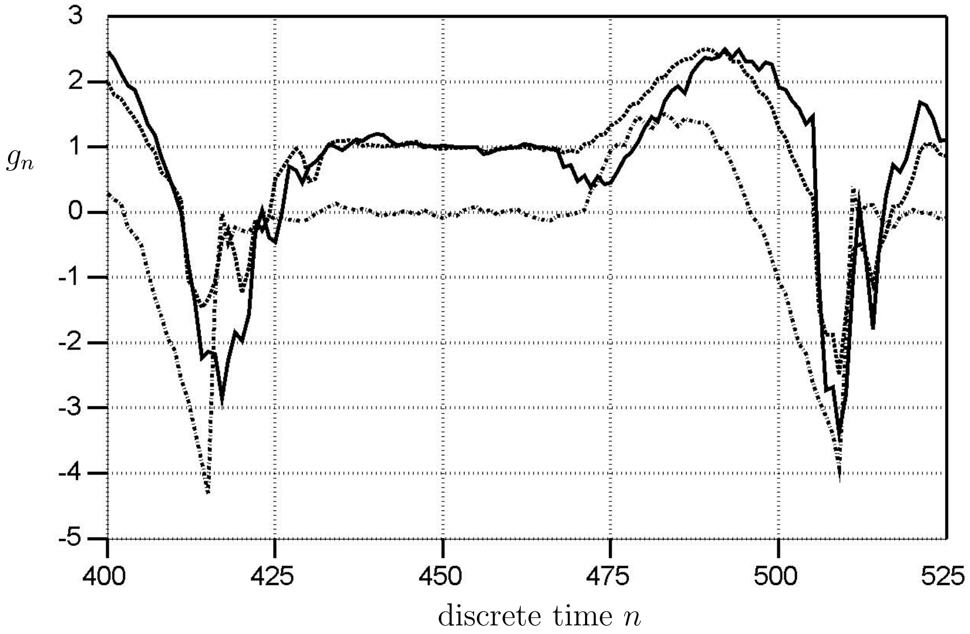

Figure 3.22: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of breathy /uw/ (Figure 3.18) from a synthesized waveform (dashed-dotted).

3.3.5 Experiment Observations

The above results show a noisier glottal waveform estimate, but did not reveal any noticable degradation of analysis performance over breathy synthesized speech. The estimates in Figures 3.19, 3.21, and 3.22 using $\lambda = 0.9$ again resulted in slower tracking than with $\lambda = 0.8$, as did the estimates using 18th order filters in Figures 3.20 and 3.22 in comparison to the 10th order cases in Figures 3.19 and 3.21, respectively. Figure 22 exhibits a dip in the estimation over the glottal opening interval like that of Figure 16 when inadequate RLS convergence was blamed, but then eventually increases and appears to start tracking the pulse.

3.3.6 Other Experiments

In Section 3.3.2 the AFRIF estimation Next, consider the AFRIF glottal estimation performance on synthesized speech with more extreme glottal waveform open quotient. The synthesized speech shown in Figures 3.23 and 3.24, corresponds to the long vowel sound /iy/ using open quotients of 64% and 5%, respectively. Figures 3.25 and 3.26 show the actual glottal waveform with the estimated glottal signals when a filter order of 18 was used to analyze the speech in Figures 3.23 and 3.24, respectively. Slightly more degradation in waveform tracking is revealed in Figures 3.27 and 3.28 when the filter order was decreased to 10. For the large open quotient case, AFRIF analysis with a forgetting factor of 0.9 is incapable of

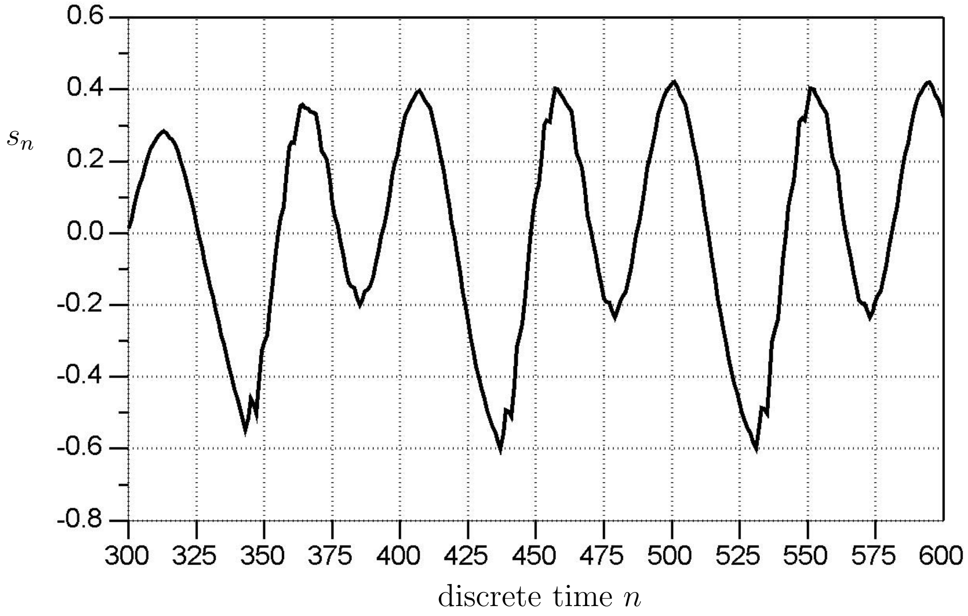

Figure 3.23: Synthesized speech of /iy/ with 64% OQ.

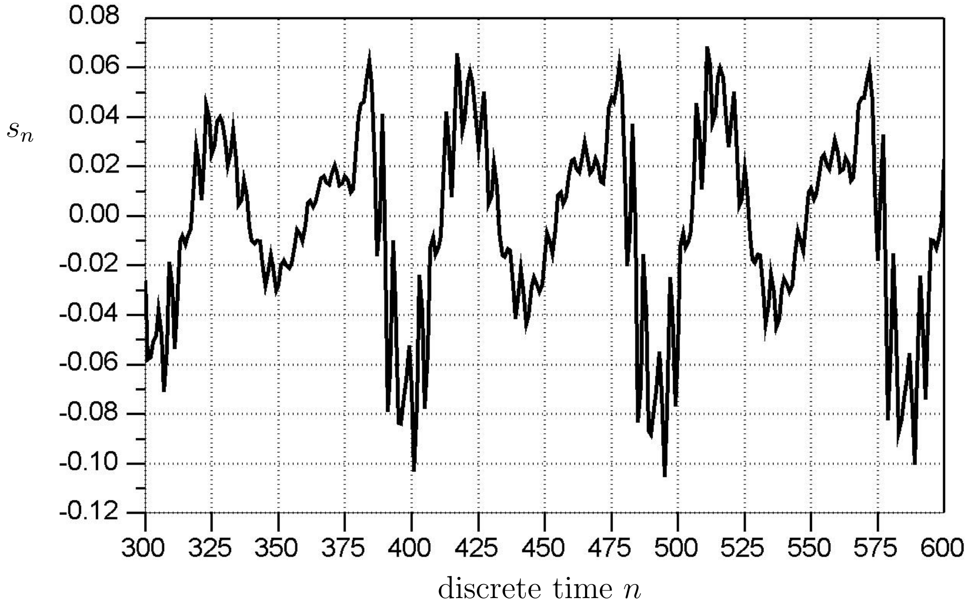

Figure 3.24: Synthesized speech of /iy/ with 5% OQ.

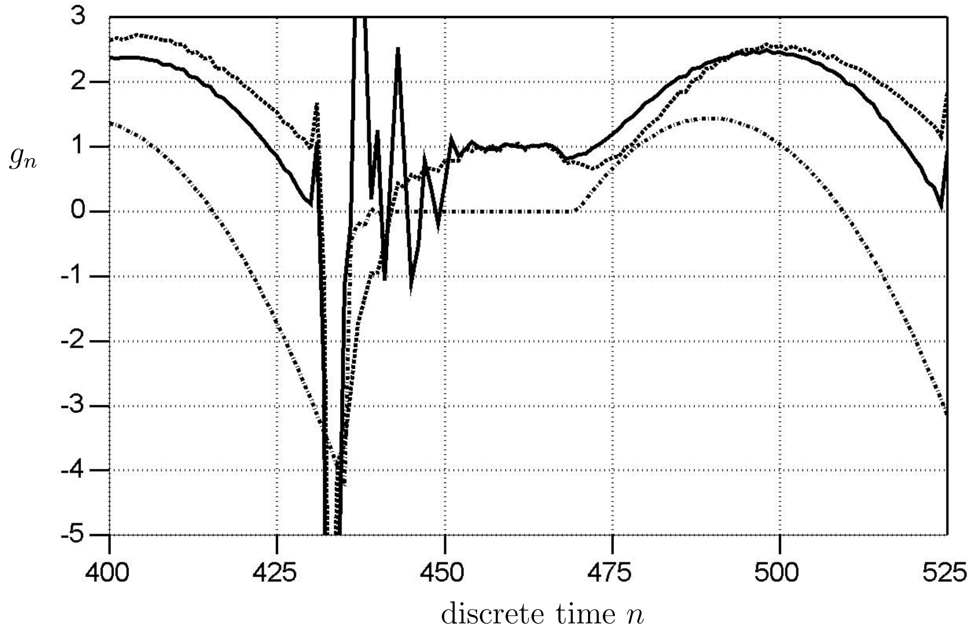

Figure 3.25: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.23) from a synthesized waveform (dashed-dotted).

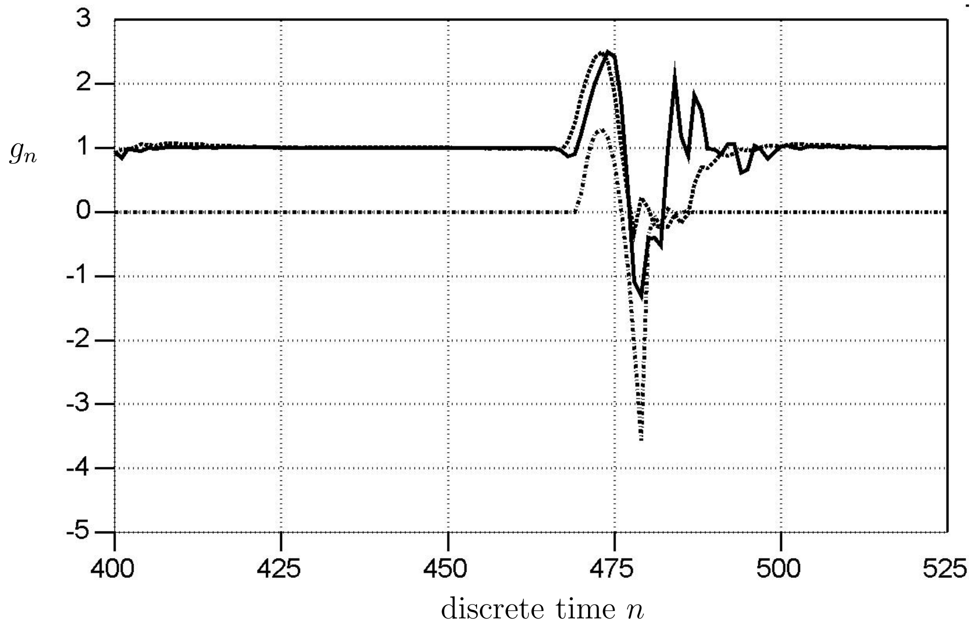

Figure 3.26: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.24) from a synthesized waveform (dashed-dotted).

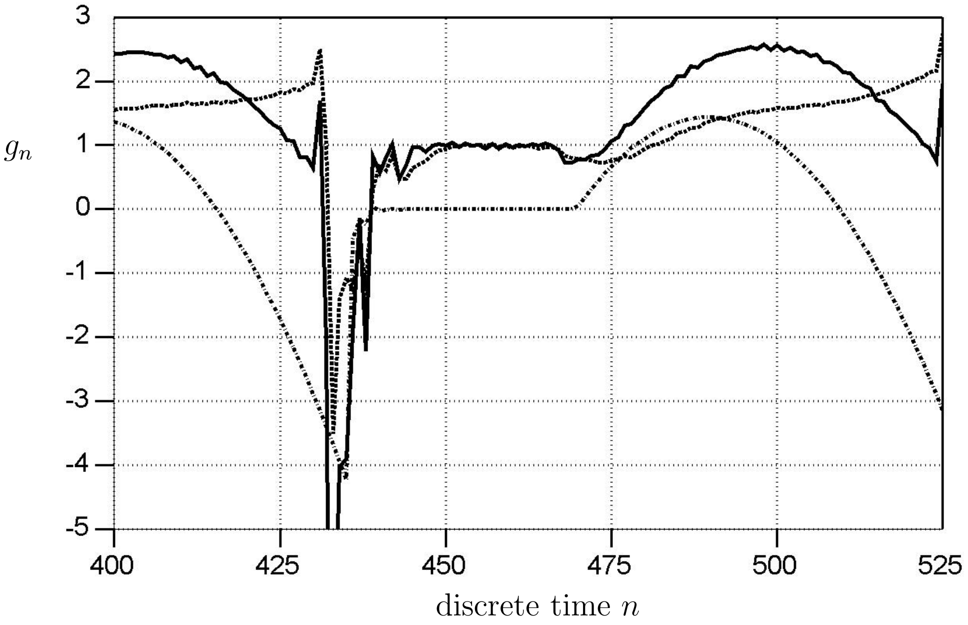

Figure 3.27: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.23) from a synthesized waveform (dashed-dotted).

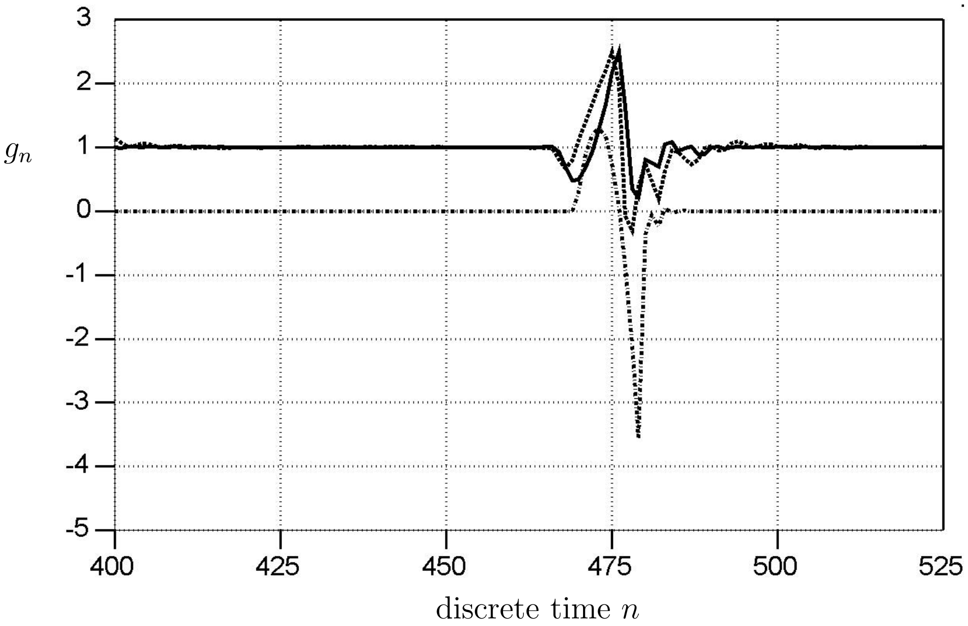

Figure 3.28: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.24) from a synthesized waveform (dashed-dotted).

modeling the glottal opening pulse, as observed in Figure 3.27. For the speech with a small open quotient, the estimates in Figure 3.28 are unable to track the smooth pulse shape but instead produce a pointed glottal opening pulse. Hence there could be problems when analyzing breathy speakers that often produce large OQs, or low pitched speakers that produce small OQs [31].

High pitched speakers present a big challenge to speech analysis systems. To test the AFRIF procedure on high pitched speech, the long vowel /iy/ was synthesized using an open quotient of 43%, but this time the pitch was doubled, as observed in Figure 3.29. The resulting glottal signal estimates using filter orders of 10, 14, and 18 are shown in Figures 3.30, 3.31, and 3.32, respectively. The plots reveal

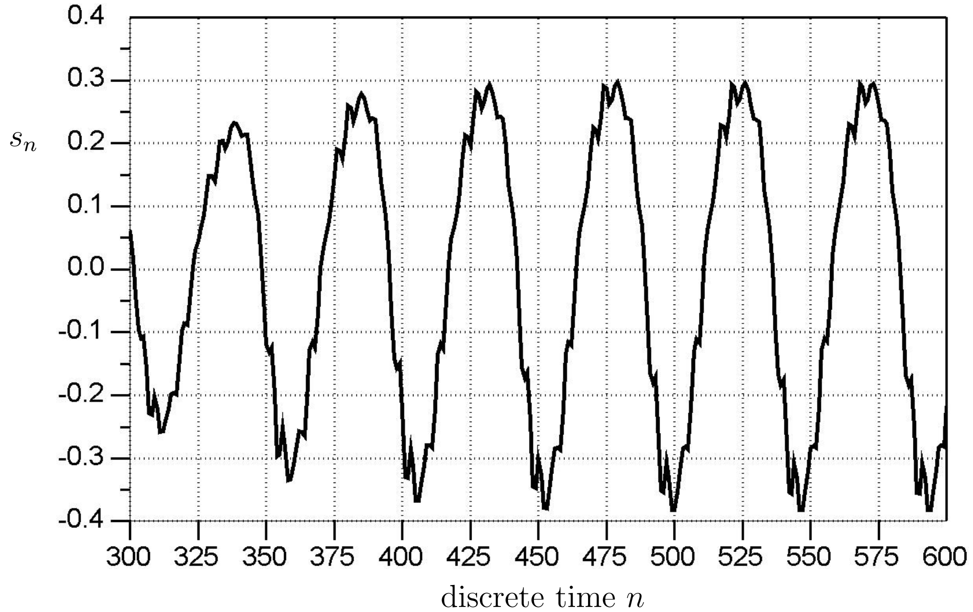

Figure 3.29: Synthesized high pitched speech of /iy/.

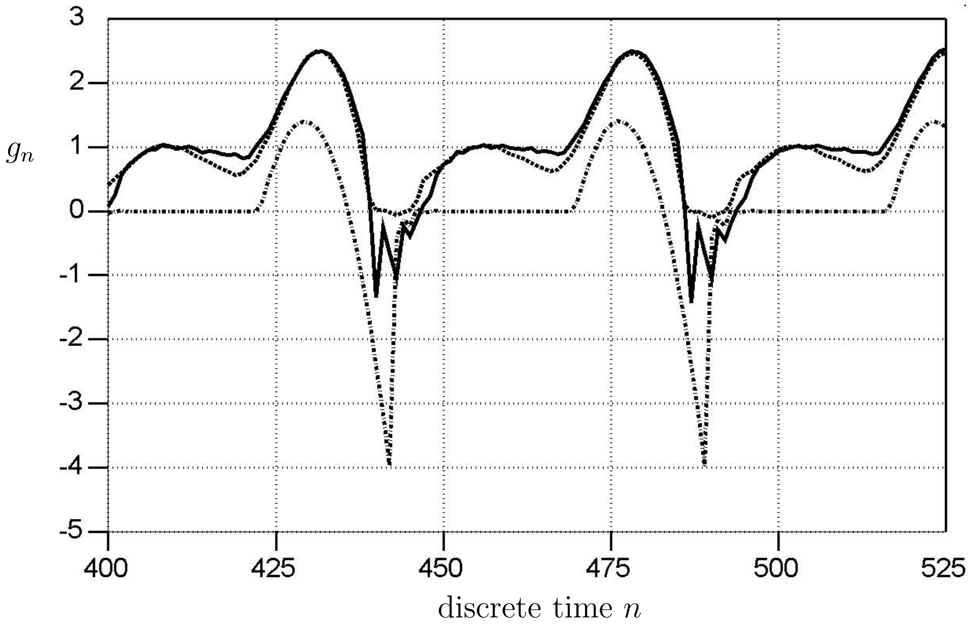

Figure 3.30: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.29) from a synthesized waveform (dashed-dotted).

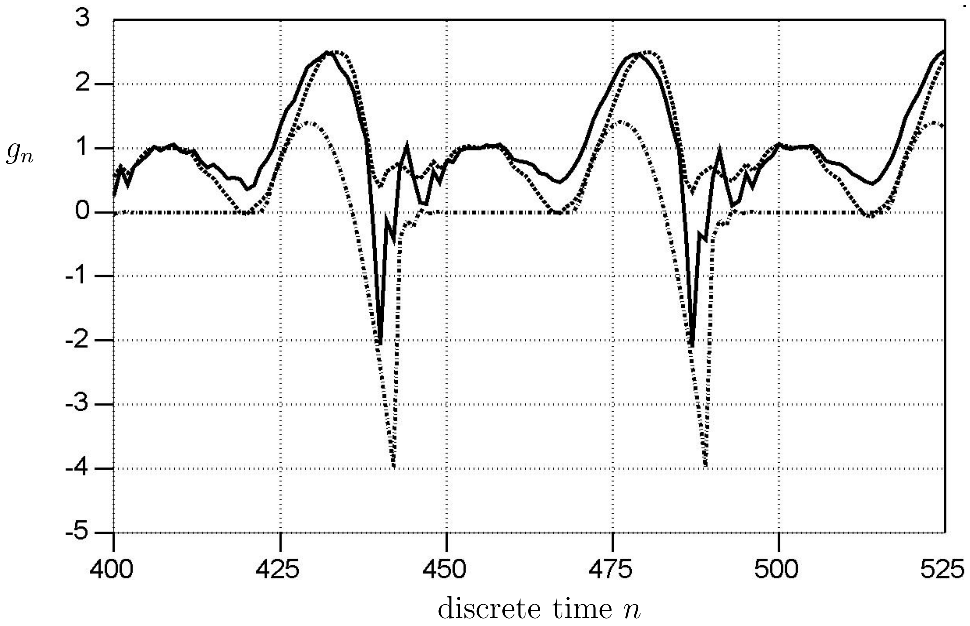

Figure 3.31: 14th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.29) from a synthesized waveform (dashed-dotted).

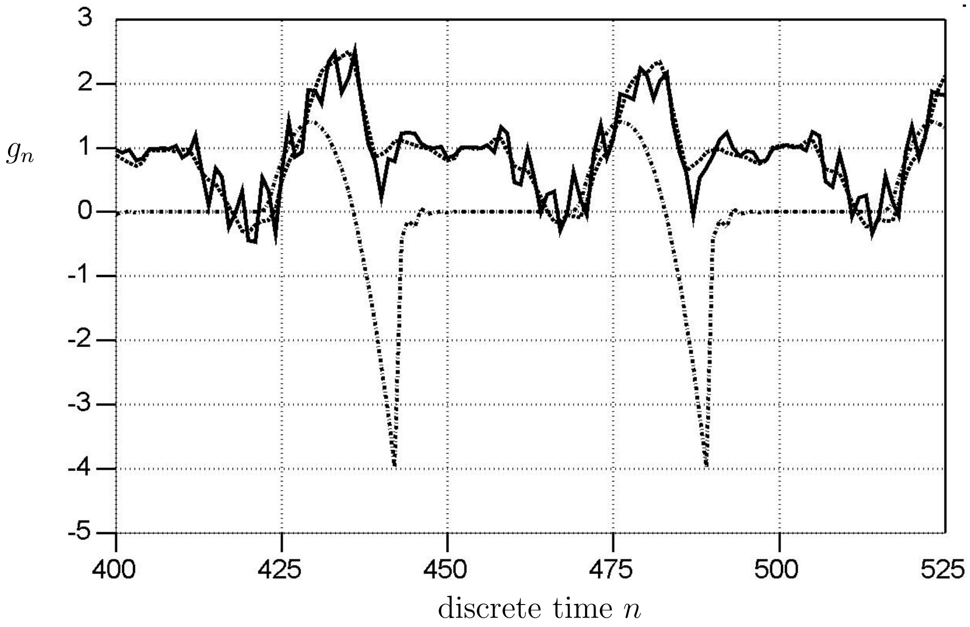

Figure 3.32: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /iy/ (Figure 3.29) from a synthesized waveform (dashed-dotted).

degrading performance with increasing analysis order. The estimates obtained when a filter order of 18 was used bear little resemblance to the actual glottal waveform. This should not be too surprising since approximately $2p = 36$ samples are required for convergence of the RLS algorithm, whereas the entire speech period in this case lasts only 45 samples.

Since the vocal cord operation primarily affects the lower frequencies of speech, the first formant in particular, a crude lowpass filter applied to the speech prior to AFRIF analysis may improve glottal waveform extraction by focusing the spectral estimation of the adaptive algorithm on the low frequencies. The synthesized vowels /ey/ (as in the word *say*) and /a/ (as in the word *pot*), shown in Figures 3.33 and in 3.34 respectively, will be analyzed with and without a lowpass filter applied.

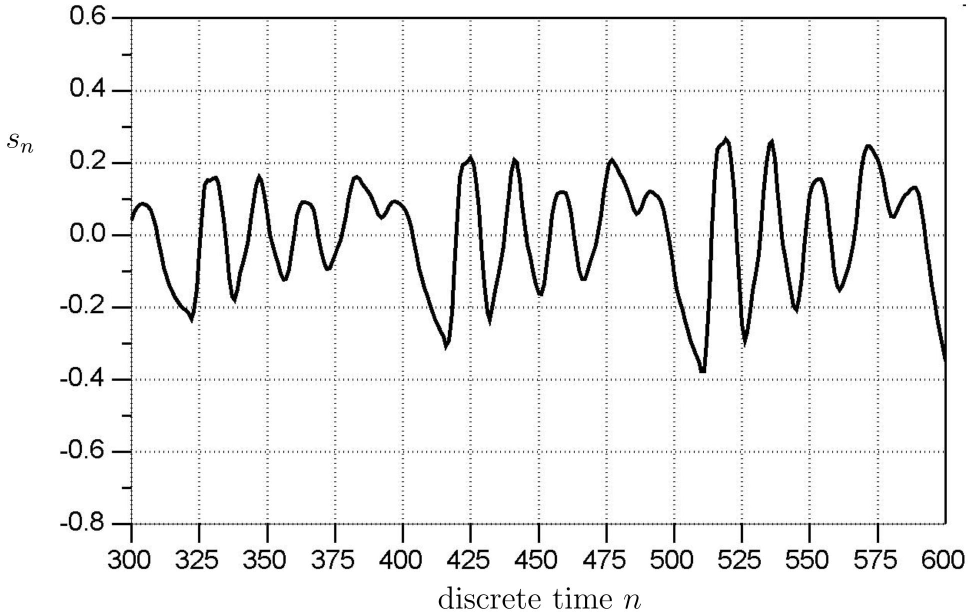

Figure 3.33: Synthesized speech of /ey/.

Glottal waveform estimates for /ey/ without prior lowpass filtering are shown in Figure 3.35 for 10th order analyses and in Figure 3.36 for 18th order analyses. The glottal waveform estimates for unfiltered /a/ are shown in Figure 3.37 for 10th order analyses and Figure 3.38 for 18th order analyses. These glottal waveform estimates for /ey/ and /a/ all exhibit a distorted glottal opening pulse compared to the synthesized glottal waveform. Applying the crude lowpass filter $H(z) = 1/(1 - 0.95z^{-1})$, resulted in a noticeable improvement in tracking performance over the open glottis pulse. These improved estimates are given in Figures 3.39 and 3.40 for the 10th and 18th order analyses of /ey/, respectively. Figure 3.42 shows the

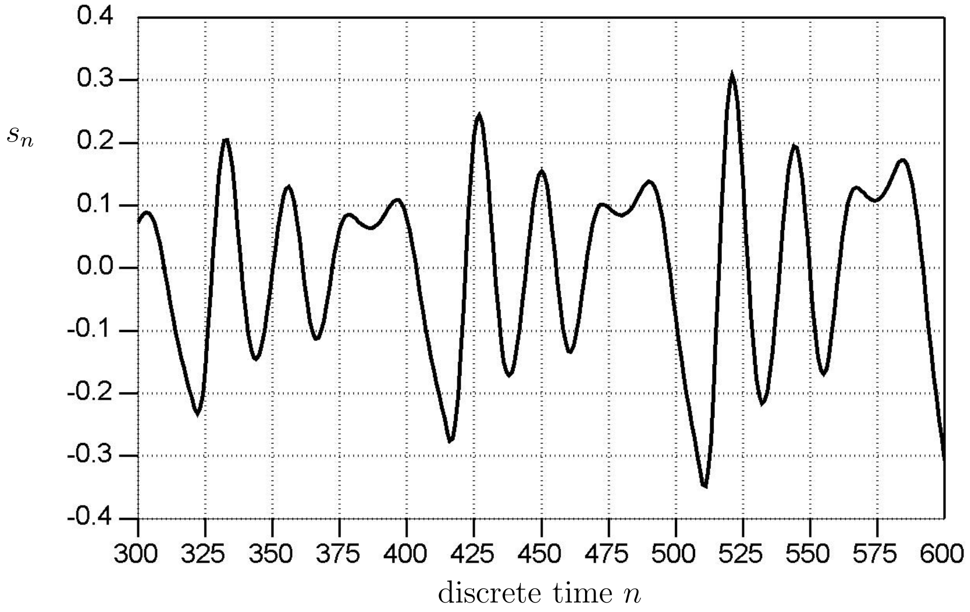

Figure 3.34: Synthesized speech of /a/.

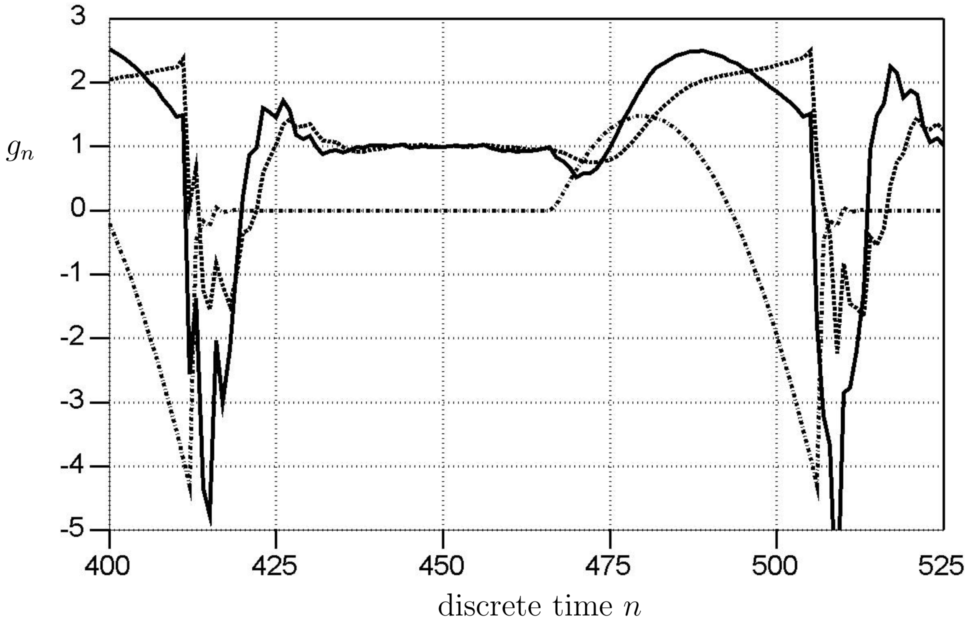

Figure 3.35: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /ey/ (Figure 3.33) from a synthesized waveform (dashed-dotted).

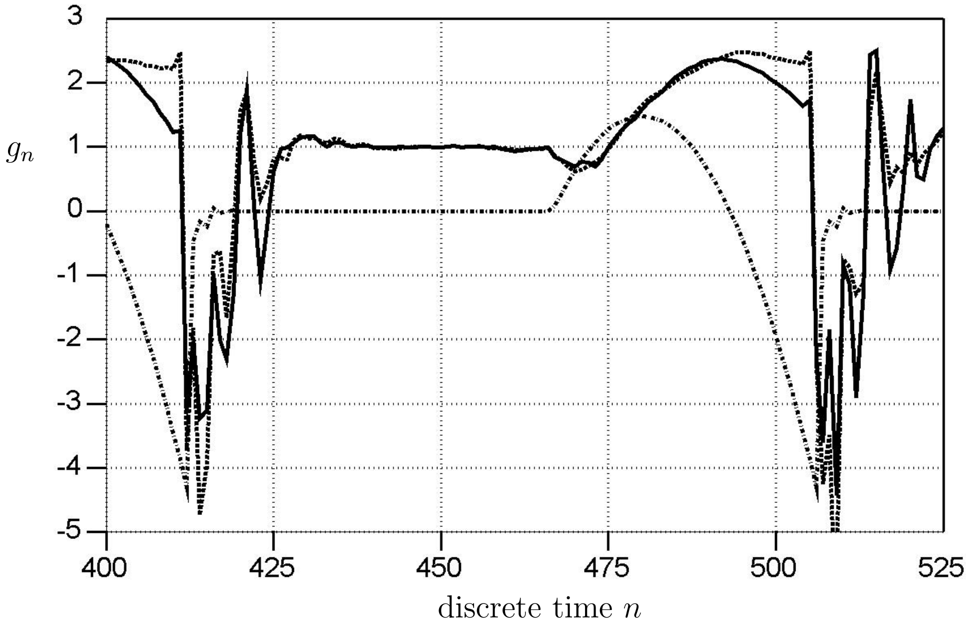

Figure 3.36: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /ey/ (Figure 3.33) from a synthesized waveform (dashed-dotted).

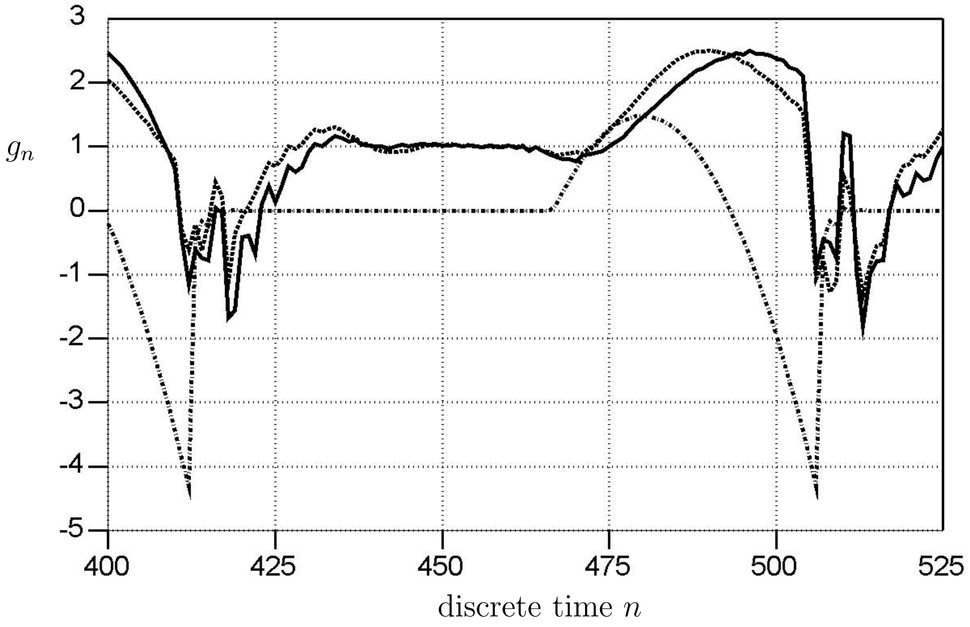

Figure 3.37: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /a/ (Figure 3.34) from a synthesized waveform (dashed-dotted).

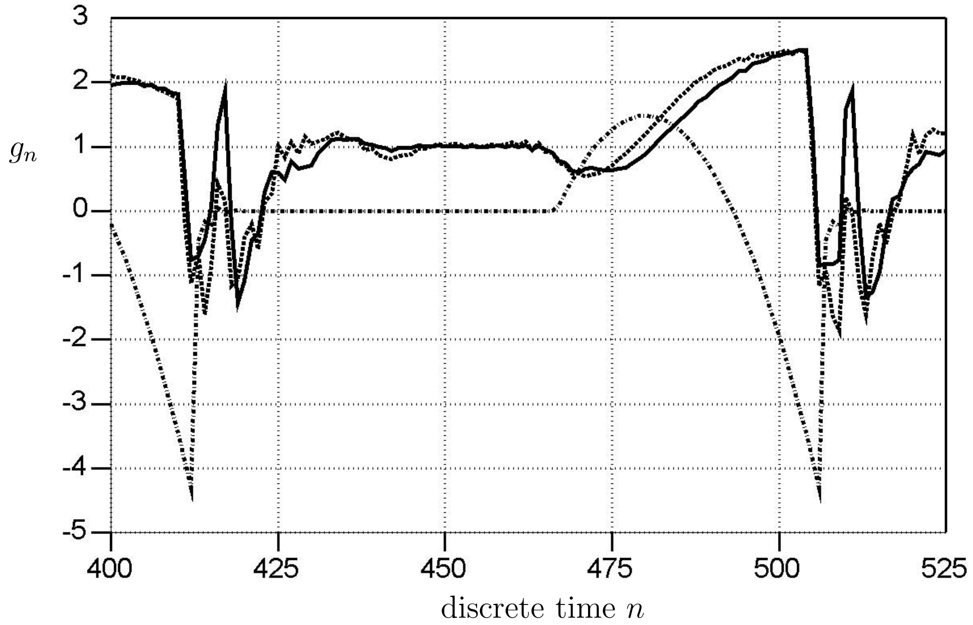

Figure 3.38: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /a/ (Figure 3.34) from a synthesized waveform (dashed-dotted).

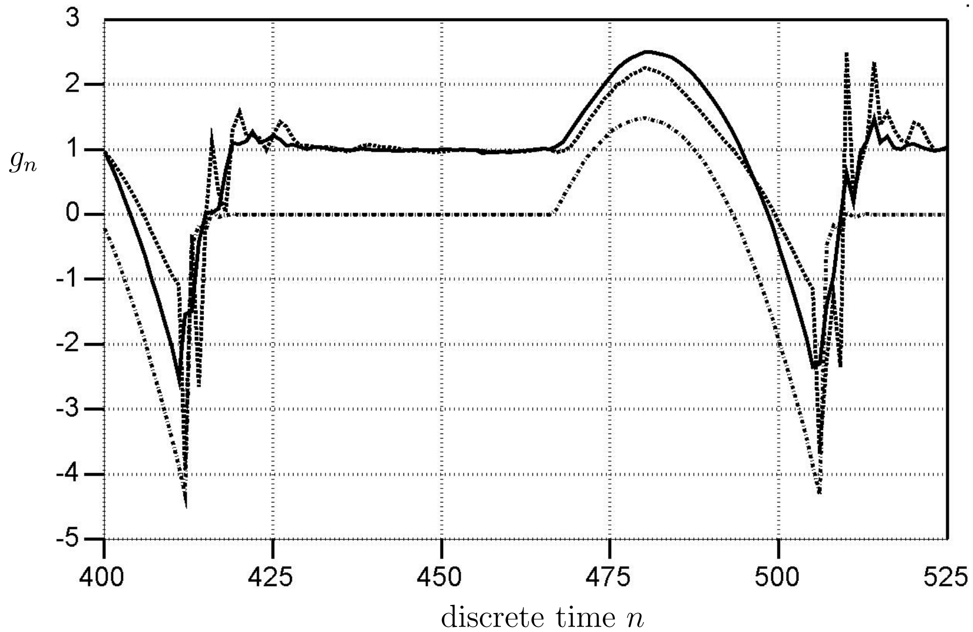

Figure 3.39: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /ey/ (Figure 3.33) from a synthesized waveform (dashed-dotted) applying LPF prior to 10th order RLS.

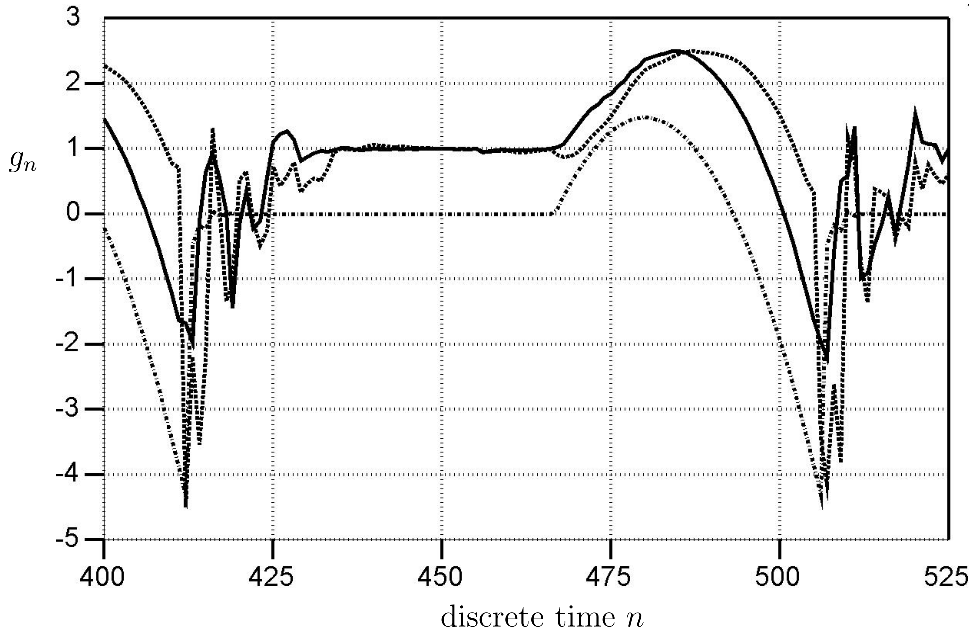

Figure 3.40: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /ey/ (Figure 3.33) from a synthesized waveform (dashed-dotted) applying LPF prior to 18th order RLS.

improved estimates for the 18th order estimates of /a/. The only exception was the 10th order case where the estimates of /a/ in Figure 3.41 show little improvement over those in Figure 3.37.

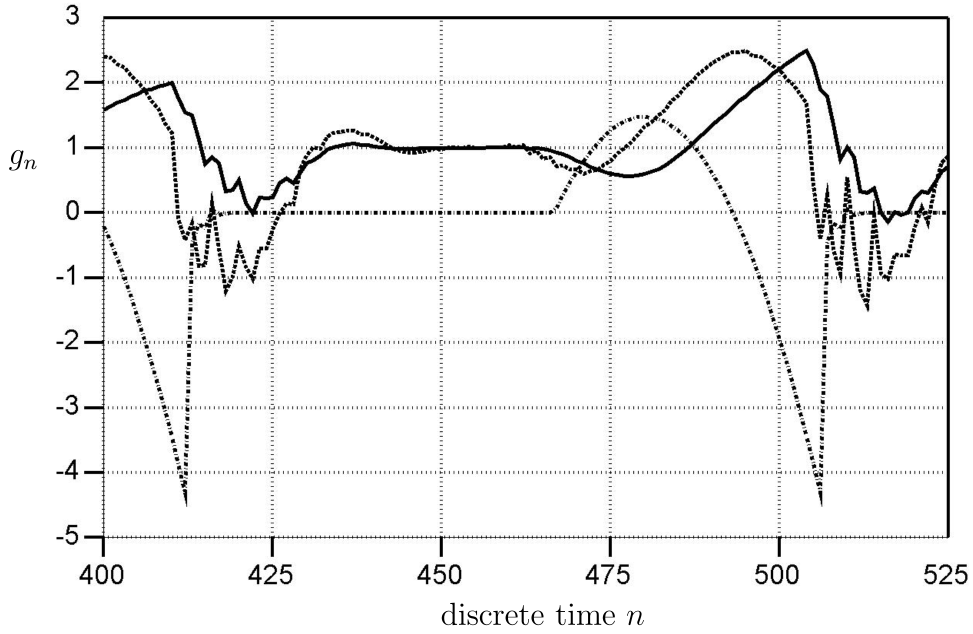

Figure 3.41: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /a/ (Figure 3.34) from a synthesized waveform (dashed-dotted) applying LPF prior to 10th order RLS.

Normally glottal waveform extraction is performed over vowels. The performance of AFRIF will now be investigated for non-vowel voiced speech. Figures 3.44 and 3.45 show glottal estimates corresponding to 10th and 18th order analysis over the synthesized phoneme /r/ shown in Figure 3.43. Note how the estimates fluctuate over the open glottis interval compared to the ideal, rounded pulse shape. The fluctuations are even greater for the estimates on the phoneme /l/ in Figure 3.46 as shown for the 10th and 18th order analyses in Figures 3.47 and 3.48, respectively. For phoneme /n/ (in Figure 3.49), this fluctuation effect is so severe that the glottal waveform estimates bear little

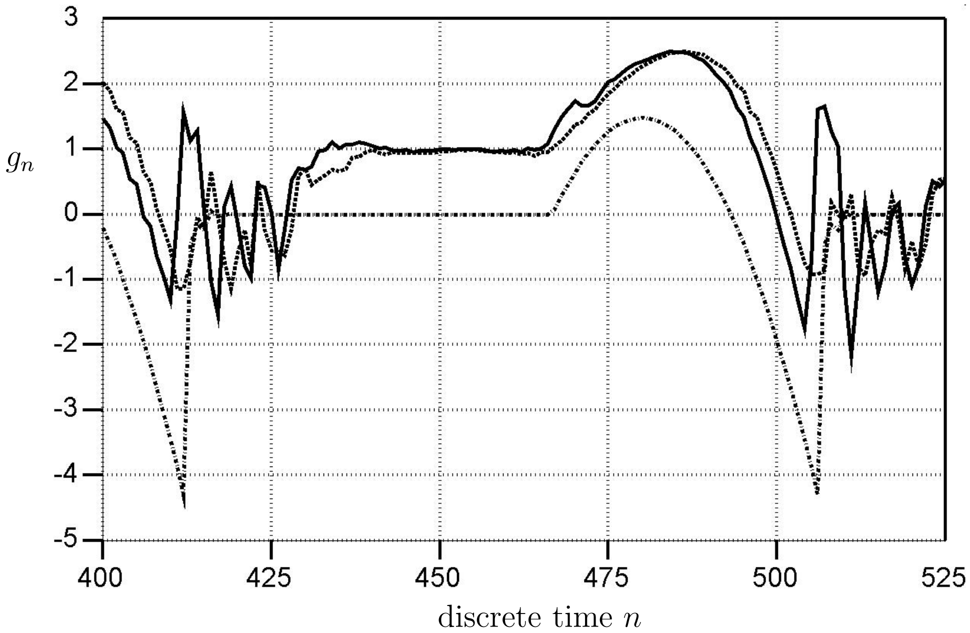

Figure 3.42: Glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /a/ (Figure 3.34) from a synthesized waveform (dashed-dotted) applying LPF prior to 18th order RLS.

resemblance to the ideal glottal waveform as shown in Figures 3.50 and 3.51.

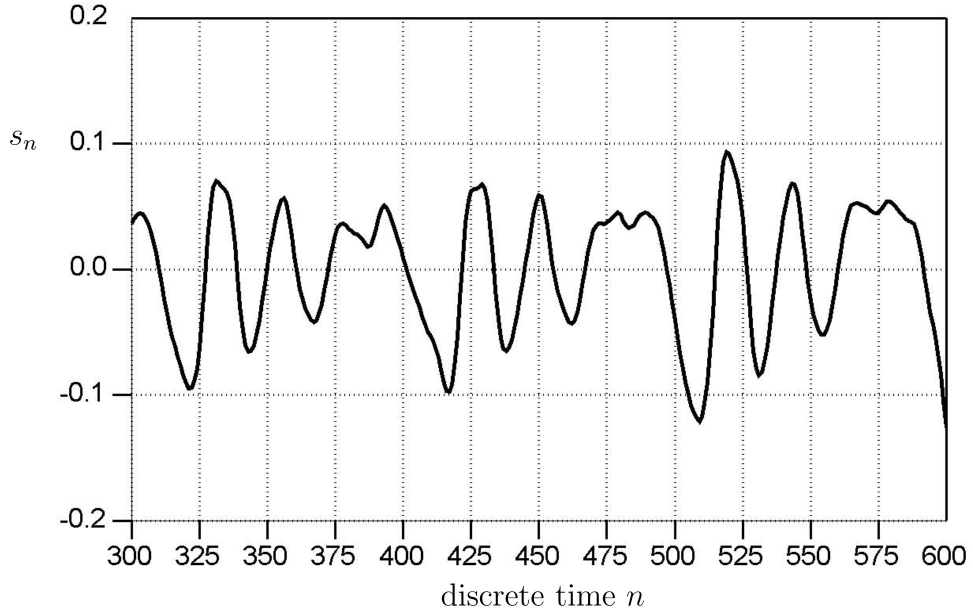

Figure 3.43: Synthesized speech of /r/.

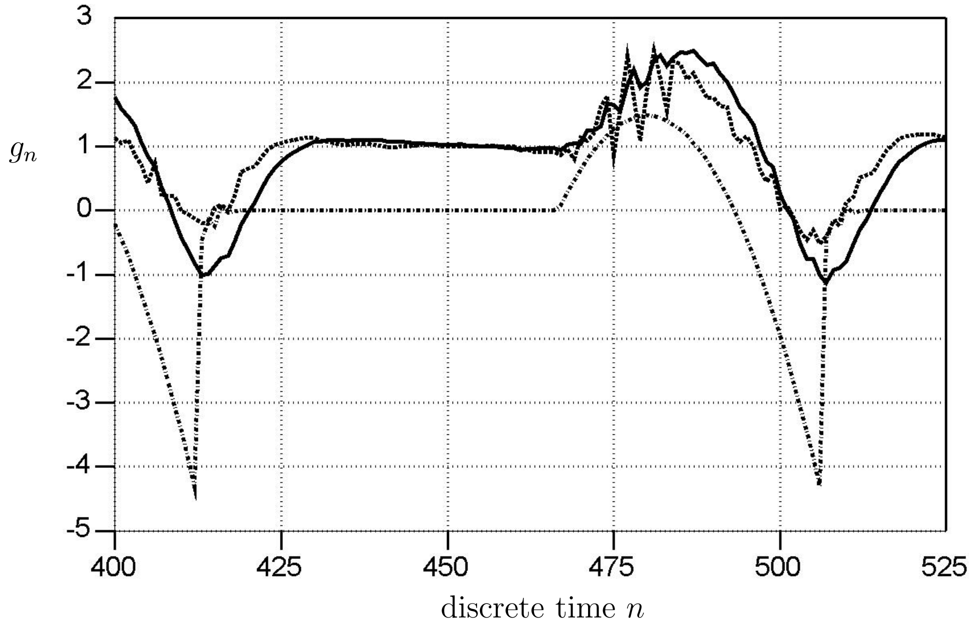

Figure 3.44: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /r/ from a synthesized waveform (dashed-dotted).

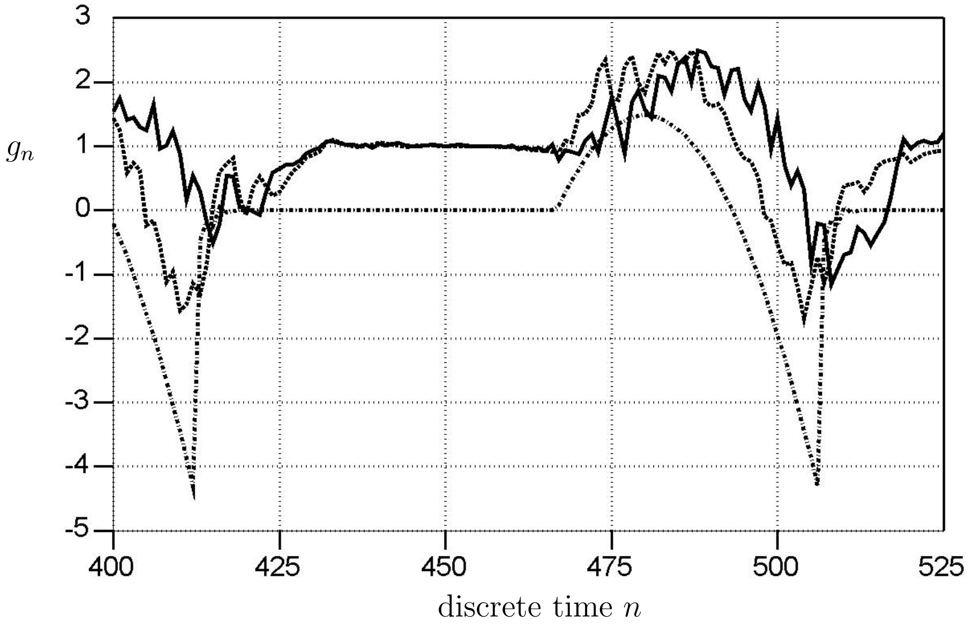

Figure 3.45: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /r/ from a synthesized waveform (dashed-dotted).

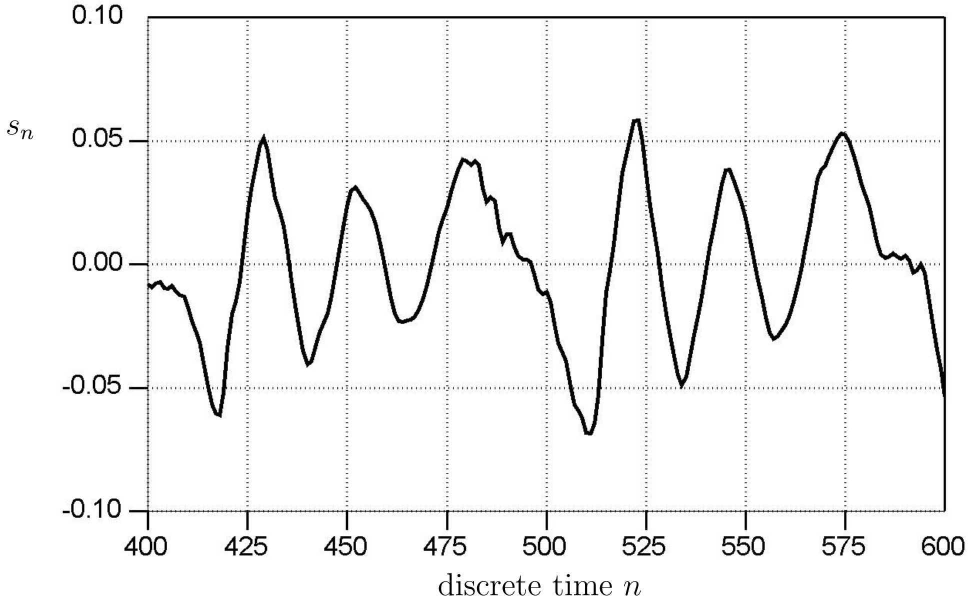

Figure 3.46: Synthesized speech of /l/.

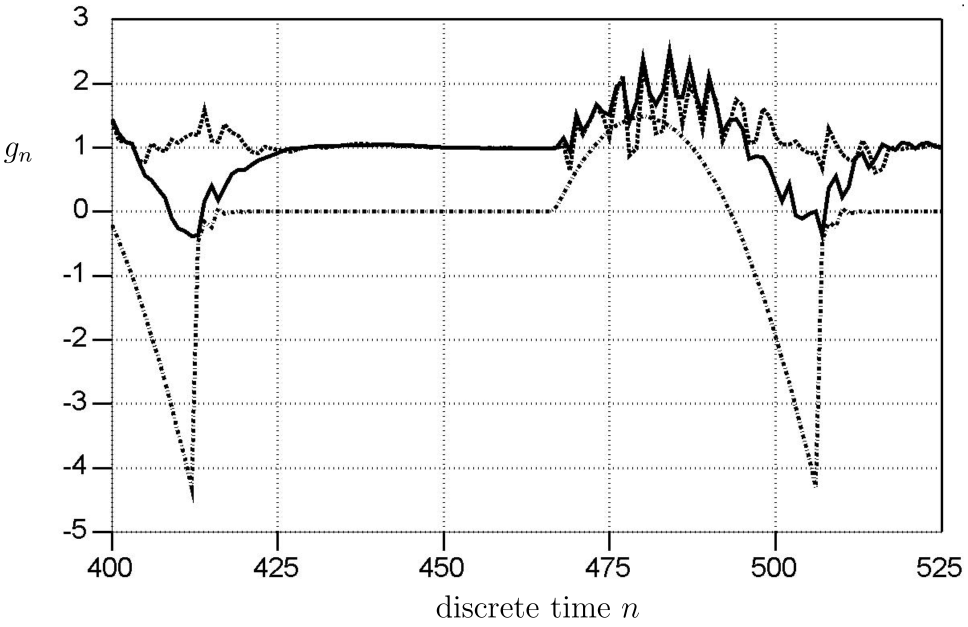

Figure 3.47: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /l/ from a synthesized waveform (dashed-dotted).

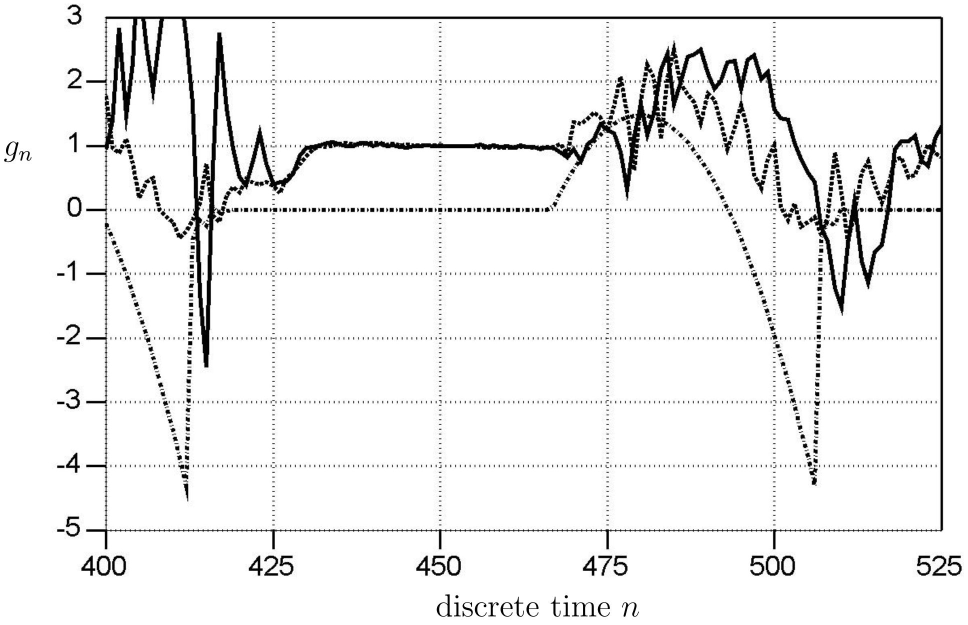

Figure 3.48: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /l/ from a synthesized waveform (dashed-dotted).

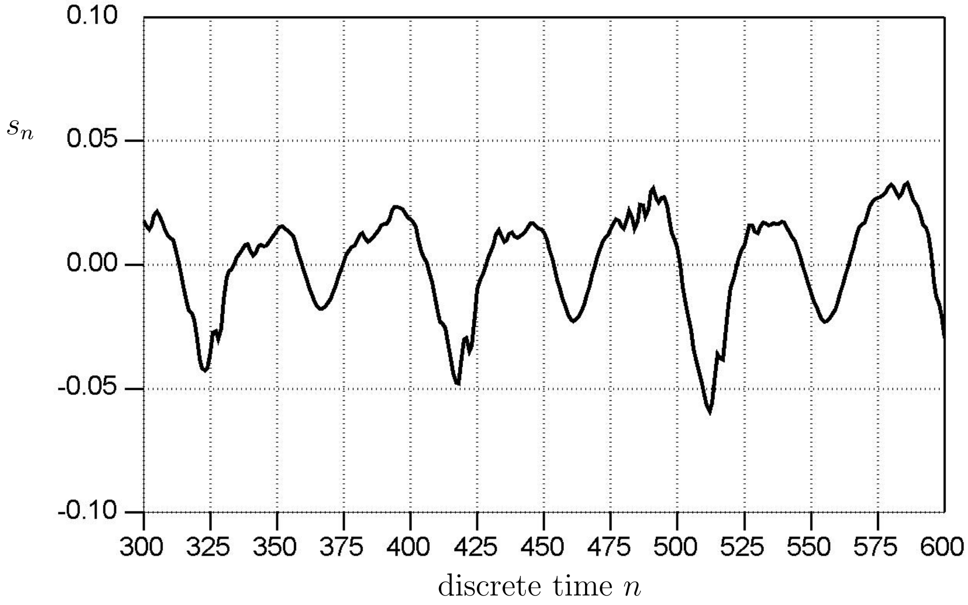

Figure 3.49: Synthesized speech of /n/.

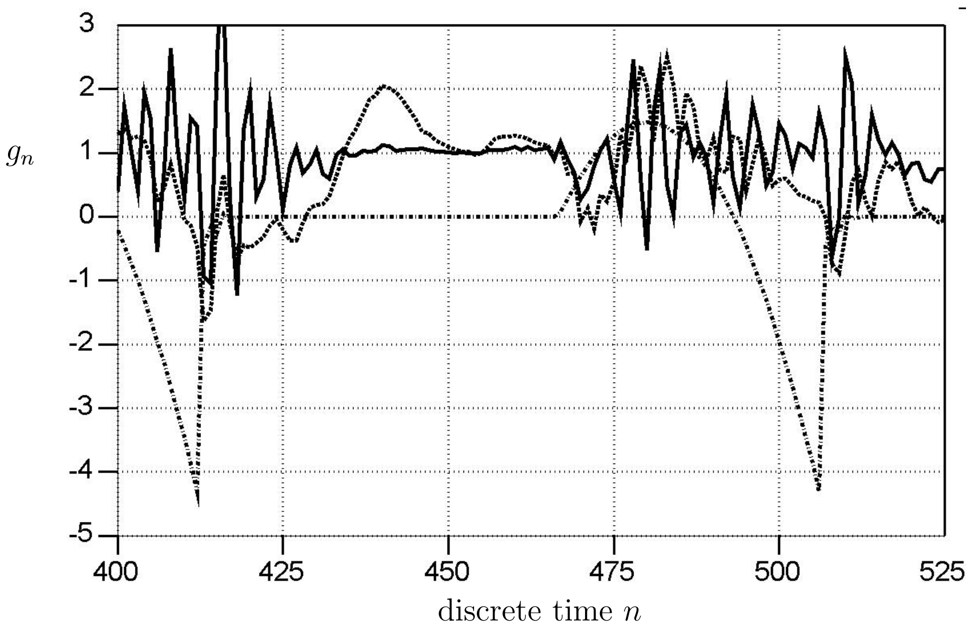

Figure 3.50: 10th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /n/ from a synthesized waveform (dashed-dotted).

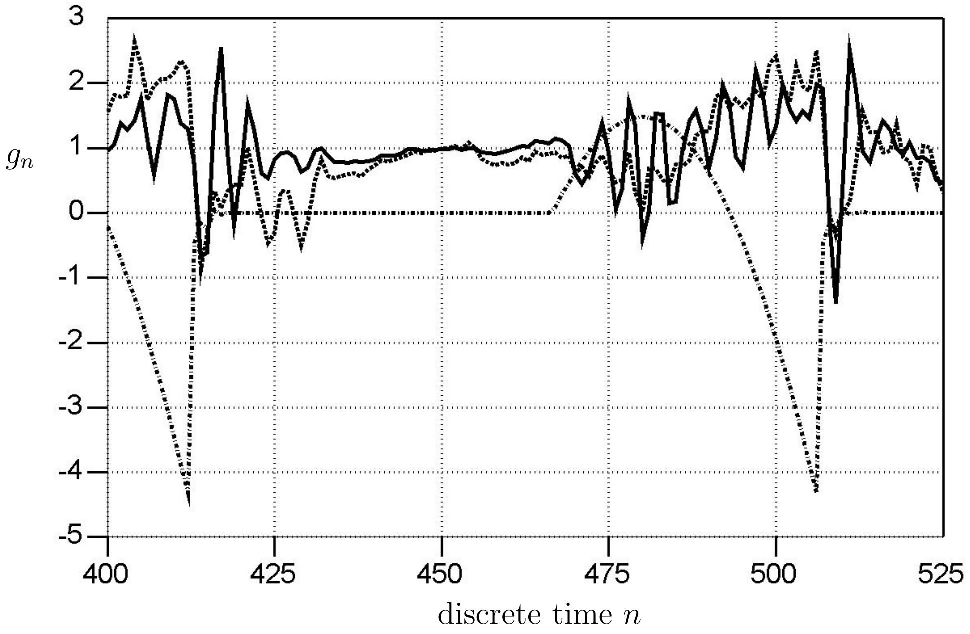

Figure 3.51: 18th order glottal analysis using $\lambda = 0.8$ (solid) and 0.9 (dashed) of /n/ from a synthesized waveform (dashed-dotted).

Although many conventional glottal waveform extraction approaches assume highly controlled recording situations, such as analysis on adult speakers using special recording chambers and equipment, the AFRIF procedure has demonstrated that successful glottal waveform modeling can be achieved in more diverse conditions. AFRIF estimation on vowels from typical speakers was shown to be quite simple, and some adjustments to the AFRIF analysis were presented if the need arose for some non-vowel analysis and for more difficult speakers. Whereas this section has successfully demonstrated the ability of AFRIF to estimate the classical glottal waveform from synthetic speech, along with several of its limitations, the next chapter will demonstrate the AFRIF glottal waveform extraction peformance on human speech.

Chapter 4

AFRIF Applications

This chapter demonstrates the AFRIF speech analysis procedure on actual human speech. First the AFRIF procedure is used for glottal waveform extraction, then it is used for speaker identity verification. General results, observations and practical issues of AFRIF are presented for both cases.

4.1 Examples of Glottal Signals from AFRIF

This section demonstrates the glottal signal extraction performance of the AFRIF procedure on actual human speech. These glottal signals were extracted from a subset of the database, described in the Appendix, made up of a diverse group of male and female adults and children. For the analysis in this section, the speech was downsampled from the original 22.050 kHz to a new sampling rate of 8 kHz. Automatic pitch, endpoint, and voiced/unvoiced detection routines again were not used in this analysis. The previous chapter revealed slightly faster tracking with $\lambda = 0.8$ than with $\lambda = 0.9$ at the cost of occasional degradations in modeling arising from noisy analysis results. Hence $\lambda = 0.8$ was used in the following analysis for achieving more localized results on the high pitched database speech as produced by the various women and children. The typical speech prediction filter order of $p = 10$ was used, since higher orders did not reveal any noticable improvements to the results in Section 3.3.

4.1.1 Practical Issues with the AFRIF Procedure

The RLS algorithm provides the ability to model stationary data as well as track nonstationary data whose statistics change slowly, but requires approximately $2p$ samples for re-convergence (where p is the order of the filter) when sudden changes to the statistics of the data occur. Glottal excitation is an example of such a sudden nonstationarity for the case of speech. Acknowledging that sudden nonstationarities result in large prediction errors and coefficient variations during RLS adaptive filtering, the location of n_c at which to extract the spectral estimate within each pitch period of AFRIF analysis is chosen around $2p$ samples past the maximum value of the residual signal. There may be instances when more samples are required to achieve complete convergence, or when the nonstationarity does not end exactly at the occurrence of the largest residual value. However, waiting more than $2p$ samples may be too long for high pitched and other speakers with short closed glottis intervals. Because of these challenges, a two-pass AFRIF verification approach was devised to help evaluate the accuracy of the spectral estimates obtained with RLS, and thus guide the choice of n_c. A description of this process is given as follows.

The verification involves performing a second AFRIF analysis on synthesized speech $\hat{s}_n$,

$$\hat{s}_n = g_n + \sum_{i=1}^{p} a_{n_c,i}\hat{s}_{n-i}, \tag{4.1}$$

to achieve estimates $\{\hat{g}_n\}$ and $\hat{a}_{n_c}$ of $\{g_n\}$ and a_{n_c}, respectively, which were obtained from AFRIF analysis on the original human speech, $\{s_n\}$. Due to the inverse filtering capabilities with linear prediction, similar results are expected from the analysis on the original speech as on the synthesized speech generated after successive analysis and synthesis operations. Although it will be shown in the following example that this invertibility test does not provide an indication of the accuracy of the specific vocal tract and glottal signal estimates, it does provide a useful verification of the estimation performance of the RLS adaptive filtering operation.

Consider the glottal waveform estimate for AFGLB, shown in Figure 4.1, which was obtained by choosing n_c at samples 4755, 4810, 4860, 4915, 4965 and 5015 corresponding to the six periods displayed. These samples are between 15 and 20 samples past the largest residual magnitude value in each period which, taken from Figure 4.2, occur at samples 4742, 4793, 4843, 4894, 4947 and 5001. Verifying the stationarity of the speech and analysis convergence at these samples, the two-pass verification result given as the dotted line shows a similar resemblance to the actual glottal signal estimate shown as the solid line. Furthermore, the time-varying pole magnitudes from the RLS analysis of the AFRIF procedure plotted in Figure 4.3 illustrate the variation of the adaptive filter vocal tract estimates over the period from sample 4755 to 4810, and indicate that they are relatively stationary by samples 4755 and 4810. Finally, this is confirmed in Figure 4.4 which compares the vocal tract estimate at $n_c = 4755$ with the other time-varying AR coefficients within the segment, using a basic Euclidean distance metric. As observed, the distances are relatively low in the interval around $n_c = 4755$, and then again at $n_c = 4810$.

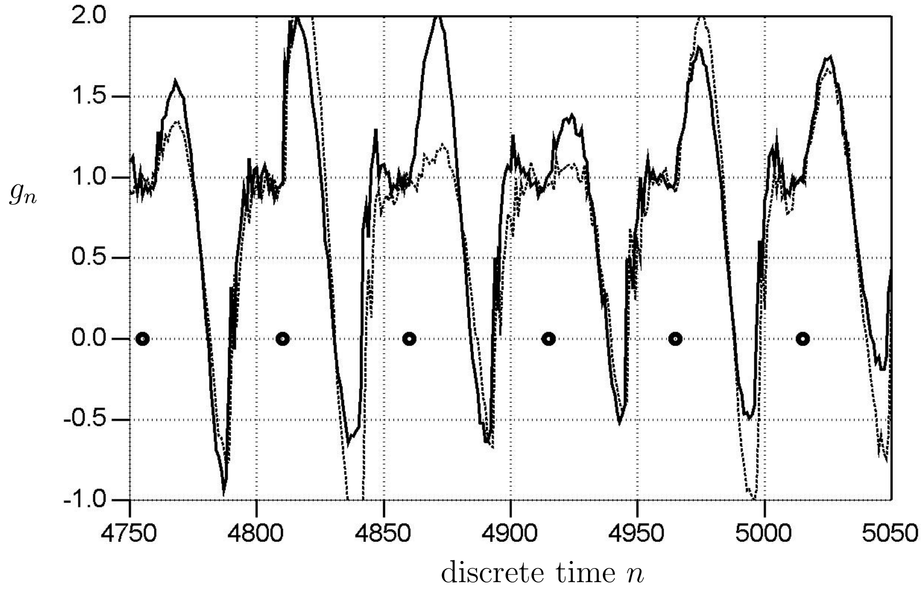

Figure 4.1: Glottal analysis for AFGLB of /iy/ (solid) and the verification result (dashed).

It is noteworthy that, although it appears that the RLS estimation has indeed

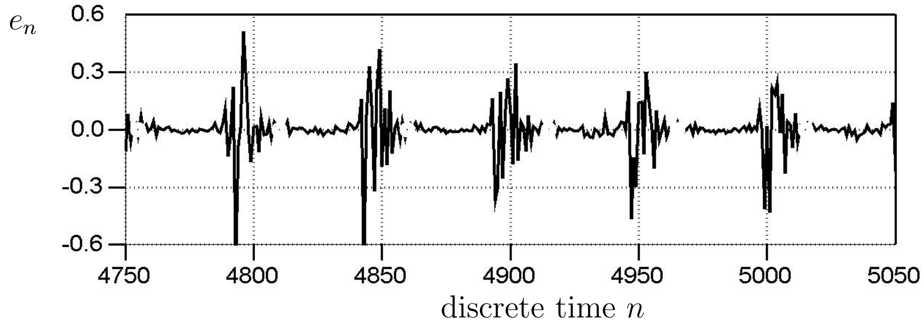

Figure 4.2: Prediction error signal for AFGLB of /iy/.

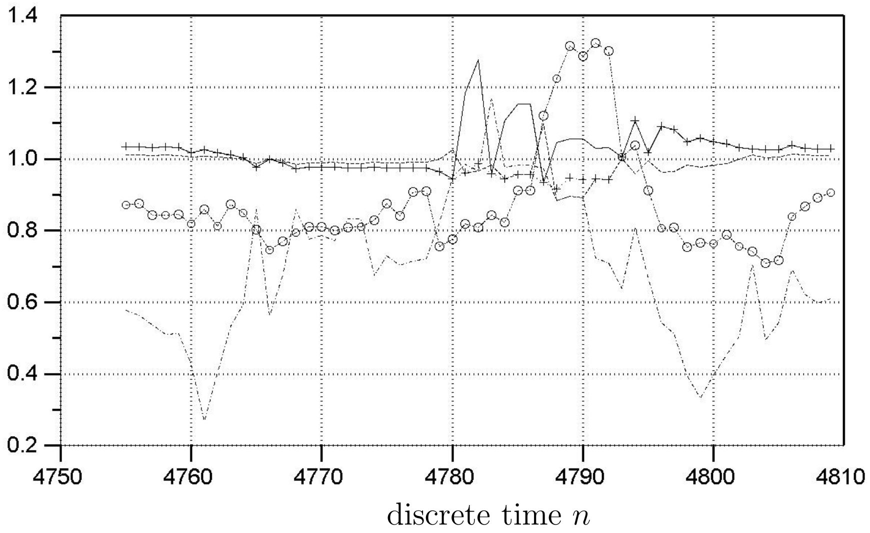

Figure 4.3: The time-varying pole magnitudes from AFRIF for speaker AFGLB.

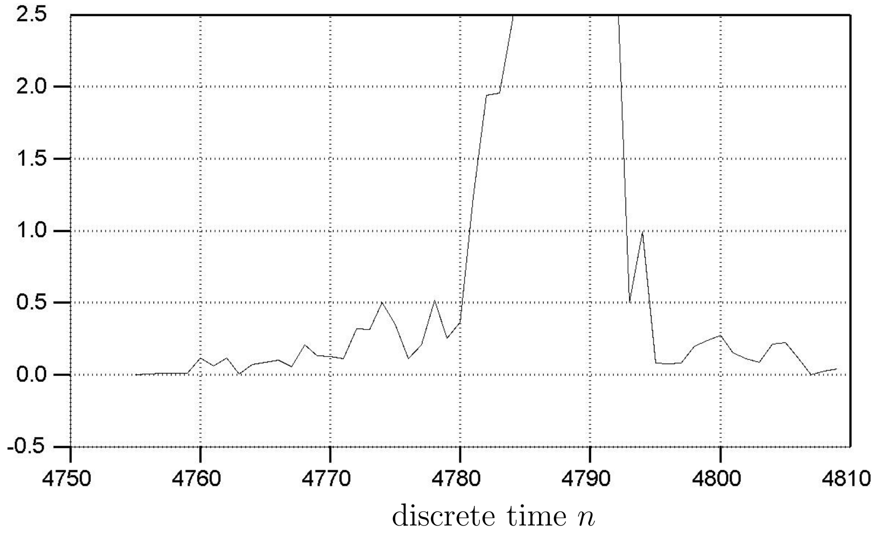

Figure 4.4: Distances between the vocal tract estimate at $n_c = 4755$ and AR parameters obtained at samples 4756 to 4805.

converged and is tracking the data around samples 4755, 4810, 4860, 4915 and 4965, the glottal opening pulses of the signals shown in Figure 4.5 tend to decrease as each of the chosen locations, n_c, is delayed (where the dotted, short dashed, long dashed, and long dashed-dotted waveforms correspond to delays of 1, 2, 3, and 4 samples, respectively). Delaying the locations by nine resulted in a glottal waveform with little or no glottal opening pulses as observed from the bold plots in Figures 4.5 and 4.6.

Although it might be argued that this particular speech record produces smaller glottal opening pulses than others, it is more likely that the location at which each n_c was chosen has been delayed past the closed glottis interval and into the open

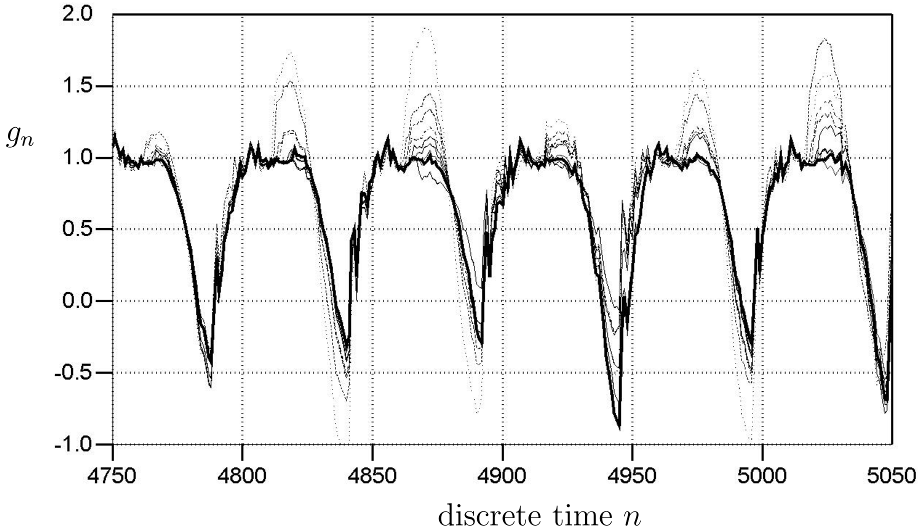

Figure 4.5: AFRIF results for AFGLB of /iy/ when n_c is delayed.

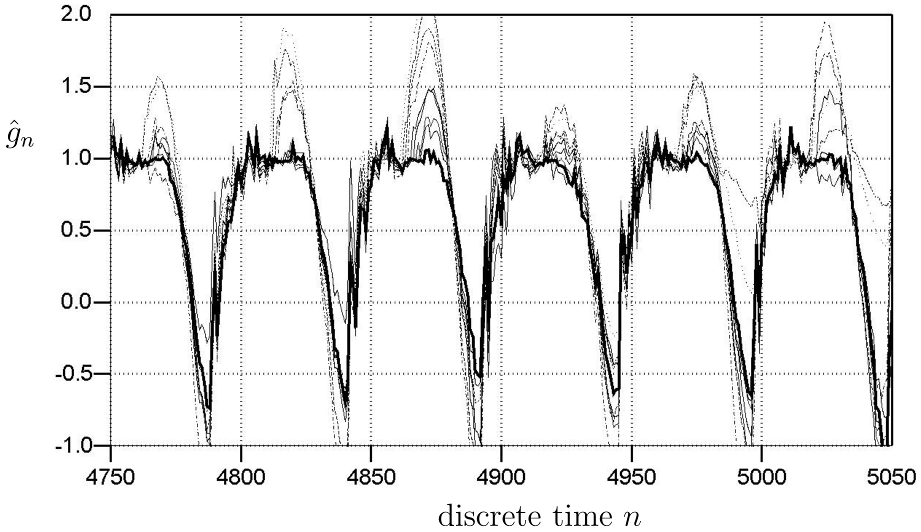

Figure 4.6: AFRIF verification results for AFGLB of synthesized /iy/ when n_c is delayed.

glottis interval, thus suppressing the opening pulse. First of all, by analyzing the time-varying pole waveforms in Figure 4.3, one can see that the largest magnitude (i.e., dominant) poles, in the interval from sample 4764 (nine samples past the original choice of n_c) on through the next 15 samples, are not as large as the other dominant poles shown in the plot. This implies lower amplitude, wider bandwidth formants in this region, consistent with the known behavior of glottal opening.

Next observe the AFRIF estimates in Figure 4.7 which were obtained while the vocal tract estimates were placed at the five samples prior to 4755, 4810, 4915, 4965 and 5015. As a result of extracting the vocal tract estimates earlier in the pitch period, the glottal opening pulses have generally increased (where the dotted, short dashed, long dashed, and long dashed-dotted waveforms correspond to choosing each n_c 1, 2, 3, and 4 samples earlier, respectively). It should be noted from the bold plot that the estimate taken when $\mathbf{n_c}$ was chosen 5 samples closer to excitation (i.e., $\mathbf{n_c}-5$) occasionally produced a distorted glottal opening pulse. The occurrence of such distortions in the glottal waveform estimate during glottal opening was frequently observed during our analysis when n_c was chosen too close to excitation so that the RLS analysis had not sufficiently converged.

Hence a general observation is that the glottal opening pulse in the estimated glottal waveform from AFRIF tends to become suppressed when n_c is chosen toward the center of glottal opening, and severely distorted when n_c was chosen too soon

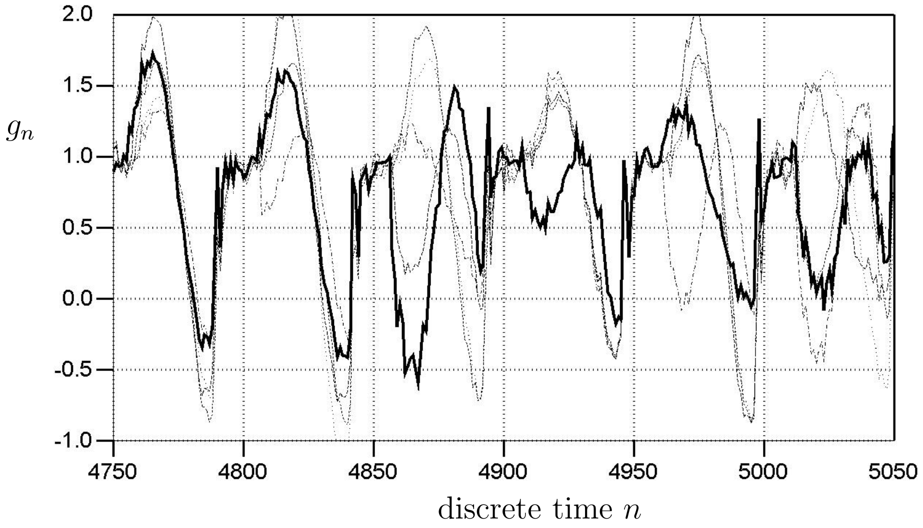

Figure 4.7: AFRIF results for AFGLB of /iy/ when n_c is set closer to excitation.

after excitation. Another observation is that the inverse filter verification did not actually help in determining whether n_c was within the glottal interval or during glottal opening since the RLS algorithm was able to track the statistics of the data in both cases. However, it was valuable in determining if n_c, the location within the speech record at which the vocal tract estimate was obtained, was too soon after excitation, prior to convergence.

Finally, Figure 4.8 shows the resulting glottal estimates if n_c is chosen so that the vocal tract estimation falls in and around the glottal closure phase. These bear little resemblance to the glottal waveform estimate obtained in Figure 4.1. Figure 4.3 revealed that the poles of the estimate at sample 4785 are either outside of, or very close to, the unit circle compared to the pole radii during the closed glottis and glottal opening intervals, accounting for the dramatic differences between Figure 4.8 and Figures 4.1, 4.5, and 4.7. Figure 4.4 showed that the distance between n_c and the other time-varying AR coefficients within the speech period gradually increased over the glottal opening phase until the glottal closure interval was reached, where the distance dramatically increased (to a distance over 25, not shown on the plot). The distance thereafter approached zero again while reaching the closed glottis interval of the next segment.

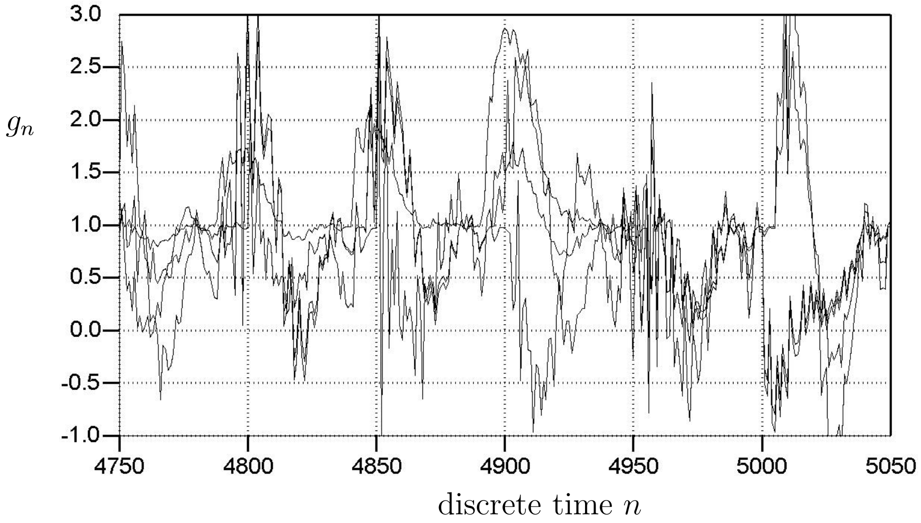

Figure 4.8: AFRIF results for AFGLB of /iy/ when n_c is chosen during glottal closures.

4.1.2 AFRIF Glottal Waveform Observations

Now that the major issues for successful AFRIF analysis performance are understood, this section presents several glottal waveform estimates analyzed from speech provided by various speakers. Locations $\mathbf{n}_c$ are determined using the prediction error signal and two-pass verification process presented in the previous section. The prediction error signals from each analysis have been inserted displaying the locations chosen for $\mathbf{n}_c$. It will be observed from the glottal signals in the previous section and those to be presented in this section, that the general glottal signal characteristics are quite universal for this yet diverse set of speakers, and resemble the common differential glottal waveform (as was used in the speech synthesis system for the analysis in Section 3.3).

Typically, the most distinctive event in these signals is produced at glottal excitation when the glottis flaps shut and causes the vibrations that produce the voiced sound waves. In the glottal signal plots, this event produces the sharp negative slopes that show up in the waveforms. For example the interval between sample 8000 to 8010 in Figure 4.9 contains one of the glottal closure events for the phoneme /iy/ spoken by AFMES. Because of its distinctiveness, the sharp glottal excitation dip can be used to help identify each pitch period.

The relatively flat interval that follows the glottal closure interval is the closed glottis interval. Recall from Section 3.1 that a glottal signal with a constant value of unity over the closed glottis interval, where $\mathbf{a}_{n_c}$ was selected, implies that the formants are stationary. Since the previously mentioned glottal closure phase for

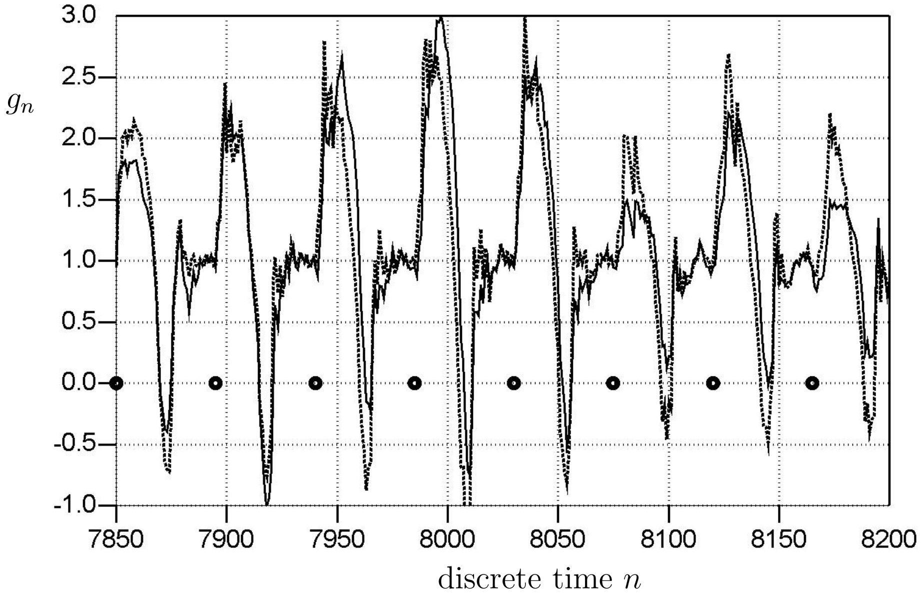

Figure 4.9: Glottal analysis for AFMES of /iy/.

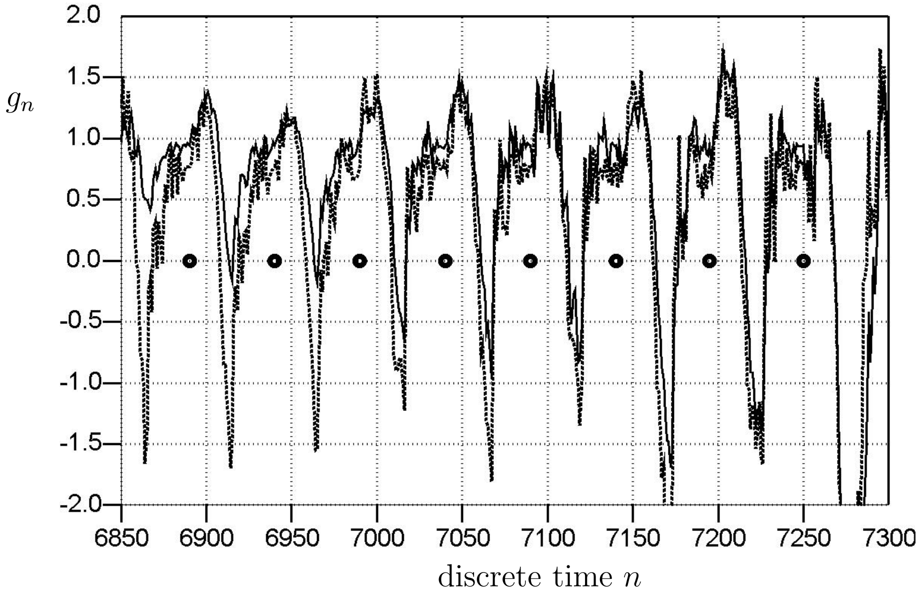

Figure 4.10: Glottal analysis for AFMES of /uw/.

AFMES in Figure 4.9 ended around sample 8010, the closed glottis interval for this speech segment extends approximately from sample 8010 to 8027. Finally, the glottal opening interval follows the closed glottis interval and generally contains the peak (i.e., the maximum value) within each period of the glottal waveform, thus spanning from approximately sample 8028 to 8032 in Figure 4.9. Similar glottal signal features, but somewhat noisier, are obtained when the same speaker produces the phoneme /uw/ as shown in Figure 4.10. The noisy characteristic might be attributed to breathier speech, which when simulated and analyzed in Section 3.3 produced similar results. Also notice how the prediction error signal from /uw/ in Figure 4.12 tends to be slightly less regular than that for /iy/ in Figure 4.11 whose excitation peaks tend to be a lot more distinct.

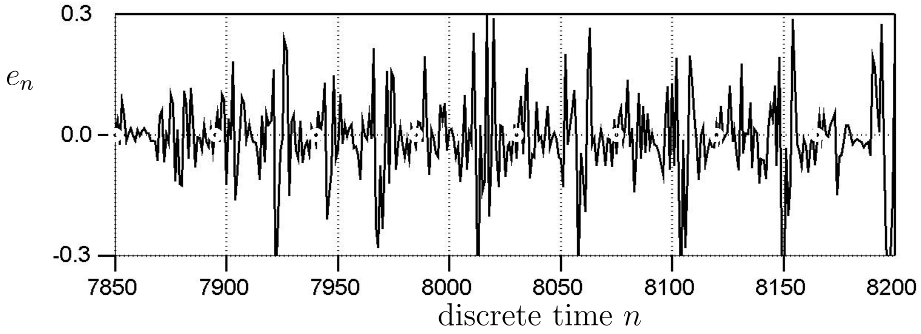

Figure 4.11: Prediction error signal for AFMES of /iy/.

Figures 4.13 and 4.14 and 4.17 and 4.18 show very similar AFRIF analysis results for /uw/ and /iy/ (respectively) spoken by SMLLV; perhaps the highest pitched adult male speaker analyzed (attributing to the relatively short pitch period).

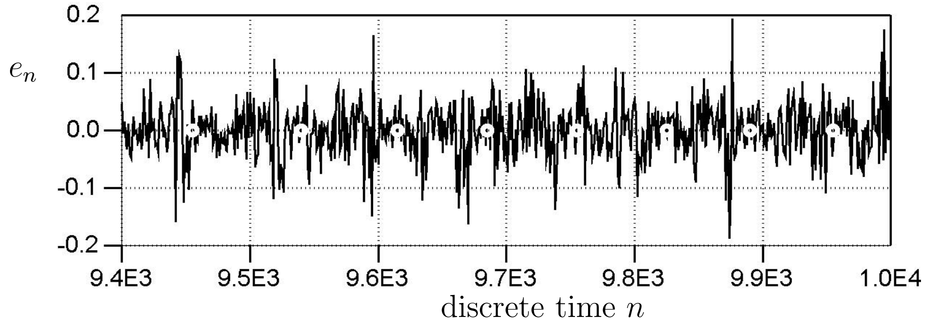

Figure 4.12: Prediction error signal for AFMES of /uw/.

The analysis for perhaps the highest pitched speaker in the entire database, CFKGB, also resulted in somewhat similar results as can be seen in Figures 4.15 through 4.20. The most unique characteristic of the glottal signals from CFKGB in comparison to the others analyzed is the shallowness (and sometimes absence) of the typically deep glottal closure dip. Rather than a characteristic of this speaker's glottal waveform, however, the shallow glottal closure depth is a result of the difficulty in tracking the glottal characteristics of high pitched speakers. Figure 4.21 shows that a deeper closure dip is obtained when the analysis is performed on the speech at the original sampling rate of 22.050 kHz (i.e., before the decimation). With more than twice as many samples available, the adaptive filter is able to better track the glottal characteristics at the cost of a higher computational cost.

Although the general characteristics of the glottal signals tend to be universal for this diverse set of speakers, some distinctive differences between the glottal signals are apparent. Perhaps the glottal waveforms that deviate from the

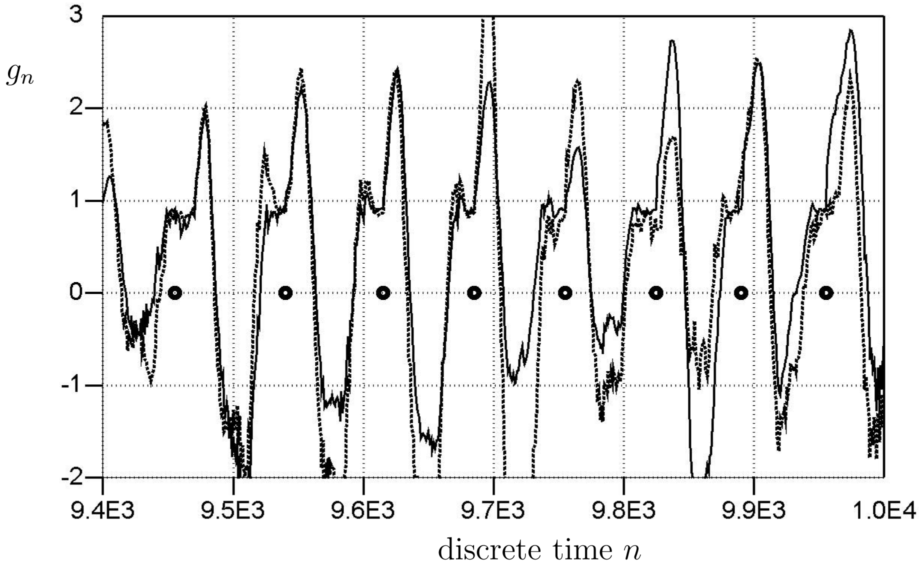

Figure 4.13: Glottal analysis for SMLLV of /uw/.

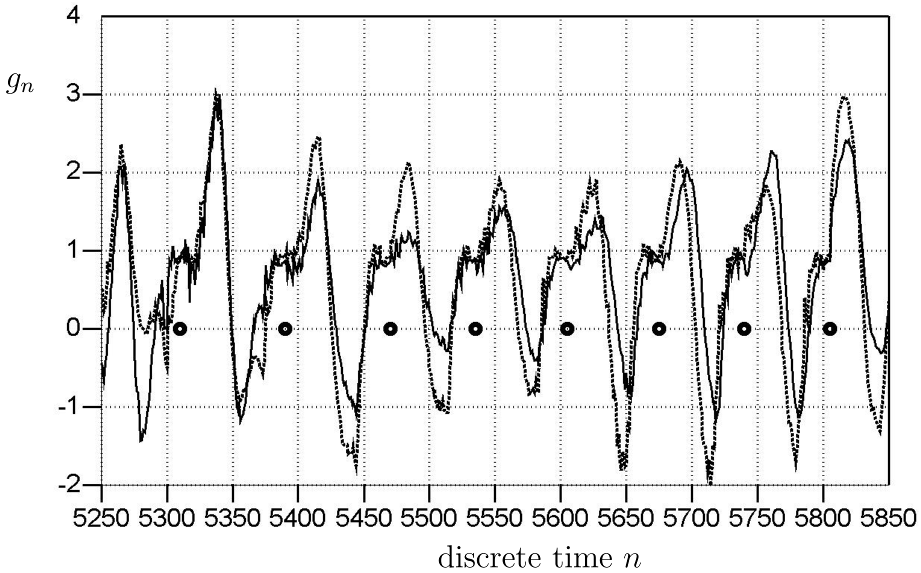

Figure 4.14: Glottal analysis for SMLLV of /iy/.

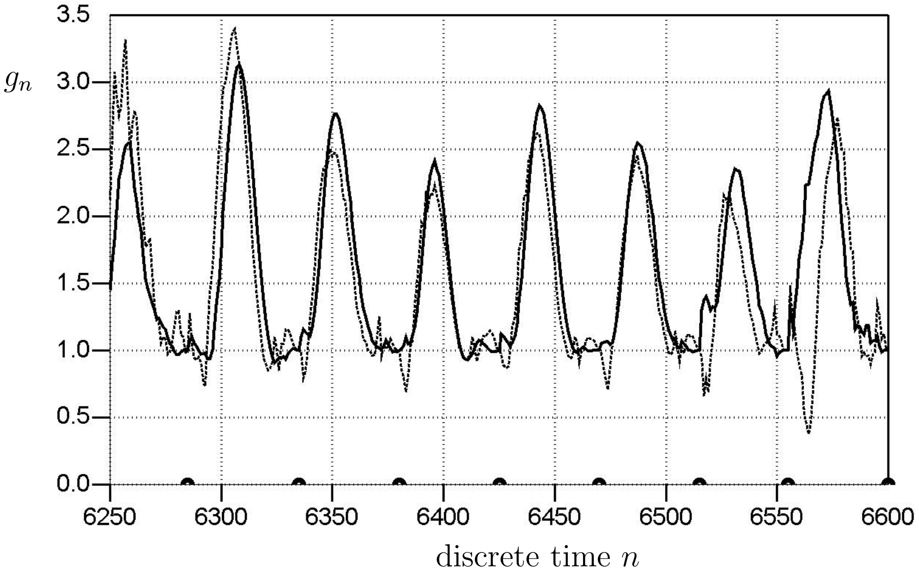

Figure 4.15: Glottal analysis for CFKGB of /uw/.

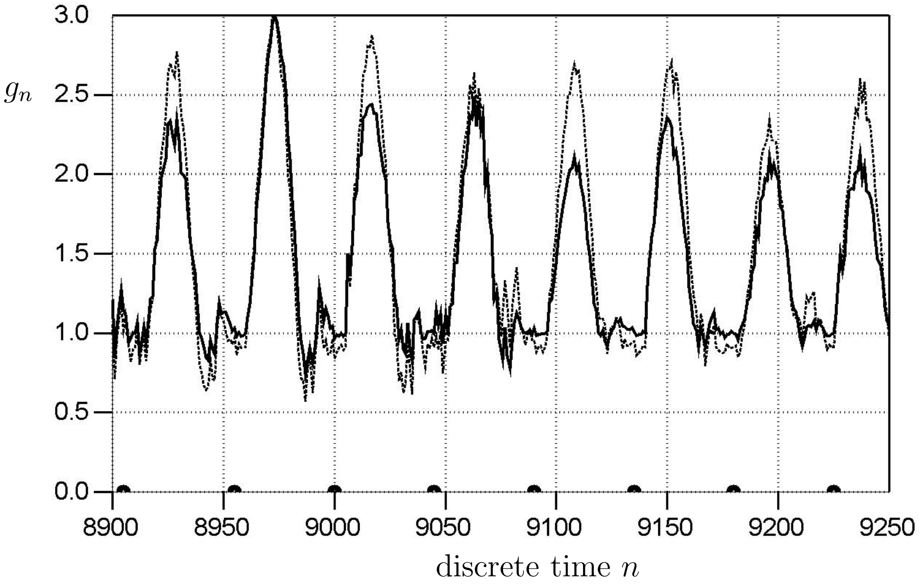

Figure 4.16: Glottal analysis for CFKGB of /iy/.

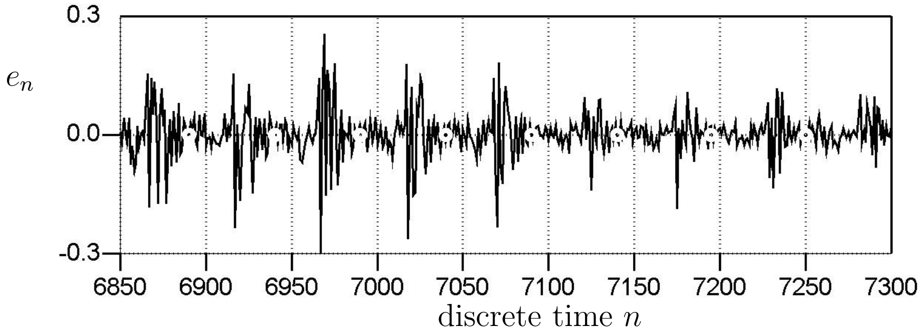

Figure 4.17: Prediction error signal for SMLLV of /uw/.

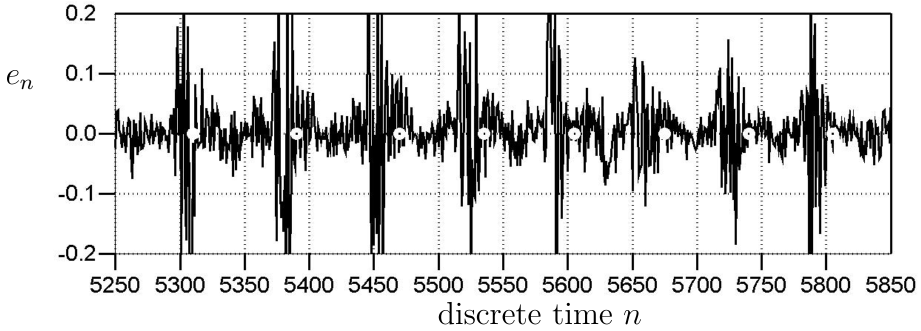

Figure 4.18: Prediction error signal for SMLLV of /iy/.

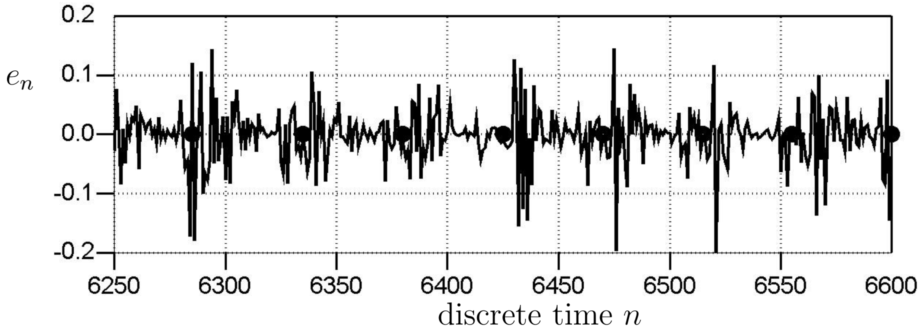

Figure 4.19: Prediction error signal for CFKGB of /uw/.

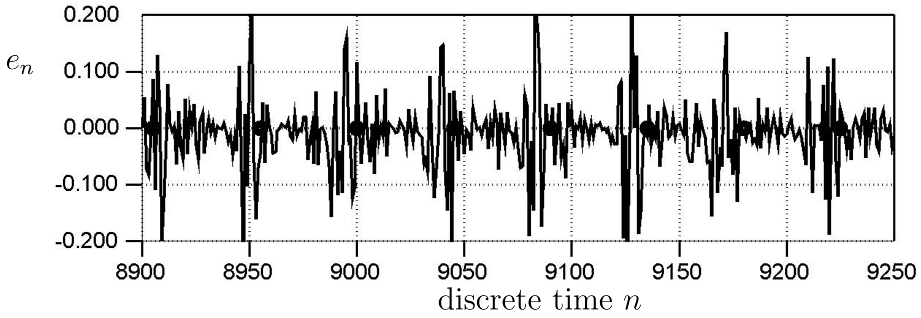

Figure 4.20: Prediction error signal for CFKGB of /iy/.

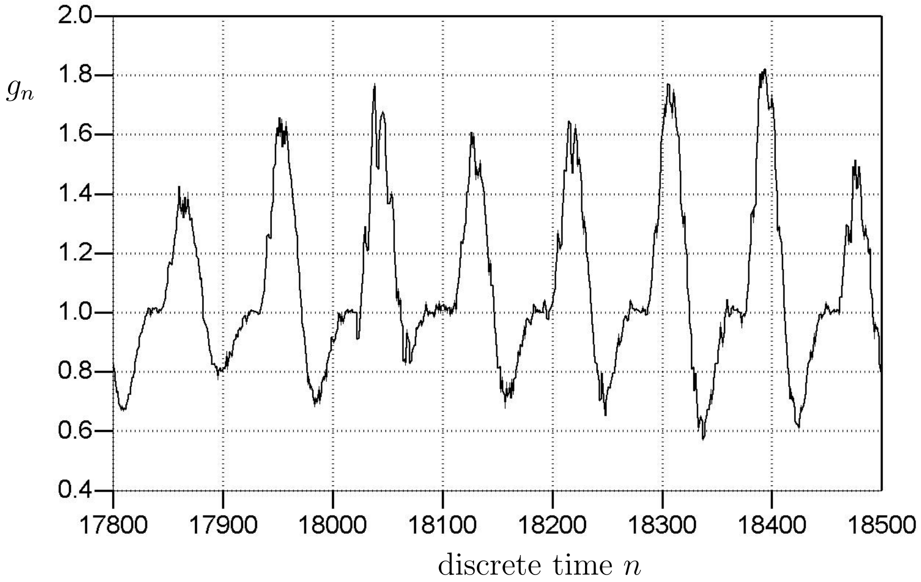

Figure 4.21: Glottal analysis for CFKGB of /iy/ sampled at 22.050 kHz.

standard glottal waveform the most are extracted from speaker SMWES in Figures 4.22 and 4.23. Rather than containing the single glottal opening pulse, these waveforms appear to contain two pulses. Glottal cycles that contain multiple open glottis phases are common for male speakers and are associated with the vocal fry register [6].

While a comparison between the glottal waveforms for SMWES with the standard glottal waveform demonstrates how glottal waveforms can vary widely amongst speakers, it is not uncommon for a speaker to vary their own glottal waveform as well. In fact, Figures 4.24 and 4.25 reveal that speaker AMDCM sometimes produces a glottal waveform with a single open glottis pulse, and other times produces a glottal waveform with two open glottis pulses. This is a common behavior referred to as diplophonic double-pulsing, where speakers that generally have a normal voice quality occasionally produce multiple excitations, particularly at the ends of utterances [31]. Figure 4.24 shows how a new pulse gradually develops just after the main pulse in each glottal cycle. This secondary glottal opening pulse can become quite large, as it was for SMWES, hence the glottal signal can change substantially for speakers such as AMDCM.

Another speech register that results in irregular glottal waveforms relates to weak voicing associated with speech breathiness. The breathy, nasal speech of AMGLS produced the glottal waveform in Figure 4.30. Although the inverse filter verification results do not reveal problems with the speech modeling performance, the glottal waveform bears little resemblance to the typical glottal waveform. The prediction-error signal in Figure 4.32 confirms a higher level of randomness and less pronounciation of excitation compared to other prediction-error signals such as the one shown in Figure 4.33 from the analysis on speech from the same speaker. Hence this register can also be controlled by certain speakers, as confirmed by the much more regular glottal signal shown in 4.31 for AMGLS on /iy/.

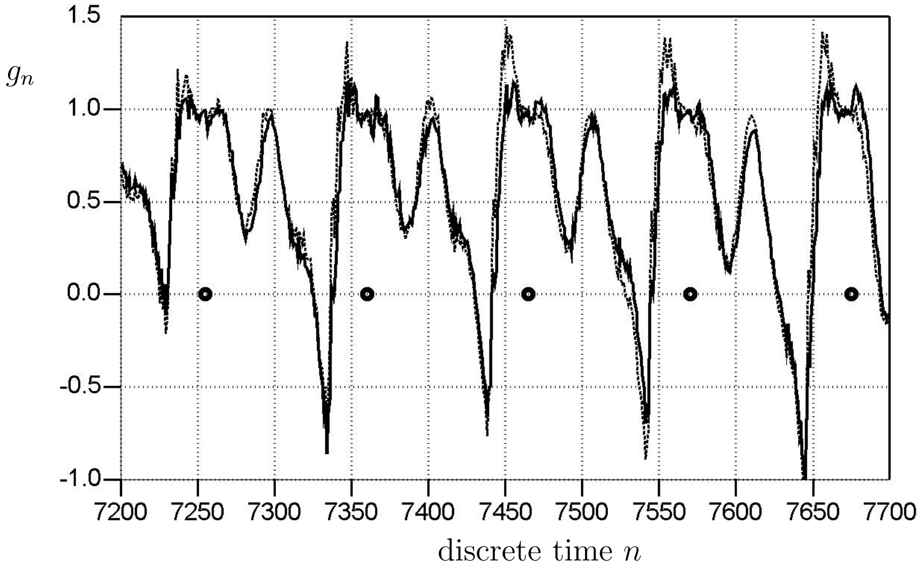

Figure 4.22: Glottal analysis for SMWES of /uw/.

Several visual comparisons have been made between the glottal waveform estimates given so far in this section. Particularly, it has been interesting observing how much each of the speaker's glottal waveform characteristics differ between the utterances /iy/ and /uw/ (intra-speaker), and how they compare to those of the

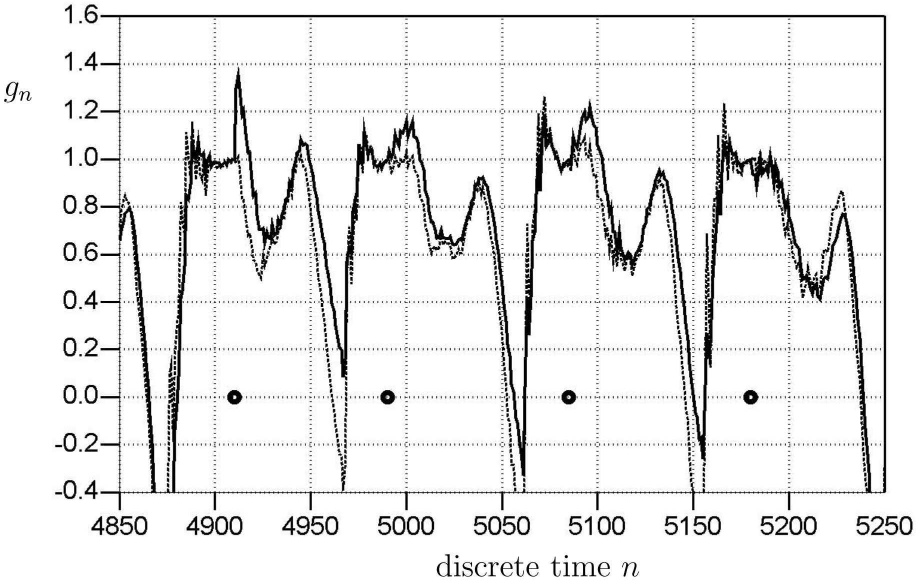

Figure 4.23: Glottal analysis for SMWES of /iy/.

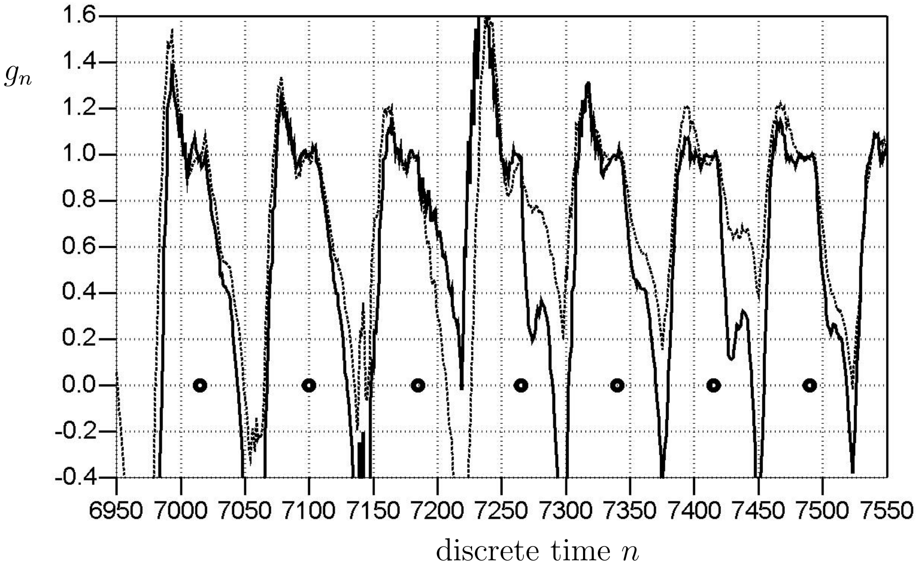

Figure 4.24: Glottal analysis for AMDCM of /uw/.

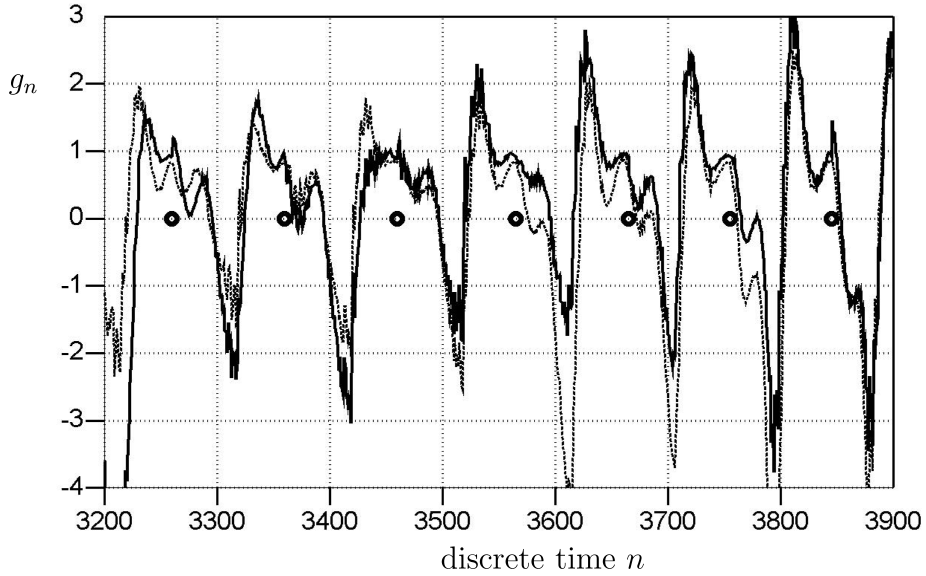

Figure 4.25: Glottal analysis for AMDCM of /iy/.

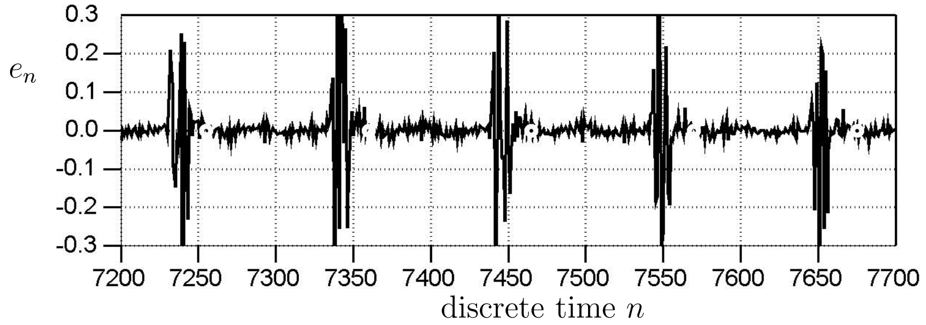

Figure 4.26: Prediction error signal for SMWES of /uw/.

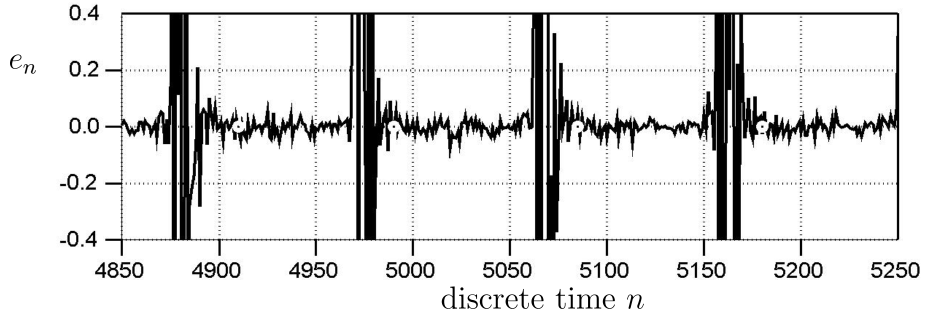

Figure 4.27: Prediction error signal for SMWES of /iy/.

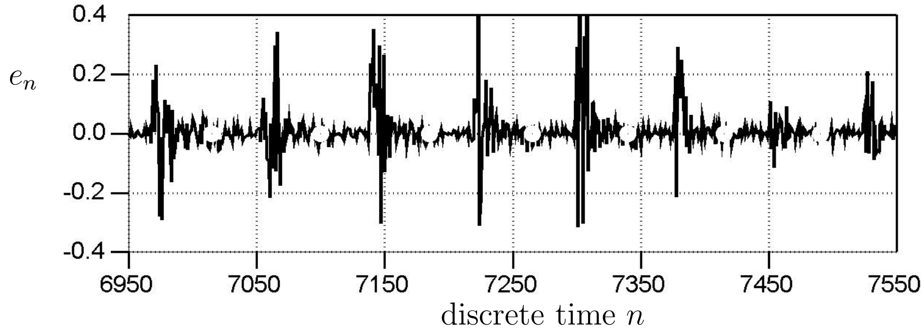

Figure 4.28: Prediction error signal for AMDCM of /uw/.

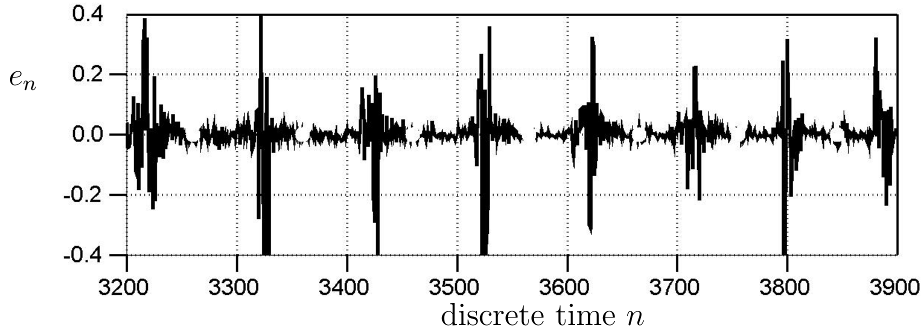

Figure 4.29: Prediction error signal for AMDCM of /iy/.

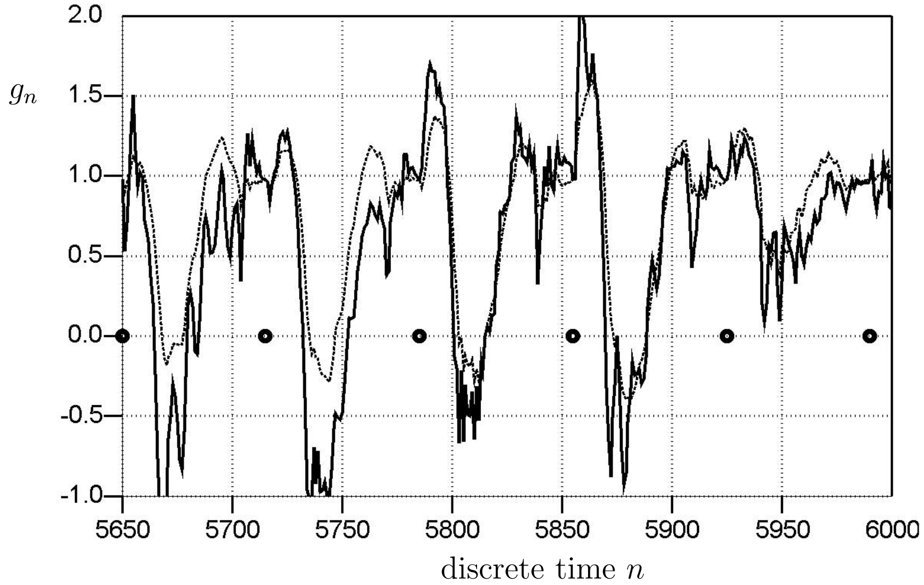

Figure 4.30: Glottal analysis for AMGLS of /uw/.

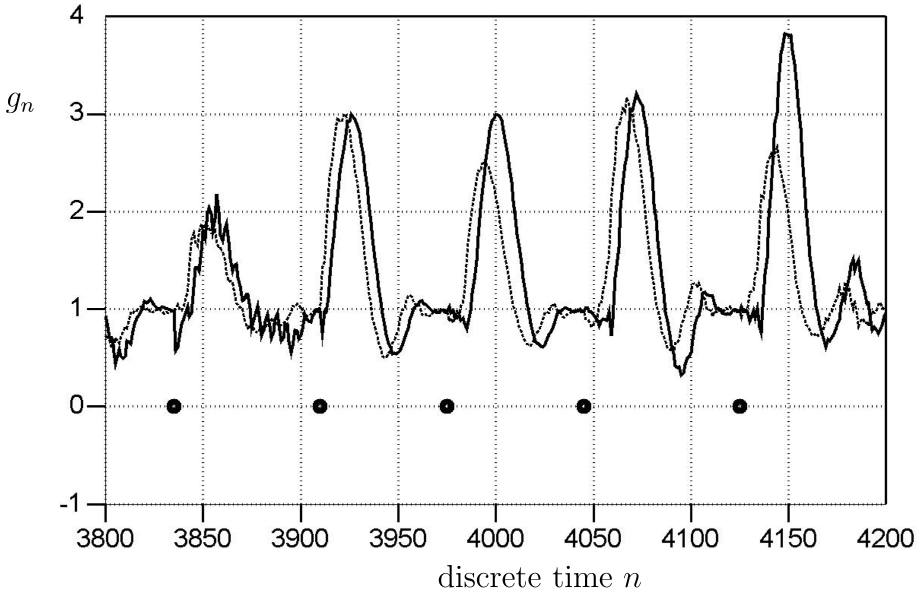

Figure 4.31: Glottal analysis for AMGLS of /iy/.

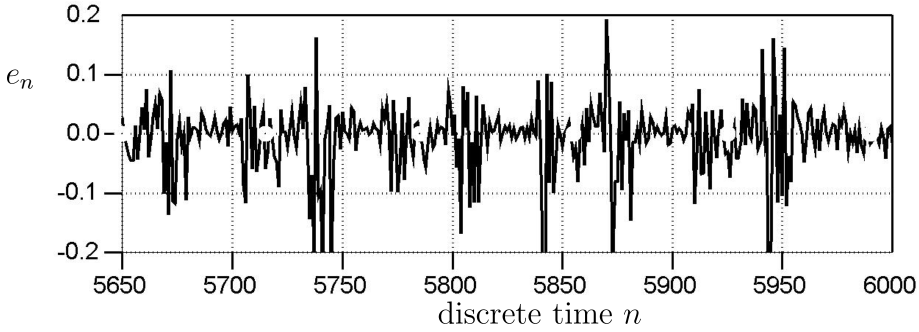

Figure 4.32: Prediction error signal for AMGLS of /uw/.

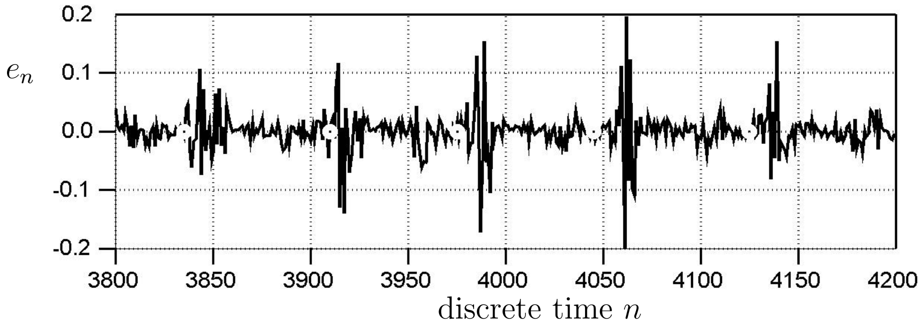

Figure 4.33: Prediction error signal for AMGLS of /iy/.

other speakers (inter-speaker). Now the Euclidean distance,

$$\mathcal{E} = \frac{1}{N} \sum_{n=0}^{N-1} |g_{1,n} - g_{2,n}|^2, \tag{4.2}$$

will be used to compare each of the glottal waveform estimates, say $g_{1,n}$ and $g_{2,n}$, over N samples. The time axes over the N samples of the two glottal waveforms are shifted so that $n = 0$ corresponds to a particular glottal closure sample n_c for each glottal waveform. N corresponds to the length of the shorter of the two glottal waveforms being compared. For example, samples 4750 through 5050 of the glottal waveform estimate in Figure 4.1 were compared to samples 6450 to 6750 in Figure 4.34.

A visual comparison between the AFRIF results for AFGLB on phoneme /uw/ in Figures 4.34 and 4.35 and phoneme /iy/ in Figures 4.1 and 4.2 reveals some noticable differences, particularly the suppressed glottal opening pulses in Figure 4.34 relative to those observed in Figure 4.1. Using the Euclidean distance, as given in Table 4.1, a distance of $3.83504(10)^{-6}$ was obtained. However in relation, comparing the glottal waveform estimates for AFGLB to the glottal waveform estimates given in this section from the other speakers resulted in an average distance of $1.06208(10)^{-5}$. The individual intra-speaker and inter-speaker comparisons for the seven speakers are broken down in Table 4.1. Other than for CFKGB, the inter-speaker distances always exceed the intra-speaker distances. Such a condition is consistent with speech features that are speaker dependent and vocabulary independent. Next, a more accurate measure of the inter-speaker and intra-speaker differences of the AFRIF results is pursued using an automated speaker identity verification (SIV) system on a much larger set of data.

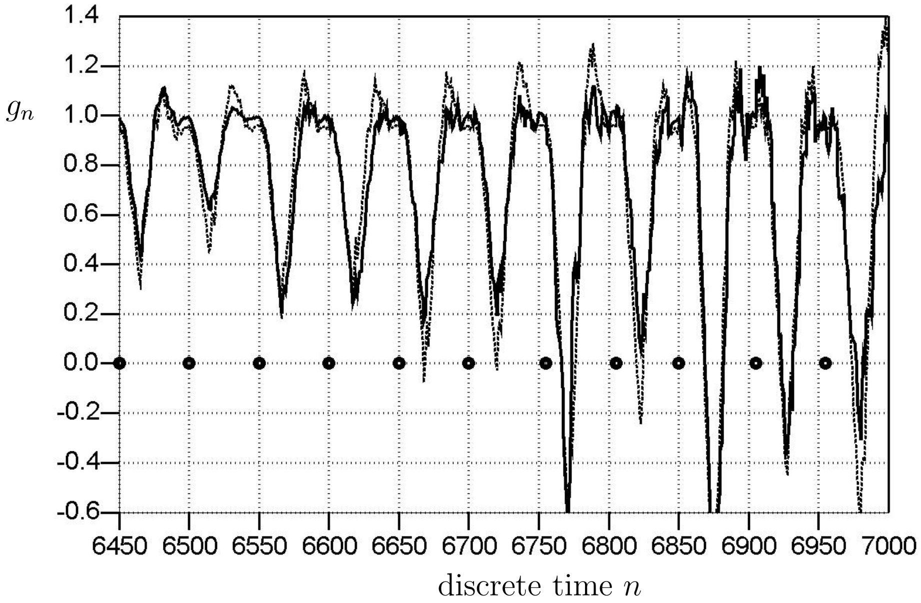

Figure 4.34: Glottal analysis for AFGLB of /uw/.

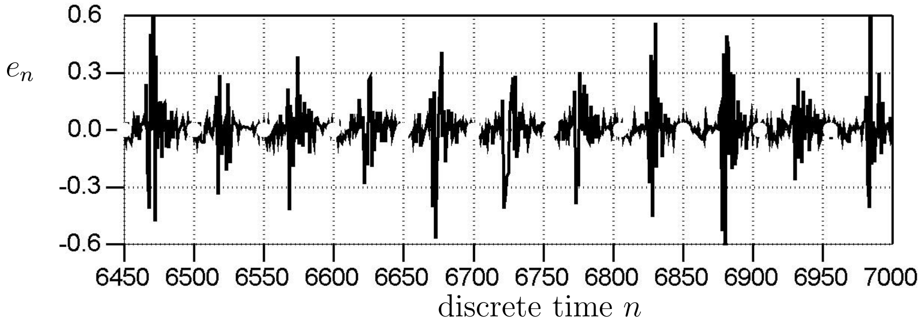

Figure 4.35: Prediction error signal for AFGLB of /uw/.

Table 4.1: Distances between glottal waveform estimates.

Subject	intra-speaker	inter-speaker
AMDCM	$3.332345(10)^{-7}$	$1.1428976(10)^{-5}$
AFGLB	$3.835039(10)^{-6}$	$1.0620845(10)^{-5}$
AMGLS	$2.142848(10)^{-4}$	$2.6148799(10)^{-4}$
CFKGB	$5.551650(10)^{-4}$	$2.8874504(10)^{-4}$
SMLLV	$1.976583(10)^{-4}$	$2.4161379(10)^{-4}$
AFMES	$4.073955(10)^{-4}$	$4.4060557(10)^{-4}$
SMWES	$1.670166(10)^{-8}$	$1.1255984(10)^{-5}$

4.2 A Simple SIV Example Using AFRIF

A simple speaker identity verification (SIV) system was used to evaluate the proposed AFRIF approach, and the effectiveness of the features it extracts, at characterizing or distinguishing speakers. The database used in the test was collected from a diverse group of male and female adults and children (refer to the Appendix), and provided a wide variation of pitch. The speech was sampled at 8 kHz and the analysis frames were set up so that a single set of coefficients (within the first detected closed glottis interval) was collected every 30 ms. The SIV approach assumed vocabulary dependence so that, given either the vocal tract or vocal cord features, always two different recordings of the same word are compared.

For the case of the vocal tract model prediction coefficients, two recordings of a word (spoken by the same or different speakers) were compared by converting them to cepstral coefficients and using the inverse variance approach explained in Section 2.5.1 for obtaining the weighted Euclidean distance. Due to its availability, DTW was used for evaluating the effectiveness of the features at characterizing speakers. A justification for choosing this matching technique is that it has been found to yield similar SIV accuracy as HMM approaches for short verification utterances when training data is limited [52]. Since we claimed that the AFRIF procedure is better at tracking transitions, and the focus of this investigation involves extracting all speaker-dependent information, a more appropriate alignment approach would be one that could better model speaker-dependent transitional information.

For the case of the vocal cord features, the actual glottal signal values derived from AFRIF are used in the verification. Instead of using the entire glottal signal, a segment of the glottal signal is chosen that provides a good representation of a typical period (or more) of the glottal signal, thus reducing storage and computations while screening out portions of the glottal signal that might contain artifacts that distort the glottal events (such as channel noises or vocal obstructions like mucous). The procedure used to find this representative glottal waveform segment always keeps two candidate segments, say $g_{A,n}$ and $g_{B,n}$; initially choosing the first two glottal waveform periods extracted by AFRIF. When the next period of the glottal waveform has been extracted, say $g_{C,n}$, it is compared to $g_{A,n}$, resulting in a distance $d_{A,C}$, and $g_{B,n}$, resulting in a distance $d_{B,C}$. The new segment $g_{C,n}$ is kept as the new glottal signal candidate if $d_{A,C}$ is smaller than $d_{B,C}$

and $d_{A,B}$ (in which case $g_{C,n}$ becomes the new candidate $g_{B,n}$) or $d_{B,C}$ is smaller than $d_{A,B}$ and $d_{A,C}$ (in which case $g_{C,n}$ becomes the new candidate $g_{A,n}$). This procedure continues until all periods have been compared to at least one other period, at which point one of the two final candidates, $g_{A,n}$ or $g_{B,n}$, is chosen as the representative glottal waveform segment. The time-domain glottal signal comparisons in this procedure used the inverse variance weighting for the Euclidean distance.

Recall that the proposed analysis approach requires some form of pitch detection in order to locate the closed-glottis interval. In real-time speech processing systems, automatic pitch detection routines are required. An automated implementation of the AFRIF procedure was used in this experiment that identifies one or more of the obvious time-domain glottal events within a 30 ms analysis frame based on the logic presented in Section 2.5.4. Hence, the glottal events that the feature extraction routine looks for are the main glottal opening and closing intervals. The extracted glottal signal, which in Section 3.3 was shown to correspond to the differential glottal waveform, is used to locate the glottal opening interval by identifying the glottal opening pulse. The glottal closing interval is located by identifying the typically sharp dip in the estimated differential glottal waveform as seen in Figure 3.4. The prediction error signal is also used to assist in the identification of the glottal closure interval. The inverse filter, $A_{nc}(z)$, is hence updated every frame with the new prediction coefficients.

Since the AFRIF procedure assumes that the speech input is voiced, an automated preprocessing routine is used that determines if the current frame consists of voiced or unvoiced speech. If the decision is “unvoiced”, standard LPC analysis is performed to obtain the vocal tract model prediction vector, but the glottal signal computation is bypassed altogether. The criterion used in the voiced/unvoiced/silence detection routine is quite standard [22], and the specific implementation used in this experiment follows that described in [3]. The main criterion is speech energy which, if large enough for a particular speech frame, results in a voiced classification. If a speech frame does not contain quite as much energy, but the number of zero crossings is low, as is the ratio of the energy in the prediction error signal frame to the speech energy, then the speech frame is also classified as being voiced. A speech frame that has been classified as voiced can be reclassified as unvoiced if the glottal event detection routine is unable to locate excitations within the speech frame.

The SIV system that incorporated the AFRIF procedure achieved an equal-error rate (EER) of 20%. As a reference, the SIV experiment was performed with a conventional linear predictive analysis algorithm (as described in [59]) resulting in an equal-error rate of 18%. Table 4.2 gives the resulting EER for each speaker. The third column in Table 4.2 gives the equal-error rate for each speaker using only the vocal tract AR parameters, $\mathbf{a}_{n_c}$ (EER $\mathbf{a}_{n_c}$). The fourth column shows the equal-error rate when the glottal signal is also used in the verification (EER $\mathbf{a}_{n_c}$ & G). For 21 of the 29 speakers, the glottal signal improved the equal-error rate as much as 2.28%, but degraded the equal-error rate by almost 14% for one of the 8 speakers whose glottal signal did not help to identify them. Hence the glottal signal contains useful speaker information for some speakers, but not for others. Breathy speakers correspond to an obvious case where the glottal information is very weak or nonexistent. In the latter case, it does not make sense to attempt to extract and use any glottal information, because the parameters are likely to be noisy. So in SIV, such noisy parameters can even degrade performance as observed with this database. One possible solution would be to determine such speakers and use only the AR parameters. So in Table 4.2, the features that produced the lower EER between the third and fourth columns would be used in the verification for each speaker. This new approach improves the EER only slightly, to 19.55%, but by almost 14% for breathy speaker AMSLH. As observed in Section 4.1.2, speakers commonly vary speech characteristics such as breathiness, therefore more extensive training may be required to determine the typical breathiness of the speakers.

A related drawback to the real-time AFRIF system was the need for automatic pitch detection, which is known to be a difficult problem that requires sophisticated logic, as discussed in Section 2.5.4. Again, breathy speakers present a challenge for pitch detection as well, as their weak voicing fails to produce the distinct glottal events used in standard pitch detection routines. As a result of not being able to precisely locate the closed-glottis intervals for breathy, high pitched, and other speakers with short (or inconsistent) closed-glottis intervals, inaccurate glottal waveforms, as compared to those obtained in Section 4.1 using manual pitch detection, were often observed. As shown in Section 4.1.2, more accurate glottal waveforms can be achieved for high pitched speakers by analyzing the speech when sampled at higher rates. Using glottal waveforms obtained from AFRIF estimation on speech from high pitched speaker CFKGB sampled at 22.050 kHz improved EER by 3.69%. Other notable improvements of 2.73% and 3.44% were achieved using the higher sampling rate for speakers TMMCM and CMRWT, respectively. As observed in the last column of Table 4.2, although the EER for most of the speakers actually degraded using the speech sampled at 22.05 kHz (resulting in an overall EER of 23%), improvements in pitch detection are expected to further improve the EER for breathy and other speakers with irregular closed-glottis intervals.

Table 4.2: Individual AFRIF SIV EER scores with and without glottal features.

Subject	Description	EER $\mathbf{a}_{n_c}$	EER $\mathbf{a}_{n_c}$ & G	EER 22.05 kHz
AMBBH	adult male	11.22	15.99	22.01
AFBJL	adult female	8.42	9.49	15.56
SMBNM	senior male	10.40	10.08	14.94
CMBSB	child male	18.03	19.08	19.15
CMBSS	child male	32.71	32.81	39.33
TFCAH	teen female	16.97	16.76	27.17
AMDCM	adult male	12.24	11.00	15.06
TMDNH	teen male	11.78	10.86	12.52
AFGLB	adult female	13.56	11.87	13.74
AMGLS	adult male	20.44	19.23	23.4
CFJLG	child female	15.04	22.98	25.81
TMJMM	teen male	20.65	18.95	19.68
CMJTA	child male	25.28	23.59	24.31
CFKGB	child female	24.9	23.09	19.4
CFKDT	child female	17.61	16.2	15.74
CMKNS	child male	17.83	16.7	14.87
CMKWA	child male	31.28	29.86	28.77
SMLLV	senior male	14.84	13.91	19.9
CMLRS	child male	35.42	33.14	35.17
SFMBM	senior female	22.82	21.04	28.26
TMMCM	teen male	46.43	47.15	44.42
AFMES	adult female	9.94	9.51	6.93
CMNEK	child male	29.47	28.54	28.94
CMPDS	child male	14.52	14.15	20.13
CMRSH	child male	18.58	17.58	23.66
CMRWT	child male	43.01	42.15	38.71
AMSLH	adult male	20.01	33.99	33.91
TFTKK	teen female	19.18	18.55	21.63
SMWES	senior male	8.45	8.67	21.21

Chapter 5

Summary

The AFRIF (Adaptive Forced Response Inverse Filtering) procedure has been proposed as an approach for extracting vocal tract and vocal cord information from a speech signal. In the AFRIF procedure, the vocal tract information is represented by autoregressive (AR) coefficients as in classical LPC analysis, and the vocal cord information is represented by an estimate of the classical glottal waveform. Unlike in speech processing systems currently used to extract this information however, an adaptive filter is used in the AFRIF procedure to derive the AR coefficients, and the glottal signal is extracted from the time-varying coefficients rather than from the prediction residual.

Although the approach also extracts a single vector of autoregressive coefficients per frame to model the vocal tract, unlike standard LPC analysis, AR coefficients are computed at every sample, and the vocal tract estimate is taken from within the closed glottis interval. To model the vocal cord behavior, the proposed system computes a glottal signal which is obtained by initially computing the unit step response through the time-varying filter, and by then filtering it through the time-invariant inverse filter defined by the autoregressive coefficients that represent the vocal tract for the respective frame. In effect, the AFRIF procedure takes advantage of the fast convergence properties of adaptive filtering to model the nonstationarities due to the vocal tract as well as the rapidly time-varying vocal cord behavior.

The AFRIF procedure was evaluated by estimating the glottal waveform from speech generated by a formant synthesizer, and was able to closely model the input glottal waveform. Glottal signals extracted from the speech of several speakers were also analyzed. The speech modeling performance was verified in this case by performing AFRIF analysis on speech that was synthesized using the glottal waveform and vocal tract estimates obtained in an initial AFRIF analysis performed on the original human speech. Glottal waveform estimates on the synthesized speech were achieved that closely matched those from the original human speech. Although many conventional glottal waveform extraction approaches assume highly controlled recording sessions, the AFRIF procedure demonstrated that successful glottal waveform modeling can be achieved in more natural conditions without the need for special rooms, recording equipment, or manual analysis intervention. Hence, the AFRIF procedure was shown to be a valuable candidate for glottal waveform modeling. Finally, the effectiveness of the features, extracted by the proposed procedure, at distinguishing speakers was evaluated using a simple speaker identity verification approach, and was found to yield reasonable performance, comparable to that achieved with standard autocorrelation analysis.

Appendix

For the analysis of the AFRIF procedure, a speech database was collected from a church congregation in Camarillo, California, and contains speech samples from a diverse group of speakers born and raised in various regions across the United States. The subjects consisted of 13 boys aged 4 to 16, 5 girls aged 7 to 17, 7 men aged 38 to 82, and 4 women aged 28 to 75 years of age. Each subject is categorized (and labeled) according to their age and sex as either a senior male (SM), senior female (SF) (for those ages 65 and over), adult male (AM), adult female (AF) (for those ages 20 to 64), teenage male (TM), teenage female (TF) (for those ages 13 to 19), male child (CM) or female child (CF) (for those under 13 years of age).

The vocabulary used for this database is given in Table 5.1. It is a subset of the phonetically diverse words used by Velius [59]. These words were recited by a given subject once per session. Subjects took part in ten recording sessions. Collection occurred over a six month period.

rich	shoes	home	full	wheat	egg	dove	toy	three	nine
zero	two	five	one	eight	six	clear	oh	seven	four

Figure 5.1: Database vocabulary.

Recordings were performed in a quiet room, with a 16 bit sound card, at a sampling rate of 22.05 kHz. The digital recordings were scaled and decimated to 8 kHz using four multirate FIR filter stages with resampling factors of 2/3, 4/3, 5/7 and 4/7. Automatic endpoint detection was performed on the decimated speech, removing most of the nonspeech sections from the recordings. The parameters of this endpoint detection routine (documented in [3]) were set conservatively to avoid the removal of speech. The endpoint detected speech was used as the input data to the analysis algorithm.

Bibliography

[1] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components" *Speech Communication*, Vol. 1 Nos. 3/4, pp. 167–184, Dec. 1982.

[2] B. Atal, V. Cuperman and A. Gersho, Speech and Audio Coding for Wireless and Network Applications, Kluwer Academic Publishers, 1993.

[3] B. L. Berg and T. C. Feustel, "A complete endpoint detection routine for use in speaker identity verification", Technical Memorandum TM-STS-021674, Piscataway, NJ, Bellcore, 1992.

[4] J. P. Campbell, Jr., "Speaker recognition: a tutorial," *Proc. IEEE*, vol. 85, pp. 1437–1462, 1997.

[5] F. Charpentier and E. Moulines, "Text-to-speech algorithms based on FFT synthesis", *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, New York, pp.667–670, 1988.

[6] D. G. Childers and C. K. Lee, “Voice quality factors: analysis, synthesis, and perception,” *J. Acoust. Soc. Am.*, Vol. 90 No. 5, pp. 2394–2410, Nov. 1991.

[7] D. G. Childers, J. C. Principe, and Y. T. Ting, “Adaptive WRLS-VFF for speech analysis,” *IEEE Trans. Speech and Audio Processing,* Vol. 3, No. 3, pp. 209–213, May 1995.

[8] R. H. Colton and J. K. Casper, Understanding Voice Problems, Williams & Wilkins, 1996.

[9] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 28, pp. 357–366, 1980.

[10] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis Discrete-Time Processing of Speech Signals, IEEE Press, 2000.

[11] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, New York, 1973.

[12] F. Fallside and W. A. Woods, Computer Speech Processing, Prentice Hall, 1985.

[13] J. L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1972.

[14] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 29, pp. 342–350, 1981.

[15] D.J. Gardner, D. DeFruiter, M. Keith, M. Dresel and R. Knapp, "Automated speech recognition as a function of formal speech training and passage of time between template training and test," *Proceedings of the Human Factors Society-29th Annual Meeting*, 1985.

[16] J. D. Gibson, "Adaptive prediction in speech differential encoding systems," *Proc. IEEE*, vol. 68, pp. 488–523, 1980.

[17] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 32, pp. 236–244, 1984.

[18] C. Hamon, E. Moulines and F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, New York, pp.238–241, 1989.

[19] W. M. Hartmann, Signals, Sound, and Sensation, Springer-Verlag, 1996.

[20] S. Haykin, Adaptive Filter Theory, Prentice Hall, 1995.

[21] H. Hermansky, B. A. Hanson and H. Wakita, “Perceptually based linear predictive analysis of speech,” *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 509–512, 1985.

[22] W. Hess, Pitch Determination of Speech Signals, Springer-Verlag, 1983.

[23] H. Hollien, “On vocal register,” *J. Phon.*, Vol. 2 pp. 125–144, 1974.

[24] J. N. Holmes, “The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer,” *IEEE Trans. Audio and Electronics,* Vol. Au-21 No. 3, pp. 298–305, June 1973.

[25] J. N. Holmes, “Low-frequency phase distortion of speech recordings,” *J. Acoust. Soc. Am.*, Vol. 58 No. 3, pp. 747–749, Sep. 1975.

[26] M. J. Hunt, J. S. Bridle and J. N. Holmes, “Interactive digital inverse filtering and its relation to linear prediction methods,” *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 15–18, 1978.

[27] M. J. Hunt, S. M. Richardson, D. C. Bateman and Alain Piau, “An investigation of PLP and IMELDA acoustic representations and of their potential for combination,” *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 881–884, 1991.

[28] B. H. Juang and S. Furui, "Automatic recognition and understanding of spoken language," *Proc. IEEE*, vol. 88, pp. 1142–1165, Aug. 2000.

[29] S. M. Kay, Modern Spectral Estimation: Theory and Application, Prentice Hall, 1988.

[30] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, Vol. 67 No. 3, pp. 971–995, Mar. 1980.

[31] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, Vol. 87, pp. 820–857, Feb. 1990.

[32] A. M. Kondoz, Digital Speech: Coding for Low Bit Rate Communication Systems, John Wiley & Sons, New York, 1999.

[33] P. Lieberman, Intonation, Perception, and Language, MIT Press, 1967.

[34] J. Makhoul, "Linear prediction: a tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[35] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition," *IEEE Signal Processing Magazine,* pp. 58–71, Sep. 1996.

[36] H. E. Miller and M. V. Mathews, "Investigation of the glottal waveshape by automatic inverse filtering," *J. Acoust. Soc. Am.*, Vol. 35 pg. 843 (A), 1963.

[37] B. C. J. Moore, An Introduction to the Psychology of Hearing Academic Press, 1997.

[38] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, Vol. 9 Nos. 5/6, pp. 453–467, Dec. 1990.

[39] —, "Diphone synthesis using a multipulse LPC technique", *Proc. FASE Int. Conf.*, Edinburgh, pp.47–54, 1988.

[40] D. O'Shaughnessy, "Speaker recognition," *IEEE ASSP Magazine,* pp. 4–17, Oct. 1986.

[41] D. O'Shaughnessy, Speech Communication: Human and Machine., IEEE Press, 2000.

[42] J. O. Smith, III, and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech and Audio Processing,* Vol. 7, No. 6, pp. 697–708, Nov. 1999.

[43] S. Parthasarathy and D. W. Tufts, "Excitation-synchronous modeling of voiced speech," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 35, pp. 561–580, 1987.

[44] P. Papamichalis, Practical Approaches To Speech Coding, Prentice-Hall, 1987.

[45] T. Parsons, Voice and Speech Processing, McGraw-Hill, 1987.

[46] J. Perkell and D. Klatt, eds., Invariance and Variability in Speech Processes, Lawrence Erlbaum Associates, 1986.

[47] N. B. Pinto, D. G. Childers and A. L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 37, No. 12, pp. 1870–1887, Dec. 1989.

[48] M. D. Plumpe, T. F. Quatieri and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Processing,* Vol. 7, No. 5, pp. 569–585, Sep. 1999.

[49] M. R. Portnoff, "Implementation of the digital phase vocoder using the Fast Fourier Transform", *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 24, pp. 243–248, June 1976.

[50] L. R. Rabiner, "A tutorial on hidden markov models and Selected applications in speech recognition," *Proc. IEEE*, pp. 257–285, 1989.

[51] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.

[52] A. E. Rosenberg, C. Lee and S. Gokcen, "Connected word talker verification using whole word hidden markov models," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 381–384, 1991.

[53] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 873–876, 1986.

[54] S. Roucos and A. Wilgus, "High quality time-scale modification for speech", *Proc. Int. Conf. Acoust., Speech, Signal Proc., Tampa*, pp.493–496, 1985.

[55] S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction", *IEEE Trans. Acoust., Speech, Signal Processing,* vol. 30, no. 4, pp. 566–577, 1982.

[56] C. Shadle, "Intrinsic fundamental frequency of vowels in sentence context," *J. Acoust. Soc. Am.*, Vol. 78 pg. 1562–1567, 1985.

[57] X. Sun, "Voice quality conversion in TD-PSOLA speech synthesis," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 2, pp. 953–956, 2000.

[58] I. R. Titze, Principles of Voice Production, Prentice Hall, 1994.

[59] G. Velius, "Variants of cepstrum based speaker identity verification," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 583–586, 1988.

[60] D. Y. Wong, J. D. Markel and A. H. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 27, pp. 350–355, 1979.

[61] L. Xu, J. Oglesby and J. S. Mason, “The optimization of perceptually-based features for speaker identification,” *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, pp. 520–523, 1989.

Vita

Brian L. Berg
29307 Quail Run
Agoura Hills, CA, 91301
(818) 706-8802
brian_berg@agilent.com

EDUCATION

Ph.D. Electrical Engineering, July 2001
Virginia Tech, Blacksburg, VA. GPA: 3.91/4.00
Funding: Bradley Fellowship–awarded by Bradley Dept. of Electrical Engineering
Dissertation: Investigating Speaker Features From Very Short Speech Records

M.S. Electrical & Electronics Engineering, August 1991
North Dakota State Univ., Fargo, ND. GPA: 3.92/4.00
Thesis: Time-Varying Modeling of Nonstationary Signals.

B.S. Electrical & Electronics Engineering,
with Computer Option March 1990
North Dakota State Univ., Fargo, ND. GPA: 3.26/4.00

Grad. Courses

Adaptive & Stochastic Filters
Neural Networks
Spectral Estimation
Information Theory
Control Systems
Real & Functional Analysis

Languages

C/C++
FORTRAN
Matlab
TMS320C54x Assembly
6800/68000 Assembly
HP 64000 Assembly

WORK

Agilent Technologies/HP, Westlake Village, CA, 1997-present

- Software radio investigator (technology and market): design of multi-mode/-band transcievers using advanced architectures such as digital IF and $\Sigma - \Delta$ modulators
- Multistage decimation/interpolation filters including CIC comb design and compensation for flexible, efficient RF downconversion for low power wireless handsets
- Cosine modulated and QMF bank (near/perfect reconstruction) design for MPEG and Dolby AC-3 audio coding and multi-channel receivers (e.g., OFDM and DSL)
- Digital filter design project originator: coordinate algorithm code (as well as write and debug), UI specification, QA testing, documentation (providing inputs and editing) and marketing
- Complex, nonlinear phase FIR filters (designing bandpass Hilbert transforms, phase compensators, and low delay systems) and multipass IIR transformations
- Digital communications pulse-shaping (Gaussian, EDGE, and raised cosine)

Hewlett Packard, Westlake Village, CA, 1994-97

- QA technical verification of wireless system simulator (e.g., GSM, TDMA, CDMA)
- Oversee S/W lifecycle (planning to obsolescence) on multiple computer platforms
- Work with world-wide sales force on customer problems, providing training and applications notes
- Worked in R&D simulation group, initially on an algorithm for computing the minimum number of executions for each node within a signal flow-graph required for a schedule period that will maintain bounded node I/O buffers for infinite length input data streams
- Algorithm investigation for parallel processor scheduling of signal flow-graph nodes using acyclic precedence graphs and a Hu-Level algorithm

1993 Summer Internship: Bell Atlantic, Silver Spring, MD

Project: developed new Speaker Identity Verification algorithm

- Changed algorithm: Autocorrelation to LMS and Fast Transversal Filter
- New feature set: from LPC coefficients to formants, pitch and glottal features
- Achievements: Extraction of tangible features, vocabulary dependent and ind.

1992 Summer Internship: Bell Communications Research, Piscataway, NJ

Project: developed Speaker Identity Verifier

- Improved endpoint detector, added adaptive thresholds and word spotting
- Produced a Technical Memorandum of the work
- Achievements: increase in accuracy, patent for home incarceration application

Graduate Teaching Assistant, 1990-91

EEE Dept., NDSU, Fargo, ND.

- Taught Signals and Circuits labs to over 100 students
- Developed new signals/communications lab and experiments
- Set up signal analysis/filter design software and TMS32010 board

PUBLICATIONS

B. L. Berg and A. A. Beex, “Investigating Speaker Features From Very Short Speech Records,”*Proc. IEEE Int. Symp. Cir. Systems,* vol. 3, pp. 102-105, 1999.

B. L. Berg and T. C. Feustel, “A Complete Endpoint Detection Routine for Use in Speaker Identity Verification”, Technical Memorandum TM-STS-021674, Piscataway, NJ, Bellcore, 1992.