2012

# COMPARING ACOUSTIC GLOTTAL FEATURE EXTRACTION METHODS WITH SIMULTANEOUSLY RECORDED HIGH-SPEED VIDEO FEATURES FOR CLINICALLY OBTAINED DATA

Sean Michael Hamlet
*University of Kentucky*, seanmhamlet@gmail.com

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained and attached hereto needed written permission statements(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine).

I hereby grant to The University of Kentucky and its agents the non-exclusive license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless a preapproved embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Sean Michael Hamlet, Student

Dr. Kevin D. Donohue, Major Professor

Dr. Zhi David Chen, Director of Graduate Studies

# COMPARING ACOUSTIC GLOTTAL FEATURE EXTRACTION METHODS WITH SIMULTANEOUSLY RECORDED HIGH-SPEED VIDEO FEATURES FOR CLINICALLY OBTAINED DATA

---

THESIS

---

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in Electrical Engineering
in the College of Engineering
at the University of Kentucky

By

Sean Michael Hamlet

Lexington, Kentucky

Director: Dr. Kevin D. Donohue, Professor of Electrical Engineering

Lexington, Kentucky

2012

ABSTRACT OF THESIS

# COMPARING ACOUSTIC GLOTTAL FEATURE EXTRACTION METHODS WITH SIMULTANEOUSLY RECORDED HIGH-SPEED VIDEO FEATURES FOR CLINICALLY OBTAINED DATA

Accurate methods for glottal feature extraction include the use of high-speed video imaging (HSVI). There have been previous attempts to extract these features with the acoustic recording. However, none of these methods compare their results with an objective method, such as HSVI. This thesis tests these acoustic methods against a large diverse population of 46 subjects. Two previously studied acoustic methods, as well as one introduced in this thesis, were compared against two video methods, area and displacement for open quotient (OQ) estimation. The area comparison proved to be somewhat ambiguous and challenging due to thresholding effects. The displacement comparison, which is based on glottal edge tracking, proved to be a more robust comparison method than the area. The first acoustic methods OQ estimate had a relatively small average error of 8.90% and the second method had a relatively large average error of -59.05% compared to the displacement OQ. The newly proposed method had a relatively small error of -13.75% when compared to the displacements OQ. There was some success even though there was relatively high error with the acoustic methods, however, they may be utilized to augment the features collected by HSVI for a more accurate glottal feature estimation.

KEYWORDS: Linear Prediction, Acoustic Signals, Glottal Features, Inverse Filtering, High-Speed Imaging

Sean Michael Hamlet

December 4, 2012

# COMPARING ACOUSTIC GLOTTAL FEATURE EXTRACTION METHODS WITH SIMULTANEOUSLY RECORDED HIGH-SPEED VIDEO FEATURES FOR CLINICALLY OBTAINED DATA

By

Sean Michael Hamlet

Dr. Kevin D. Donohue
Director of Thesis

Dr. Zhi David Chen
Director of Graduate Studies

December 4, 2012

To Regina

ACKNOWLEDGMENTS

I would like to thank my Thesis Advisor, Dr. Kevin Donohue for his knowledge and guidance throughout the research and thesis writing process.

I would like to thank Dr. Rita Patel of the UK Clinical Voice Center and Hari Unnikrishnan for their intellectual contributions and for providing the data from the subjects.

I would also like to thank Dr. Sen-ching "Samson" Cheung and Dr. Laurence Hassebrook for being on my Defense Committee.

I wish to thank my family, especially my wife, for helping me through the tough times and for always being there.

This thesis project was supported in part by the Lexmark Fellowship and Dr. Robert D. Hayes Fellowship. I would like to acknowledge that the data collected for this experiment was also supported by NIH/NIDCD R03DC11360-01.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter One: Introduction

## Introduction to Speech

Speech is a communication tool that is utilized by many human cultures in the world. Communication is a fundamental way in which humans interact with one another. This essential role that speech has made in society has lead to strong scientific research interest in how humans anatomically produce sounds, especially in the fields of phonetics, phoniatrics, cognitive neuroscience and engineering [1]. And as time has progressed, speech processing techniques have sufficiently developed for explicit extraction of information in the speech waveform features related to the production. Therefore, features related to speech production may be more accurate and more distinguishable among different speakers. Several different models and methods have been proposed and studied to understand speech production and extract features to properly determine speech production functionality. These features may aid in the fundamental understanding of the underlying structure of speech production and may lead us to determine what is normal and what may be related to vocal disorders or pathologies in terms of extracted information.

## Speech Production

Speech production is a result of airflow from the lungs moving through 3 main processes: glottal excitation, vocal tract filtering, and lip radiation effects [2]. The glottal excitation or glottal source is essentially the pulsating air-flow waveform produced by the lungs controlled by the abduction and adduction of the vocal folds. This airflow becomes quasi-periodic due to the periodic motion of vocal fold vibration, which is the controlling factor in voiced speech. However, unvoiced speech can also occur when the vocal folds stop vibrating and allow forced air through the vocal tract, producing turbulence, with the unvoiced sounds being controlled mainly

by the tongue, teeth, and lips [3]. The vocal tract filter allows for the airflow waveform to resonate throughout the cavity and an airflow fundamental frequency to equal that of the glottal source at the vocal folds. As the air is expelled through the lips, a direction effect and gain are applied to the output acoustic waveform. The shape of the oral and nasal cavity, as well as the oscillation of the vocal folds has a strong effect on vocal quality and clarity [4].

## Speech Modeling

It is well known that speech is a non-Linear Time-Invariant process since the vocal tract, nasal cavity and oral cavities constantly change shape during speech. However, simplification of the production into a Linear Time-Invariant Source-Filter Model, which is a reasonable approximation over short time intervals, has greatly increased the feasibility and ease of separating the components of the speech model for better understanding of their functionality [3]. The acoustic speech signal $s[n]$ is a result of the convolution of the glottal source waveform $g[n]$, which is the volume of the lung-produced airflow across the vocal folds, with the impulse-response of the filter created by the vocal tract, which represents the resonating formant frequencies. The simplification of the model, which is known to be time-varying, to a time-invariant model allows for basic Digital Signal Processing and linear systems techniques to be utilized. A visual representation of speech modeling is shown in Figure 1.1.

The most popular method of speech source estimation is inverse filtering, which utilizes the output acoustic waveform as an input to a system that removes vocal tract component and lip radiation component effects to achieve the glottal source. Even though there are many ways to properly characterize the vocal tract, past studies have relayed the difficulty of determining the accuracy of such methods. Nevertheless, properly characterizing features in the glottal source waveform is very

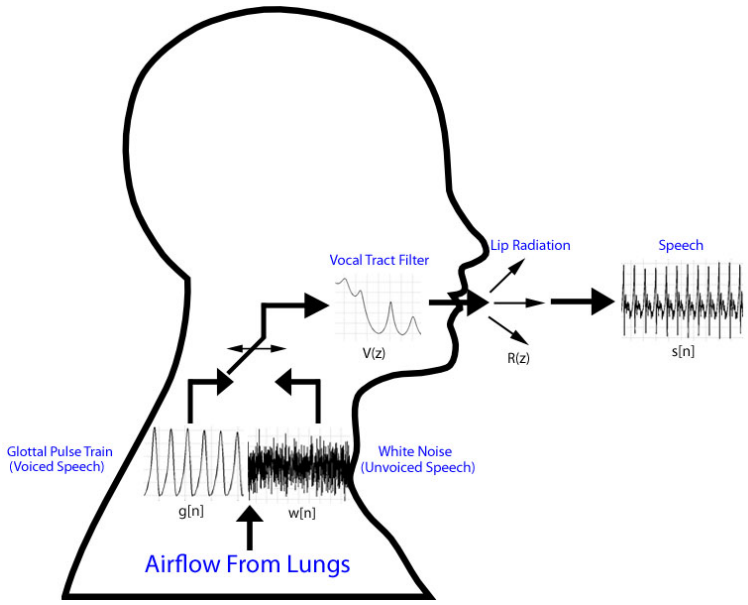


Figure 1.1: A Linear Time-Invariant Source-Filter Model of Speech Production

important to understanding the fundamental structure of voice production.

The time domain and Z-domain representation of the Linear Time Invariant Model of speech production are

$$s[n] = g[n] * v[n] * r[n] \tag{1.1}$$

$$S(z) = G(z)V(z)R(z) \tag{1.2}$$

where $S(z)$, $G(z)$, $V(z)$, $R(z)$, are the output speech, input glottal source, transfer function of the vocal tract filter, and transfer function of the lip radiation, respectively. During voiced speech, the source is essentially a train of quasi-periodic glottal air pulses, $g[n]$, however, during unvoiced speech, the source is better modeled by white noise $w[n]$. The glottal source is essentially the volume velocity of airflow, which occurs in these quasi-periodic pulses, across the vocal folds over time. This airflow is produced by the lungs and as it passes across the vocal folds, the vocal folds vibrate to control the airflow velocity variation and change the output pressure waveform that exits the mouth and/or nose. The lip radiation $R(z)$ can be

treated as a gain and direction application to the expelling air. The speech waveform $s[n]$, can be determined by converting $S(z)$ to the time-domain.

The fundamental frequency $f_0$, or pitch, of the speech waveform $s[n]$ is computed from $f_0 = 1/T_0$, which can be calculated using the autocorrelation method whose results are shown in Figure 1.2. By referencing the figures in Figure 1.2a, 1.2c, and 1.2e, it can be easily demonstrated how the glottal source waveform and output acoustic waveform have the same fundamental period $T_0$, which is understandable since the output acoustic waveform is ultimately created by the glottal source. In Figure 1.2b, the spectrum of $s[n]$ is shown with its estimated formant frequencies (dashed lines) that relate to the vocal tract filter $V(z)$. The $V(z)$ spectrum is shown in Figure 1.2d and has peaks that correspond to the formant frequencies illustrated in Figure 1.2b.

## Glottal Source Modeling

The glottal source is the driving force behind speech production and therefore is a focus behind many medically-based speech production research including laryngeal pathology detection and healthy versus pathological voice distinction [5]. All of the characteristics of the glottal source affect speech clarity and quality, which make it a desirable waveform to realize. Therefore, these characteristics are sought after to understand normal and abnormal production of speech, which may be found in patients with certain vocal pathologies or vocal disorders. One of the most popular models of the glottal source is the Liljencrants-Fant model shown in Figure 1.3, which illustrates the time-domain characteristics previously discussed. In the Glottal Source waveform, in Figure 1.3, the symbol $U_0$ represents the maximum amplitude velocity of airflow the glottal source achieves, and $T_p$, $T_c$, and $T_o$ represent the peak time instant, closure time instant, and opening time instant, respectively.

The glottal source $g[n]$ period begins with vocal fold abduction, in which an

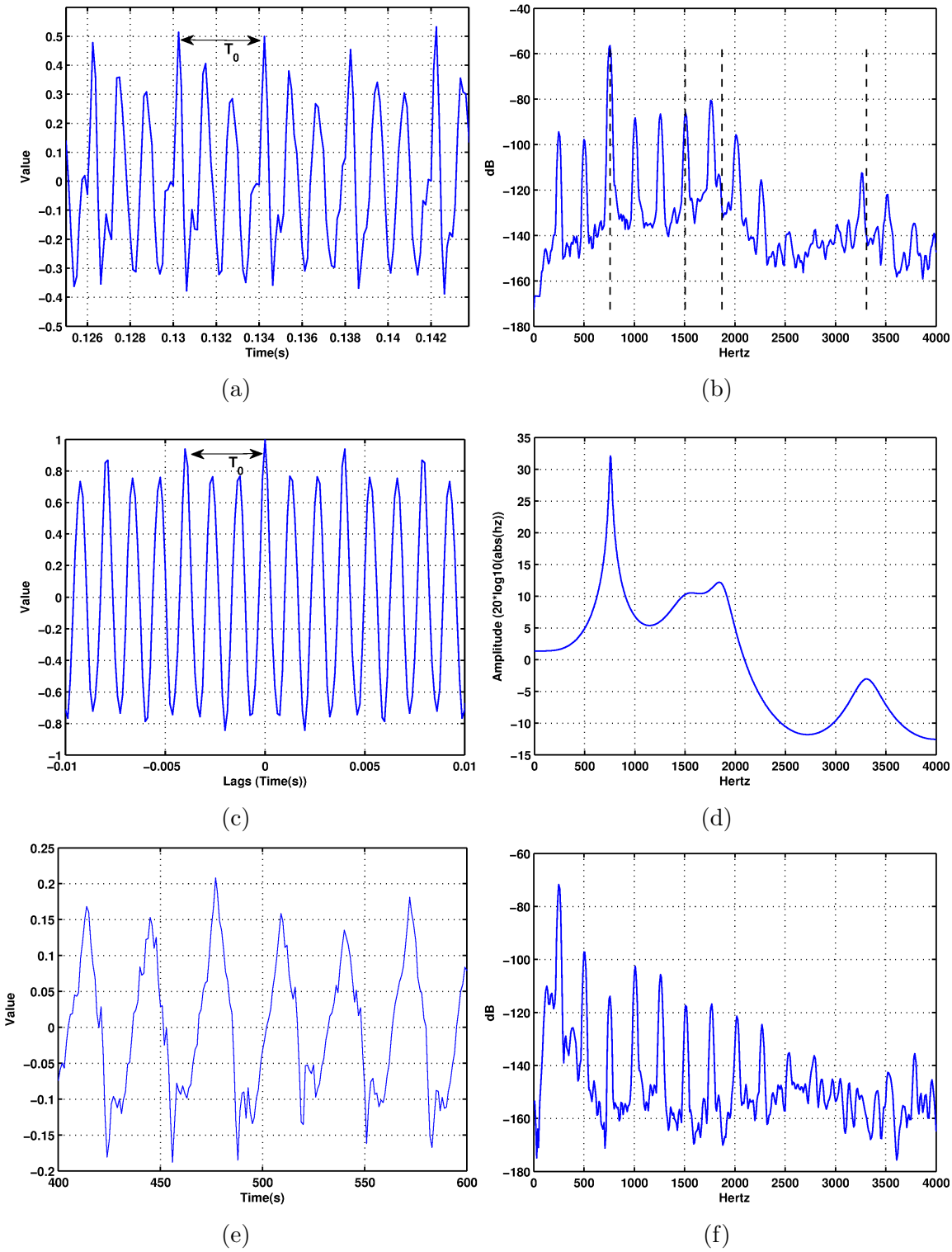


Figure 1.2: Speech, Glottal Source and Vocal Tract Component Examples During Phonation: (a) Acoustic Waveform $s[n]$, (b) Acoustic Waveform Spectrum $S(e^{jw})$, (c) Autocorrelation of Acoustic Waveform, (d) Vocal Tract Filter Spectrum $V(e^{jw})$, (e) Glottal Source Waveform $g[n]$, (f) Glottal Source Spectrum $G(e^{jw})$

increase in airflow occurs until the maximum velocity airflow at $T_p$. The elasticity of the vocal folds then causes them to adduct resulting in a strong negative peak in the glottal derivative $g'[n]$ immediately before the glottal closure instant (GCI) at time $T_c$. With these characteristics, time-domain features can be calculated, such as open quotient (OQ), which is the glottal source's open phase time divided by the source's total period $T$, and speed quotient (SQ), which is the glottal source's opening phase time divided by its closing phase time. These time-domain features, illustrated by Equations 1.3 and 1.4 may relate to the development of vocal fold pathologies or disorders, which may be of interest to voice research [6].

$$OQ = \frac{T_c - T_o}{T} \tag{1.3}$$

$$SQ = \frac{T_p - T_o}{T_c - T_p} \tag{1.4}$$

## Introduction to Methods Studied

Due to the previously described time-domain features possibly affecting speech production, ample research has been made in the field of feature extraction of the glottal source with the latest research being in the area of High-Speed Video Imaging (HSVI). Earlier studies have focused on feature extraction by more indirect methods, mainly due to the fact that the current method of HSVI was just too computationally extensive at the time. These indirect methods involve determining the dynamics of the glottis by utilizing the recorded acoustic signal. There have been many different methods proposed utilizing pressure masks, physical hardware, and even a reflectionless tube with focuses on glottal source extraction from the acoustic waveform, but one of the first and most popular is the Inverse Filtering method [1]. The two previously proposed methods that this thesis discusses are the

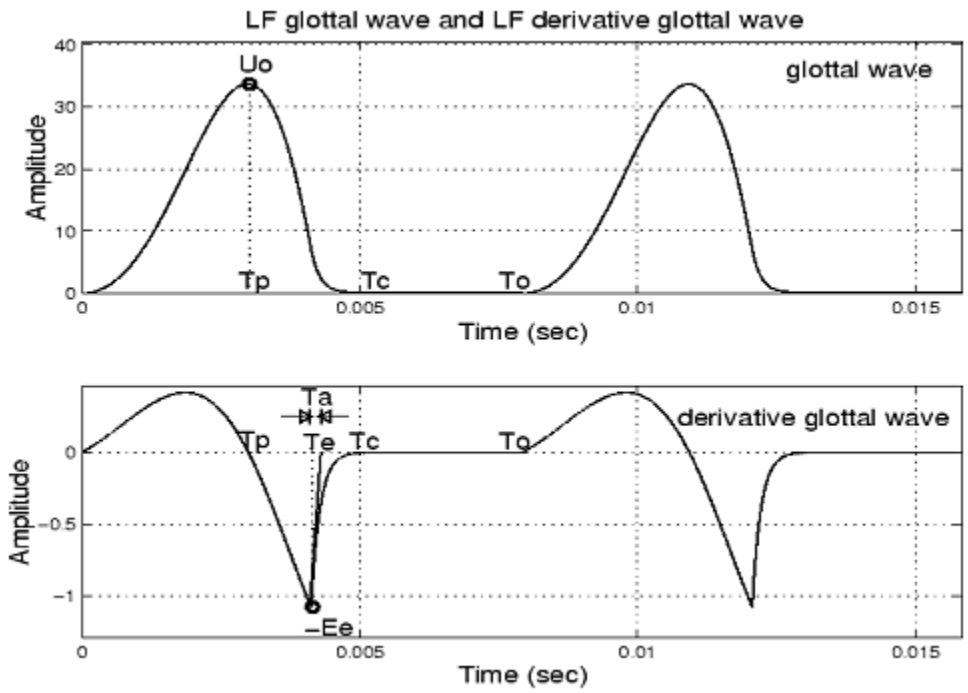


Figure 1.3: Liljencrants-Fant Glottal Model: Glottal Source (top) and Glottal Derivative (bottom). Adapted from "Recent Developments in Musical Sound Synthesis Based on a Physical Model" by Julius O. Smith III, 2003. [7]

Iterative Adaptive Inverse Filtering (IAIF) method and OQ estimation using linear prediction with glottal source modeling. Both methods employ the inverse filtering method, which aims to determine the effects of the vocal tract filter and lip radiation by inverse filtering the output speech waveform to derive the glottal source waveform. From this waveform, the OQ estimation can be computed. For this thesis' research, both these methods were utilized to calculate the open quotient of the glottal source dynamics and were compared with the simultaneous recorded HSVI data to determine their OQ estimation accuracy.

This twelve step process described by the first studied method, IAIF, removes the vocal tract effects with an iterative procedure by estimating these effects with LPC analysis and with these estimated effects, inverse filters the original speech signal and integrates the results to obtain the glottal source [2]. The IAIF method computes the glottal contribution and vocal tract transfer function

with an iterative structure that is repeated twice. The OQ estimation using linear prediction is a similar model to the previous method, however it does not perform the glottal source estimation with an iterative structure, like IAIF. Similarly though, this second previously studied method also treats the lip radiation as a differentiator and inverse filters the integrated voice signal to obtain the glottal source. A second order LPC analysis is then performed on the glottal source and the resulting two coefficients are used in an equation based on a modeling of speech production for OQ calculation [8]. A simple percent error was computed with the simultaneously recorded area and displacement waveforms to determine the accuracy of this and the previous method's OQ calculation.

There are limitations to the previous work. First, because these methods were proposed earlier in the research, they originally could not be compared to a standard value to access the accuracy of the glottal flow. For example, these methods were mostly evaluated using synthetic speech that was created with a specific glottal source and then, as the algorithm was utilized, the estimated glottal source was determined and compared with original. There are however some problems with just utilizing synthetic speech. A large problem is the fact that synthetic speech is more "mechanical" than natural speech and does not have the artifacts that occur during natural speech. Even if the previous work did compare its algorithm's accuracy using natural speech, there was not a very definitive way to access the actual glottal dynamics without HSVI. Moreover, only a few subjects, who sometimes were trained singers, were used to determine the accuracy. This, however, would not be best for determining accuracy of these methods in a clinical setting, where the subjects would most likely not be trained singers and may even suffer from vocal disorders or pathologies. Also, a comparison to HSVI data gives a visual standard to compare the previous method's and the following newly proposed method's OQ estimation.

After the results of the first two methods were obtained, observations lead to the creation of a new simpler method for determining the OQ from the recorded acoustic signal. Many research groups have examined the acoustically-extracted glottal source waveform or its derivative to understand the vocal fold dynamics. However, the utilization of linear prediction is strongly influenced by the pressure differential during speech phonation. During the glottal open phase, the linear prediction's error will be minimized due to the glottal dynamics and oscillating air in the vocal tract resulting in less change relative to the open and closing events. Moreover, since the largest pressure differential occurs right before glottal closure and immediately after glottal opening, large values in the prediction error will occur. These large errors will appear as strong spikes at the closure and opening time instants of the glottis in the LPC error waveform and will be periodic matching the fundamental frequency of the acoustic signal and therefore, the glottal source. By knowing the fundamental frequency of the acoustic signal, this simpler method focuses on tracking the error spikes closure time instants and then finds the opening time instant spike in between each period. From this, the OQ can be calculated from the known open and closure time instants.

It is very important to be able to extract information regarding glottal fold dynamics because the glottal source will determine the quality and clarity of speech. Therefore, this thesis hypothesis states that important time-domain features can be extracted from the recorded acoustic waveform that relate directly to glottal fold dynamics. The following section outlines this thesis' particular contribution.

## Thesis Contribution

Even though there has been much research in the area of glottal feature extraction with the recorded acoustic waveform, many of the experiments performed were carried out with the aid of other waveforms, like the Electroglottography

(EGG) signal, in order to help define glottal closure and opening instants [9] [8]. Even the experiments that strictly utilized the acoustic waveform for glottal feature extraction only compared the results with the features extracted from a synthetic model of the glottal source waveform or area waveform or a very small number of real subjects [10] [8]. This thesis's contribution is made by the fact that an extensive analysis of the IAIF and OQ Estimation using LP method's accuracy on real clinically-obtained data has not been performed. Also, since the obtained data also includes the simultaneous recorded HSVI data, we can use that as the standard for an objective comparison of the two methods of IAIF and OQ Estimation using LP, and the newly proposed method, to determine the accuracy of the methods on real clinical data.

Even though HSVI has become popular recently in determining voice features, there still may be advantages to observing and determining these features strictly from the acoustic signal. One advantage is the fact that HSVI along with other video techniques like, stroboscopy and kymography, require invasive procedures to extract their information, as well as costly equipment [1]. Another advantage is due to the fact that the acoustic signal can be recorded at a sampling rate much higher than the video recording frame rate. And since both the OQ and SQ features are time-dependent and relate explicitly to the glottal movement themselves, a higher sampling rate may yield better time estimates of key events and may lead to a more accurate measurement or could be used in complement with the HSVI data to achieve a better understanding of the glottal source dynamics.

In order to fully understand how well time-domain features are extracted from recorded acoustic signals, a comparison between clinically extracted features from High-Speed Video Imaging of the vocal fold vibrations and the extracted features from strictly the acoustic recording was necessary. Forty-six subjects, male and female, child and adult, were utilized over a range of fundamental frequencies

and ages to assure proper comparison and to restrict biasing due to type of speaker. An error analysis of the results aided in determining the validity and extremity of the relationship between the high speed video feature extraction and the acoustic signal feature extraction for each method. Utilizing the area between the vocal folds of the HSVI data as a waveform, a time-domain feature, open quotient, was calculated for each period of each subject over 30 cycles of vocal fold phonation. The mean open quotient values for the displacement of the vocal folds from the HSVI data were also utilized and because these features were calculated from a visual source that can be easily verified, they were used as the basis or standard for the comparison test.

## Thesis Organization

In this chapter, a simple explanation of speech production has been provided as well as a description of speech modeling, an introduction to the methods that were examined and compared and lastly, a new method of OQ estimation that has been proposed. Chapter 2 describes previous work and a more in depth look of how the examined glottal methods were derived. Chapter 3 describes the experimental design and equipment utilized to gather the original acoustic and video data, apply each method, and test, compare, and analyze the results. Chapter 4 emphasizes the comparison of the results from each method related to its corresponding video data results. Chapter 5 draws conclusions derived from the data as well as any limitations considered when utilizing each method.

# Chapter Two: Previous Work and Methods

## Previous Work

Before HSVI was available as a tool for kinematic vocal fold parameter extraction, researchers utilized a recorded acoustic waveform. There have been several acoustic glottal feature extraction methods, and the majority of them involve inverse filtering (IF). There are many applications to understanding the vocal fold parameters or glottal source parameters and each have lead research in a specific direction.

At Bell Laboratories, Schroeter proposed deriving model parameters for speech encoders from just the input speech signal in order to properly model and synthesize voice accurately. The reason for this particular research was that, at the current time, synthesized speech below bit rates of 4.8kb/s utilized a vocoder, which sounds unnatural and speaker identification was difficult to the listener [11]. Schroeter's method was essentially separated into three parts: an acoustic analysis system, a codebook of vocal tract and chord features and related acoustic characteristics, and an optimization of a voice synthesizer [11]. To extract the necessary glottal source features, a simple 10th order autocorrelation LPC analysis was performed on the input speech signal, as well as a pitch estimation and voicing parameter extraction that was necessary for the vocoder [11]. With the LPC coefficients, amplitude $A_{g0}$, speech energy $P_s$, and mass/spring scaling factor $q$, an extensive comparison with the codebook yielded the vocal fold and tract shape at that time segment.

When the closest model was chosen, its glottal source derivative waveform was synthesized and compared to the inversely filtered speech. A comparison in order to address the accuracy of the model parameters extracted from the speech and determine how well the vocoder models the input speech was performed.

However, for this method, no direct error percentages between the synthesized glottal source derivative $\dot{u}_g$ and inversely filtered speech $\hat{\dot{u}}_g$ were calculated. Only an optimization was performed by simply calculating a minimum distance between the two waveforms given by:

$$d_{u_g}(k) = 1 - \frac{\langle \dot{u}_g(i)\hat{\dot{u}}_g(i-k)\rangle - \langle \dot{u}_g(i)\rangle\langle \hat{\dot{u}}_g(i-k)\rangle}{[\langle \dot{u}_g^2(i)\rangle - \langle \dot{u}_g(i)\rangle^2]^{1/2}[\langle \hat{\dot{u}}_g^2(i)\rangle - \langle \hat{\dot{u}}_g(i)\rangle^2]^{1/2}} \tag{2.5}$$

However, without any error calculations, besides perceptually determining the improvement of this method by listening to the analyzed and corresponding synthesized speech signal, it is difficult to understand how much improvement was made.

A second method that has been proposed for obtaining glottal features, particularly glottal closure instants (GCI), is the Line of Maximum Amplitude (LoMA) method. This method observes the tree patterns that are associated with the time-scale domain and after a wavelet transform is performed for each period of the input speech signal, the maximum amplitude of each wavelet transform is linked together [9]. Several features can be determined from the LoMA method including: glottal closure instants, open quotient because of its relation to the LoMA phase delay, and glottal source amplitude related to cumulative amplitude of the LoMA [9]. It was determined by a comparison with the EGG signal that this was an "effective method for the detection of GCIs" [9]. Four subjects, 2 male and 2 female, were used in this study and the simultaneous acoustic and EGG signals were recorded with the help of the derivative of the EGG as the standard for GCI detection using thresholding and peak detection [9]. Three different algorithms were compared to the EGG signal in determining GCIs including: LoMA, LPC analysis of 18th order prior to LoMA to remove vocal tract effects, and the DYPSA algorithm for GCI detection.

Table 2.1: Miss Rates (MR) and False Alarm Rates (FA) for Male (M) and Female (F) Subjects for the LPC-LoMA (LPC), DYPSA (DYP), and LoMA (LOM) Algorithms compared to EGG GCIs [9]

| Method | MR T | MR M | MR F | FA T | FA M | FA F |
|---|---|---|---|---|---|---|
| LPC | 12.95% | 10.25% | 12.84% | 0.53% | 0.60% | 0.50% |
| DYP | 4.25% | 1.33% | 5.21% | 0.52% | 0.63% | 0.48% |
| LOM | 2.88% | 3.03% | 2.83% | 0.50% | 0.59% | 0.47% |

The objective comparison for this method is the derivative of the EGG waveform (DEGG) so false alarm is defined as GCI detected by the algorithm, but not present in DEGG and miss rate is defined as DEGG detected GCI, but not detected in the algorithm. The 4 subjects, 2 male and 2 female, read 3 short stories at normal, high, and low pitch and also recorded sustained vowels and spontaneous speech [9]. For the 4 subjects, miss rate (MR) and false alarm rate (FA) for the male (M), female (F), and total (T) voices for each algorithm is shown in Table 2.1. It appears from this table that the LoMa method is fairly accurate in determining the glottal closure instants when compared to EGG and when the comparison of the phase delay of the LoMA in calculating the open quotient and the EGG signal were analyzed it was determined they were "highly correlated" [9]. Again, it is difficult to evaluate the open quotient accuracy when an error percentage was not calculated.

A third method presented by Plumpe focused on utilizing extracted features from the glottal flow derivative in speaker identification. This derivative is extracted from inverse filtering utilizing the Closed glottis interval Covariance LPC analysis (CC-LP) method during the glottal closure segments of the glottal flow. These glottal closure time segments were identified through differences in formant frequency modulation during opening and closing phases of the glottal flow source [12]. Using the Liljencrants-Fant (LF) model, the course structure of the glottal flow derivative was determined and with the utilization of energy and perturbation components helped determine the fine structure [12]. From these two structures, the

glottal model parameter's could be utilized in speaker identification (SID) because of their components containing very speaker-specific identification information [12]. The glottal source derivative features that were utilized in SID were time-domain features such as open quotient, close quotient, and return quotient. It was determined by a large data set of male and female subjects that the course structure was about 60% accurate in SID, the fine structure was about 40% accurate in SID, and the combined components were about 70% accurate in correct speaker identification [12].

This Closed glottis interval Covariance Linear Predictive (CC-LP) method is also seen in other research, where covariance LPC analysis was applied to glottal closure phases of the acoustic signal that were indicated by the EGG waveform [13]. However, there may be some limitations to utilizing the CC-LP method, because if the closed glottis phase is overestimated in time, the analysis interval will include some of opening phase and the formant results will be affected. Also, if the CC-LP underestimates the closed interval, there will not be enough information to accurately determine the needed parameters. Assuming proper closed-phase determination, once these parameters were obtained, they were also, as in the previously discussed study, applied to the LF model and along with the fundamental period $T_0$, the glottal parameters could be calculated, such as open quotient, speed quotient or skewing [13].

This previous study performed this analysis on two natural speakers and on two synthetic speech waveforms but, this study did not objectively compare the parameter results for natural speech waveforms with any accepted parameters, so only synthetic speech was compared. The two methods compared for synthetic speech were AUDIO2LF, which derives the LF model parameters directly from the audio waveform, and the Formant Bandwidth Tracker method, which utilizes Formant Bandwidth pairs and then applies the AUDIO2LF method to derive the

LF model parameters. Even though the two methods were compared against each other using synthetic speech, in order to compare the results to the synthetic glottal source derivative, the same CC-LP method was performed on the synthetic speech to obtain the glottal source derivative, just as would be done to the natural speech. It was determined through slope, intercept, amplitude detection, and a correlation coefficient that the second method was more accurate when compared to the reference model by consistently having a higher coefficient than the first method. In detecting the glottal closure instant and glottal peak instant, the second method had a correlation coefficient of 0.76 and 0.62, respectively [13].

A fourth study and another method that utilizes inverse filtering was proposed by Paavo Alku and is known as Iterative Adaptive Inverse Filtering (IAIF). This method is first proposed in 1991 and utilizes LPC analysis and integration in an iterative manner to adaptively remove the vocal tract filter and lip radiation (differentiation) effects in order to achieve an estimation of the glottal source waveform [14]. This estimation can then be used to calculate time-domain parameters such as OQ, SQ, spectral tilt and skewness. One of the studies, which utilized pitch synchronous IAIF, calculated the glottal source using synthetic speech waveforms for a male and female. In this study, IAIF was utilized with autocorrelation LPC analysis and closed phase covariance LPC analysis to determine the glottal source waveforms. A noticeable limitation for the closed-phase LPC technique when breathy speech waveforms were utilized. Since a breathy waveform doesn't have a very explicit glottal closure time instant, it is difficult to determine the closed-phase interval [14]. This study only performed the algorithm on a couple synthetic speech waveforms and did not empirically address the accuracy of the results, just the improvement in relation to the CC-LP method. Another study that addresses IAIF utilizes Hidden Markov Models to help generate natural sounding synthetic speech [10]. However, this study doesn't objectively

compare the results with any time-domain features from the acoustically-extracted glottal source waveform.

A previously proposed study by Gang Chen, involved objectively comparing its acoustically-extracted glottal source waveform results to a simultaneously recorded HSVI extracted glottal flow source. This study aimed to determine how robust and accurate three separate methods were at estimating the glottal flow source from a voiced signal, with an emphasis on how noise affects the results. To compare the results, the HSVI area waveform was utilized and converted to a glottal flow source signal [15]. Synchronous acoustic and video data were utilized from six subjects, three females and three males, of which none had any vocal disorders [15]. In order to extract the video, a laryngoscope was required and since it was placed invasively in the throat, the attempted /i/ phonation's quality and clarity was affected [15].

For the nine recordings from each speaker, an attempt at pitch $F_0$ variation (low, normal, high) and voice quality variation (pressed, normal, breathy) was made. The video was recorded at 3000 frames/s and the area was obtained over 150 video frames from an edge-detection algorithm and visually verified before an open quotient calculation was computed [15]. The area OQ was noted by marking the time-instant of glottal opening and time-instant of maximum closure or minimum area if closure was not fully completed, divided by the cycle period $T$ [15]. The glottal source extracted from the area waveform was obtained by a Matlab toolkit LeTalker, which utilizes a three-mass vocal fold model [15]. And by setting the vocal fold shape for the /i/ phonation, using the default parameters and the area waveform, the corresponding glottal source was estimated. The estimation of the glottal source from the acoustic waveform was characterized by a model from a glottal flow codebook, which linked parameters of the inverse filtered acoustic waveform to the shape and duration of the estimated glottal source [15]. The

Table 2.2: OQ Estimation Error for Each Phonation Type (Breathy (B), Normal (N), Pressed (P)), Pitch Level (Low (L), Normal (N), High (H)), and Gender for Clean and SNR of 5dB [15]

| | | B | N | P | L | N | H |
|---|---|---|---|---|---|---|---|
| clean | Male | .035 | .072 | .107 | .025 | .053 | .082 |
| | Female | .083 | .049 | .155 | .045 | .098 | .148 |
| 5dB | Male | .064 | .092 | .120 | .035 | .063 | .084 |
| | Female | .092 | .108 | .207 | .104 | .123 | .161 |

codebook was generated using specific parameters such as open quotient and asymmetry coefficient to realize the glottal source waveform output for various values of those parameters for synthetic speech [15]. This codebook could then be compared to the algorithm output by minimizing the mean squared error to estimate parameters such as open quotient.

And lastly, just for comparison, the Aparat software toolkit was utilized to extract the IAIF glottal source estimation. A comparison of all three methods and their results of the OQ estimation compared to the reference waveform in different pitch and vocal quality variations are shown below in Table 2.2, where it is noted that estimation error ranges from 0 to 1. One limitation involved in this method is the fact that an assumption is made that the LeTalker is producing an accurate glottal source waveform from the area data. Also, it is noted that limitations on accurately detecting glottal dynamic time-domain features are strongly affected by Gaussian white noise at a signal-to-noise ratio of 5dB, vocal quality, and pitch, which can be observed in Table 2.2 where, on average, pressed phonation and phonation for females yielded higher OQ estimation errors [15].

In all of the previous studies and methods performed, even if methods were compared across studies, a very detailed and extensive analysis of the accuracy of specific methods on clinically obtained data across a wide range of fundamental frequencies $F_0$ and ages, was not ever performed. Noting that time-domain features,

such as open quotient or speed quotient relate directly to the kinematic vocal fold movement, this thesis's experimental study focused on extracting the OQ of the estimated glottal source and compared it to the HSVI extracted area and medial-line displacement waveform OQ values.

# Chapter Three: Methods Studied and Compared

## Methods Studied and Compared

The methods compared for this thesis' research are based on an understanding of how speech is accurately modeled. Speech can be treated as a Linear Time-Invariant Source-Filter Model, which is a reasonable approximation over short time intervals of the vocal tract. Therefore, over short time intervals, speech can be modeled as a linear system which is Time-Invariant as the vocal tract and lip radiation shape will be approximately constant. We can approximate speech production into this time-invariant model because speech phonation is utilized, which is not regular speech, which is time-varying. From the model, we can utilize inverse filtering to obtain the glottal source waveform. Three methods were compared in this study and are Iterative Adaptive Inverse Filtering, OQ Estimation using Linear Prediction with Glottal Source Modeling, and a newly proposed method for this thesis, Linear Prediction Error Waveform Analysis with Peak Detection.

## Iterative Adaptive Inverse Filtering

Inverse filtering is a popular technique used to extract the glottal source or glottal volume airflow. It can be demonstrated from equation 1.2 and the LTI Source-Filter model of speech production in Figure 3.1 that the acoustic speech waveform can be modeled by a convolution of the glottal source with the vocal tract filter impulse response and lip radiation, which is treated as a first-order differentiator. The IAIF method utilizes a twelve step process of vocal tract filter and lip radiation effect calculations by Linear Prediction analysis, inverse filtering of the speech waveform, and integration to remove the effects of the lip radiation $R(z)$,

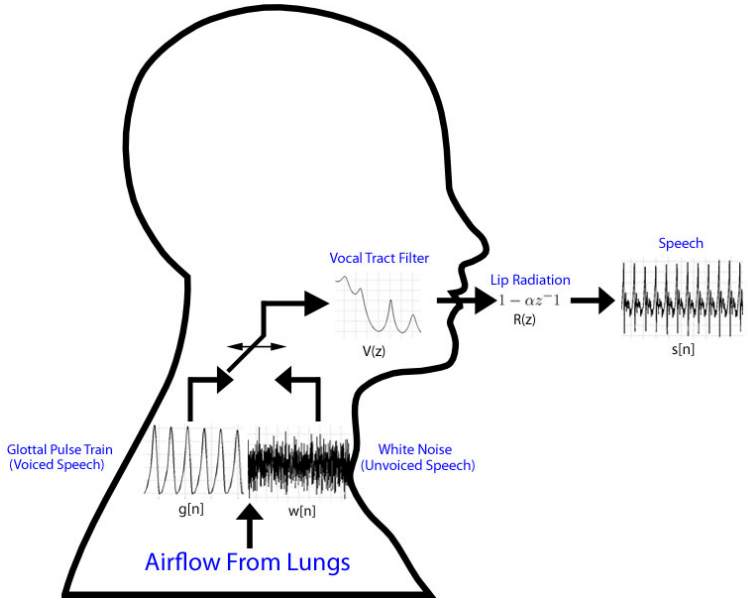


Figure 3.1: A Linear Time-Invariant Source-Filter Model of Speech Production for the IAIF and OQ Estimation using Linear Prediction Methods

which, in this method, is treated as a first-order differentiator of the expelling air, or

$$R(z) = 1 - \alpha z^{-1} \tag{3.6}$$

where we can assume in most cases that $\alpha \approx 1$. The difference from the previously described model's lip radiation component is illustrated in Figure 3.1, whereas the earlier model is shown in Figure 1.1.

The glottal source can then be extracted by inverse filtering and canceling effects of the vocal tract filter and lip radiation shown in equation 3.7. If the vocal tract filter effects are known, this process becomes simplified. However, the accuracy of the results depend on how well the vocal tract filter is estimated and is severely dependent on the quality of the input acoustic waveform [14].

$$G(z) = \frac{S(z)}{V(z)R(z)} \tag{3.7}$$

This method utilizes Linear Prediction analysis to model the vocal tract as a

filter and the lips as a differentiator and iteratively reduces the effects from the vocal tract and the lips by inverse filtering the acoustic waveform at different LPC orders and integrating the results.

A linear *pth* order prediction system is defined by equation 3.8.

$$\hat{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{3.8}$$

where $\hat{s}[n]$ is the predicted speech signal and the error $e[n]$ of the signal is of the form

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \tag{3.9}$$

The coefficients of the *pth* order prediction system are chosen so as to minimize the prediction error $e[n]$. LPC analysis is utilized in the twelve steps of the IAIF method because of its accuracy of modeling the speech spectrum and therfore can be applied iteratively to remove vocal tract filtering effects [2]. The IAIF method is outlined in a block diagram illustrated in Figure 3.2.

In stage one, the recorded acoustic waveform $s[n]$ is high-pass filtered in order to remove low frequency room noise or reverberations that may be recorded by the microphone, which would affect the results of the LPC analysis. The high-pass filtered speech signal is then analyzed by a first-order LPC in stage two, which yields $H_{g1}(z)$, a precursory estimate for the combined glottal flow and lip radiation effects [2]. Using this obtained first-order LPC filter, the high-pass filtered speech signal is inverse filtered in stage three, canceling the effects estimated by $H_{g1}(z)$. The output of this is analyzed by an LPC *pth* order system in stage four and the resulting estimate of the filtering effects, indicated as $H_{vt1}z$, is used to again inverse filter the high-pass filtered speech signal to reduce the vocal tract effects in stage five. The order $p$ for LPC analysis in stage four is usually between the values of 8 and 12. From stage five's output we have an estimate for the speech waveform

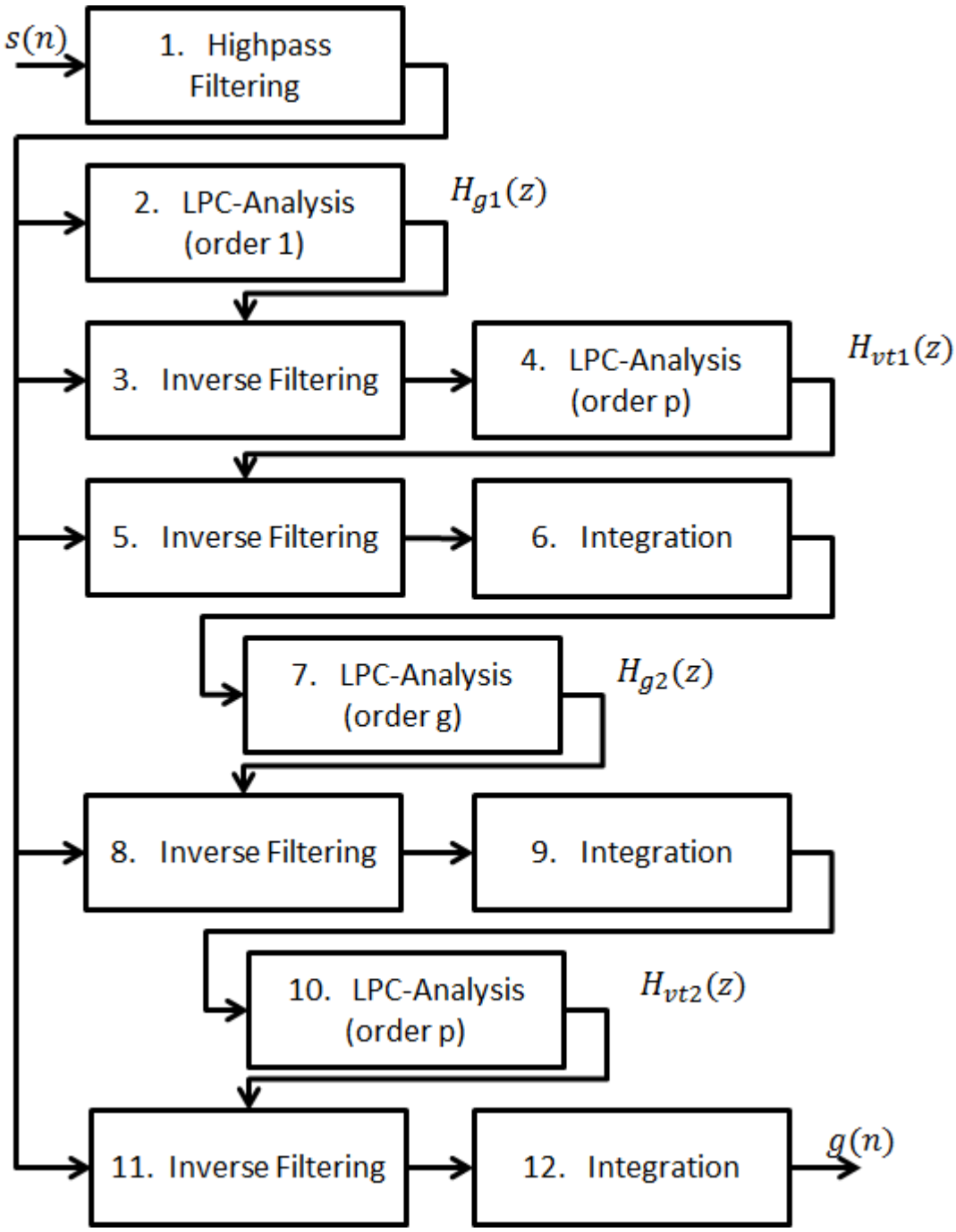


Figure 3.2: A Block Diagram of the IAIF Method [2]

with canceled vocal tract effects. Stage six integrates the output from stage five to achieve the first estimate of the glottal source by canceling the lip radiation effects.

In stage seven, the second iteration begins as a newer estimate for the glottal flow effects is determined as $H_{g2}(z)$ by utilizing an LPC analysis of order $g$, where

LPC order $g$ is usually between the values of 2 and 4. Stage seven's output basically estimates the glottal excitation. The output of stage seven is used in stage eight to inverse filter the high-pass filtered speech waveform and to cancel the effects of the glottal contribution, with the output of stage eight integrated in stage nine to further reduce the lip radiation effects. The final estimate for the vocal tract effects is computed by another $pth$ order LPC analysis in stage ten to yield $H_{vt2}(z)$. This vocal tract filter estimate $H_{vt2}(z)$ is utilized to inversely filter the high-pass filtered acoustic waveform a final time in stage eleven and then integrated to yield the final estimate of the glottal source waveform, $g[n]$, by removing the lip radiation effects, in stage twelve.

## Open Quotient Estimation using Linear Prediction with Glottal Source Modeling

A second method proposed by Nathalie Henrich, titled "Glottal OQ estimation using Linear Prediction", also incorporates inverse filtering, but does not perform it in two iterative passes like the IAIF method utilizes [8]. This OQ estimation method makes one estimation for the vocal tract filter and lip radiation effects to inverse filter the effects out of the speech waveform. However, a strong difference between the IAIF and this method is the reliability of this method's OQ calculation on previously defined glottal flow waveform models. This method assumes abrupt glottal closures and then treats the glottal volume airflow waveform as the impulse response of a two-pole anticausal filter [8]. The details of this method are discussed in the following paragraphs.

According to Henrich, like with the previous method, the glottal flow waveform can be obtained by inverse filtering the vocal tract and lip radiation effects of the speech waveform [8]. There has been many time-domain models of the glottal flow waveform developed to describe the source with all being relatively close

and requiring a few parameters. Parameters utilized in OQ Estimation using Linear Prediction include: $A_v$, the maximum amplitude of the glottal flow, $T_0$, the fundamental period of the glottal flow waveform, $OQ$, the open quotient which helps define the glottal closure instant relative to $T_0$, $TL$, the spectral tilt factor, which is linked to the abruptness of glottal closure, and the asymmetry coefficient $\alpha_m$, relating the glottal opening phase and closing phase [8].

The parameters all have their own effects on the output speech waveform, since the speech waveform is derived from the glottal source flow. The amplitude $A_v$ controls the amplitude of glottal source and therefore, the output speech waveform. The fundamental period $T_0$ controls the pitch of the voiced part of the speech waveform. $OQ$ has a strong relationship with the amount of effort that is imposed during voiced phonation [8]. For example, a *pressed* phonation usually corresponds to a small $OQ$ and a *relaxed* or *breathy* phonation typically corresponds to a large $OQ$ [8]. From this, it can be noted that large $OQ$ values may dictate that a complete glottal closure never occurred. There also exists a relationship between the glottal asymmetry $\alpha_m$ and the glottal open quotient $OQ$. Typically, a small $OQ$ corresponds to a large $\alpha_m$ and therefore, a large $OQ$ corresponds to a small $\alpha_m$. These two parameters dictate most of the shape of the glottal flow and therefore are used in most glottal flow models [8].

Overall, this method is based on a spectral representation of the glottal flow and is essentially modeled as a truncated impulse response of a two-pole anticausal filter [8]. From this, the open quotient can be calculated from a second-order LPC analysis of the estimation of the glottal flow. This model functions properly whenever abrupt glottal closure occurs, which also indicates minimum spectral tilt and an unambiguous open quotient. From previous glottal flow model studies, it was determined that the glottal flow pulse is shaped like a time-shifted second-order low-pass filter that is time-reversed and time-limited [8]. And using the

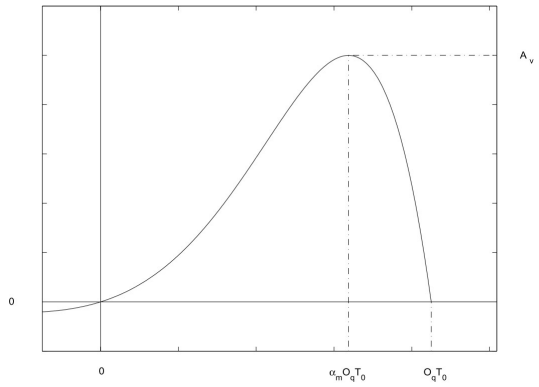


Figure 3.3: General Form of the Linear Glottal Flow Model $G(t)$. Adapted from "Glotal Open Quotient Estimation using Linear Prediction" by Nathalie Heinrich, 1999. [8]

time-domain parameters discussed previously, the filter can be described. The impulse response of a second-order causal filter is

$$h_c(t) = Ae^{-Bt}sin(Ct)u(t) \tag{3.10}$$

where $u(t)$ is the unit step function defined by:

$$u(t) = \begin{cases} 0, \text{ if } t < 0 \\ 1, \text{ if } t > 0 \end{cases} \tag{3.11}$$

and therefore the anticausal equivalent of the filter is:

$$h_a(t) = Ae^{Bt}sin(-Ct)u(-t) \tag{3.12}$$

In order for the filter to open at time 0 and close at time $OQT_0$ then $h_a(t)$ is shifted by a factor $\gamma = OQT_0$ demonstrated by:

$$G(t) = Ae^{B(t-\gamma)}sin(-C(t-\gamma))u(1-\frac{t}{\gamma}) \tag{3.13}$$

and shown in Figure 3.3. The constants $A$, $B$, and $C$ can be computed from the

described glottal flow waveform, where $G(0) = 0$ and $G'(\alpha_m OQT_0) = 0$ at $\alpha_m OQT_0$, which defines the time instant of the maximum amplitude, so we know $G(\alpha_m OQT_0) = A_v$ [8]. The constant equations are then given by:

$$A = \frac{A_v e^{\frac{\pi(\alpha_m - 1)}{tan(\pi\alpha_m)}}}{sin(\pi\alpha_m)} \tag{3.14}$$

$$B = -\frac{\pi}{\gamma tan(\pi\alpha_m)} \tag{3.15}$$

$$C = \frac{\pi}{\gamma} \tag{3.16}$$

From these constant definitions, we can derive the linear glottal flow model for one period temporal length as:

$$G(t) = A_v \frac{sin(\frac{\pi t}{\gamma})}{sin(\pi\alpha_m)} e^{\frac{\pi(\alpha_m - \frac{t}{\gamma})}{tan(\pi\alpha_m)}} u(1 - \frac{t}{\gamma}) \tag{3.17}$$

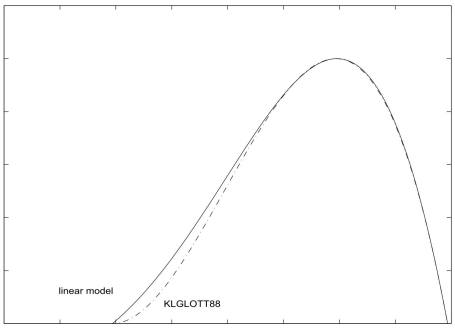


Figure 3.4: Comparison of the KLGLOTT88 Model (dotted lines) with the Linear Model ($\alpha_m = 0.7$). Adapted from "Glotal Open Quotient Estimation using Linear Prediction" by Nathalie Heinrich, 1999. [8]

The model to compare the linear model to is shown in Figure 3.4 and is known as the KLGLOTT88 model. The equation of the transfer function of this model can be shown by:

$$\tilde{G}(z) = \frac{G_1 z^{-N+1}}{1 + b_1 z + b_2 z^2} \tag{3.18}$$

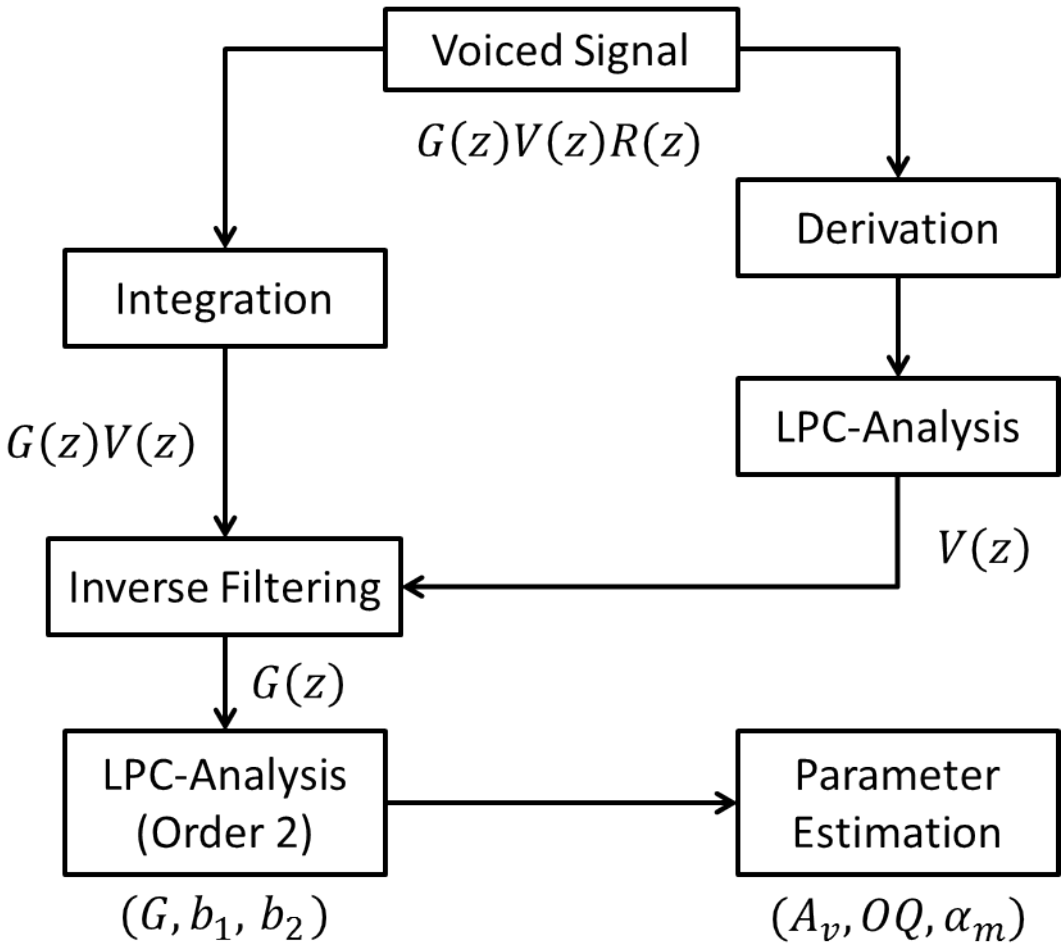


Figure 3.5: Block Diagram of OQ Estimation with Linear Prediction Inverse Filtering [8]

and with sampling period of $T_e$, the constants in the KLGLOTT88 are related to the linear glottal flow model by:

$$G_1 = A_v \frac{sin(\frac{\pi T_e}{\gamma})}{sin(\pi \alpha_m)} e^{\frac{\pi(\alpha_m - 1 + \frac{T_e}{\gamma})}{tan(\pi \alpha_m)}} \tag{3.19}$$

$$b_1 = -2cos(\frac{\pi T_e}{\gamma}) e^{\frac{\pi T_e}{tan(\pi \alpha_m)}} \tag{3.20}$$

$$b_2 = e^{\frac{2\pi T_e}{tan(\pi \alpha_m)\gamma}} \tag{3.21}$$

The block diagram of the algorithm to obtain the glottal flow for OQ Estimation using Linear Prediction is found in Figure 3.5. After the glottal flow is estimated

using the method illustrated in Figure 3.5, a second-order LPC analysis is performed and the coefficients of the filter are obtained. Since the method is time-invariant, the autocorrelation method for the LPC analysis was utilized. The open quotient $OQ$ can then be calculated using the following equation:

$$OQ = \frac{\pi T_e}{T_0} \frac{1}{cos^{-1}(-\frac{b_1}{2\sqrt{b_2}})} \tag{3.22}$$

## Linear Prediction Error Waveform Analysis with Peak Detection

Each of the previous theories and methods have utilized linear prediction to extract the glottal source or flow waveform from the recorded acoustic waveform. Calculating the open quotient of the vocal fold dynamics with the previous methods may lead to large errors if the vocal tract effects and lip radiation effects are not properly estimated. Therefore, a new method has been proposed that utilizes LPC analysis in a simpler fashion based on a glottal source derivative dynamic feature.

It is well known that linear prediction is a tool that predicts the next sample of a particular waveform based on a specific number of previous samples (order p) and their weights (coefficients). Therefore, if a signal is fairly sinusoidal, it can be accurately estimated by linear prediction, i.e. the error will be minimized. However, sudden amplitude changes over short periods of time will not be able to be predicted as easily and will result in a larger error.

During phonation, vocal folds vibrate as air is moved across them and the resulting waveform has a fundamental frequency equal to the vocal fold's. When the vocal folds are in the open phase of their period, air is moved transversely across with slight pressure differentials due to the abduction or adduction of the folds. When the vocal folds are closed, we can consider the air pressure waveform to resonate through the vocal tract cavity represented by a cylinder-type shape that is

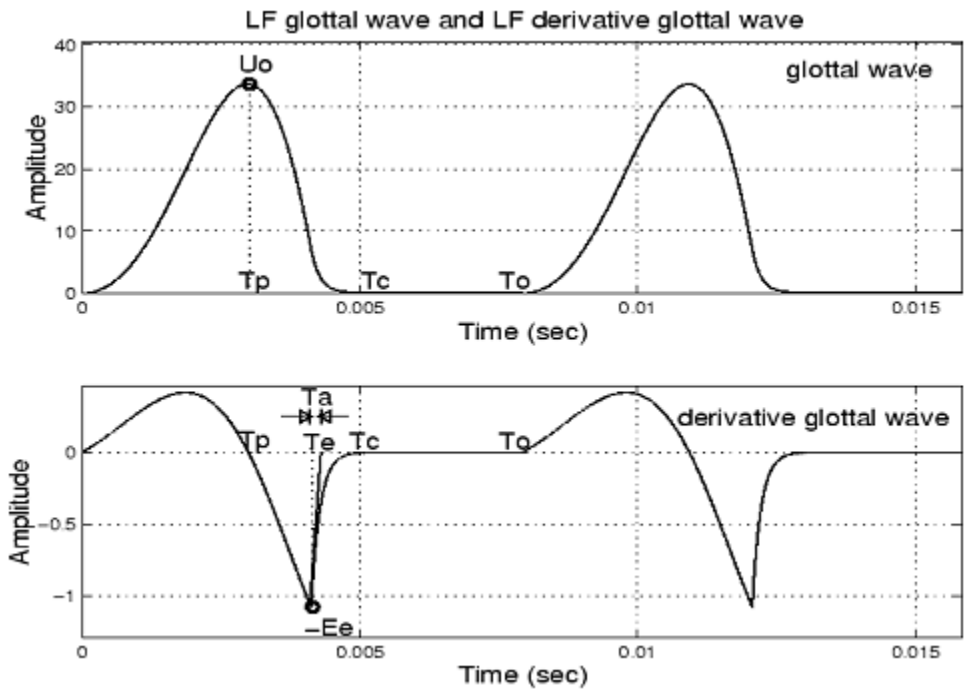


Figure 3.6: Liljencrants-Fant Glottal Model: Glottal Source (top) and Glottal Derivative (bottom). Adapted from "Recent Developments in Musical Sound Synthesis Based on a Physical Model" by Julius O. Smith III, 2003. [7]

open on one end.

In each case of being open or closed, the air pressure waveform will resonate in a more sinusoidal-type fashion, and therefore, will be more accurately predicted by LPC analysis and will have a smaller error waveform. However, the largest prediction error will occur immediately near the vocal fold closure and opening time instants, where the greatest pressure differentials occur. This effect can be seen in Figure 3.6, which represents the glottal source derivative and pressure change that occurs, where the large negative spike in the glottal source derivative represents this pressure change immediately before closure.

So, with the LPC error waveform, the greatest events of nonstationarity that cannot be as accurately predicted by LPC analysis can be matched up with the glottal dynamics to help estimate the time-domain features, such as open quotient. By simplifying the process of feature extraction, we can examine the error waveform

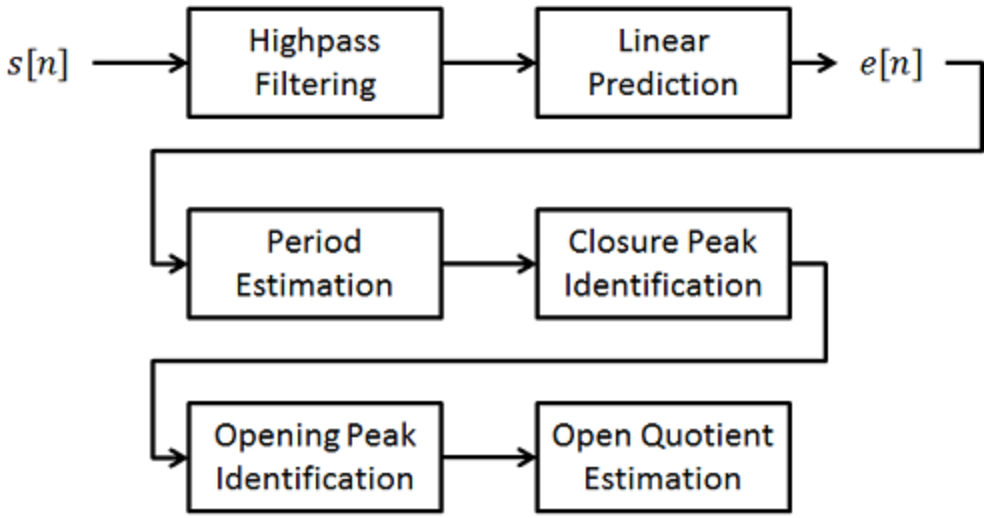


Figure 3.7: Block Diagram of Linear Prediction Error Waveform Analysis with Peak Detection

of the LPC analysis to aid in determining these critical points. And even though the recorded outputted acoustic waveform is filtered with the vocal tract and radiated with the lips, it will still contain the pressure differential features initiated by the vocal fold opening and closure.

The block diagram of this method is shown in Figure 3.7. Assuming the vocal folds are vibrating at a particular fundamental frequency $F_0$, the output acoustic signal $s[n]$ will also have this fundamental frequency. Because of this, the glottal open and closure points or greatest pressure differentials will be periodic and display in the linear prediction error waveform $e[n]$ as strong amplitude swings or peaks. These peak locations of closure can then be detected by knowing the fundamental frequency or period of the acoustic waveform and since the second largest pressure differential will occur at opening, the maximum peaks in between the closure time instant peaks can be detected as well. Knowing all of these closure and opening time instants, the open quotient can be easily calculated using

$$OQ = \frac{t_{c1} - t_{o0}}{t_{c1} - t_{c0}} \tag{3.23}$$

where times $t_{c0}$ and $t_{c1}$ correspond to consecutive glottal closure time instants and $t_{o0}$ corresponds to the glottal open time instant between them.

## Comparison Measures

Once the glottal source waveform is derived, it can be compared to the HSVI data to understand how accurately each acoustic extraction method estimates open quotient. As defined before, the open quotient is the time at which the glottis is open divided by the total period of the glottal cycle. The ideal situation is that glottal opening instant (GOI) for OQ estimation is strongly defined. However, for breathy speech waveforms and some area waveforms, which may not have explicitly defined GOI's, it may be easier to define an OQ threshold value. This was performed in a study by Sapienza in 1997 and is used as a way to compare the HSVI-extracted glottal area waveform and estimated glottal source waveform open quotient values [16]. Sapienza defined thresholds of 20% and 50% of the peak-to-peak value of each period to calculate the 20%OQ and 50%OQ values. As seen in the example in Figure 3.8, sometimes the GOI is not explicitly defined, which is why the threshold values will aid in determining the accuracy of this glottal source extraction method for open quotient estimation to the HSVI area waveform open quotient values.

The second HSVI data that can be used to compare the open quotient estimation for the three studied methods against is the medial-line displacement waveform of the vocal folds. This waveform follows the left and right vocal fold movement on a adaptive medial-line defined by the glottis. This displacement waveform's open quotient will be related to the average open quotient of the glottal cycle.

For the first method, IAIF, the 20% and 50% OQ values of the glottal source were compared to their corresponding 20% and 50% OQ values of the area waveform. These threshold levels and open quotients would be calculated for each

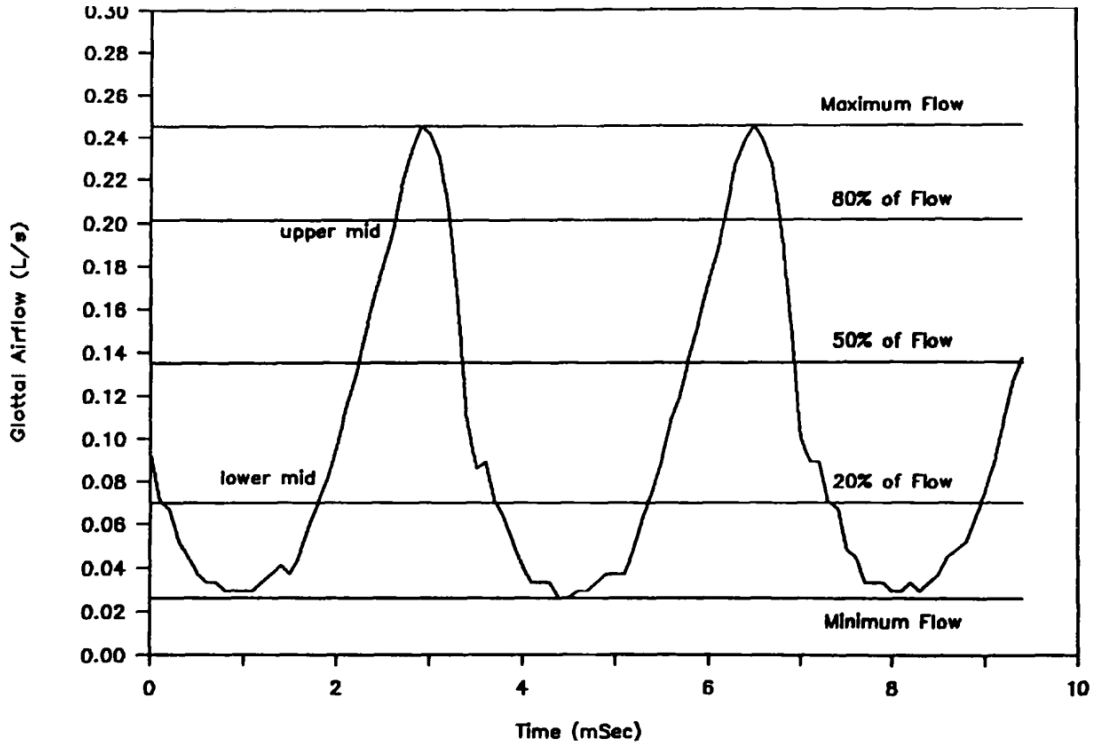


Figure 3.8: OQ 20%, 50%, 80%, and Maximum Flow Threshold Levels for Two Cycles of a Glottal Airflow Waveform. Adapted from "Approximations of Open Quotient and Speed Quotient from Glottal Airflow and EGG Waveforms: Effects of Measurement Criteria and Sound Pressure Level" by Christine M. Sapienza, et al., 1998. [16]

period and averaged for 30 cycles. It is assumed that the open quotient of the glottal source, which is essentially a ratio, would equal the corresponding open quotient of the area waveform. To compare the IAIF method with the displacement waveform, only the 20% glottal source open quotient was utilized due to the fact that the 50% threshold would underestimate the actual open quotient value.

The second method compared, OQ estimation using linear prediction with glottal modeling, utilized an equation that calculated the open quotient value for an estimated glottal source waveform based on the 2nd order LPC coefficients. This 2nd order LPC analysis was also applied to the corresponding area waveform to extract the open quotient value for comparison. And since the displacement waveform had explicitly defined GOI's and GCI's, the open quotient could be calculated easily for comparison with the estimated glottal source. From all of this,

threshold levels weren't needed because single OQ values were calculated for each of the glottal source, area and displacement waveforms.

The third method compared, LPC error waveform analysis with peak detection, had explicitly defined GOI's and GCI's for its error waveform and a mean 30 cycle open quotient value was calculated. This value was compared to a corresponding area 7%OQ threshold value based on a simple error analysis determining it was the best threshold for comparison. The error waveform mean OQ value was also compared to the corresponding 30 cycle mean displacement OQ and the error was calculated.

There are limitations with some of these methods, especially when determining their accuracy with an objective comparison. The first method, IAIF realizes a glottal source estimate, which doesn't always have explicitly defined glottal closure and opening time instants. This limitation leads to the use of the threshold values to help define an open quotient for comparison between the estimated glottal source and area waveform. The open quotient thresholds were also used to compare against the error waveform open quotient from the third method, linear prediction error waveform with peak detection. The issue with these thresholds is an inconsistency or variability of the threshold time instants over multiple periods, which may yield erroneous results. The use of a more consistent measurement with greater repeatability, i.e. the displacement waveform open quotient, is necessary. The displacement waveform has explicitly defined glottal closure and opening time instants and therefore will yield more consistent and reliable open quotient values for an objective comparison of the acoustic feature extraction methods.

# Chapter Four: Experiment

## Recording System

Previous to this study of the comparison between the discussed methods, acoustic and video data was extracted in a clinical setting from volunteered subjects. A total of 46 volunteer participants were recruited for the study after signing an institutional review board approved informed consent / assent forms, at the University of Kentucky, Vocal Physiology and Imaging Laboratory. Participants without voice disorders were included in the study if they met the following criterion: had negative histories of vocal pathology, not being professional voice users, and perceptually judged to have normal voice by a certified speech language pathologist specializing in voice disorders. Participants going through puberty as identified via case history were excluded. Selection criteria for adult controls were similar to those of pediatric group, except that the adult controls had a negative history of smoking. The volunteer participants were instructed to phonate the vowel sound /i/ at normal pitch and loudness. The High-Speed Video Imaging of the vocal folds was recorded for 4 seconds at a sampling rate of 4000 fps at 512 x 256 pixel resolution using a KayPENTAX Color High-Speed Video System, Model 9710. The black and white camera head was used instead of the color to increase the sensitivity of the recorded video. The endoscope was placed in the subject's mouth, while the the tongue was held by clinician in order to prevent it from obstructing the HSVI recording. The audio was recorded simultaneously at a sampling rate of 100 (or 80)kHz using a clip-on lapel microphone placed near the subject's collar on his or her shirt. The sustained oscillation phases were identified by viewing the audio envelope as well as by using the custom play back software developed by KayPENTAX.

The visually and auditory perceptually best quality data was utilized for this

Figure 4.1: Cropped Video Frame from HSVI, Determined Region of Interest, and Detected Edge Contour. Adapted from "Analysis of high-speed digital phonoscopy pediatric images" by Harikrishnan Unnikrishnan et al., 2012. [17]

thesis study of the accuracy of the three methods. For this thesis' research, 46 subjects, male and female, child and adult, were compared. The ages of the subjects ranged from 5 to 48 in order to allow for a proper comparison across multiple frequency ranges, where the adult males would be the lowest followed by the adult females and then the children.

In order to extract the area and displacement data from the video, a robust edge detection algorithm was applied to video frames where a thresholding level and region of interest could be set [17]. The area waveform was then defined as the number of the pixels contained within the contour edge of the vocal folds over time. An example of a video frame, region of interest and detected edge contour of the algorithm is shown in Figure 4.1.

The medial-line for the displacement waveform was determined when the vocal folds were maximally abducted for each period of the glottis [17]. An example is shown in Figure 4.2, where $\vec{O}(x, y)$ is the medial line reference point and $\vec{R}(x, y)$ is the right-fold point. After denoising and interpolation, the left and right vocal fold displacements can be used to calculate the total displacement waveform (in pixels), which was utilized in this study to calculate open quotient. Further

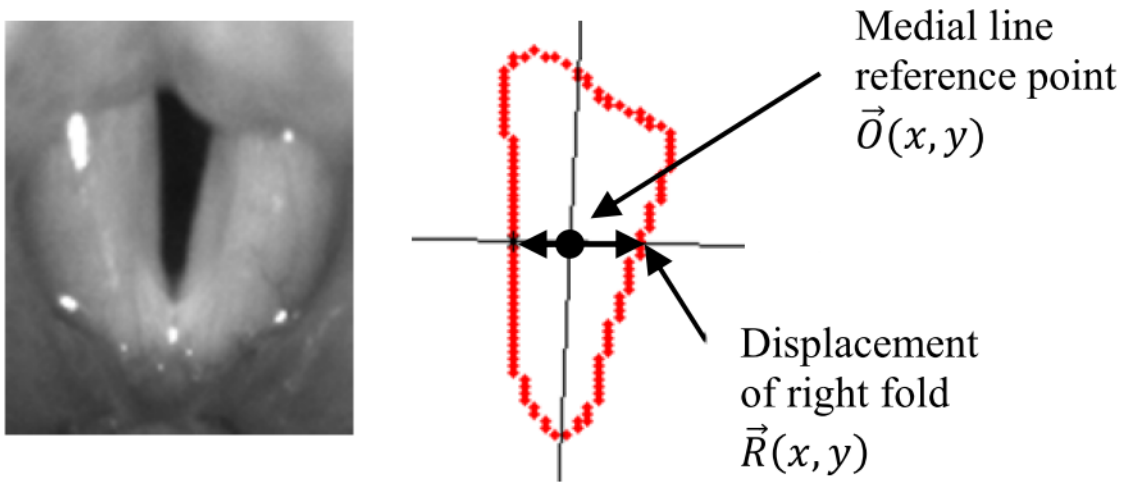


Figure 4.2: Medial-Line Definition for a Cropped Video Frame. Adapted from "Analysis of high-speed digital phonoscopy pediatric images" by Harikrishnan Unnikrishnan et al., 2012. [17]

description of the edge detection algorithm and displacement calculation can be reviewed in "Analysis of high-speed digital phonoscopy pediatric images" [17].

## Analysis

After recording, the video and acoustic waveforms needed to be synchronized properly in order to compare the correct time segments for OQ estimation. When the recordings were made, a Camera Information Header data file was automatically generated that recorded how many video samples the audio was out of sync with synchronization pulse, where the synchronization pulse occurred exactly on a video frame number. It also recorded the frame rates of the video and acoustic recordings and how many video samples long the recording was. Since the video was recorded at a much lower sampling rate than the acoustic waveform, it could be lined up

properly by knowing the video start frame and utilizing the following equations:

$$offset = \frac{S_{frame} f_s}{vf_s} \tag{4.24}$$

$$v_{sync} = \frac{v_{ax} f_s}{vf_s} + offset - idx(0) \tag{4.25}$$

where $S_{frame}$ is the start frame of the video, $f_s$ is the acoustic sampling rate, $vf_s$ is the video frame rate, $v_{ax}$ is the original video axis values and $idx$ is the acoustic waveform indice values. The axis to properly plot the video synchronized in time with the acoustic waveform is $v_{sync}$. Then $v_{sync}(0)$ and $v_{sync}(end)$ can be used as the start and stop cropping indices for the acoustic waveform which yields a new cropped acoustic waveform that is the same temporal-length as the video signal.

However, in order to properly compare them, they needed to be resampled to the same sampling frequency so that the waveforms could be lined up sample-to-sample. The video and acoustic waveforms were resampled to 8kHz using Matlab's *resample* command, which performs the resampling backwards and forwards across the signal in order to preserve the temporal elements of the waveform and to not time-shift the waveform. The waveforms were then high-pass filtered with a cutoff frequency of 100Hz in order to remove any low frequency recording noise or room reverberations prior to applying any glottal source estimation algorithms.

The algorithm glottal source estimations of the 46 subjects were obtained. To line up the area waveform with the glottal source waveform for proper comparison, a few adjustments were made. Matlab's *xcorr* command was utilized to crosscorrelate the acoustic signal LPC error waveform of order 10 and area waveform to synchronize the glottal source estimation and area waveform in period. This lines up the two waveforms in period because the sound pressure has maximum differential just before glottal closure. This crosscorrelation method will line up the

quasi-periodic peaks from the error waveform of the acoustic signal, which occur at glottal closure since they can't be predicted as well, and the area signal max peaks, which correspond to max glottal opening. The maximum lag value computed from crosscorrelating is the delay between the two crosscorrelated signals, which was then used to shift the area signal by the necessary amount to line it up with the glottal source to within a period. This will line up the signals properly to within a period.

However, one more adjustment needed to be made to line up the glottal source estimation and area waveforms in phase and to scale their amplitudes properly. Prior to this step, the means for each area waveform and glottal source estimation were subtracted from its corresponding waveform to zero the mean. The Ordinary Least Squares (OLS) method was then applied to two waveforms to adjust the glottal source by shifting and scaling to fit the area waveform utilizing the equation:

$$Y = \beta X - \alpha \tag{4.26}$$

where $Y$ is the area waveform and $X$ is the glottal source estimation. For the OLS method, the square of the residual must be minimized by choosing a proper $\beta$ and $\alpha$ given by:

$$e^2 = [(\beta X - \alpha) - Y]^2 \tag{4.27}$$

where $e$ is the difference between $Y$ and the $\beta$-scaled $\alpha$-shifted $X$. For each subject's area and estimated glottal source, $\beta$ and $\alpha$ can be easily calculated using:

$$\beta = \frac{cov(X, Y)}{var(X)} \tag{4.28}$$

$$\alpha = \bar{Y} - \beta \bar{X} \tag{4.29}$$

where $cov(X, Y)$ is the covariance between waveforms $X$ and $Y$, $var(X)$ is the variance of $X$, $\bar{Y}$ and $\bar{X}$ are the means of $Y$ and $X$, respectively. After the OLS

method lined up the waveforms in phase, the Pearson Correlation Coefficient $\rho$ was computed between the two waveforms and any slight manual adjustments were made, within a period, for $\alpha$ to maximize this coefficient.

### Iterative Adaptive Inverse Filtering

The Iterative Adaptive Inverse Filtering method, shown in the block diagram in Figure 3.2, was coded in Matlab and applied to all the acoustic recordings. However, prior to applying this algorithm, the proper LPC order for stages four ($8 < p < 12$), seven ($2 < g < 4$), and ten ($8 < p < 12$) had to be determined. In order to do so, an IAIF including only stages one through four was applied to each subject. Stage one's LPC order is always constant at one, but for each subject, the LPC analysis at stage four was applied for orders $8 < p < 10$. For each order $p$ the energy $E_e$ of the LPC error $e[n]$ was calculated using:

$$E_e = \sum_{n=1}^{N} |e[n]|^2 \tag{4.30}$$

where $N$ is the length of the error waveform. Using an Akaike Criteria, which states that the order should be chosen to minimize the energy of the residual, the sum of the energies was computed and the order $p$ corresponding to the minimum total energy was selected, which resulted in a selection of LPC order $p = 10$. Since $p$ was also used in stage ten, only one LPC order, order $g$ in stage seven, was left to be determined. This time, an IAIF including only stages one through seven was applied to each subject. Stage one's LPC order is always constant at one and stage four's LPC order was constant at $p = 10$, but for each subject, the LPC analysis at stage seven was applied for orders $2 < g < 4$. The energy was computed for each order $g$ for each subject and the LPC order corresponding to the minimum total energy was determined, which resulted in a selection of LPC order $g = 2$. The

Table 4.1: Iterative Adaptive Inverse Filtering Algorithm Parameters For Thesis Experiment

| | |
|---|---|
| Resample $f_s$ | 8kHz |
| Highpass Cutoff $f_c$ | 100Hz |
| Stage 1 LPC Order | 1 |
| Stage 4 LPC Order $p$ | 10 |
| Stage 7 LPC Order $g$ | 2 |
| Stage 10 LPC Order $p$ | 10 |

resulting parameters for the IAIF method are demonstrated in Table 4.1.

With these parameters, the IAIF method was applied to each of the 46 subjects and the resulting glottal source estimation was obtained. After alignment, each waveform was then cropped to thirty glottal cycles for comparison due to the fact that the displacement waveform OQ data was computed for thirty glottal cycles. To separate the glottal cycle periods, a zero crossing Matlab algorithm was used to find the length of the period of each glottal cycle. Then, the 20% and 50% peak-to-peak amplitude was calculated for each period of the thirty cycles. Using the zero crossing algorithm for each period, with the level set to 20% and 50% peak-to-peak, the time-instants for each corresponding level were determined and from these known time-instants, the OQ can be calculated for the estimated glottal source and area waveforms. After the OQ was calculated for each of the thirty cycles, the mean of the entire segment was calculated and this mean value was used to compare the glottal source estimation and area waveform. A simple percent error was calculated, using the area waveform as the accepted value, to access the accuracy of the IAIF glottal source estimation on OQ calculation. A second percent error was also calculated using the displacement waveform OQ mean as the accepted value.

### OQ Estimation using Linear Prediction with Glottal Source Modeling

For this method, the waveforms that have already been filtered and synchronized were utilized. The waveforms were then cropped to the same thirty cycles that were determined for the IAIF method. To estimate the glottal source, the inverse filtering algorithm illustrated in the block diagram in Figure 3.5.

In order to calculate the OQ using Equation 4.31, the following parameters need to be determined, $T_e$, the sampling period of the waveform, $T_0$, the fundamental period of the waveform, and $b_1$ and $b_2$, the 2nd order LPC coefficients.

$$OQ = \frac{\pi T_e}{T_0} \frac{1}{cos^{-1}(-\frac{b_1}{2\sqrt{b_2}})} \tag{4.31}$$

The sampling period $T_e$ is already known to be 8kHz from resampling. The fundamental period $T_0$ can be determined using Matlab's *xcorr*, which can compute the autocorrelation of a waveform. The strongest peak compared to the zero lag peak corresponds to the fundamental period $T_0$. All of these values and the LPC 2nd order coefficients were stored for each subject and used to calculate the open quotient.

To compare the area properly, the last step of the algorithm was applied to find the two 2nd order LPC coefficients for the area waveform. Then, along with the knowing the sampling frequency and the fundamental frequency of the area waveform, which was also calculated using Matlab's *xcorr* command, the OQ could calculated. The algorithm was compared also with itself before and after lowpass filtering at 1700Hz. This lowpass filtering was used to filter out any high frequency components that were unnecessary and may have had an effect on the 2nd order LPC analysis. A simple percent error was used to determine the accuracy of this method with the area waveform OQ values and the already computed displacement waveform OQ values.

### Linear Prediction Error Waveform Analysis With Peak Detection

Two different types of LPC analysis were applied to the cropped acoustic waveforms to determine the error waveforms. One of them involved applying stages one through four of the IAIF algorithm, since the fourth stage yields the first vocal tract filter estimate, and the other involved just an LPC analysis of order 10. Since the time-instants of the strong spikes of each of these error waveforms were needed to calculate the glottal open quotient, a peak find algorithm needed to be utilized. The algorithm only needed to search for strong peaks every period, therefore the period was determined from autocorrelating the error waveform and extracting the fundamental period $T_0$. The needed time-instants could then be determined.

First, the glottal closure instants were determined from searching for the strong negative spikes, corresponding to the largest pressure differential occurring immediately prior to closure. The next positive spike after the strong negative spike was determined to be the glottal closure instant. To find the negative spikes, the error waveform was multiplied by $-1$ to flip the strong negative spikes to positive, so that Matlab's *findpeaks* command could be utilized. The time instants for the strong spikes were collected and applied to the original non-negatively scaled error waveform. Each period of the thirty glottal cycles was defined from glottal closure instant to next glottal closure instant. The algorithm windowed the error waveform, period-by-period and searched for the largest spike within each period, which was determined to be the glottal opening instant spike, and those time-instants were also recorded. Knowing the GCI and GOI time-instants allowed for an easy open quotient calculation for each period. The mean OQ for the thirty cycles was used to compare against the 7%OQ of the area and the displacement waveform's OQ means. A simple percent error was calculated to determine the accuracy of the newly proposed LPC error waveform analysis method.

# Chapter Five: Results and Discussion

## Iterative Adaptive Inverse Filtering

To fully determine how accurate the IAIF method was in realizing the Glottal Source waveform time-domain features, a simple percent error calculation was performed with the corresponding area waveform for thirty glottal cycles for the 20% and 50% open quotient threshold levels. From this calculation, the following subject, a 42 year-old female, had a small error of 8.17% when comparing the area 20%OQ and estimated glottal source 20%OQ. An example of the female's acoustic, glottal area and IAIF-estimated glottal source waveforms is shown in Figures 5.1 and 5.2. This subject's acoustic recording was determined to have a fundamental frequency of 250Hz. As seen in the glottal source waveform, the second periodic peak, not all the formant's of the vocal tract were filtered out properly with the IAIF method. However, the overall shape of the glottal source was still resolved. The power spectral density (PSD) of the acoustic waveform denoted strong frequency spikes and the acoustic waveform itself was perceived to be clean and not noisy.

An 11 year-old male child had a large 142% error when comparing the area 20%OQ and estimated glottal source 20%OQ. An example of the male child's acoustic, glottal area and IAIF-estimated glottal source waveforms is shown in Figures 5.3 and 5.4. This subject's acoustic recording was determined to have a fundamental frequency of 285Hz. As seen in this glottal source waveform, and other subjects with large errors, the large negative pressure differential has not been filtered out enough, affecting the 20% glottal open quotient threshold level. Because of this, the time-instant for the 20% threshold level or the 20%OQ value itself does not equal that of the corresponding area waveform and this effect can be seen in the Figure 5.4. After listening, the acoustic waveform was perceived to be slightly muffled.

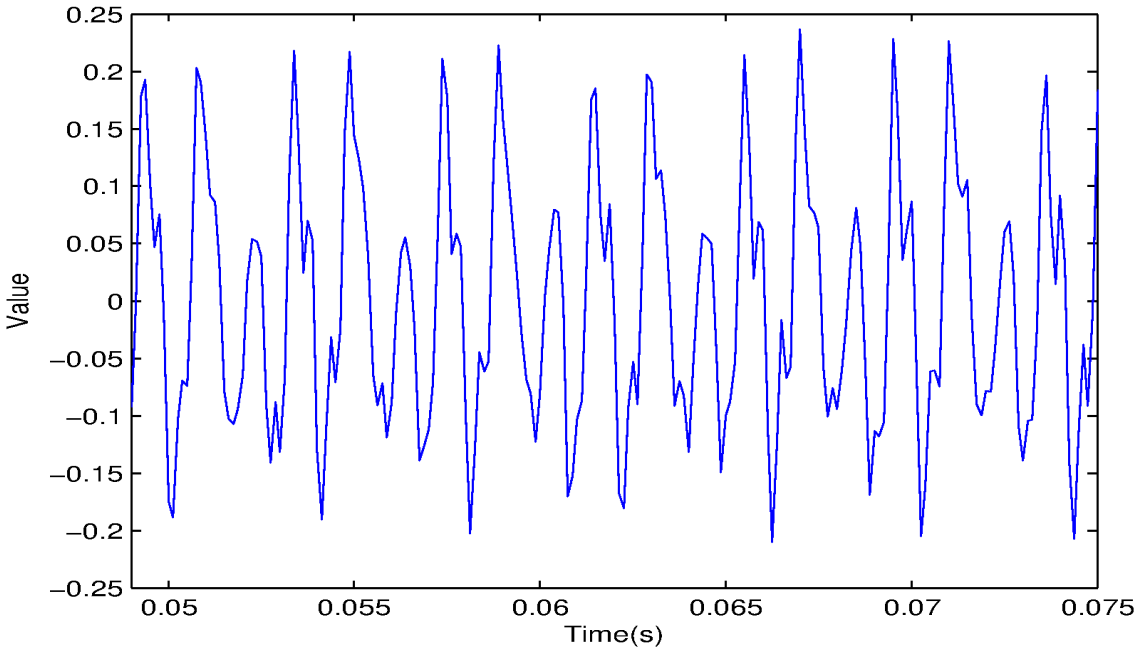


Figure 5.1: IAIF Method Result: Acoustic Waveform for a 42 year-old Female With 20%OQ Error of 8.17% Between Area and IAIF-Estimated Glottal Source.

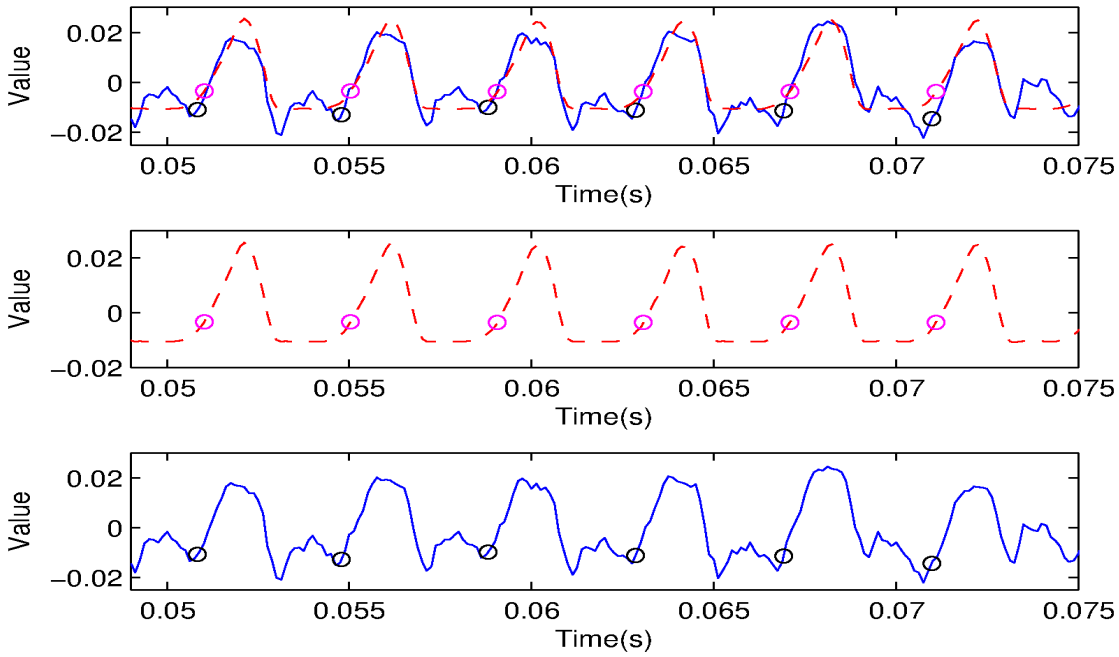


Figure 5.2: IAIF Method Result: Comparison Of 20%OQ Threshold for the Area and Estimated Glottal Source for a 42 year-old Female With Error of 8.17%. Glottal Source Waveform (solid blue line) with 20% Threshold for Each Glottal Source Period (black circles). Area Waveform (dashed red line) and 20% Threshold for Each Area Period (magenta circles).

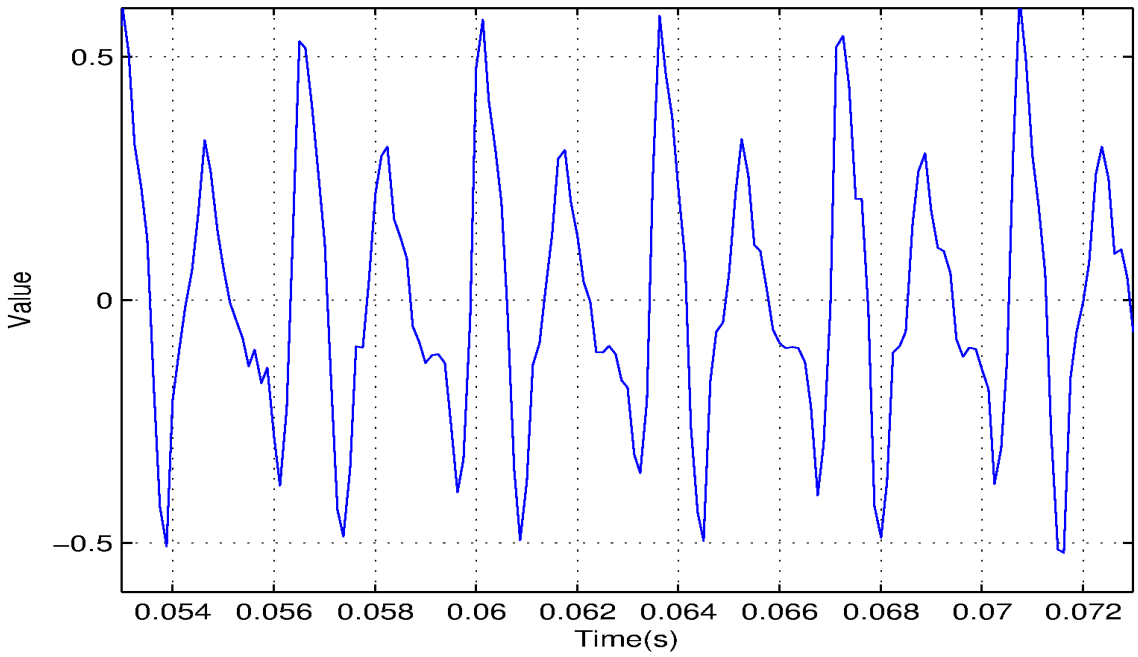


Figure 5.3: IAIF Method Result: Acoustic Waveform for an 11 year-old Male Child With Corresponding 20%OQ Error of 142% Between Area and IAIF-Estimated Glottal Source.

A 27 year-old female had a small -2.36% error when comparing the area 50%OQ and estimated glottal source 50%OQ. An example of the female's acoustic, glottal area and IAIF-estimated glottal source waveforms is shown in Figures 5.5 and 5.6. This subject's acoustic recording was determined to have a fundamental frequency of 296Hz. And even though a second peak can be seen in each period of the glottal source, the overall timing of the threshold is still similar to the area waveform, which leads to a close open quotient value. After listening, the acoustic waveform was perceived to be very clean and the visual appearance of the waveform is not noisy and quasi-periodic.

A 9 year-old male child had a 141% error when comparing the area 50%OQ and estimated glottal source 50%OQ. An example of the male child's acoustic, glottal area and IAIF-estimated glottal source waveforms is shown in Figures 5.7 and 5.8. This subject's acoustic recording was determined to have a fundamental frequency of 320Hz. The large error in this case still seems to be resulting from the

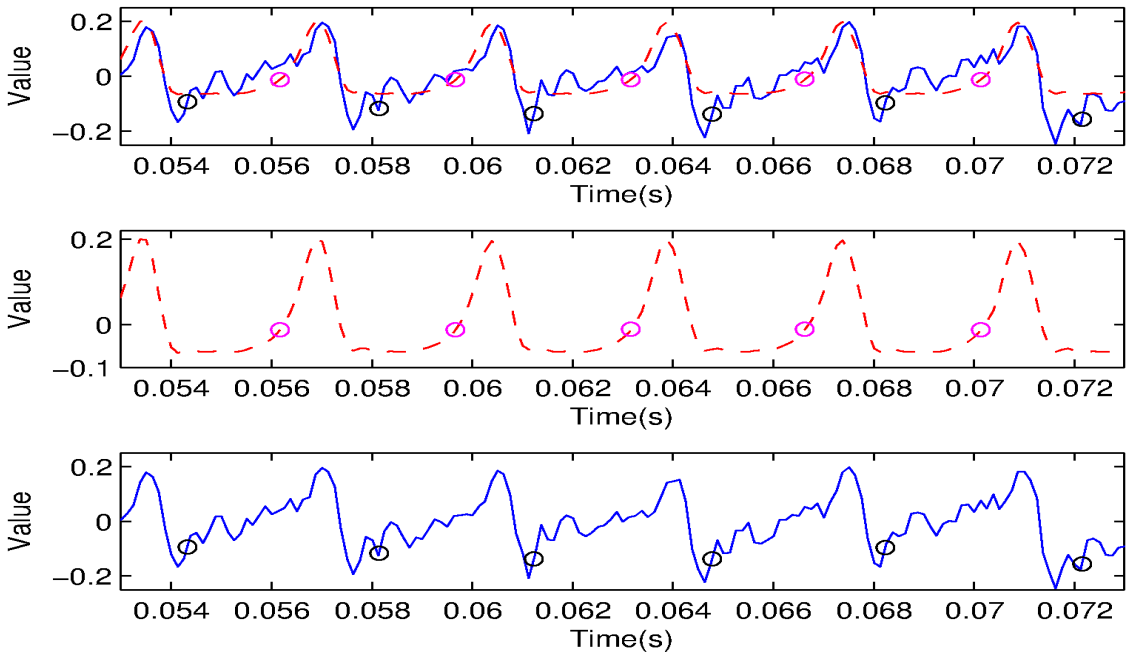


Figure 5.4: IAIF Method Result: Comparison Of 20%OQ Threshold for the Area and Estimated Glottal Source for an 11 year-old Male Child With Error of 142%. Glottal Source Waveform (solid blue line) with 20% Threshold for Each Glottal Source Period (black circles). Area Waveform (dashed red line) and 20% Threshold for Each Area Period (magenta circles).

large negative pressure differential (negative spike) not being filtered out by the inverse filtering method. This negative spike results from the linear prediction filter not accurately predicting the large pressure differential right before an abrupt glottal closure leads to a large negative spike in the linear prediction error waveform, in which the glottal source is derived. The negative spike dominates this particular glottal source estimate and directly impacts the 50% threshold time-instant. After listening, the acoustic waveform was perceived to be somewhat clean and not noisy.

A 21 year-old male had a -0.16% error when comparing the displacement OQ and estimated glottal source 20%OQ. An example of the male's acoustic, glottal displacement and IAIF-estimated glottal source waveforms is shown in Figures 5.9 and 5.10. This subject's acoustic recording was determined to have a fundamental frequency of 151Hz. The resulting estimated glottal source waveform has an appropriate shape, noting the longer opening phase and shorter closing phase

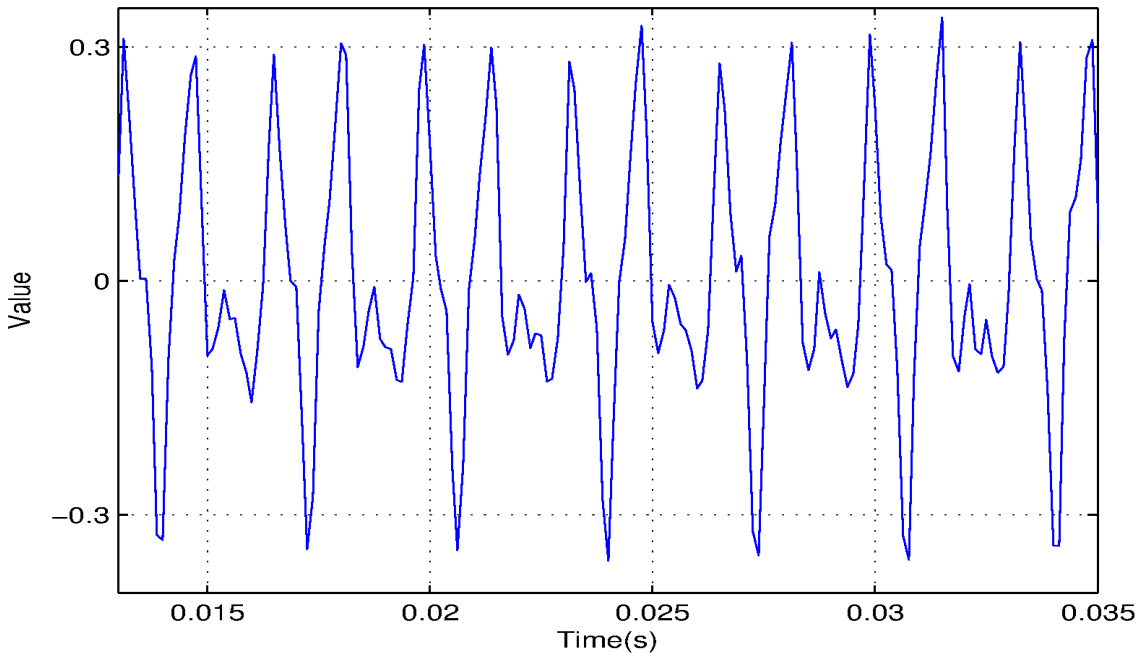


Figure 5.5: IAIF Method Result: Acoustic Waveform for a 27 year-old Female With Corresponding 50%OQ Error of -2.36% Between Area and IAIF-Estimated Glottal Source.

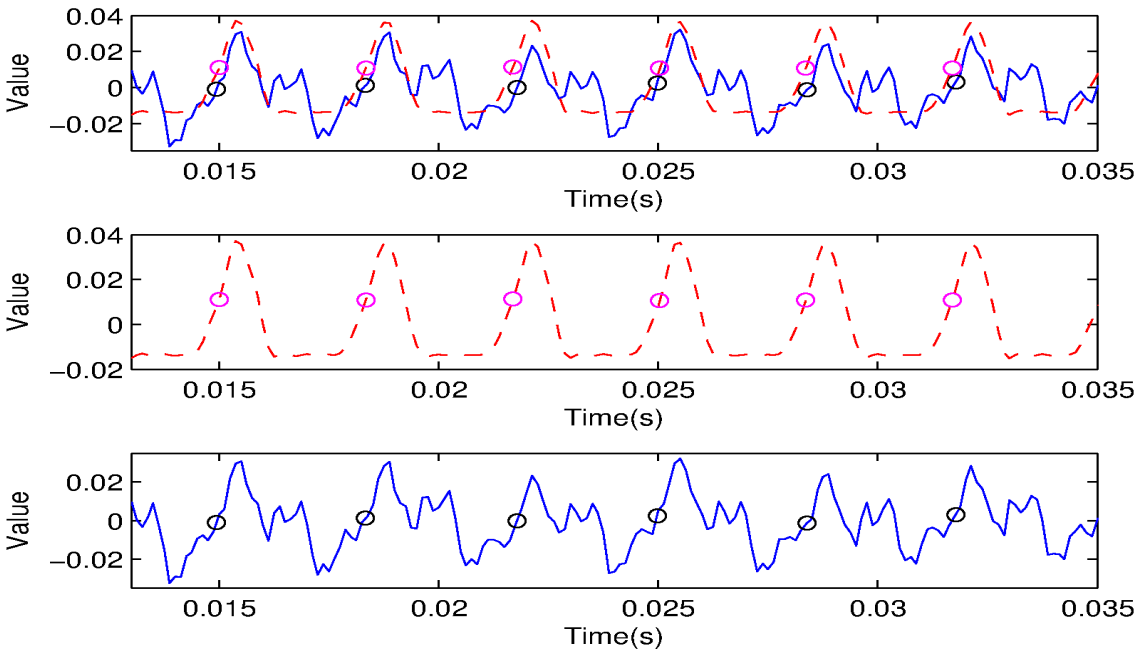


Figure 5.6: IAIF Method Result: Comparison Of 50%OQ Threshold for the Area and Estimated Glottal Source for a 27 year-old Female With Error of -2.36%. Glottal Source Waveform (solid blue line) with 50% Threshold for Each Glottal Source Period (black circles). Area Waveform (dashed red line) and 50% Threshold for Each Area Period (magenta circles).

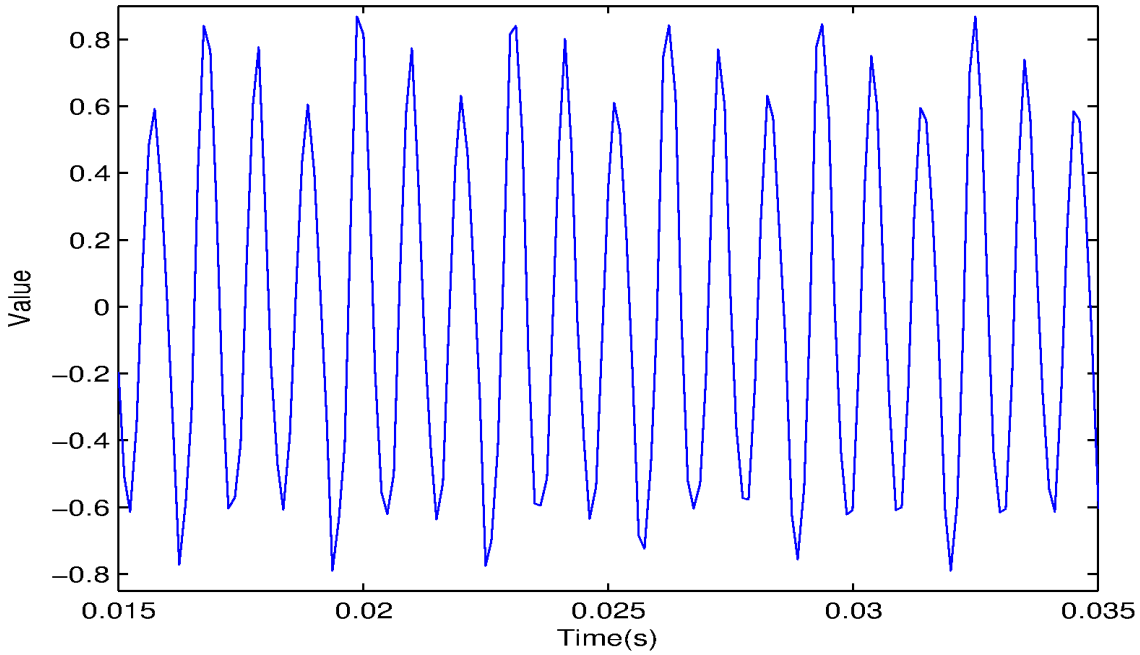


Figure 5.7: IAIF Method Result: Acoustic Waveform for a 9 year-old Male Child With Corresponding 50%OQ Error of 141% Between Area and IAIF-Estimated Glottal Source.

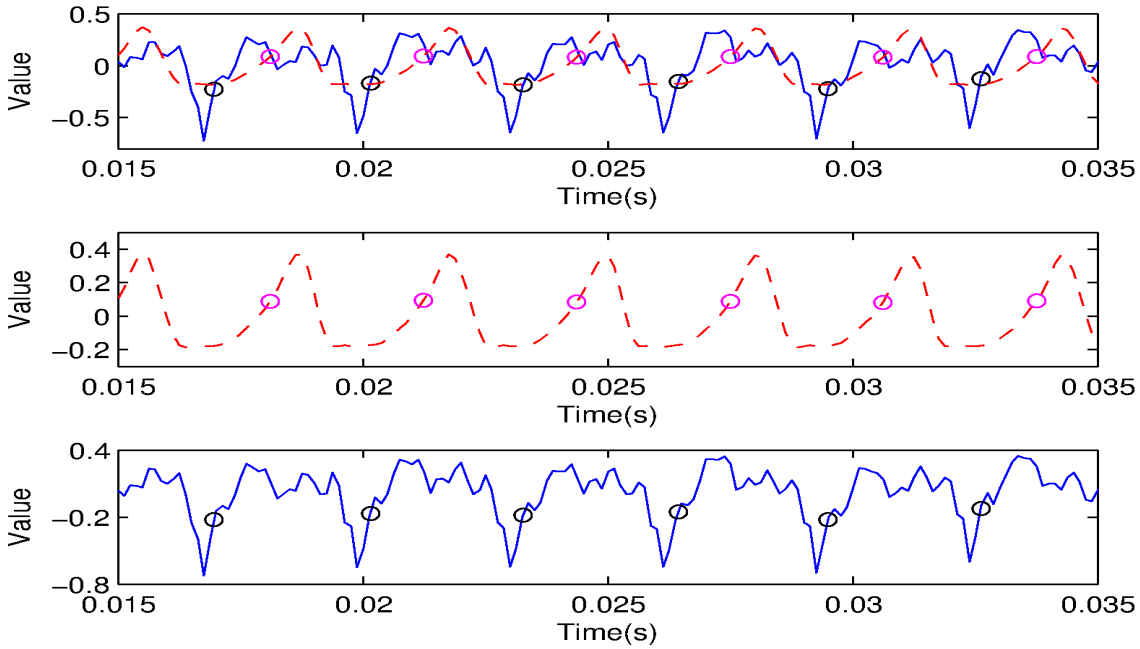


Figure 5.8: IAIF Method Result: Comparison Of 50%OQ Threshold for the Area and Estimated Glottal Source for a 9 year-old Male Child With Error of 141%. Glottal Source Waveform (solid blue line) with 50% Threshold for Each Glottal Source Period (black circles). Area Waveform (dashed red line) and 50% Threshold for Each Area Period (magenta circles).

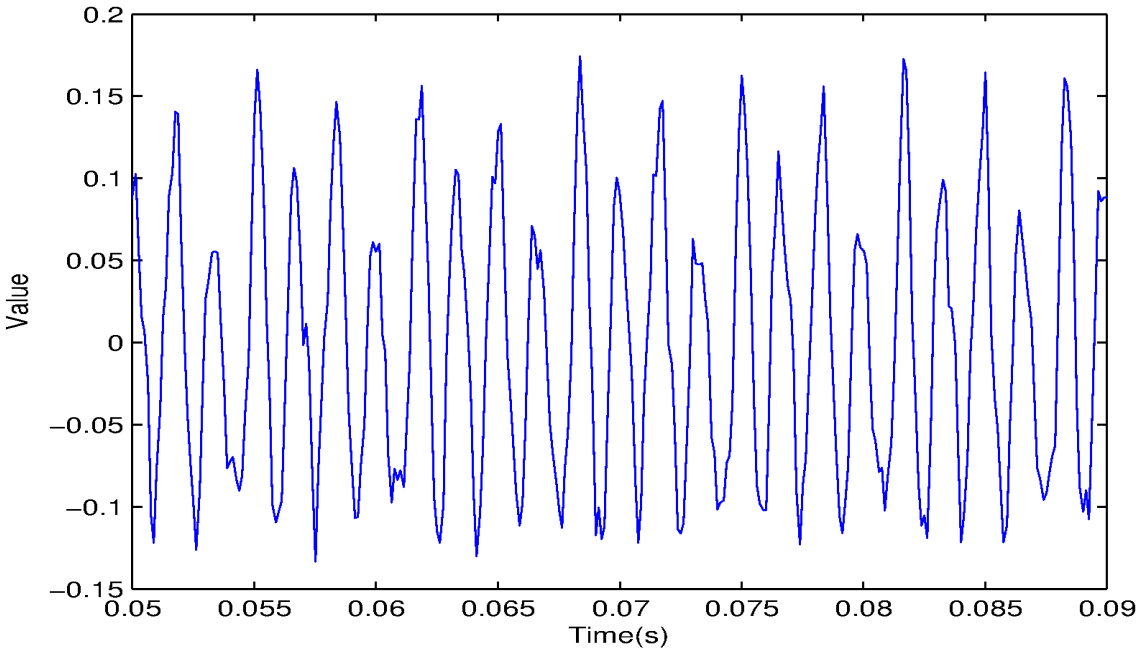


Figure 5.9: IAIF Method Result: Acoustic Waveform for a 21 year-old Male With Corresponding 20%OQ Error of -0.16% Between Displacement and IAIF-Estimated Glottal Source.

causing the form to skew to the right. The estimated glottal source may not have the exact glottal closure and opening time-instants of the corresponding displacement waveform, however, the open quotient estimation of the glottal source averages to be almost exactly the same displacement waveform's. After listening, the acoustic waveform was perceived to be slightly noisy and the resulting PSD shows some white noise across the spectrum. However, since the fundamental frequency dominated the spectrum, the noise did not strongly affect the results.

An 11 year-old male child had a 64.2% error when comparing the displacement OQ and estimated glottal source 20%OQ. An example of the male child's acoustic, glottal displacement and IAIF-estimated glottal source waveforms is shown in Figures 5.11 and 5.12. This subject's acoustic recording was determined to have a fundamental frequency of 286Hz. The resulting estimated glottal source waveform has an appropriate shape, noting the right skewness of each glottal pulse, however, the strong negative spike that was not filtered out directly affects the 20%

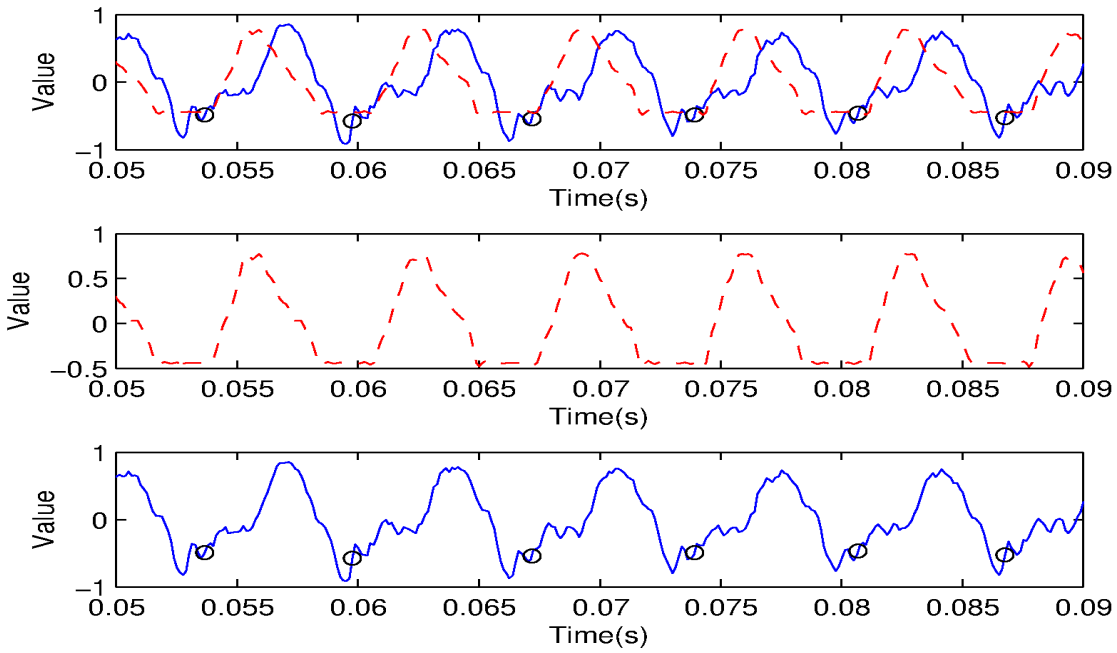


Figure 5.10: IAIF Method Result: Comparison Of 20%OQ Threshold Estimated Glottal Source and Glottal Displacement for a 21 year-old Male With Error of -0.16%. Glottal Source Waveform (solid blue line) with 20% Threshold for Each Glottal Source Period (black circles) and Displacement Waveform (dashed red line).

threshold level, and therefore the 20% time-instant and OQ. This leads to an overestimated glottal open quotient by the IAIF-estimated glottal source when compared to the corresponding displacement waveform. After listening, the acoustic waveform was perceived to be muffled, which may have been affected by the microphone placement or clothing around the microphone during the recording.

The IAIF algorithm was applied to 46 different subjects and the open quotient was estimated and compared to its corresponding area open quotient for the 20% and 50% threshold levels. Due to the limited supply of displacement waveform data, only 43 subjects were compared with their corresponding displacement waveform open quotient for the 20% threshold. The 50% threshold level of the estimated glottal source was not compared to its corresponding displacement waveform due to the fact that it should not be very close to the displacement waveforms open quotient value. It is assumed that the 50% threshold

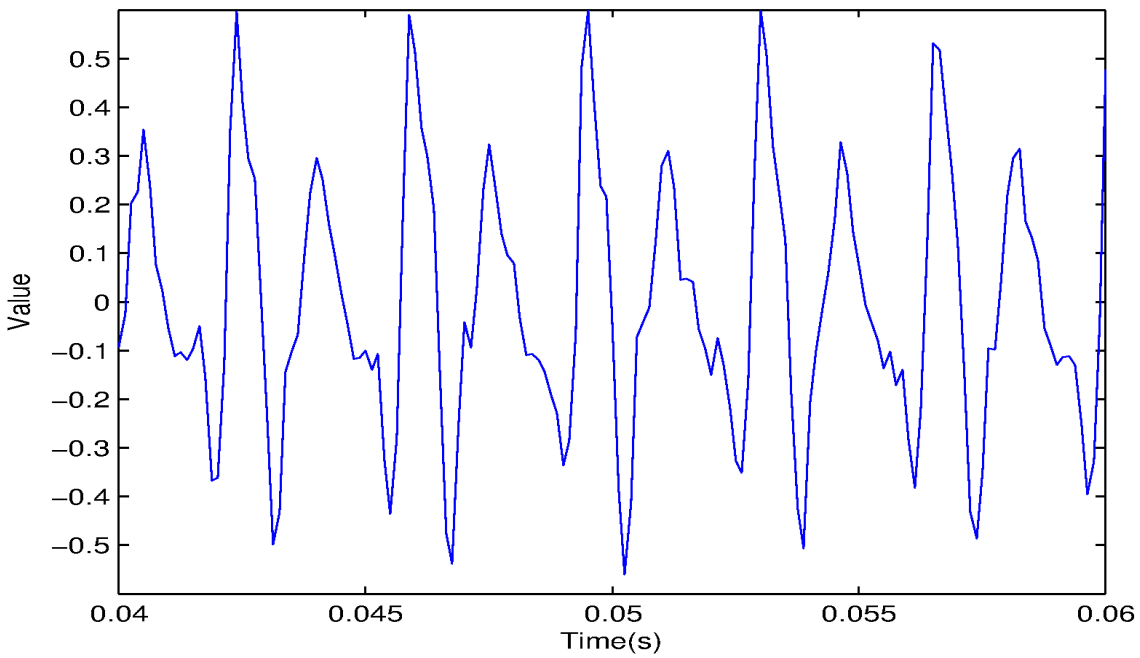


Figure 5.11: IAIF Method Result: Acoustic Waveform for an 11 year-old Male Child With Corresponding 20%OQ Error of 64.2% Between Displacement and IAIF-Estimated Glottal Source.

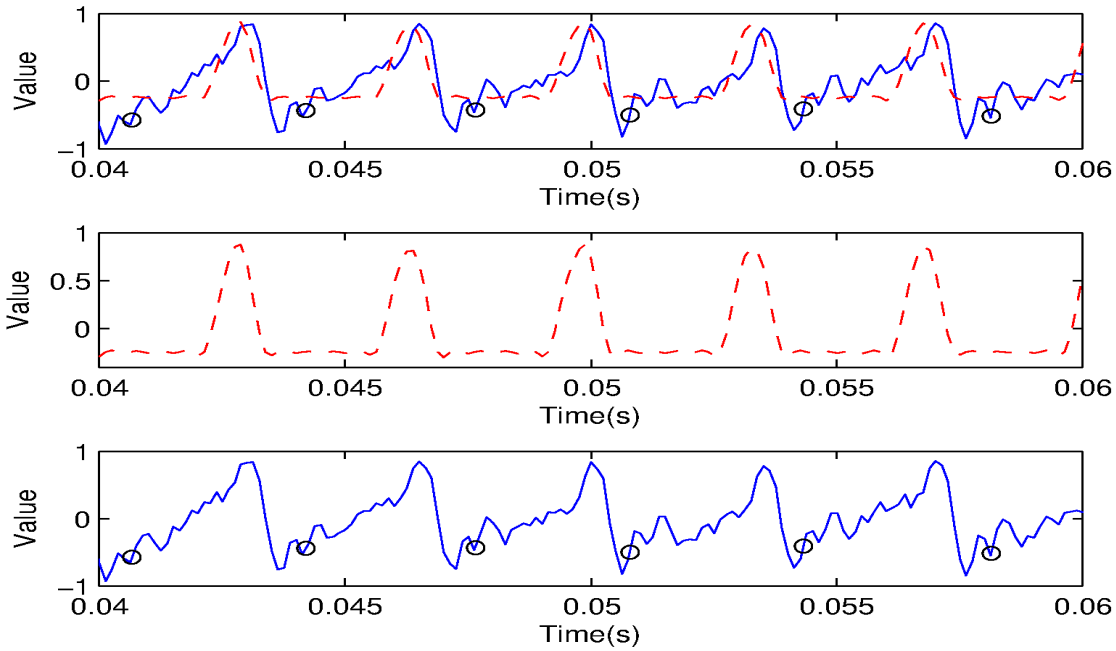


Figure 5.12: IAIF Method Result: Comparison Of 20%OQ Threshold Estimated Glottal Source and Glottal Displacement for an 11 year-old Male Child With Error of 64.2%. Glottal Source Waveform (solid blue line) with 20% Threshold for Each Glottal Source Period (black circles) and Displacement Waveform (dashed red line).

Table 5.1: Percent Error Mean (M) and Standard Deviation (SD) for the Total Data Set and Non-Anomalous (NA) Data along with the Percent Anomalous (PA) Data for the IAIF-Estimated Glottal Source Waveform Open Quotient Separated By Area 20% OQ Threshold, Area 50% OQ Threshold and Displacement With Glottal Source 20% OQ Threshold.

| | Tot. M | Tot. SD | PA | NA M | NA SD |
|---|---|---|---|---|---|
| Area 20% | 49.01% | 31.57% | 84.78% | 8.46% | 11.07% |
| Area 50% | 37.14% | 34.47% | 71.74% | 5.66% | 11.36% |
| Disp GS 20% | 8.90% | 22.34% | 30.23% | 1.39% | 8.85% |

time-instant would correspond to always underestimating the glottal open quotient and that the 20% threshold time-instant would more closely follow the glottal opening time-instant and therefore more accurately represent the glottal open quotient. To compare the results of all of the methods, the errors were separated. Any errors exceeding 20% in value of the actual open quotient, deemed to be from the area or displacement waveforms, would be considered as a result of the algorithm not properly filtering out key components, especially the large negative spike, affecting the threshold level time-instants. Any errors less than 20% would be more of the result of slight estimation errors from inverse filtering or open quotient calculation. The percent anomalous detections were given by:

$$P_A = \frac{N_A}{N_T} \tag{5.32}$$

where $N_A$ is the number of subjects whose error exceeded 20% and $N_T$ is the total number of subjects in the set. The percent error mean and standard deviation of the entire set and the non-anomalous set were calculated and are shown along with the percent anomalous for the IAIF-estimated open quotient in Table 5.1. It can be easily seen from Table 5.1 that, for the entire data set, compared to the area 20% OQ and area 50% OQ, the IAIF-estimated glottal source OQ did not strongly agree. An error of 49.01% for the 20% and a standard deviation of 31.57% meant that the

20% threshold time-instant was not consistent throughout all of the data compared to its corresponding area 20% threshold which may have been inconsistent as well, shown previously in the example figures. However, overall, when the entire data set was compared to the displacement waveform's OQ, the mean error dropped significantly and the overall standard deviation dropped as well. The percent anomalous was large for the area 20% and 50% thresholds and became significantly less for the displacement waveform. For the non-anomalous errors, it is, again, consistent throughout the table that the displacement waveform yielded a smaller mean of 1.39% when compared to the area 20% and 50% as well as a smaller standard deviation of 8.85%.

The displacement waveform may have lead to a smaller standard deviation and mean overall because it is a more consistent comparison because of the glottal opening and glottal closure time-instants being very explicit. These explicit time-instants lead to an open quotient calculation that may be a more consistent and non-ambiguous comparison than the area open quotient. An issue with comparing to the area is the non-explicitly defined glottal closure and opening time-instants, which lead to the use of the 20% and 50% threshold levels. However, these ambiguous threshold levels may lead to larger errors and larger standard deviation across those errors as shown in Table 5.1. This results in the conclusion that the displacement waveform's glottal features may be more accurate to compare with the acoustically-extracted glottal time-domain features.

## OQ Estimation Using Linear Prediction with Glottal Source Modeling

To fully determine how accurate the OQ Estimation Using Linear Prediction with Glottal Source Modeling method was in realizing the glottal source waveform time-domain features, a simple percent error calculation was performed. The

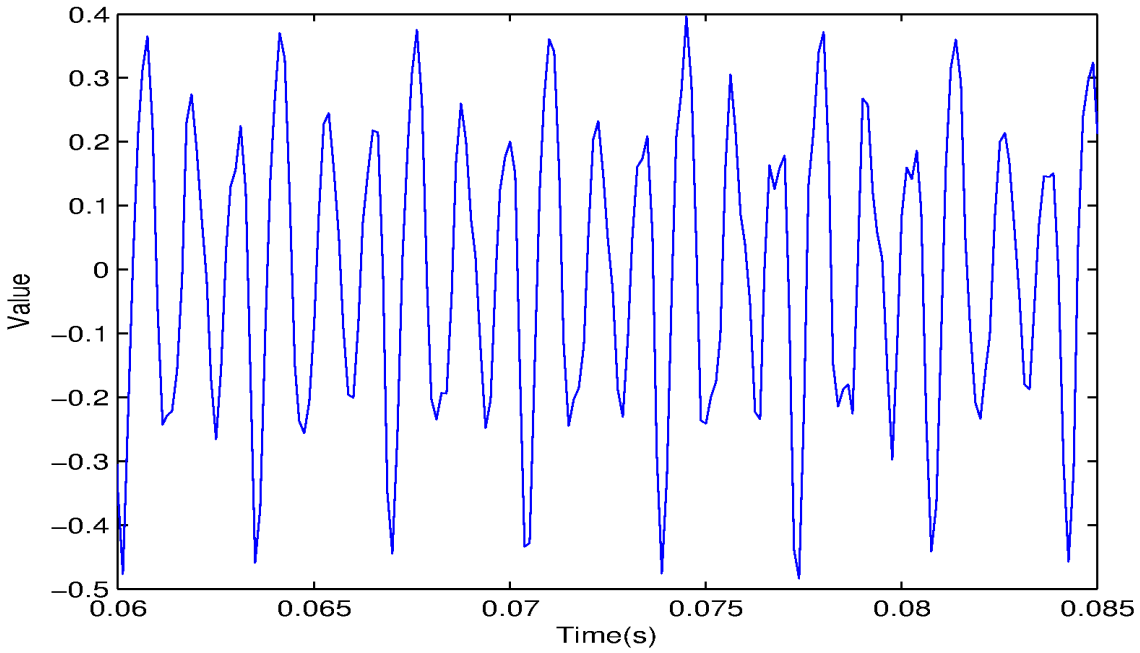


Figure 5.13: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for a 9 Year Old Male Child With Corresponding OQ Error of -61.9% Between Area and Non-Filtered Glottal Source.

corresponding area and displacement waveforms for thirty glottal cycles were compared with the nonfiltered and filtered glottal source estimation. A 9 year-old male child had a -61.9% error when comparing the area OQ and nonfiltered estimated glottal source OQ calculated from the glottal model equation. An example of the male child's acoustic, glottal area and estimated glottal source waveforms is shown in Figures 5.13 and 5.14. This subject's acoustic recording was determined to have a fundamental frequency of 286Hz. The resulting estimated glottal source waveform has an appropriate shape, noting the right skewness of each glottal pulse, however, the waveform is not very smooth and an almost "dual peak" pulse occurs, resulting from the algorithm not filtering the formant frequencies properly. Because of this, the 2nd order LPC coefficients will be affected and the estimated glottal source open quotient value will not match the corresponding area waveform. After listening, the acoustic waveform was perceived to be muffled, which may have affected the outcome of the results.

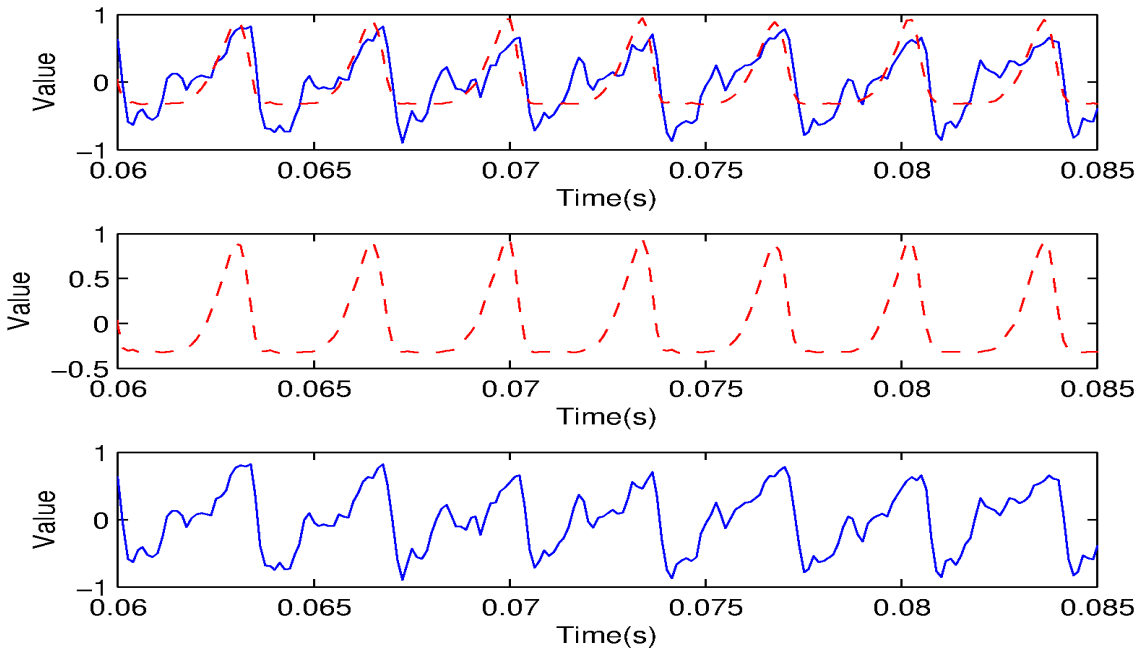


Figure 5.14: OQ Estimation Using Glottal Modeling Result: Comparison of OQ for the Area and Estimated Nonfiltered Glottal Source for a 9 Year Old Male Child With Error of -61.9%. Non-Filtered Glottal Source Waveform (solid blue line) and Area Waveform (dashed red line).

A 27 year-old female had a -91.5% error when comparing the area OQ and nonfiltered estimated glottal source OQ calculated from the glottal model equation. An example of the female's acoustic, glottal area and estimated glottal source waveforms is shown in Figures 5.15 and 5.16. This subject's acoustic recording was determined to have a fundamental frequency of 222Hz. The resulting estimated glottal source waveform appears to be not smooth due to the algorithm not properly filtering out the necessary vocal tract formants. In this case, the inverse filtering algorithm failed to accurately estimate the vocal tract filter and left the resulting glottal source with higher frequency peaks that will affect the 2nd order LPC coefficients, and thus the open quotient. After listening, the acoustic waveform was perceived to be slightly muffled, but clearly audible nonetheless, which leads to the conclusion that the inverse filtering algorithm failed to properly estimate the glottal source.

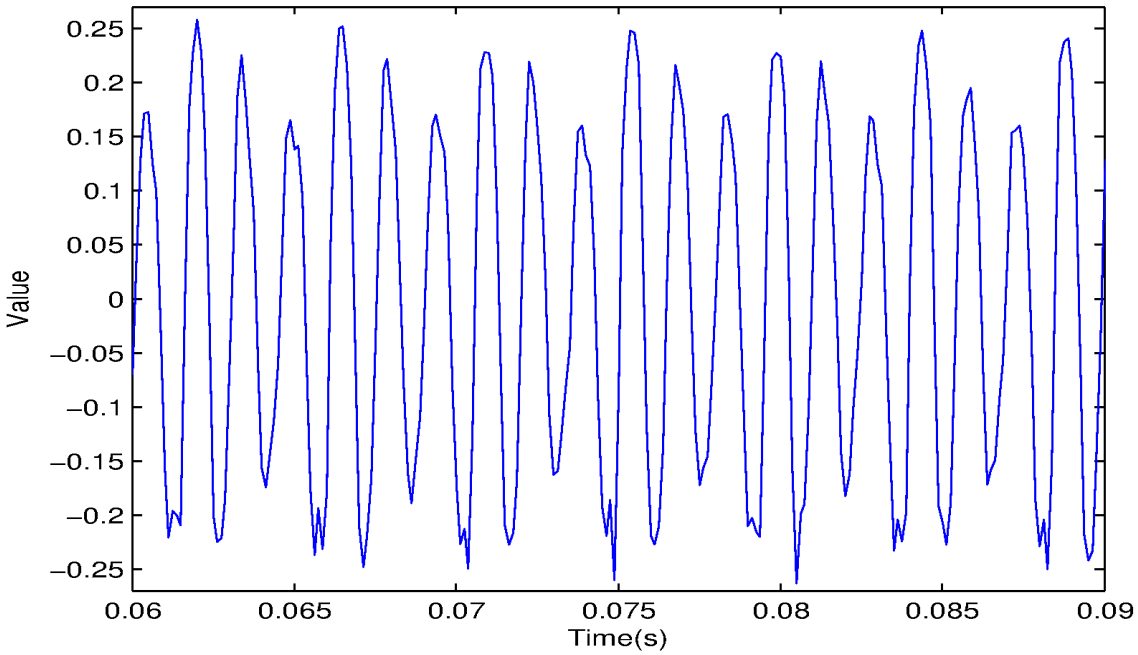


Figure 5.15: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for a 27 year-old Female With Corresponding OQ Error of -91.5% Between Area and Non-Filtered Glottal Source.

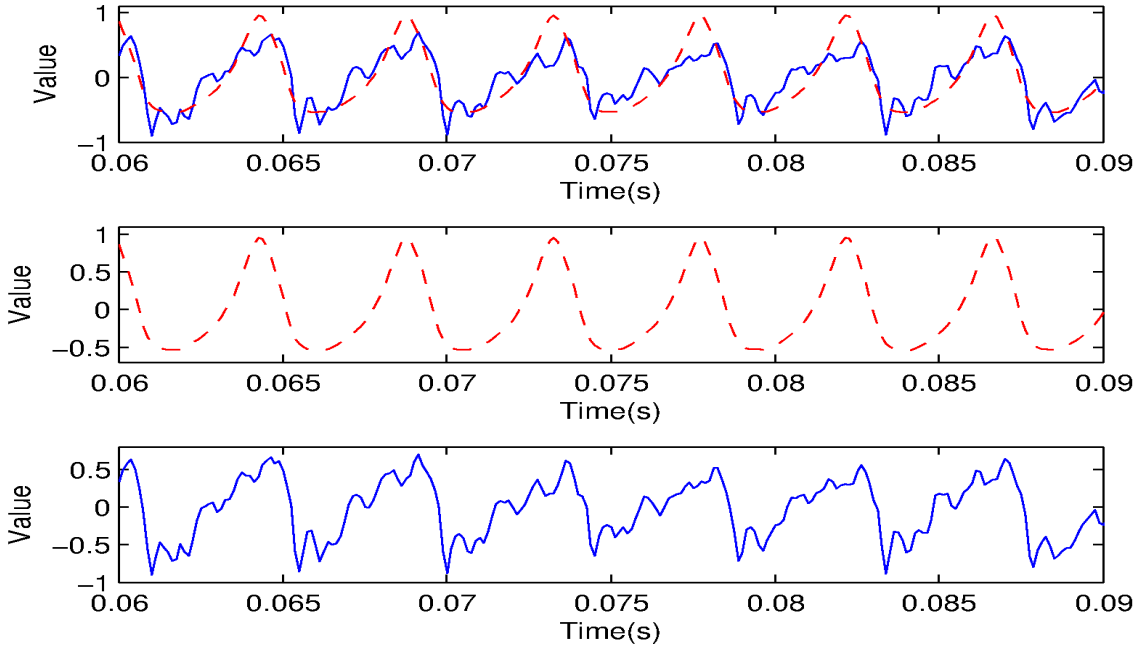


Figure 5.16: OQ Estimation Using Glottal Modeling Result: Comparison of OQ for the Area and Estimated Nonfiltered Glottal Source for a 27 year-old Female with Error of -91.5%. Non-Filtered Glottal Source Waveform (solid blue line) and Area Waveform (dashed red line).

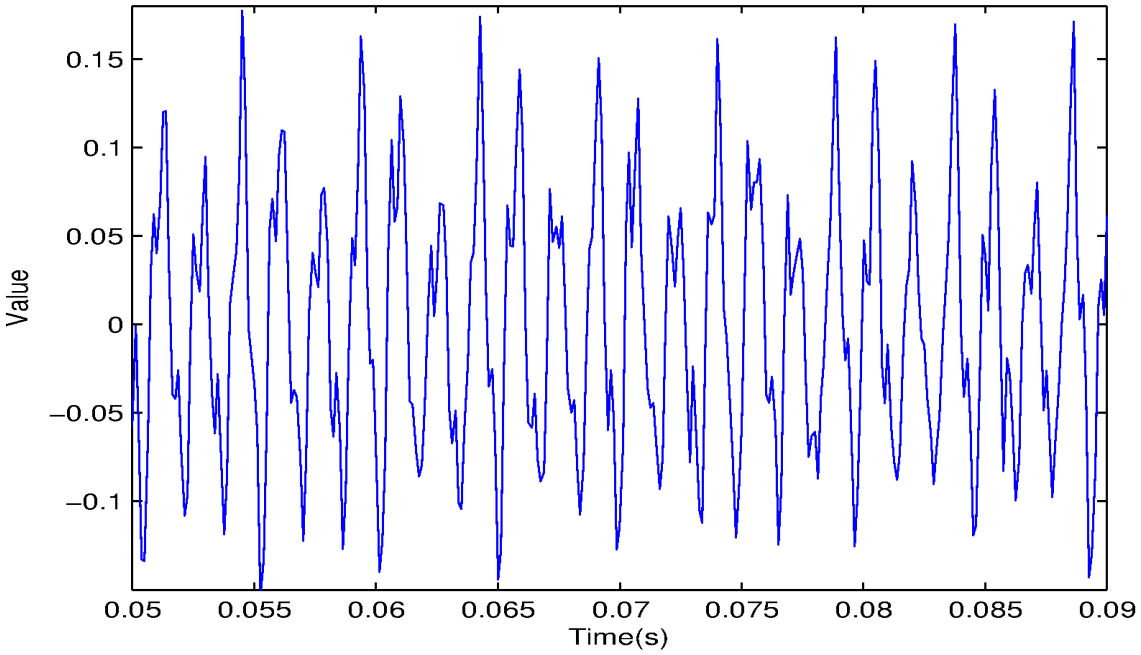


Figure 5.17: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for a 21 year-old Female with Corresponding OQ Error of 2.18% Between Area and Lowpass Filtered Glottal Source.

A 21 year-old female had a 2.18% error when comparing the area OQ and lowpass filtered estimated glottal source OQ calculated from the glottal model equation. An example of the female's acoustic, glottal area and estimated glottal source waveforms is shown in Figures 5.17 and 5.18. This subject's acoustic recording was determined to have a fundamental frequency of 205Hz. The resulting estimated glottal source waveform appears to be very smooth due to the algorithm properly filtering out the necessary vocal tract formants. In this case, the glottal source appears to be very similar to the area waveform yielding the 2nd order LPC coefficients to be similar and the open quotient error to be very small for the glottal source estimation. The lowpass filtering after estimating appeared to have some impact on smoothing the waveform and filtering out the unnecessary high frequency components. After listening, the acoustic waveform was perceived to have some slight noise in the recording, however it appeared to be white noise in the PSD of the acoustic waveform and did not ultimately affect the results.

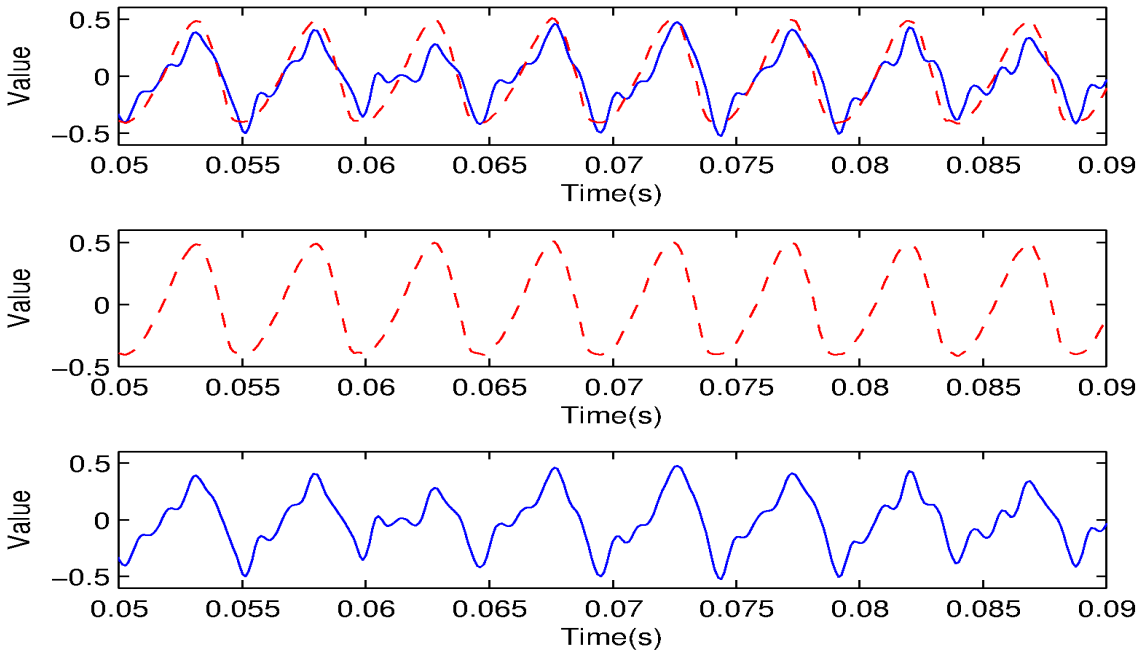


Figure 5.18: OQ Estimation Using Glottal Modeling Result: Comparison of OQ for the Area and Lowpass Filtered Estimated Glottal Source for a 21 year-old Female With Error 2.18%. Lowpass Filtered Glottal Source Waveform (solid blue line) and Area Waveform (dashed red line).

A 27 year-old female had a -91.4% error when comparing the area OQ and lowpass filtered estimated glottal source OQ calculated from the glottal model equation. An example of the female's acoustic, glottal area and estimated glottal source waveforms is shown in Figures 5.19 and 5.20. This subject's acoustic recording was determined to have a fundamental frequency of 222Hz. The resulting estimated glottal source waveform appears to attempt to be the correct glottal pulse shape but does not accurately dictate the same time-instants as the area waveform. The lowpass filtering after estimating appeared to have some impact on smoothing the waveform, but due to the fact that the pre-filtered glottal source estimate did not have the same timing instants when compared to its corresponding area waveform, and left dominant higher frequency "ripples" in the glottal source waveform, the lowpass filtering could not aid in resolving this problem. After listening, the acoustic recording was perceived to be slightly muffled, and it may

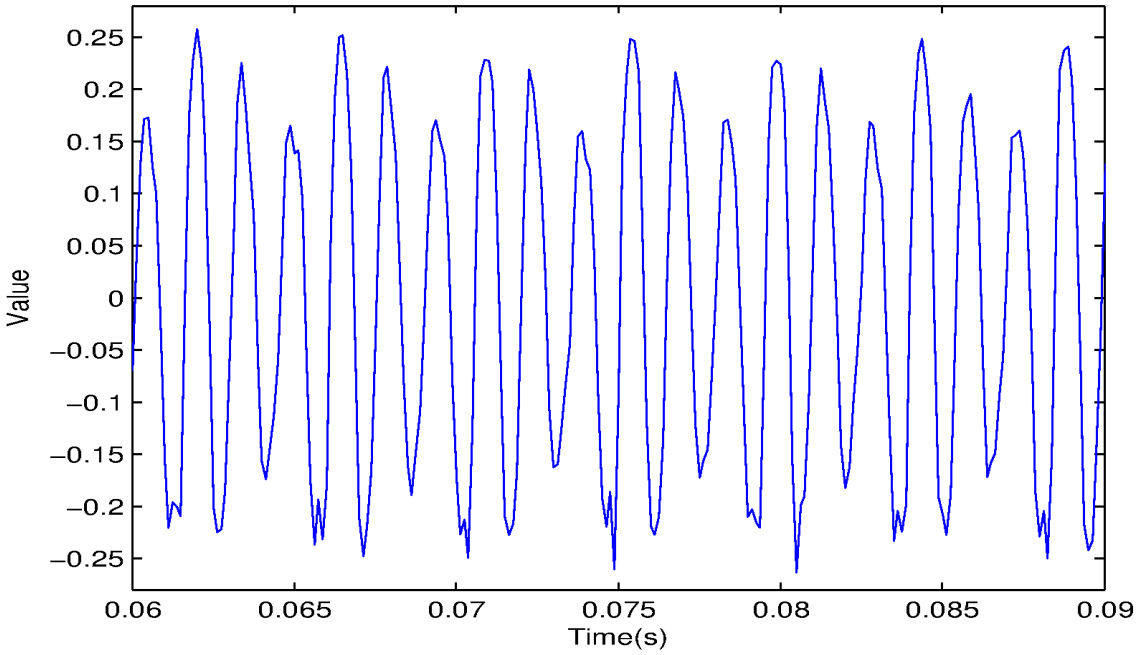


Figure 5.19: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for 27 year-old Female With Corresponding OQ Error of -91.4% Between Area and Lowpass Filtered Glottal Source.

have had an effect on the results.

A 6 year-old male child had a -24.7% error when comparing the displacement OQ and nonfiltered estimated glottal source OQ calculated from the glottal model equation. An example of the male child's acoustic, glottal displacement and estimated glottal source waveforms is shown in Figures 5.21 and 5.22. This subject's acoustic recording was determined to have a fundamental frequency of 333Hz. The resulting estimated glottal source waveform appears to attempt to be the correct glottal pulse shape but does not accurately dictate the same time-instants as the area waveform. The amplitude of waveform seems to change overtime as well as the time-instants, yielding an inconsistent open quotient calculation over multiple periods. Because of this issue, the 2nd order LPC coefficients for the glottal source and area will be different and the corresponding open quotient's will not agree as well. After listening, the acoustic recording was perceived to be slightly muffled, and it may have had an effect on the results.

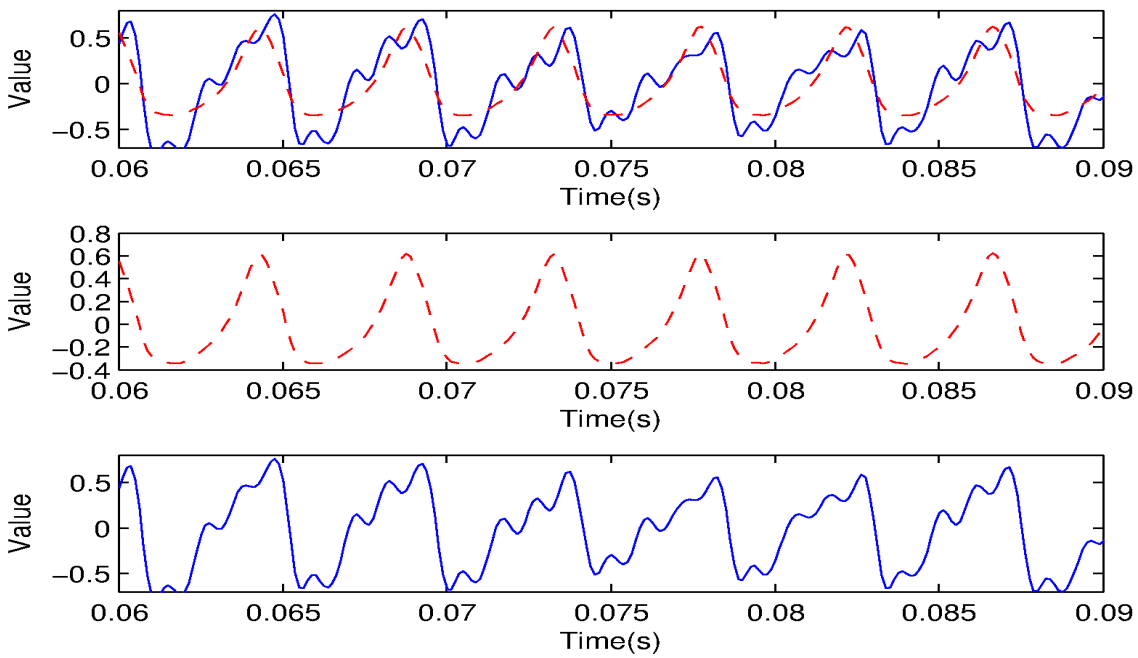


Figure 5.20: OQ Estimation Using Glottal Modeling Result: Comparison Of OQ For The Area and Lowpass Filtered Estimated Glottal Source For 27 Year Old Female With Error -91.4%. Lowpass Filtered Glottal Source Waveform (solid blue line) and Area Waveform (dashed red line).

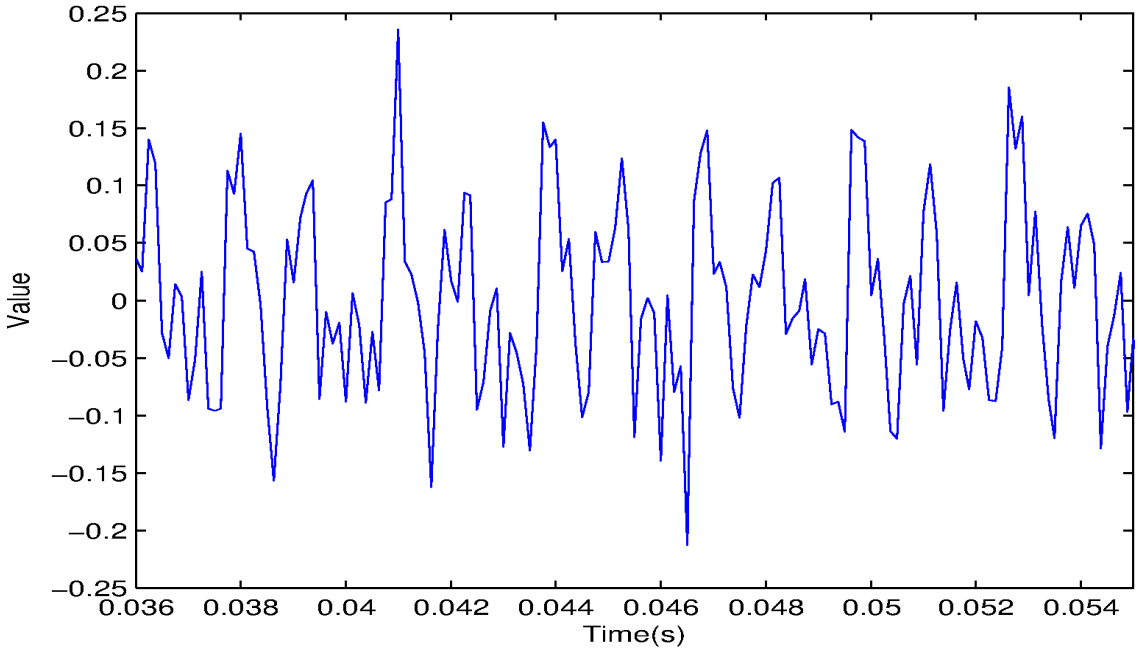


Figure 5.21: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for a 6 year-old Male Child With Corresponding OQ Error of -24.7% Between Displacement and NonFiltered Glottal Source.

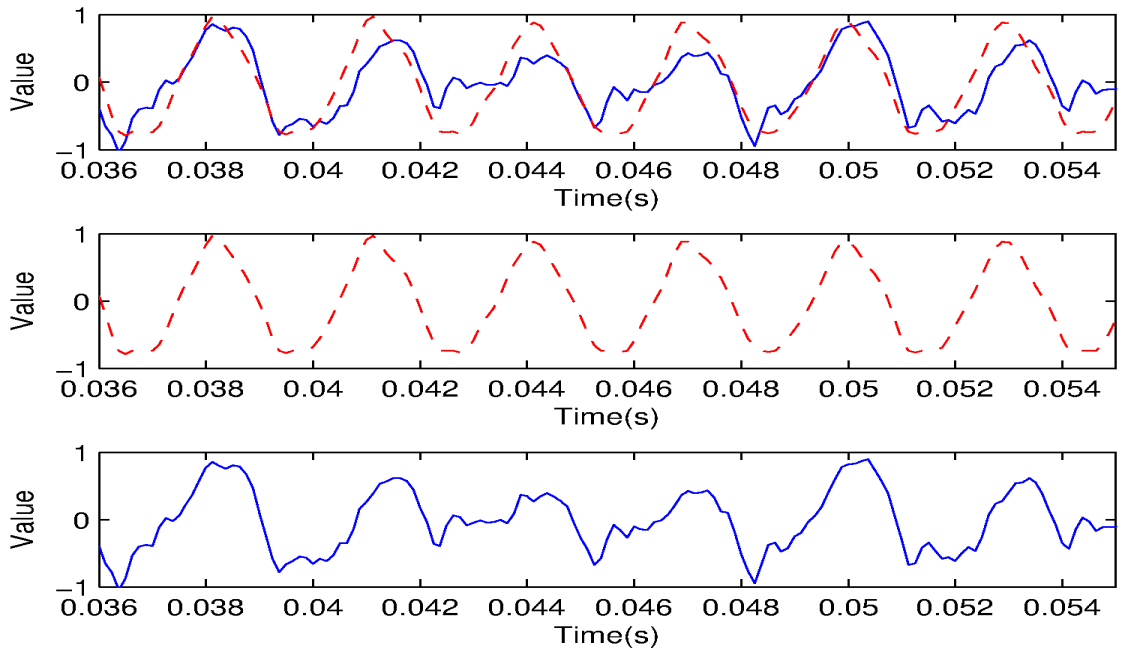


Figure 5.22: OQ Estimation Using Glottal Modeling Result: Comparison Of OQ For The NonFiltered Estimated Glottal Source and Glottal Displacement for a 6 year-old Male Child With Error of -24.7%. NonFiltered Glottal Source Waveform (solid blue line) and Displacement Waveform (dashed red line).

A 27 year-old female had a -91.1% error when comparing the displacement OQ and nonfiltered estimated glottal source OQ calculated from the glottal model equation. An example of the female's acoustic, glottal displacement and estimated glottal source waveforms is shown in Figures 5.23 and 5.24. This subject's acoustic recording was determined to have a fundamental frequency of 222Hz. The resulting estimated glottal source waveform appears to attempt to be the correct glottal pulse shape but does not accurately filter out the higher frequency components. The negative pressure differential before glottal closure can be seen in a few of the periods and will affect the timing of the glottal source pulses. And although the waveform appears to be quasi-periodic, the time instants aren't explicitly defined and will affect the 2nd order LPC coefficients of the glottal source and result in a larger error when comparing its open quotient to the displacement waveform's. After listening, the acoustic recording was perceived to be slightly muffled, and it

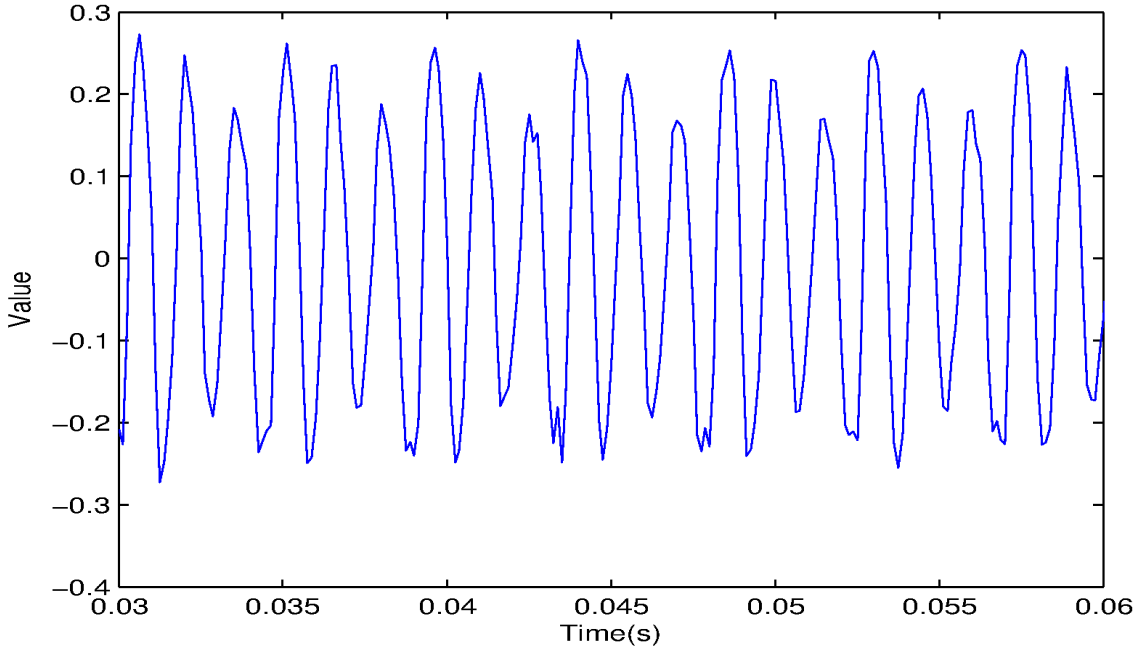


Figure 5.23: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for a 27 year-old Female With Corresponding OQ Error of -91.1% Between Displacement and NonFiltered Glottal Source.

may have had an effect on the results.

A 9 year-old male child had a -53.6% error when comparing the displacement OQ and lowpass filtered estimated glottal source OQ calculated from the glottal model equation. An example of the male child's acoustic, glottal displacement and estimated glottal source waveforms is shown in Figures 5.25 and 5.26. This subject's acoustic recording was determined to have a fundamental frequency of 258Hz. The resulting estimated glottal source waveform appears to be the correct glottal pulse shape but also did not filter out a strong vocal tract formant and that resulted in a second peak in the glottal source during the glottal closure phase defined by the displacement waveform. The waveform is clearly more smooth due to lowpass filtering, however, the underlying issue is the second peak during the glottal closure phase that will have an effect on the 2nd order LPC coefficients and therefore, the open quotient value because, other than the second peak, the glottal pulse appears to be very similar to the displacement pulse. After listening, the acoustic recording

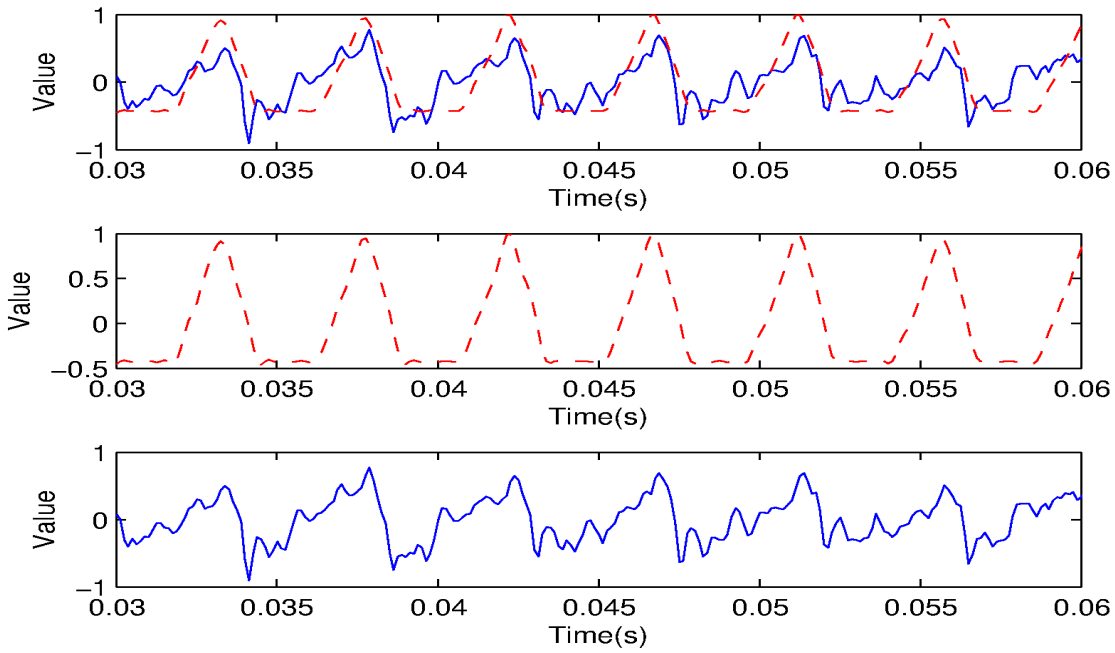


Figure 5.24: OQ Estimation Using Glottal Modeling Result: Comparison of OQ For The Estimated Glottal Source and Glottal Displacement for a 27 year-old Female with Error of -91.1%. NonFiltered Glottal Source Waveform (solid blue line) and Displacement Waveform (dashed red line).

was perceived to be slightly noisy.

A 27 year-old female had a -91.0% error when comparing the displacement OQ and lowpass filtered estimated glottal source OQ calculated from the glottal model equation. An example of the female's acoustic, glottal displacement and estimated glottal source waveforms is shown in Figures 5.27 and 5.28. This subject's acoustic recording was determined to have a fundamental frequency of 222Hz. The resulting estimated glottal source waveform appears to attempt to be the correct glottal pulse shape, but did not filter out some higher frequency components of the vocal tract. The waveform is clearly more smooth due to lowpass filtering and some glottal closure phases can be almost visually perceived, however, the timing instants are clearly different for the glottal source and displacement waveforms. This has a strong effect on the open quotient calculation. After listening, the acoustic recording was perceived to be slightly muffled.

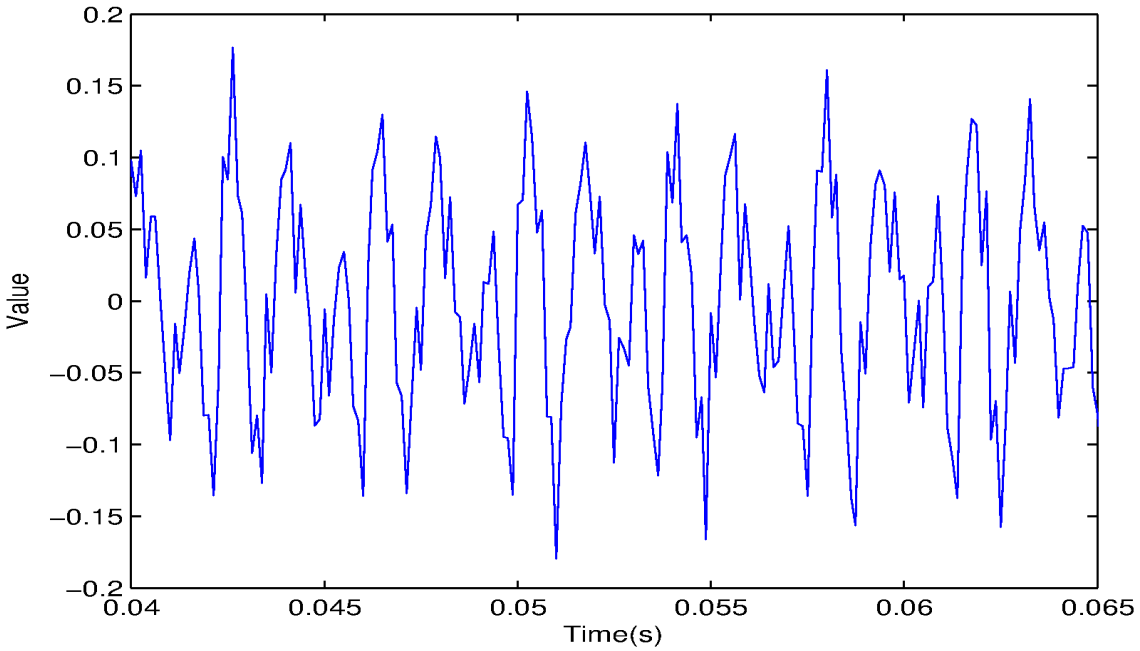


Figure 5.25: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for a 9 year-old Male Child With Corresponding OQ Error of -53.6% Between Displacement and Lowpass Filtered Glottal Source.

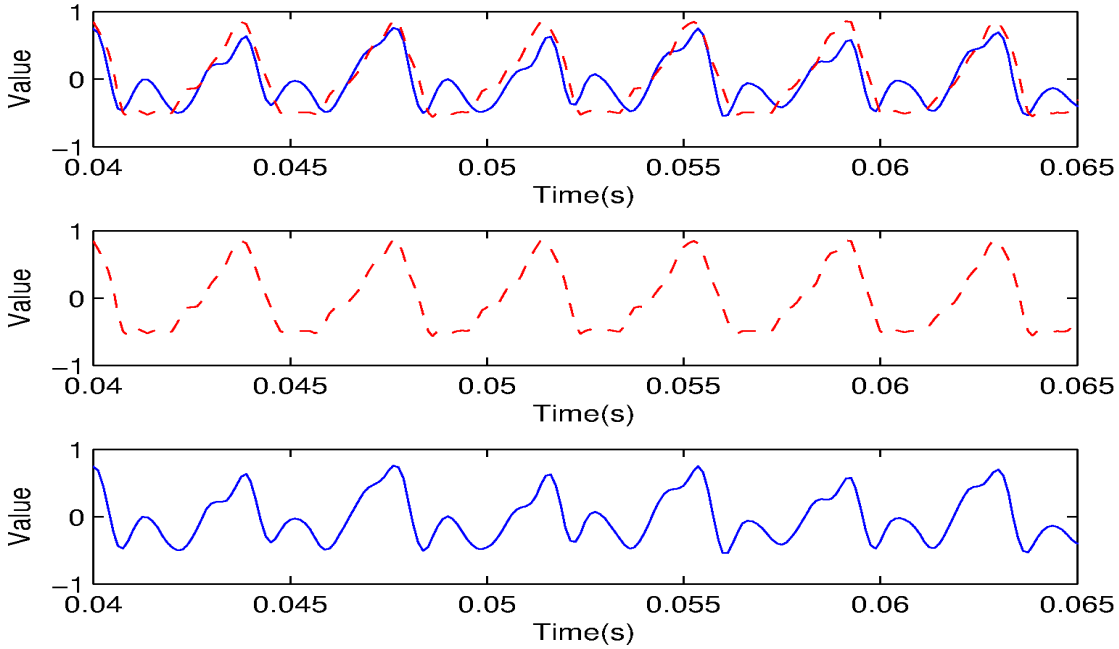


Figure 5.26: OQ Estimation Using Glottal Modeling Result: Comparison Of OQ For The Lowpass Filtered Estimated Glottal Source and Glottal Displacement for a 9 year-old Male Child With Error -53.6%. Lowpass Filtered Glottal Source Waveform (solid blue line) and Displacement Waveform (dashed red line).

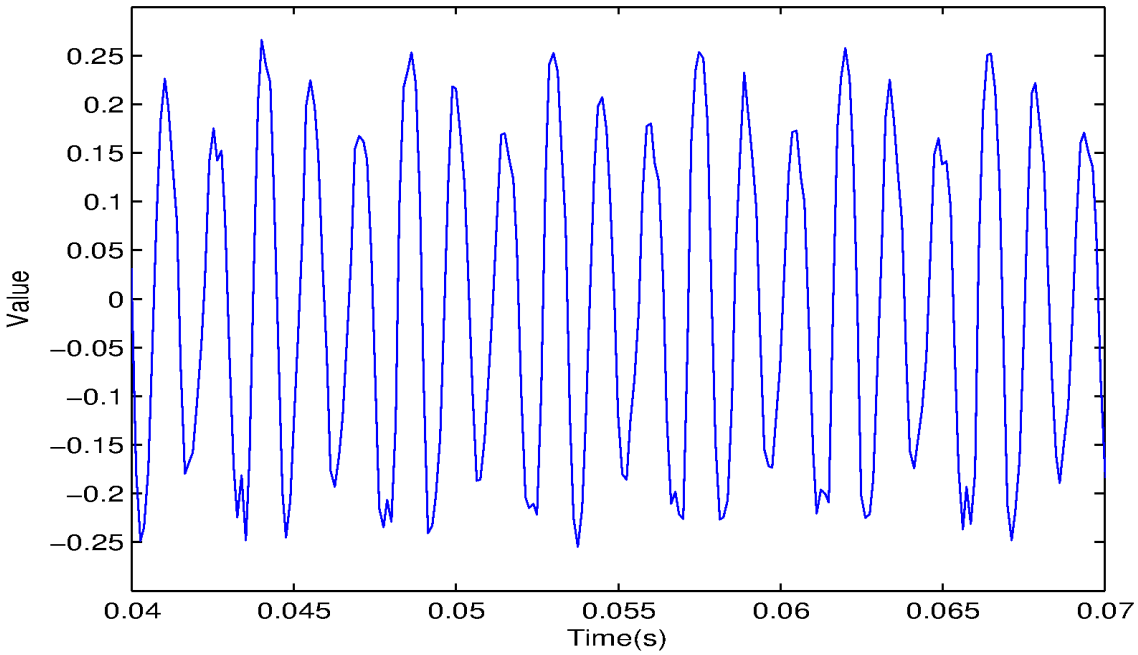


Figure 5.27: OQ Estimation Using Glottal Modeling Result: Acoustic Waveform for 27 year old Female With Corresponding OQ Error of -91.0% Between Displacement and Lowpass Filtered Glottal Source.

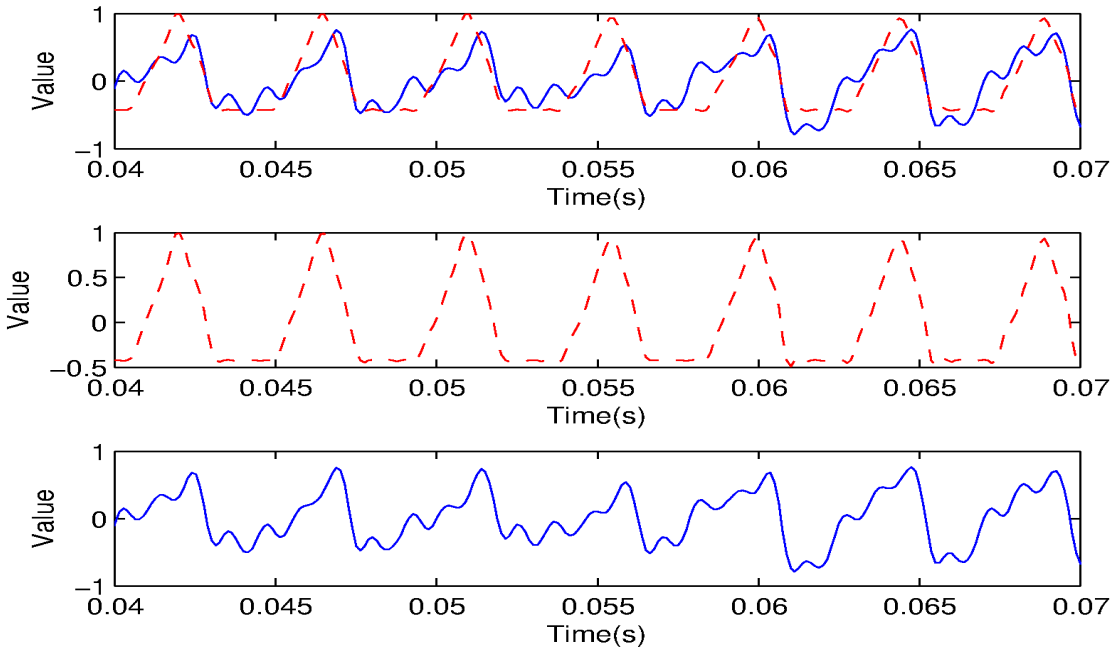


Figure 5.28: OQ Estimation Using Glottal Modeling Result: Comparison of OQ For The Estimated Glottal Source and Glottal Displacement for 27 Year Old Female With Error -91.0%. Lowpass Filtered Glottal Source Waveform (solid blue line) and Displacement Waveform (dashed red line).

Table 5.2: Percent Error Mean (M) and Standard Deviation (SD) for the Total Data Set and Non-Anomalous Data (NA) along with the Percent Anomalous (PA) For The OQ Estimation With Glottal Modeling Glottal Source Waveform Separated By Area With NonFiltered Glottal Source (Area NF), Area With Filtered Glottal Source (Area F), Displacement With NonFiltered Glottal Source (Disp NF) and Displacement With Filtered Glottal Source (Disp F).

| | Tot. M | Tot. SD | PA | NA M | NA SD |
|---|---|---|---|---|---|
| Area NF | -63.14% | 28.39% | 94.87% | -9.89% | 4.09% |
| Area F | -45.71% | 40.50% | 85.00% | -6.28% | 8.19% |
| Disp NF | -66.51% | 29.11% | 100.00% | - | - |
| Disp F | -51.70% | 36.35% | 83.78% | -0.94% | 11.58% |

The OQ estimation using linear prediction with glottal source modeling algorithm was applied to 46 different subjects and the open quotient was estimated and compared to its corresponding area open quotient. Any outliers (open quotient calculations that yielded values greater than 100%) were immediately removed, leaving 40 subjects to compare. Due to the limited supply of displacement waveform data and the previous outlier removal, only 36 subjects were compared with their corresponding displacement waveform open quotient. To compare the results of all of the methods, the errors were separated as previously with the IAIF method. Any errors exceeding 20% in value of the actual open quotient, deemed to be from the area or displacement waveforms, would be considered as a result of the algorithm not properly filtering out key components, especially the large negative spike or higher frequency components for this algorithm's case. Any errors less than 20% would be more of the result of slight estimation errors from inverse filtering or open quotient calculation. The percent error mean and standard deviation of the total data set and the non-anomalous data set were calculated along with the percent anomalous for this method's estimated open quotient and are shown in Table 5.2. This table was separated by comparison with the area and displacement as well as the glottal source with lowpass filtering and non-filtering.

It can be easily seen from Table 5.2 that, for the entire data set, this

algorithm did not seem to work well in calculating the open quotient from the estimated glottal source when compared to the area and displacement waveform open quotient values. The large percent anomalous for all the comparisons dictate that this algorithm does not accurately estimate the glottal source, or at least accurately enough for a 2nd order LPC analysis to be used to calculate the open quotient. The means for all of the different comparisons are all greater than 50%, which is even more emphasized by the fact that the percent anomalous is 80% or higher for all of the comparisons. However, one noticeable trend is the fact that filtering decreases the mean error for the total group and at least for the area comparison for the non-anomalous error. The displacement filtering versus nonfiltering cannot be compared due to the fact that the displacement nonfiltering data was 100% anomalous, meaning that all the errors were greater than 20%. Even though filtering did decrease the mean overall for the comparisons, it also increased the standard deviation meaning that there was more variability in the error after filtering.

## Linear Prediction Error Waveform Analysis with Peak Detection

Because this method did not derive a glottal source waveform estimate, a way to compare to the area open quotient had to be determined. A test set of twelve subjects were chosen and their corresponding area OQ was calculated for 1% to 30% threshold levels. These threshold levels were then compared to the open quotients calculated from the two different methods being applied: an LPC order of 1 applied to the acoustic waveform and then and LPC order of 10 being applied to the corresponding error waveform from LPC order 1, and just a 10th order LPC analysis being applied to the acoustic waveform. The percent error was calculated between each of the threshold levels for the area from 1% to 30% and their

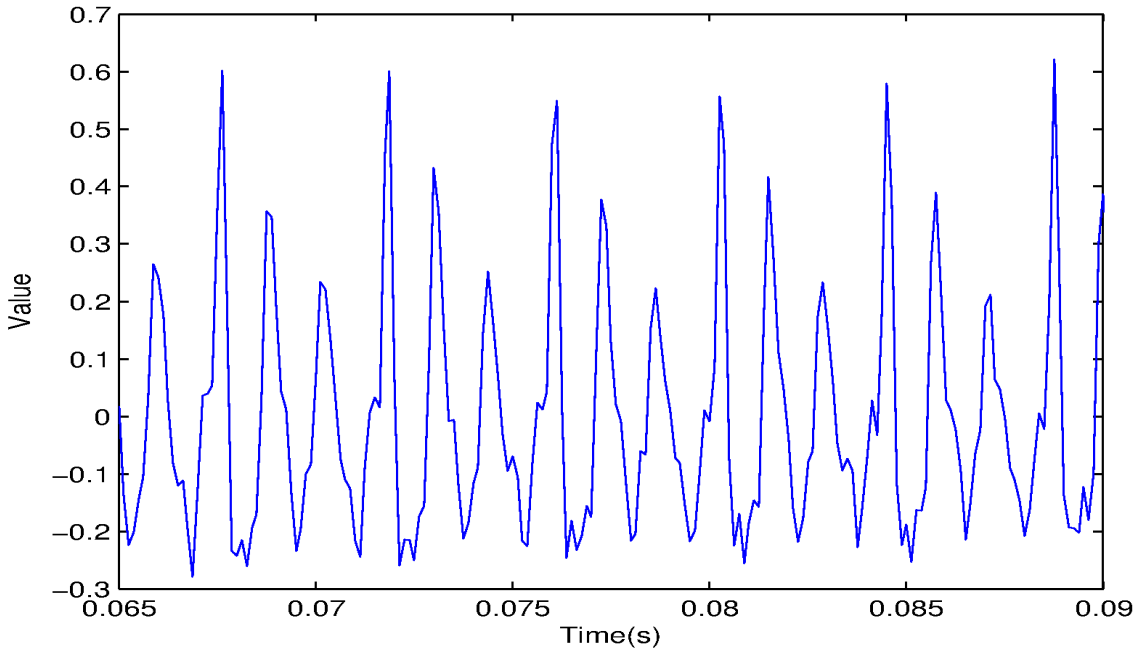


Figure 5.29: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for a 19 year-old Female With Corresponding OQ Error of 0.0096% Between Area and Linear Prediction Error Waveform.

corresponding error waveform for the open quotient. The smallest mean percent error for the training set yielded an open quotient threshold of 7%. This was then used as the standard to compare the open quotient results against for the area.

A 19 year-old female had a 0.0096% error when comparing the area 7%OQ and an LPC 10th order error waveform OQ. An example of the female's acoustic, glottal area with 7% threshold level and LPC 10th order error waveform is shown in Figures 5.29 and 5.30. This subject's acoustic recording was determined to have a fundamental frequency of 236Hz. As discussed previously, the glottal closure instants are defined as the positive peak following the strongest negative peak, which is denoted by the x's in the figure and the opening instants are defined by the strongest positive peak between the glottal closure peaks, denoted by the black circles. In this case, it can be shown that the time-instants for the area 7% threshold OQ and error waveform time-instants agree. After listening, the acoustic recording was perceived to be very clean.

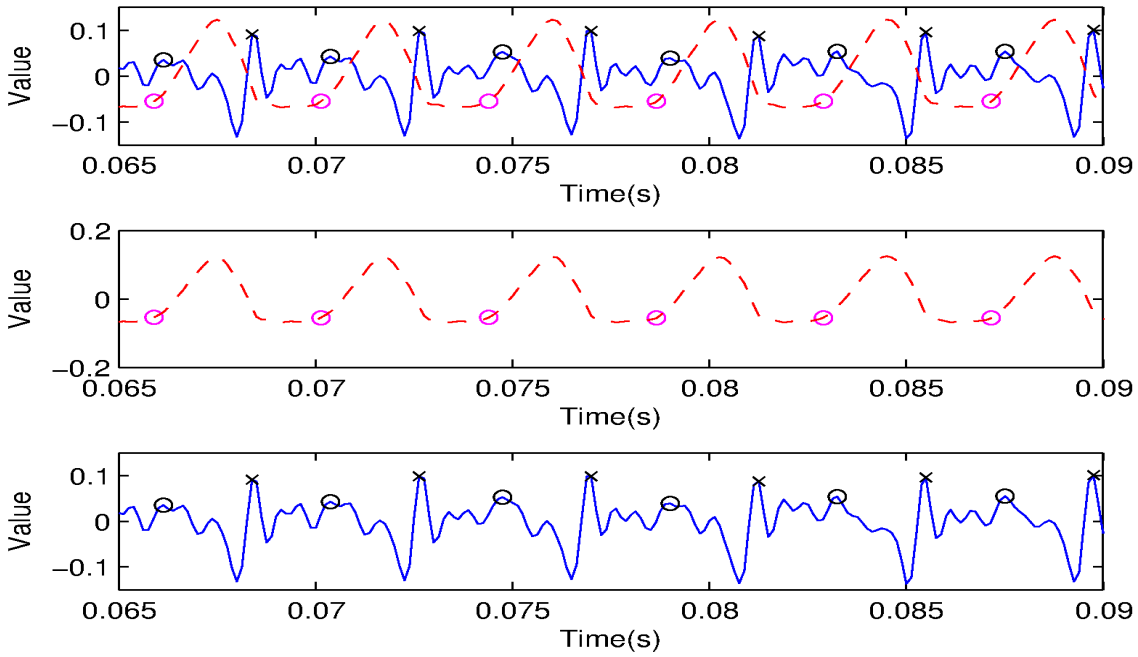


Figure 5.30: Linear Prediction Error Waveform With Peak Detection Result: Comparison of 7%OQ Threshold for the Area and OQ of Error Waveform for a 19 year-old Female With Error 0.0096%. LPC 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period. Area Waveform (dashed red line) with 7% Threshold for Each Area Period (magenta circles).

A 9 year-old male child had a 75.6% error when comparing the area 7%OQ and an LPC 10th order error waveform OQ. An example of the female's acoustic, glottal area with 7% threshold level and LPC 10th order error waveform is shown in Figures 5.31 and 5.32. This subject's acoustic recording was determined to have a fundamental frequency of 286Hz. It can be easily seen that the algorithm may have identified the incorrect peaks for glottal closure and opening. This is evident when the algorithm identifies the strong negative peak and then does not immediately identify the next peak as the glottal closure, which was an error in Matlab's *findpeak* command. This will have a effect on the open quotient calculation and therefore, disagree with the 7% area open quotient. After listening, the acoustic recording was perceived to be slightly muffled.

A 8 year-old male child had a 0.39% error when comparing the area 7%OQ

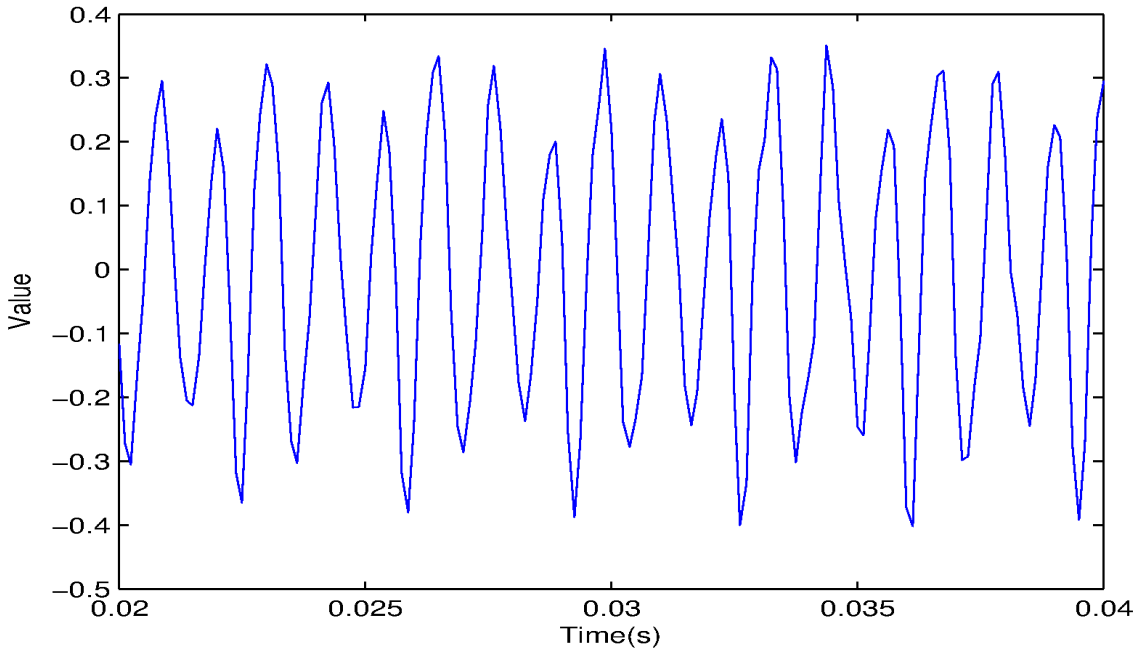


Figure 5.31: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for 9 year-old Male Child With Corresponding OQ Error of 75.6% Between Area and Linear Prediction Error Waveform.

and an LPC 1st order, followed by a 10th order, error waveform OQ. An example of the male child's acoustic, glottal area with 7% threshold level and LPC 1st order, followed by a 10th order, error waveform is shown in Figures 5.33 and 5.34. This subject's acoustic recording was determined to have a fundamental frequency of 258Hz. It can be easily seen that the algorithm identified the glottal closure and glottal opening peaks correctly, without mis-identifying closure or opening time-instants. Because of this, there was a strong agreement between the area 7% open quotient values and the open quotient values calculated from the error waveform. After listening, the acoustic recording was perceived to be slightly muffled.

A 20 year-old male had a 49.3% error when comparing the area 7%OQ and an LPC 1st order, followed by a 10th order, error waveform OQ. An example of the male's acoustic, glottal area with 7% threshold level and LPC 1st order, followed by a 10th order, error waveform is shown in Figures 5.35 and 5.36. This subject's

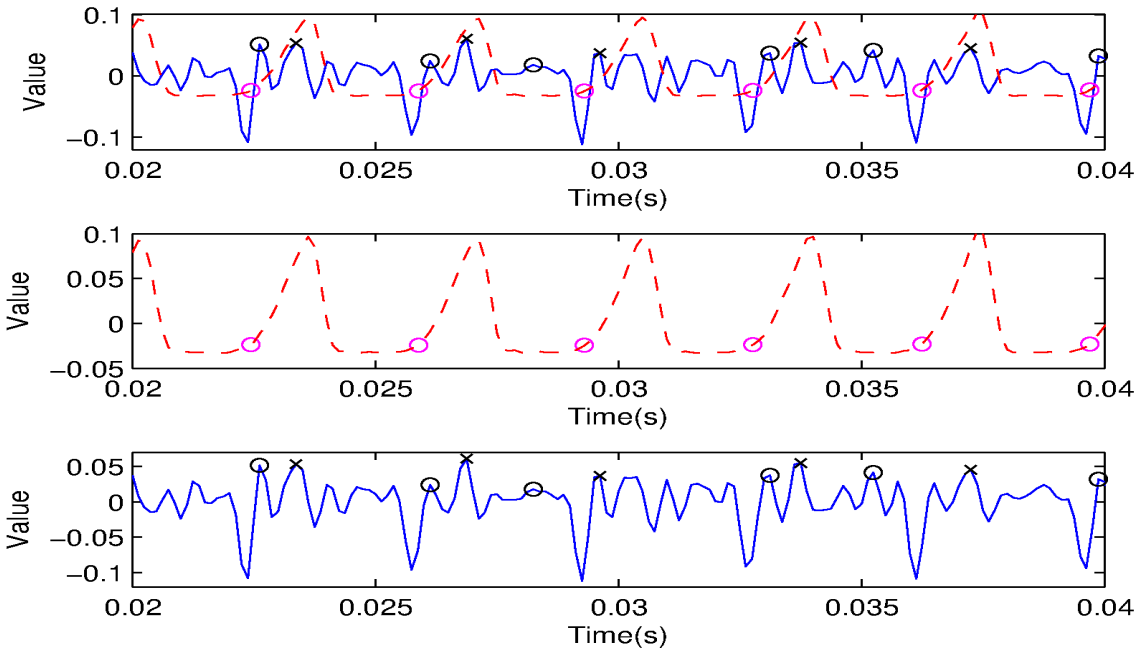


Figure 5.32: Linear Prediction Error Waveform With Peak Detection Result: Comparison of 7%OQ Threshold for the Area and OQ of Error Waveform for a 9 year-old Male Child With Error 75.6%. LPC 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period. Area Waveform (dashed red line) with 7% Threshold for Each Area Period (magenta circles).

acoustic recording was determined to have a fundamental frequency of 116Hz. It can be easily seen that the algorithm identified the glottal closure instants correctly. However, due to the strongest peak between the glottal closure instants changing in phase every period, the glottal opening instant changed in phase every period. This effect had an impact on the glottal open quotient estimation from the error waveform and caused disagreement between the error waveform open quotient and glottal area open quotient. After listening, the acoustic recording was perceived to be very noisy, and after observing the PSD of the acoustic waveform, the fundamental frequency spikes were slightly buried in the noise floor. Because of the the fundamental frequency not being as prominent, it would have been harder to detect with the LPC analysis and therefore the error waveform would have been affected.

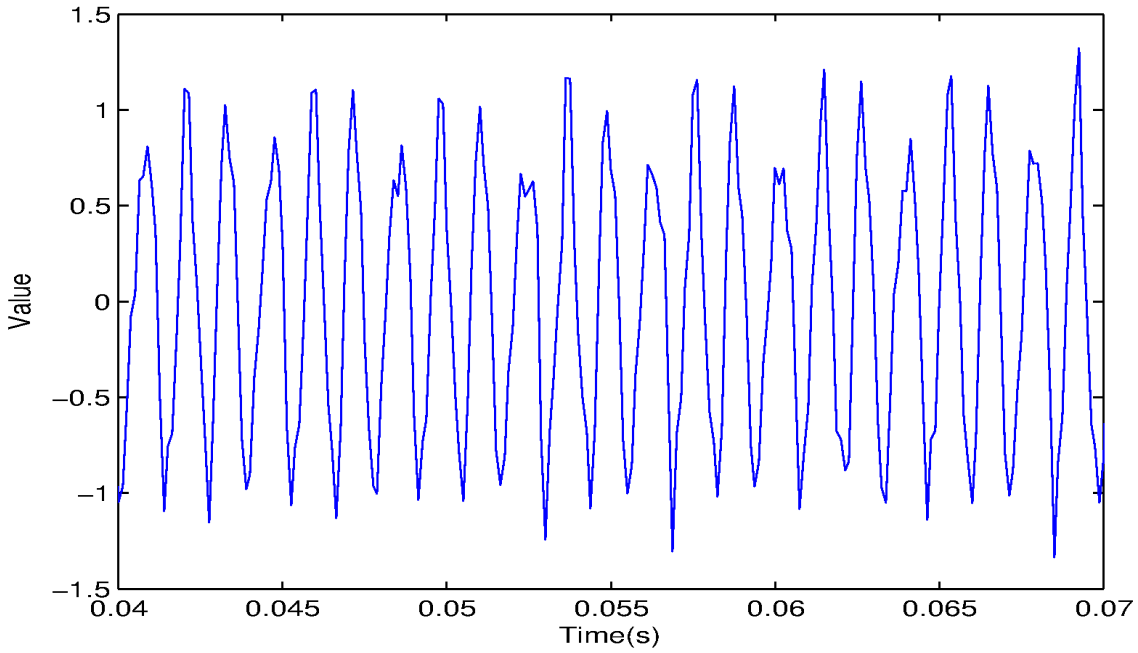


Figure 5.33: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for an 8 year-old Male Child With Corresponding OQ Error of 0.39% Between Area and Linear Prediction Error Waveform.

A 42 year-old female had a -1.62% error when comparing the displacement and an LPC 10th order error waveform OQ. An example of the female's acoustic, glottal displacement and LPC 10th order error waveform is shown in Figures 5.37 and 5.38. This subject's acoustic recording was determined to have a fundamental frequency of 250Hz. It can be easily seen that the algorithm identified the glottal closure and opening instants correctly and also had an agreement with the displacement waveform. Because of this agreement, the glottal open quotient calculated from the error waveform is very similar to the open quotient calculated from the corresponding displacement waveform. After listening, the acoustic recording was perceived to be slightly muffled and noisy, but the PSD reflected that the fundamental frequency components power was much stronger than that of the noise, which is why the algorithm still functioned properly.

A 19 year-old female had a -20.9% error when comparing the displacement and an LPC 10th order error waveform OQ. An example of the female's acoustic,

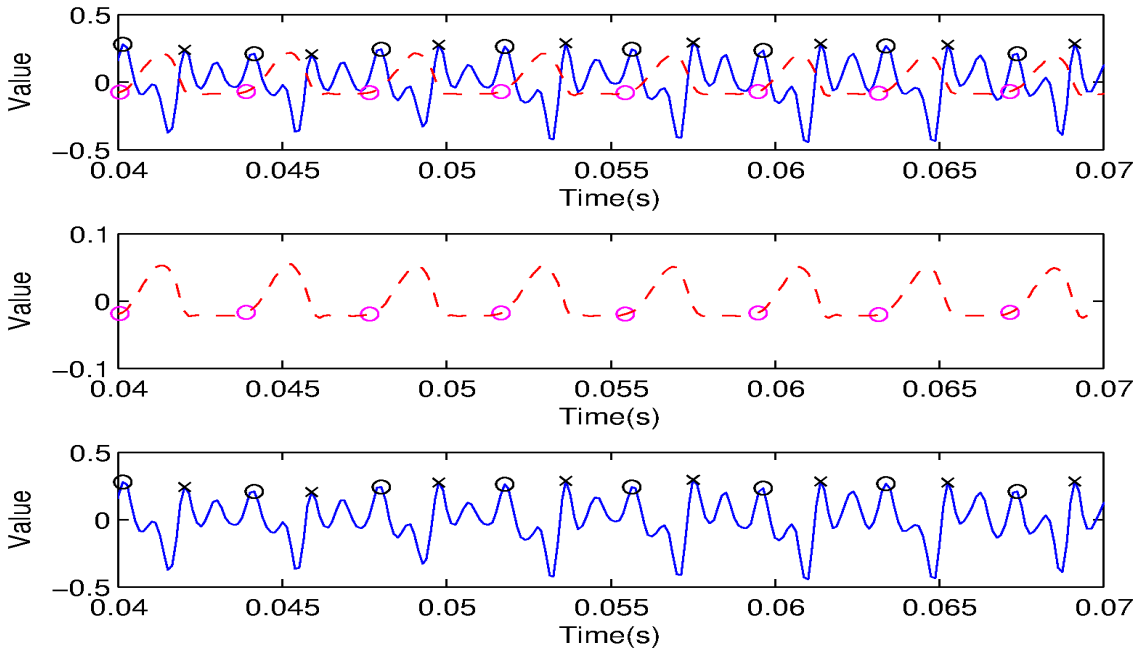


Figure 5.34: Linear Prediction Error Waveform With Peak Detection Result: Comparison of 7%OQ Threshold for the Area and OQ of Error Waveform for an 8 year-old Male Child With Error 0.39%. LPC 1st Order Followed by A 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period. Area Waveform (dashed red line) with 7% Threshold for Each Area Period (magenta circles).

glottal displacement and LPC 10th order error waveform is shown in Figures 5.39 and 5.40. This subject's acoustic recording was determined to have a fundamental frequency of 208Hz. It can be easily seen that the algorithm identified the glottal closure and opening instants correctly, but the error waveform itself and the glottal displacement waveform had a slight disagreement as to where the time-instants occurred. This disagreement directly affected the results of the open quotient and resulted in a slight error when compared to the displacement waveform's. After listening, the acoustic recording was perceived to be slightly muffled.

A 38 year-old male had a -1.67% error when comparing the displacement and an LPC 1st order, followed by a 10th order, error waveform OQ. An example of the male's acoustic, glottal displacement and LPC 1st order, followed by a 10th order, waveform is shown in Figures 5.41 and 5.42. This subject's acoustic recording was

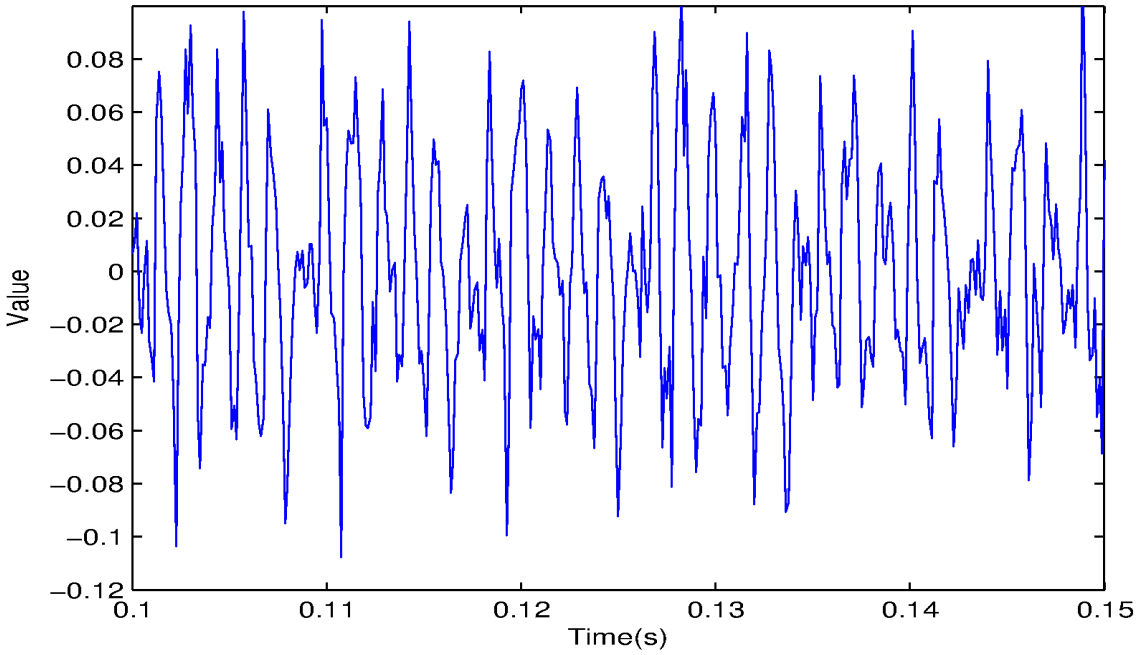


Figure 5.35: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for a 20 year-old Male Subject With Corresponding OQ Error of 49.3% Between Area and Linear Prediction Error Waveform.

determined to have a fundamental frequency of 250Hz. It can be easily seen that the algorithm identified the glottal closure and opening instants correctly, but the error waveform itself and the glottal displacement waveform were in close agreement as to where the time-instants occur. This agreement lead to the algorithm yielding a very small error of -1.67% when compared to the displacement waveform's open quotient. After listening, the acoustic recording was perceived to be muffled.

A 29 year-old female had a -17.7% error when comparing the displacement and an LPC 1st order, followed by a 10th order, error waveform OQ. An example of the female's acoustic, glottal displacement and LPC 1st order, followed by a 10th order, error waveform is shown in Figures 5.43 and 5.44. This subject's acoustic recording was determined to have a fundamental frequency of 216Hz. It can be easily seen that the algorithm identified the glottal closure and opening instants correctly, but the error waveform itself did not have the glottal opening instants appear in approximately the same locations for each period. These opening

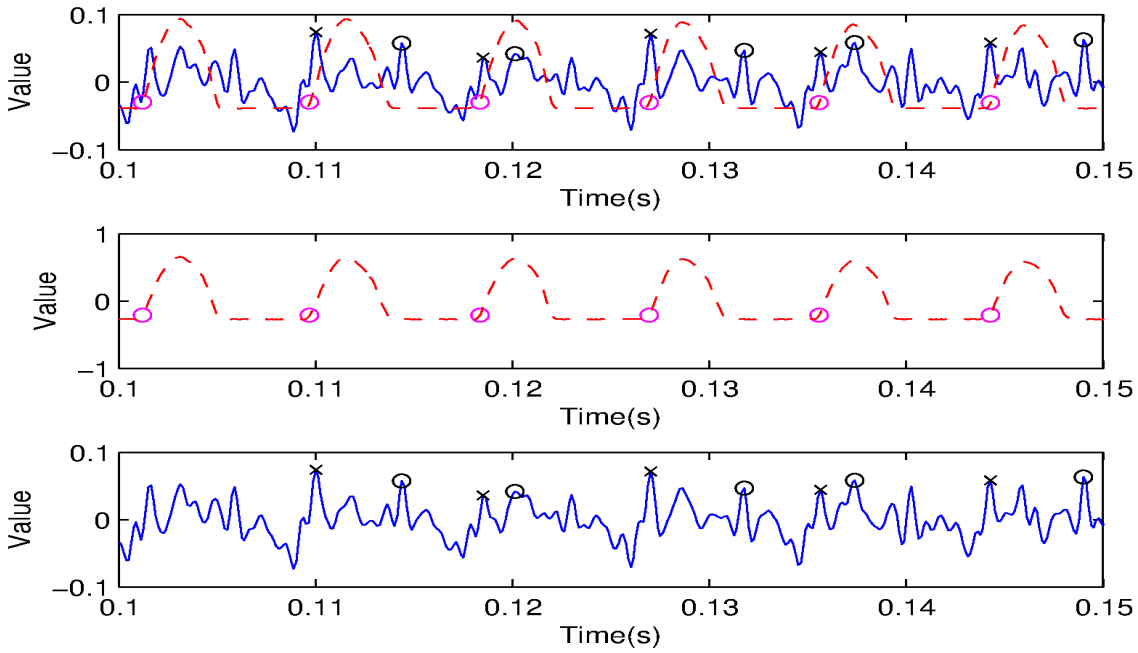


Figure 5.36: Linear Prediction Error Waveform With Peak Detection Result: Comparison of 7%OQ Threshold for the Area and OQ of Error Waveform for a 20 year-old Male With Error 43.9%. LPC 1st Order Followed by a 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period. Area Waveform (dashed red line) with 7% Threshold for Each Area Period (magenta circles).

time-instant fluctuations had an effect on the open quotient calculation and created a slight error. After listening, the acoustic recording was perceived to be noisy, which is visually evident in the acoustic waveform.

The LPC error waveform analysis with peak detection algorithm was applied to 46 different subjects and the open quotient was estimated and compared to its corresponding area open quotient. Due to the limited supply of displacement waveform data, only 43 subjects were compared with their corresponding displacement waveform open quotient. To compare the results of all of the methods, the errors were separated as previously with the IAIF method. Any errors exceeding 20% in value of the actual open quotient would be considered as a result of the algorithm not properly identifying the peaks. Any errors less than 20% would be more of the result of slight estimation errors. The percent error mean and standard

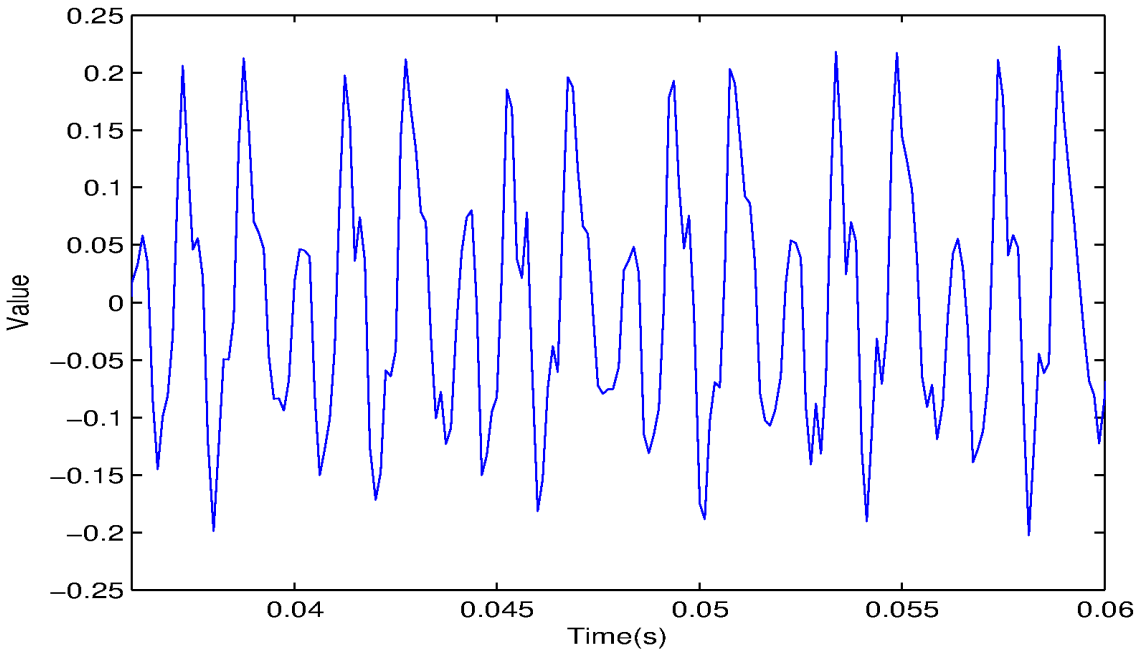


Figure 5.37: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform of a 42 year-old Female With Corresponding OQ Error of -1.62% Between Displacement and Linear Prediction Error Waveform.

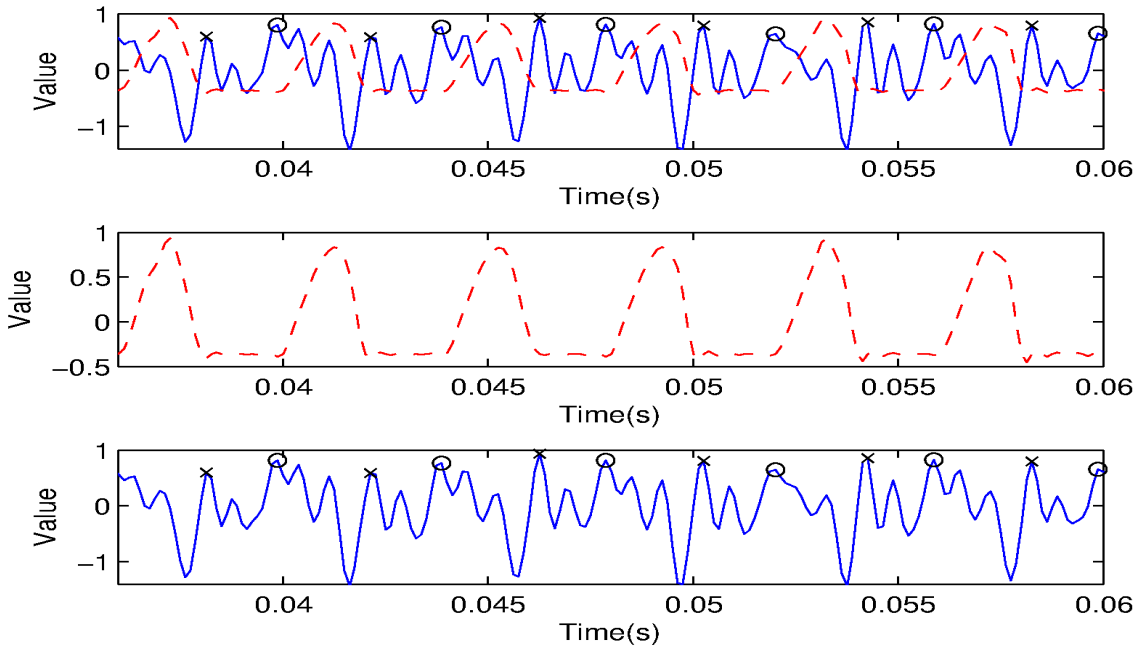


Figure 5.38: Linear Prediction Error Waveform With Peak Detection Result: Comparison of OQ for the Displacement and LPC Error Waveform for a 42 year-old Female With Error -1.62%. LPC 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period and Displacement Waveform (dashed red line).

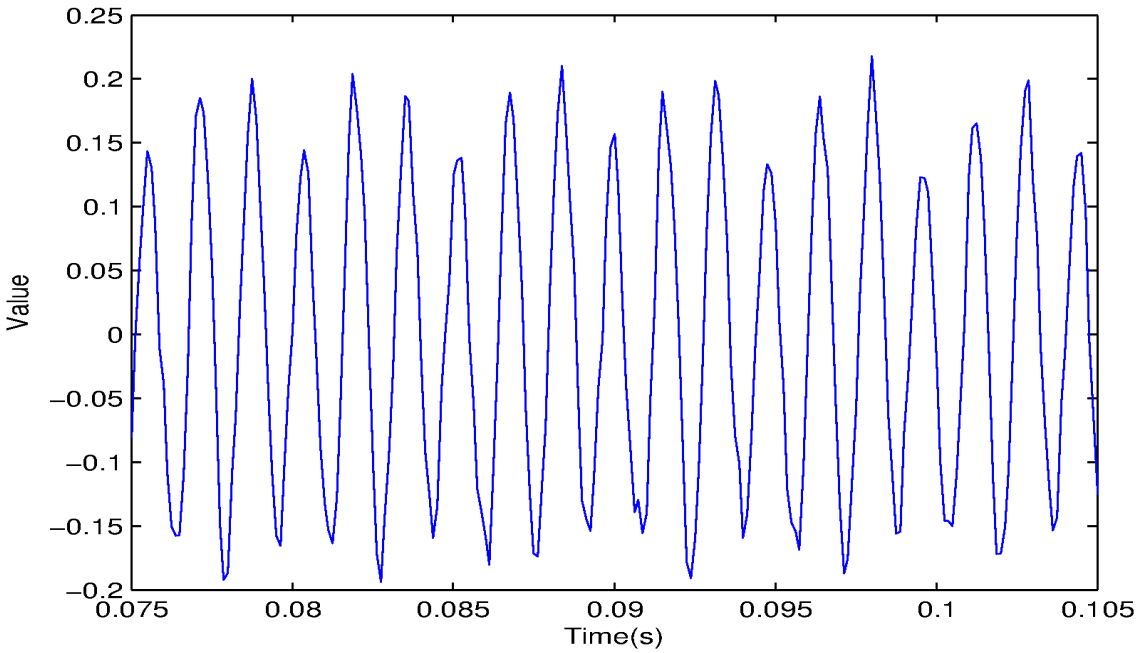


Figure 5.39: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for a 19 year-old Female With Corresponding OQ Percent Error of -20.9% Between Displacement and Linear Prediction Error Waveform.

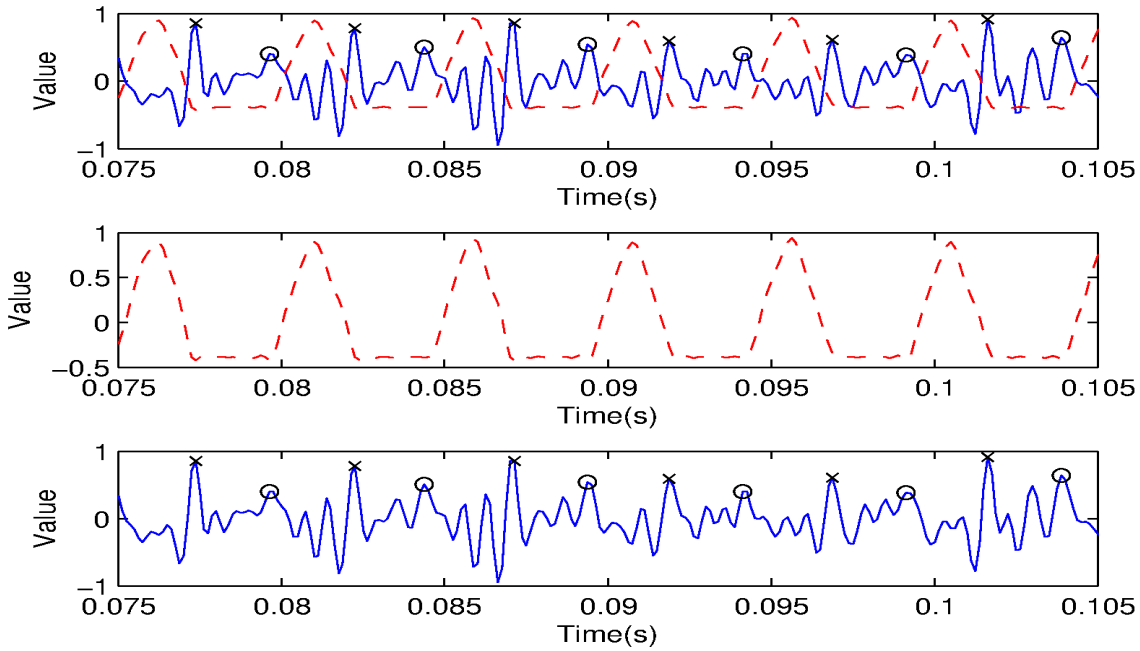


Figure 5.40: Linear Prediction Error Waveform With Peak Detection Result: Comparison of OQ for the Displacement and LPC Error Waveform for a 19 year-old Female With Error -20.9%. LPC 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period and Displacement Waveform (dashed red line).

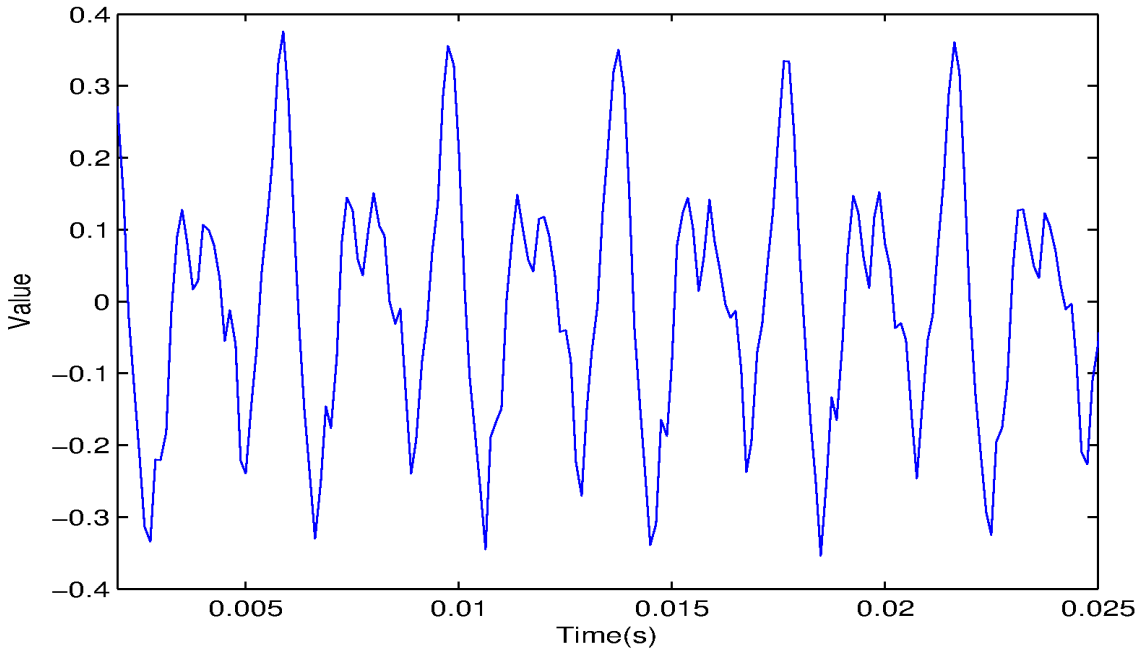


Figure 5.41: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for a 38 year-old Male With Corresponding OQ Percent Error of -1.67% Between Displacement and Linear Prediction Error Waveform.

deviation of the total data set and the non-anomalous set were calculated and are shown along with the percent anomalous in Table 5.3 for this method's open quotient estimation. This table was separated by comparison with the area and displacement as well as the error waveform for an LPC analysis of order 10 and an LPC analysis of order 1, followed by an order 10.

It can be easily seen from Table 5.3 that, for the entire data set, this algorithm was fairly accurate in detecting the glottal opening and closure instants. The small means overall along with fairly small standard deviations indicate that this algorithm is more consistent throughout the comparisons (area and displacement) and approaches (LPC 10 and LPC 1-10). It is also consistent that the displacement waveform has smaller standard deviation for the total set and may be a more accurate method for comparison compared to the area waveform. The percent anomalous for the area was approximately 40% and 50% for the displacement. An overall observation of the means for the non-anomalous data is

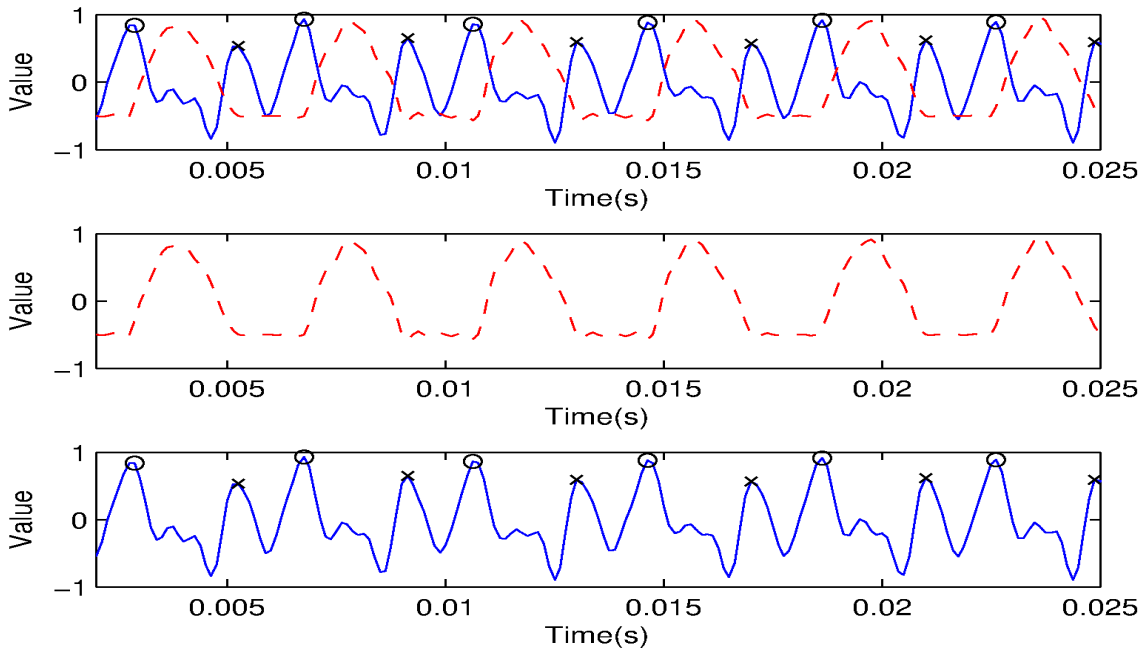


Figure 5.42: Linear Prediction Error Waveform With Peak Detection Result: Comparison of OQ for the Displacement and LPC Error Waveform for a 38 year-old Male With Error -1.67%. LPC 1st Order Followed by a 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period and Displacement Waveform (dashed red line).

that this error waveform analysis method usually underestimates the open quotient, however with much more consistency since the standard deviation of the non-anomalous error is smaller than the total set and percent anomalous.

## Overall Discussion

Table 5.4 shows the results for all the methods for easier comparison. The first method, Iterative Adaptive Inverse Filtering, doesn't seem to perform as well in comparison to the area at the threshold levels of 20% and 50%. However, this method does seem to perform well in comparison to the displacement waveform. The small error of 8.90% for the total set and the smaller error of 1.39% for the non-anomalous data, which makes up about 70% of the total data, dictates that this algorithm works better than compared to the area waveform comparison due to its

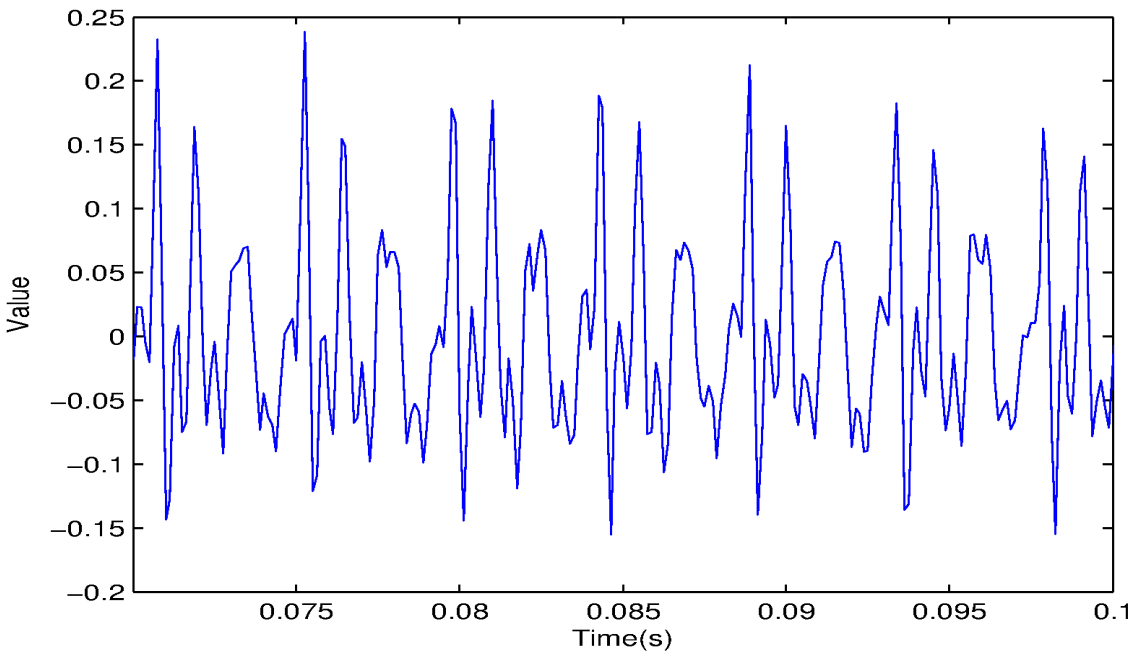


Figure 5.43: Linear Prediction Error Waveform With Peak Detection Result: Acoustic Waveform for a 29 year-old Female With Corresponding OQ Percent Error of -17.7% Between Displacement and Linear Prediction Error Waveform.

Table 5.3: Percent Error Mean (M) and Standard Deviation (SD) for the Total Set and Non-Anomalous (NA) Data along with Percent Anomalous (PA) For The LPC Error Waveform Analysis With Peak Detection Open Quotient Estimation Separated By Area with 7% OQ Threshold (Area 7%), Displacement (Disp), an 10th Order LPC Error Waveform (LPC 10) and a 1st Order, then 10th Order LPC Error Waveform (LPC 1-10).

| | Tot. M | Tot. SD | PA | NA M | NA SD |
|---|---|---|---|---|---|
| Area 7% LPC 1-10 | -0.69% | 23.10% | 43.48% | -2.44% | 10.13% |
| Area 7% LPC 10 | -1.43% | 26.34% | 41.30% | -2.46% | 10.72% |
| Disp LPC 1-10 | -13.74% | 21.39% | 51.16% | -5.77% | 13.05% |
| Disp LPC 10 | -13.76% | 23.97% | 58.14% | -6.90% | 9.20% |

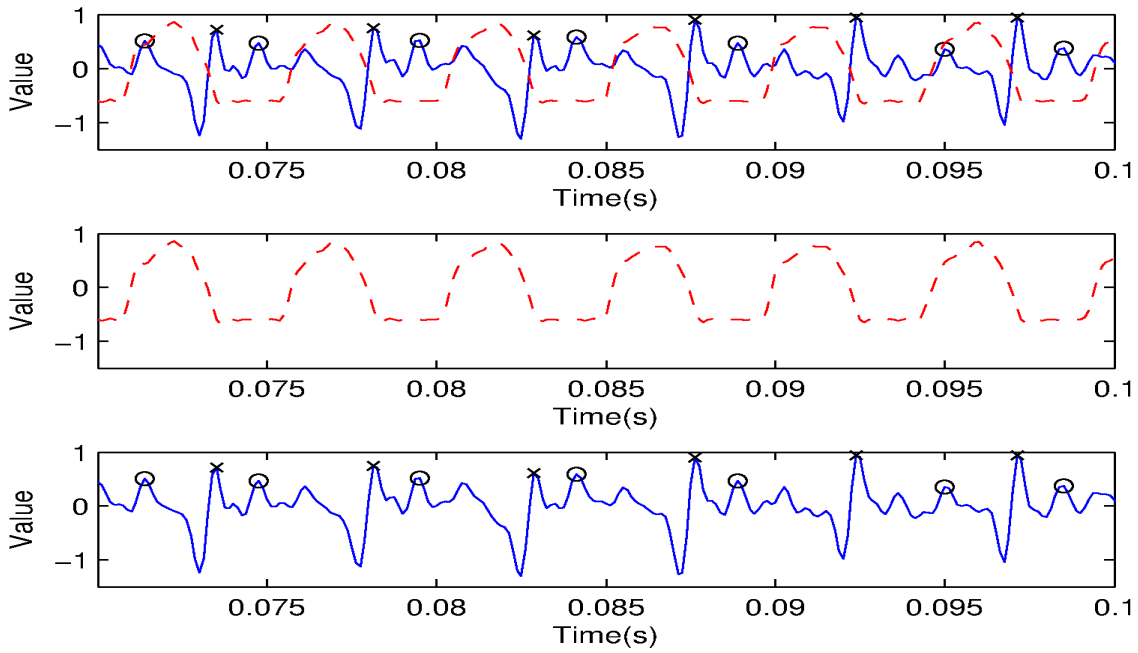


Figure 5.44: Linear Prediction Error Waveform With Peak Detection Result: Comparison of OQ for the Displacement and LPC Error Waveform for a 29 year-old Female With Error -17.7%. LPC 1st Order Followed By a 10th Order Error Waveform (solid blue line) with Indicated Glottal Closure Instants (black x's) and Glottal Opening Instants (black circles) for Each Glottal Source Period and Displacement Waveform (dashed red line).

robustness. It is easily seen that the OQ estimation using linear prediction with glottal source modeling method did not seem to work very well at all. It severely underestimated the actual glottal source, given by the corresponding area and displacement waveforms, and most of the data fell in the percent anomalous category. Noting some of the cases that had small error, it appeared that if the glottal source waveforms were not smooth and somewhat noisy, then this algorithm would easily fail to calculate the open quotient correctly. Moreover, even after filtering the glottal source waveform, the open quotient calculation would still be affected by vocal tract components the inverse filtering failed to filter out. Even though the non-anomalous data's mean and standard deviation are very small, the non-anomalous data makes up such a small percentage of total set and therefore, would not be a good indication of how well this algorithm worked. The third and

newly proposed method, LPC error waveform analysis with peak detection, seems to work decently well in estimating the glottal source features in comparison to the area at 7% threshold and the displacement waveform. This algorithm having a mean error that is negative dictates that it underestimates the actual open quotient feature. This may have to do with the fact that the actual glottal closure or opening instant my not be directly on the peak of the largest error spike, but somewhere on the trajectory towards it. For example, this algorithm would first detect the most negative error spike and identify the next positive error peak as the glottal closure instant. However, the actual glottal closure instant might have occurred on the trajectory from the negative peak value to the positive peak value. This is a phenomenon that would need to be further researched in order to properly determine if it is accurate. Nonetheless, this method, compared to the first two methods, did have overall smaller standard deviations for the total set, meaning it was a more consistent method for the open quotient calculation.

The main focus of this thesis is the comparison of the acoustic feature extraction methods with the HSVI data, area and displacement. A notable visual observation of the area waveforms is that, in most cases, an definite glottal closure or opening instant was not explicitly defined. To account for this issue for comparison, the 20% and 50% peak-to-peak threshold values were determined for the glottal source and area. It was assumed that these glottal source threshold time-instants would be a good approximation and compare properly to the corresponding time-instants of the area. During observation, it was determined this time-instants may be somewhat ambiguous and were affected by notable characteristics of the glottal source estimation. For example, it was noted in the IAIF method that the large negative spike prior to closure would sometimes still be prevalent in the glottal source due to the algorithm not properly filtering out that vocal tract component. Because of the prevalence of this spike, the overall amplitude of the 20% and 50%

thresholds would be affected and therefore affect the corresponding 20% and 50% threshold time-instants. The area waveforms did not have these large negative spikes, but a comparison with a glottal source waveform estimate with a prevalence of large negative spikes would yield erroneous results. From this problem, it was determined that the medial-line displacement waveform is a good indication for comparison due to its explicitly defined glottal closure and opening instants. This displacement waveform comparison truly works under the assumption that the actual vocal folds are opening and closing in a zipper-like fashion. This displacement waveform tracks the left and right vocal fold total displacement from the vertical medial line of the glottis over the horizontal medial line of the glottis. Even under the zipper-like opening and closing assumption, the fact that the standard deviations for the methods were less for displacement than area indicated that the displacement waveform may be a slightly more robust comparison waveform.

It has been noted in the literature and in this thesis' research that all acoustically-extracted glottal feature methods are affected by the quality of the input speech signal. For example, in some cases, presented in the results chapter, too much noise in the speech waveform will affect each algorithm's ability to correctly estimate the vocal tract components and therefore won't correctly inversely filter those components out of the signal. The noise may have been a result of the patient moving during the endoscope recording which may have been due to the fact that the patient was uncomfortable. Since the lapel mic was placed on the patient's clothes, fabric could have rubbed against the microphone, creating noise. Also, an endoscope in a patients vocal cavity during phonation may have affected the vocal quality and clarity of the recorded acoustic waveform due its slight obstruction of the cavity. This slight obstruction with the endoscope could have also affected the vocal tract estimation with LPC analysis, which assumes a open cylinder-like tract, but in this case would have an object in the tract. Another

limitation noted in the literature and this research was the fact that, since these features rely on knowing the GCI and GOI, a speech waveform that is produced by a glottis that does not fully close will introduce some ambiguity in the identification of the GCI and GOI. For example, a breathy speech waveform will not have as strong pressure differentials prior to closure and immediately after opening due to the fact that the glottis never completely closed. The estimation of the vocal tract component is based on knowing the resonating formants of the tract and could leave remnants of these formants in the estimated glottal source. This was seen in several cases where the negative pressure differential was left in the glottal source waveform and directly affected the 20% and 50% threshold values for comparison. Also, for the newly proposed method, if the open spike time-instants aren't as dramatic, the peak detection may have difficulty identifying the correct one. Other factors may have had an impact on the calculated errors. For example, quantization of the video and acoustic waveforms may introduce some disagreement in the calculation of time features. Further research needs to be completed in order to determine how much these limitations affect these particular methods.

It is interesting to understand how well these acoustically-extracted glottal features compare to the HSVI-extracted glottal features. Even though HSVI is an accurate method, acoustic methods are still popular. A main reason for this is the fact that the acoustic waveform can be recorded at a much higher sampling rate, in this case 100kHz, compared to the video, which is sampled at 4kHz. This is especially important for children who usually phonate at higher frequencies. For example, a child may phonate at 300 Hz and at 4kHz the video resolution is 13 samples per glottal cycle, while the 100kHz acoustic resolution is 330 samples per glottal cycle. The 100kHz acoustic waveform would be able to essentially extract more "information" per time segment than the corresponding video recording at 4kHz. The acoustic waveform basically has more time resolution and used in

Table 5.4: Percent Error Mean (M) and Standard Deviation (SD) for the Total Set and Non-Anomalous (NA) Data along with Percent Anomalous (PA), For All the Compared Methods' Open Quotient Estimation: For The IAIF-Estimated Glottal Source Waveform Separated By Area 20% OQ Threshold, Area 50% OQ Threshold and Displacement With Glottal Source 20% OQ Threshold. For The OQ Estimation With Glottal Modeling Glottal Source Waveform Separated By Area With NonFiltered Glottal Source (Area NF), Area With Filtered Glottal Source (Area F), Displacement With NonFiltered Glottal Source (Disp NF) and Displacement With Filtered Glottal Source (Disp F). For The LPC Error Waveform Analysis With Peak Detection Separated By Area with 7% OQ Threshold (Area 7%), Displacement (Disp), an 10th Order LPC Error Waveform (LPC 10) and a 1st Order, then 10th Order LPC Error Waveform (LPC 1-10)

| | Tot. M | Tot. SD | PA | NA M | NA SD |
|---|---|---|---|---|---|
| Area 20% | 49.01% | 31.57% | 84.78% | 8.46% | 11.07% |
| Area 50% | 37.14% | 34.47% | 71.74% | 5.66% | 11.36% |
| Disp GS 20% | 8.90% | 22.34% | 30.23% | 1.39% | 8.85% |
| Area NF | -63.14% | 28.39% | 94.87% | -9.89% | 4.09% |
| Area F | -45.71% | 40.50% | 85.00% | -6.28% | 8.19% |
| Disp NF | -66.51% | 29.11% | 100.00% | - | - |
| Disp F | -51.70% | 36.35% | 83.78% | -0.94% | 11.58% |
| Area 7% LPC 1-10 | -0.69% | 23.10% | 43.48% | -2.44% | 10.13% |
| Area 7% LPC 10 | -1.43% | 26.34% | 41.30% | -2.46% | 10.72% |
| Disp LPC 1-10 | -13.74% | 21.39% | 51.16% | -5.77% | 13.05% |
| Disp LPC 10 | -13.76% | 23.97% | 58.14% | -6.90% | 9.20% |

conjunction with the simultaneously recorded video data may yield better and more accurate results.

# Chapter Six: Conclusion

## Conclusion

A lot of research has been done in the field of speech production, particularly with regards to glottal vocal fold dynamics. The ability to understand the kinematics of the vocal folds in terms of glottal features, will aid in defining normal and abnormal characteristics, in which abnormal may relate to vocal fold pathological development or disorders. Previously, acoustic recordings were utilized and currently, since computational power has developed, high-speed video imaging is being focused on. An extensive analysis was performed for 46 subjects, male and female, adult and child, for this thesis' research on the accuracy of the acoustic-extraction methods for the glottal source features compared to the features of the simultaneously recorded high-speed video. Since the previous studies only performed analyses on a small number of subjects or synthetic speech, it was interesting to see the performance of the methods with regards to a wide range of ages of real clinically obtained data. Using the HSVI data as the accepted value, the error was computed to understand how accurate the acoustic-extraction methods were. The main goal of this thesis research is to understand the fundamental underlying structure and functionality of the vocal source. Since speech is utilized in communication, it impacts people on an everyday basis. The further we understand the basis of speech, the better we can understand how it is truly developed and used.

The results in the previous chapter dictate that some methods may be more accurate, consistent, or robust than others. The first method, IAIF, seemed to function better than the second, indicated by the total mean values. IAIF is based on determining some parameters, like the LPC model orders for the different stages, to properly model what the glottal source is for a particular input speech waveform. The data seemed to fit this IAIF model well, which is somewhat based on

parameters. Because this model represented the data well, it estimated the characteristic, open quotient, accurately compared the objective video comparison. This IAIF model also utilized multiple orders of LPC instead of one to estimate the vocal tract filter effects. It would be interesting to perform an experiment to see if using multiple LPC orders in stages has a direct advantage over making one LPC estimation with a higher order of components like the vocal tract filter.

The results from the second method, OQ estimation using linear prediction with glottal source modeling, yield that this method grossly underestimates the actual open quotient value. A strong reason for this gross underestimation of the actual glottal source open quotient dictated by the video displacement waveform is that the clinically obtained data did not fit this method's model well. Since OQ estimation using linear prediction with glottal source modeling is based on defining specific parameters, these parameters need to properly represent the characteristics of the data. If the parameters don't, then the data won't fit and the model won't estimate accurate features.

The third, and newly proposed method, LPC error waveform analysis with peak detection, had the smallest overall total mean and overall standard deviation of the three methods. This method may be more robust than the previous two especially since the GCI and GOI can be more efficiently identified with the error waveform and therefore will be easier to compare with the corresponding displacement waveform's GCI and GOI. This method is not based on some parameters that define a model for the calculation of characteristics for the data. The open quotient is calculated one period at a time from a peak detection algorithm. However, since this method is not based on a model, it is more susceptible to fluctuations in the data and an improper peak identification could have an impact on the open quotient estimation, since it is an average.

An interesting conclusion that can be made from this research is that the

displacement waveform, which is based on glottal edge tracking, may be a much more non-ambiguious approach to for comparison of the glottal source open quotient estimation. Utilizing the area, while considered an appropriate comparison because considered an average, may not be consistent enough. This is evident in the fact that a threshold time-instant needed to be used to properly compare glottal source estimates and area waveforms. Even though the area is considered an average of the glottal fold dynamics, the medial-line glottal displacement can also be considered an average, and in most cases, has explicitly defined GCI's and GOI's which are key to correctly estimating glottal time features. Because of the area's ambiguity of the threshold level time-instants, it should be noted that a displacement waveform should be a more consistent and accurate comparison measure. From this research, it appears that the displacement waveform should be a standard comparison.

Future work that can be completed as a result of this research is that the error percent may be decreased by tweaking the algorithm for the linear prediction error waveform with peak detection method. The error waveform analysis algorithm can be made more robust by adding constraints and writing a better peak detection algorithm to better realize the peaks. The current algorithm uses Matlab's *findpeaks* command and therefore is limited to its capability so tweaking the algorithm to better reflect this purpose may yield better results. Another area that should be researched is related to the fact that since the endoscope is placed in the mouth during recording, it may affect the vocal tract estimation from the acoustic waveform. An experiment needs to be performed to see the difference between the estimation of the vocal tract would be different with the endoscope placed in and out of throat. If the vocal tract estimate is different, it would be interesting to see if acoustically estimated features such as open quotient change from algorithm's calculations.

It should be understood that even though HSVI has been developed recently,

acoustic methods should not be ignored, mainly for the reason that acoustic waveforms can be sampled at a frequency of up to 100kHz which is 25 times higher than the video frame rate of 4kHz. The highest vibrating vocal folds, seen in children, phonate at 300Hz to 400Hz and are limited in how many samples or how much information we can extract every second. Because of this, upsampling may need to be used as a way to "fill in the gaps". But, because the acoustic recording is sampled at 100kHz, a lot more information is extracted per second and may help understand even more intricately how speech production occurs. Utilizing the acoustic waveform in conjunction with the video data may aid in the fundamental understanding of speech production and lead to proper characterization of these glottal source features. These features will then lead to understanding underlying functionality of speech production and may lead to advancements in clinical applications.

# Bibliography


[1] Alku, P. "Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications." *Indian Academy of Sciences*, 36(5), 623–650, 2011.

[2] Alku, P., Tiitinen, H., Naatanen, R. "A method for generating natural-sounding speech stimuli for cognitive brain research." *Clinical Neurophysiology*, 110, 1329–1333, 1999.

[3] Quatieri, T.F. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2001.

[4] Titze, I.R. *Principles of Voice Production*. Prentice-Hall, Inc., 1994.

[5] Gavidia-Ceballos, L., Hansen, J.H.L. "Direct speech feature estimation using an iterative em algorithm for vocal fold pathology detection." *IEEE Transactions on Biomedical Engineering*, 43(4), 373–383, 1996.

[6] Drugman, T., Dubuisson, T., Dutoit, T. "On the mutual information between source and filter contributions for voice pathology detection." In "Interspeech Conference Papers: Interspeech 2009," 2009.

[7] O. Smith, III, J. "Recent developments in musical sound synthesis based on a physical model." In "Presentation Overheads: Stockholm Musical Acoustics Conference," 2003.

[8] Henrich, N., Doval, B., d'Alessandro, C. "Glottal open quotient estimation using linear prediction." In "In Proc. Intern. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications," pages 12–17. 1999.

[9] d'Alessandro, C., Sturmel, N. "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude." *Indian Academy of Sciences*, 36, 601–622, 2011.

[10] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P. "Hmm-based speech synthesis utilizing glottal inverse filtering." *IEEE Transactions on Acoustics Speech and Signal Processing*, 19(1), 153–165, 2011.

[11] Schroeter, J., Larar, J., Sondhi, M. "Speech parameter estimation using a vocal tract/cord model." In "IEEE Proceedings International Conference Acoustics Speech and Signal Processing," pages 308–311. 1987.

[12] Plumpe, M.D., Quatieri, T.F., Reynolds, D.A. "Modeling of the glottal flow derivative waveform with application to speaker identification." *IEEE Transactions on Speech and Audio Processing*, 7(5), 569–586, 1999.

[13] Strik, H., Jansen, J., Boves, L. "Comparing methods for automatic extraction of voice source parameters from continuous speech." *Proceedings International Conference Spoken Language Processing*, 1, 121–124, 1992.

[14] Alku, P. "Glottal wave analysis and pitch synchronous iterative adaptive inverse filtering." *Speech Communication*, 11, 109–118, 1992.

[15] Chen, G., Shue, Y.L., Kreiman, J., Alwan, A. "Estimating the voice source in noise." In "Interspeech Conference Papers: Interspeech 2012," 2012.

[16] Sapienza, C.M., Stathopoulos, E.T., Dromey, C. "Approximations of open quotient and speed quotient from glottal airflow and egg waveforms: Effects of measurement criteria and sound pressure level." *Journal of Voice*, 12(1), 31–43, 1998.

[17] Unnikrishnan, H., Donohue, K.D., Patel, R.R. "Analysis of high-speed digital phonoscopy pediatric images." volume 8207, page 82071Q. SPIE, 2012. doi:10.1117/12.916752.

## VITA

- Date and Place of Birth
  - August 9, 1988, Born in Louisville, KY, U.S.A.
- Educational Institutions
  - 2006-2010, B.S. in Electrical Engineering
    Western Kentucky University, Bowling Green, KY.
- Professional Positions Held
  - 2009-2010, Research Assistant, Applied Physics Institute
    Western Kentucky University.
  - Fall 2010, Teaching Assistant
    Department of Electrical Engineering
    Western Kentucky University.
  - Spring 2011, Teaching Assistant
    Department of Electrical and Computer Engineering
    University of Kentucky, Lexington, KY.
  - Summer 2011, EE Hardware Student, Power and Engine Group,
    Lexmark International, Inc, Lexington, KY.
  - 2011-2012, Teaching Assistant
    Department of Electrical and Computer Engineering
    University of Kentucky, Lexington, KY.
  - Summer 2012, EE Hardware Student, Power and Engine Group
    Lexmark International, Inc, Lexington, KY.
- Scholastic and Professional Honors
  - Tau Beta Pi
  - Eta Kappa Nu
- Professional Publications
  - Hamlet SM, Devore W. Design of a Programmable Logic Controller Trainer. Poster session presented at: WKU Engineering Day; 2010 May 12; Bowling Green, KY.
  - Hamlet SM, Simpson M, Morrison T, Berry A. Redesign of Electrical System for Remotely Controlled ATV Platform. Poster session presented at: 40 th Annual WKU Student Research Conference; 2010 Feb 27; Bowling Green, KY.

  - Hamlet SM, Simpson M, Morrison T, Berry A. Re-Engineering of a User-Controlled Robotic ATV Platform. Poster session presented at: 20 th Annual Argonne Symposium for Undergraduates; 2009 Nov 13; Lemont, IL.

- Typed name of student on final copy
  - Sean Michael Hamlet