UNIVERSITY OF CALIFORNIA

Los Angeles

The Voice Source in Speech Production: from Models to Applications

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

by

Gang Chen

2014

Abstract of the Dissertation

The Voice Source in Speech Production: from Models to Applications

by

Gang Chen

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2014

Professor Abeer Alwan, Chair

The voice source contains important lexical and non-lexical information. The non-lexical information can convey, for example, prosodic events, emotional status, as well as cues pertaining to the uniqueness of the speaker's voice. A better understanding, and eventually a better model of the voice source, would benefit various speech applications, such as speech recognition, speech synthesis, speaker identification, age/gender classification, as well as clinical assessments.

This dissertation has three main goals. The first is to better understand the voice source through analyzing images of the vocal folds using laryngeal high-speed videoendoscopy (HSV) recordings. A new automatic method is proposed to compactly summarize the overall spatial synchronization pattern of vocal fold vibration for the entire laryngeal area from HSV data. Additionally, a new measure is proposed to adequately capture perceptually-important variations in glottal area pulse shapes, which are extracted from HSV data.

The second goal is to study the acoustic consequence of a physiological vocal-fold vibration pattern—the glottal gap effect, and apply our findings to a gender classification task of children's voices. Voice source related measures are found to improve classification accuracy, especially for younger (10-15 year old) speakers.

The third goal is to propose new voice source models and evaluate them in different applications. In the first application, a new source model and a noise-robust automatic source estimation algorithm are proposed to estimate the voice source from speech signals. Results in both clean and noisy conditions show that the proposed model and algorithm are robust in accurately estimating the voice source signal. The second application is to use the proposed source model for vowel synthesis. Perceptual listening experiments show that the proposed model provides a better perceptual match to the target voice than do traditional models.

The dissertation of Gang Chen is approved.

Jody Kreiman

Paulo Tabuada

Gregory Pottie

Abeer Alwan, Committee Chair

University of California, Los Angeles

2014

To my family . . .

my mom, dad, and wife.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisor, Professor Abeer Alwan for her intellectual guidance and gracious support during my study in UCLA. Her understanding, patience, and wisdom made it easier for me to navigate past the many research obstacles I faced. My sincere gratitude also goes to my committee members, professors Gregory Pottie, Paulo Tabuada, and Jody Kreiman for their interest in my work.

I was fortunate and privileged to have been able to work on a interdisciplinary project, with the guidance from professors Jody Kreiman and Patricia Keating. I have greatly benefited from their knowledge on voice quality and linguistics. Special thanks must be given to Professor Jody Kreiman, who introduced me to the physiology of speech production and statistical analysis methods. Her constructive suggestions and comments, as well as careful revision and proofreading, have been a tremendous help in my paper writing.

I would also extend my gratitude to Dr. Yen-Liang Shue, who was instrumental in helping me start and advance my research. His enthusiasm and optimism encouraged me to explore the amazing world of voice source.

I also would like to thank collaborators in UCLA Linguistics and Head and Neck Surgery departments, professors Robin Samlan, Marc Garellek, Bruce R. Gerratt, Juergen Neubauer, Zhaoyan Zhang, and Dinesh Chhetri. My work would not have been possible without their tremendous effort on facilitating the collection of invaluable high-speed glottal imaging data. I am thankful to Michael Döllinger and the Department of Phoniatrics and Pediatric Audiology of the University Hospital, Erlangen, Germany, for generously supplying the GlotAnTools software for high-speed image segmentation.

I am very thankful to my family–father, mother, and my wife. They have always been there by my side throughout my PhD study. Without their love,

support and encouragement, this dissertation would not have been possible.

Thanks go my SPAPL labmates, Dr. Shizhen Wang, Dr. Wei Chu, Dr. Jonas Borgstrom, Lee Ngee, Harish, Jom, and many others for their help and friendships.

This research was supported in part by NSF Grant No. IIS-1018863 and by NIH/NIDCD Grant Nos. DC01797 and DC011300. Parts of this dissertation have appeared in the publications listed under “Publications and Presentations”.

Vita

2004-2008	B.S. with Honors (Electronics Engineering) Tsinghua University, Beijing, China
2006	Exchange Student Department of Electrical and Electronic Engineering University of Hong Kong, Hong Kong
2007	Summer Intern EMC, Beijing/Hong Kong
2008-2010	M.S. (Electrical Engineering) University of California, Los Angeles
2008–present	Research Assistant Electrical Engineering Department, UCLA.
2010	Teaching Assistant Electrical Engineering Department, UCLA.
2010	Summer Intern 3M.Cogent, Pasadena, CA, USA
2011	Summer Intern Disney Research, Pittsburgh, PA, USA
2012	Summer Intern Starkey Lab, Eden Prairie, MN, USA
2013	Summer Intern Qualcomm, San Diego, CA, USA

Publications

G. Chen, J. Kreiman, and A. Alwan, "The glottaltopogram: a method of analyzing high-speed images of the vocal folds," Computer Speech and Language, 2013, in press, http://dx.doi.org/10.1016/j.csl.2013.11.006

G. Chen, J. Kreiman, B. R. Gerratt, J. Neubauer, Y.-L. Shue, and A. Alwan, "Development of a glottal area index that integrates glottal gap size and open quotient," Journal of the Acoustical Society of America, Vol. 133, Issue 3, March 2013, pp. 1656-1666.

J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. R. Gerratt, J. Neubauer, and A. Alwan, "Relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," Journal of the Acoustical Society of America, Volume 132, Issue 4, 2012, pp. 2625-2632.

G. Chen, M. Garellek, J. Kreiman, B. R. Gerratt, A. Alwan, "A perceptually and physiologically motivated voice source model," Interspeech 2013, pp. 2001-2005.

G. Chen, R. A. Samlan, J. Kreiman, A. Alwan, "Investigating the relationship between glottal area waveform shape and harmonic magnitudes through computational modeling and laryngeal high-speed videoendoscopy," Interspeech 2013, pp. 3216-3220.

G. Chen, Y.-L. Shue, J. Kreiman, and A. Alwan, "Estimating the voice source in noise", Interspeech 2012, pp. 1600-1603.

G. Chen, J. Kreiman, and A. Alwan, “The glottaltopograph: a method of analyzing high-speed images of the vocal folds,” ICASSP 2012, pp. 3985-3988.

G. Chen, J. Kreiman, Y.-L. Shue, and A. Alwan, “Acoustic Correlates of Glottal Gaps,” Interspeech 2011, pp. 2673-2676.

Y.-L. Shue, G. Chen, and A. Alwan, “On the Interdependencies between Voice Quality, Glottal Gaps, and Voice-Source related Acoustic Measures,” Interspeech 2010, pp. 34-37.

G. Chen, X. Feng, Y.-L. Shue, and A. Alwan, “On Using Voice Source Measures in Automatic Gender Classification of Children’s Speech,” Interspeech 2010, pp. 673-676.

CHAPTER 1

Introduction

1.1 Overview and motivation

The human speech production system allows a speaker to produce a vast range of sounds. The system consists of many organs intervening in the phonation process, which can be generally divided into three parts: (1) the system below the larynx (subglottal system), (2) the larynx and its surrounding structures, and (3) the structures and the airways above the larynx (supraglottal system). These three components of the speech production system are illustrated in Figure 1.1. From a physiological point of view, the subglottal system provides energy in the airflow during the production of most sounds. The airflow is then pushed through a constricted region called the glottis, which is located within the larynx. The glottis physically divides the subglottal and supraglottal systems. The larynx, a structure made of cartilage and muscle, is where the airflow is modulated. Within the larynx, the vocal folds (also known as the vocal cords) control the amount of airflow that passes through by vibration, which converts airflow to acoustic energy [DAA14]. The glottal modulation provides either a periodic or a noisy excitation source signal to the vocal tract. The supraglottal system (the vocal tract) consists of the oral, nasal, and pharyngeal resonant cavities. Above the larynx, the pharynx forms the vertical portion of the vocal tract system. The oral cavity follows the pharynx, with the lip forming the anterior end of the vocal tract system. The vocal tract further shapes the spectrum of the airflow signal, and the airflow is then radiated by the lips. The variation of the sound pressure travels through media and is perceived by the listener.

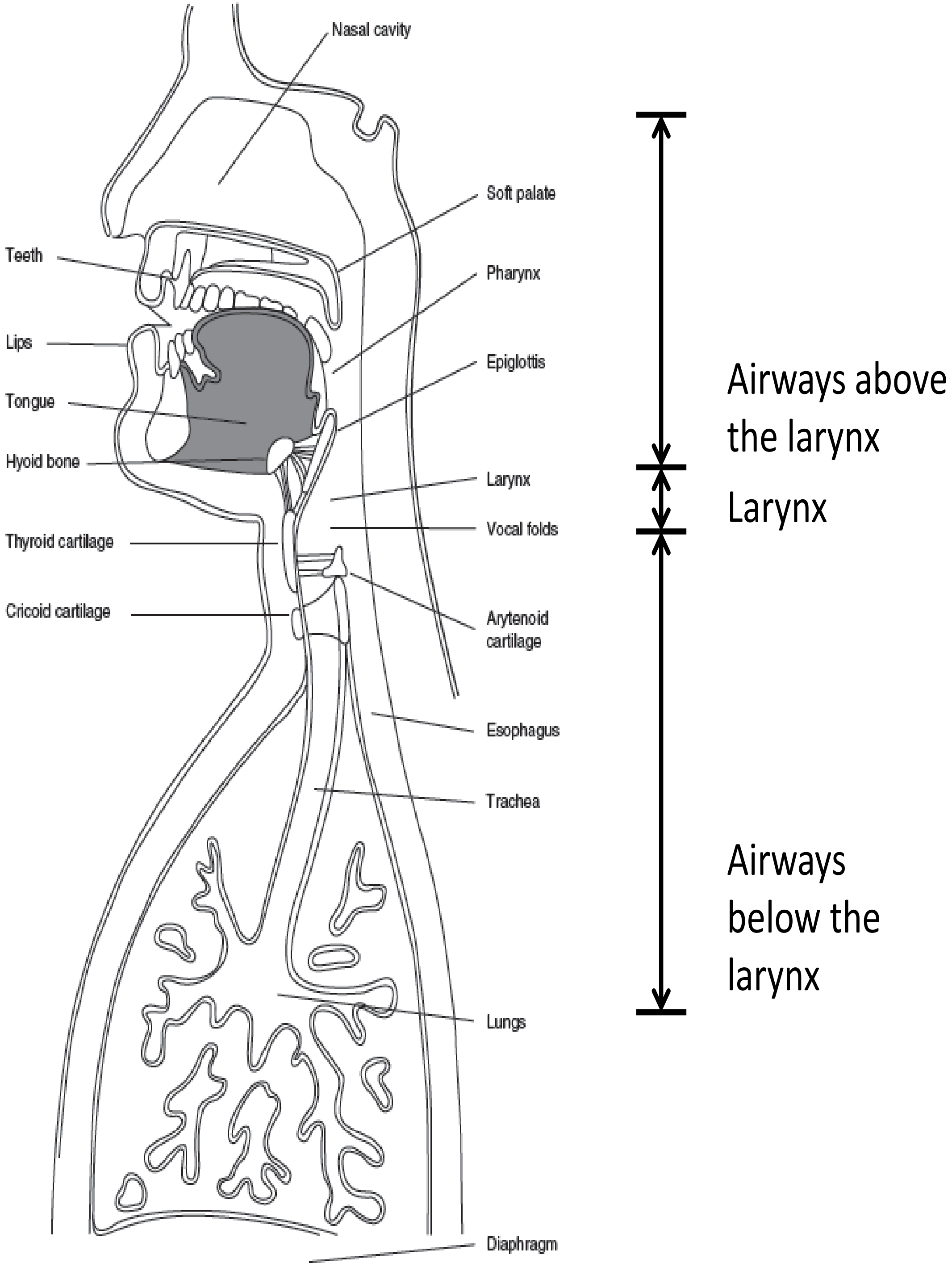


Figure 1.1: Schematic representation of the three components of the speech production system (from [KS11])

In voiced speech, the oscillation of the vocal folds periodically interrupts the airflow from the lungs and creates changes in air pressure [Ber58]. Thus, the glottal airflow (air volume velocity) is converted into a train of flow pulses which is referred to as the “voice source” excitation signal [DAA14]. When no sound is being phonated, the vocal folds are usually open. To produce unvoiced sounds, the vocal folds are held apart, allowing the airflow to pass through the glottis. A noise excitation signal is generated due to flow turbulence. To produce voiced sounds, the adductor muscles bring the vocal folds together and provide resistance to the air pressure from the subglottal system. The air pressure builds up below the closed vocal folds and then forces the vocal folds to open, allowing airflow to pass through the glottis [DAA14]. The two factors that contribute to the closing of the glottis are (1) elasticity of the tissue, which forces the vocal folds to regain their original configuration near the midline, and (2) aerodynamic forces [KS11]. One such force is described by the Bernoulli principle, under which the pressure near the edges of the vocal folds reduces because air travels faster near the edges of the vocal folds than it does at the midline [KS11]. Another aerodynamic force occurs when vortices form in the airflow as it exits the glottis. Vortices along the superiormedial surface of the folds create an additional negative pressure between the vocal folds, further contributing to the closing of the glottis [McG88, Zha08]. Once the vocal folds are closed, the air pressure below them builds up again, and the vocal folds are blown open once again. This cycle is repeated many times during one second and the cycle duration is called the “fundamental period” (T_0). Its frequency is referred to as the “fundamental frequency” (F_0). Figure 1.2 shows consecutive high-speed images of a complete glottal cycle. In this dissertation, the discussion focuses on the voice source characteristics during the production of voiced sounds.

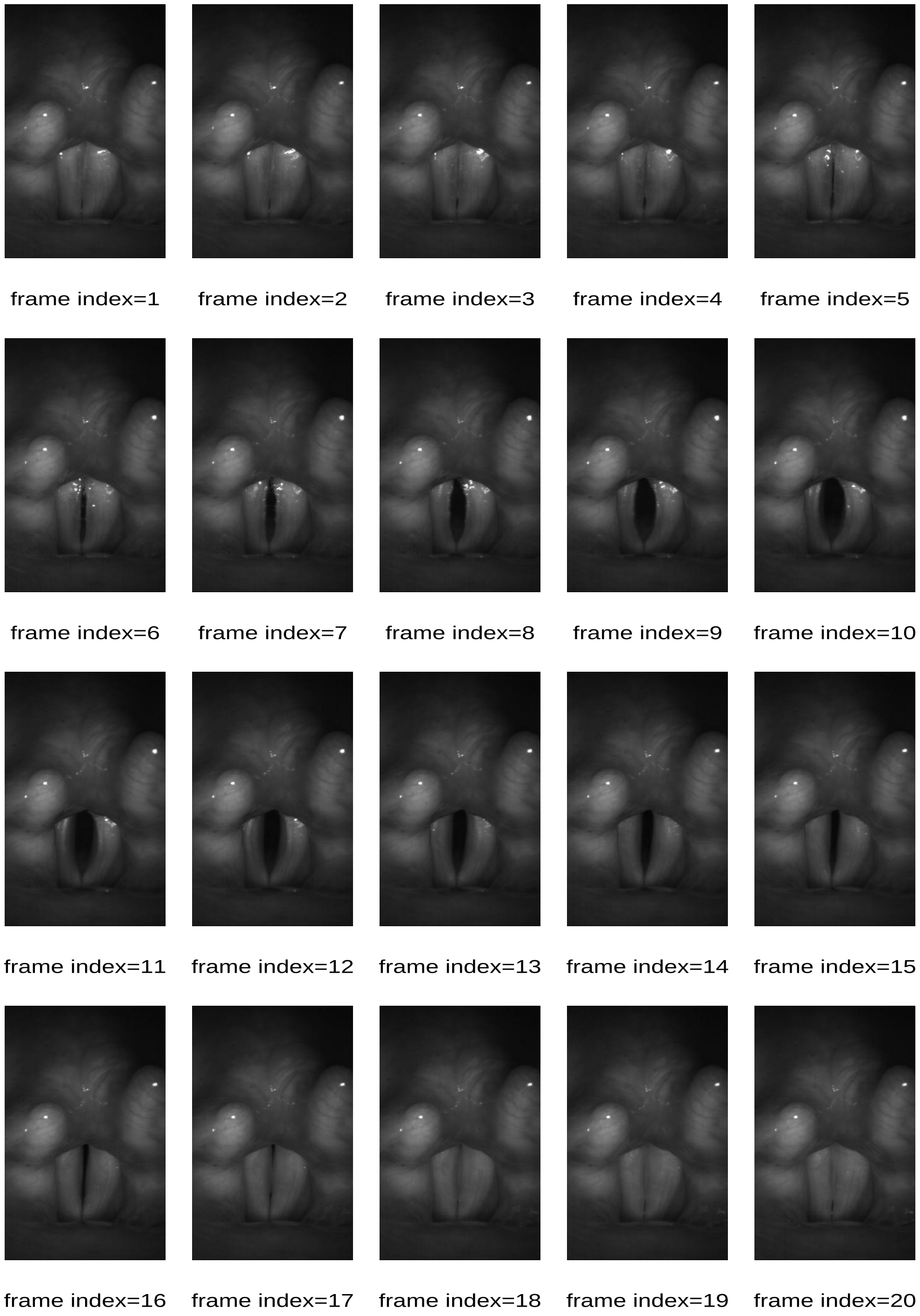


Figure 1.2: High-speed images of a female speaker's modal voice showing a complete glottal cycle (closed-open-closed). The plots are sequential from left to right and top to bottom, according to cycle index numbers. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

The subglottal system provides energy in the airflow, and the laryngeal and supraglottal systems modulate the airflow to produce audible sounds. Changes in the vocal fold vibration pattern, as well as changes in the configuration of the vocal tract above the larynx (the tongue, jaw, soft palate, and lips), will change the sound produced. The voice source, manipulated by vocal fold vibrations, controls “voice quality”, which is the perceptual characteristic of a speaker’s voice. It can differ from one speaker to another, and vary within a speaker from occasion to occasion [KS11]. Section 1.3 provides a review of the voice qualities studied in this dissertation.

The voice source contains important lexical and non-lexical information. The non-lexical information can convey, for example, prosodic events, emotional status, as well as cues pertaining to the uniqueness of the speaker’s voice. The study of the voice source would improve our knowledge of how the voice is generated physiologically and how it affects the resultant acoustic characteristics of human speech, as well as how these acoustic aspects are perceived by listeners. A better understanding, and eventually a better model of the voice source, would benefit various speech applications, such as speech recognition, speech synthesis, speaker identification, age/gender classification, as well as clinical assessments.

This dissertation has three main goals.

- The first is to better understand the voice source through analyzing images of the vocal folds using laryngeal high-speed videoendoscopy (HSV) recordings. A new automatic method is proposed to compactly summarize the overall spatial synchronization pattern of vocal fold vibration for the entire laryngeal area from HSV data. Additionally, a new measure is proposed to adequately capture perceptually-important variations in glottal area pulse shapes, which are extracted from HSV data.

- The second goal is to study the acoustic consequence of a physiological vocal-fold vibration pattern—the glottal gap effect—and apply our findings to a gender classification task of children's voices. Voice source related measures are found to improve classification accuracy, especially for younger (10-15 year old) speakers.

- The third goal is to propose new voice source models and evaluate them in different applications. In the first application, a new source model and a noise-robust automatic source estimation algorithm are proposed to estimate the voice source from speech signals. Results in both clean and noisy conditions show that the novel approach is robust in accurately estimating the voice source signal. The second application is to use the proposed source model for vowel synthesis. Perceptual listening experiments show that the proposed model provides a better perceptual match to the target voice than do traditional models.

1.2 The linear speech production model

Speech production is a highly complicated non-linear time-variant process. However, during a short time period (e.g., 10–20 ms), the system can be approximated as a cascade of linear systems involving a source function (voice source), a transfer function (representing the vocal tract), and a differentiator (simulating lip radiation effects). This theory is known as the linear source-filter model of speech production [Fan70], as shown in Figure 1.3. In a signal processing point of view, the speech signal, voice source, vocal tract, and lip radiation are denoted as $s(t)$, $u(t)$, $v(t)$, and $r(t)$. Then the system can be expressed in the time domain as:

$$s(t) = u(t) * v(t) * r(t) \tag{1.1}$$

and in the frequency domain as:

$$S(w) = U(w) \cdot V(w) \cdot R(w) \tag{1.2}$$

The lip radiation process can be simulated by a derivative operation, and it is common to move the derivative operator to the voice source signal (Figure 1.3, bottom panels). Therefore, the *glottal flow derivative* is commonly referred to as the "excitation to the voice tract" [Fan70]. Note that source-filter interactions and nonlinearity effects have been observed and studied in [Tit08b, TRP08], which are not represented in this model. However, this linear source-filter model has been extensively used for over four decades, and is still the basis in numerous speech research areas.

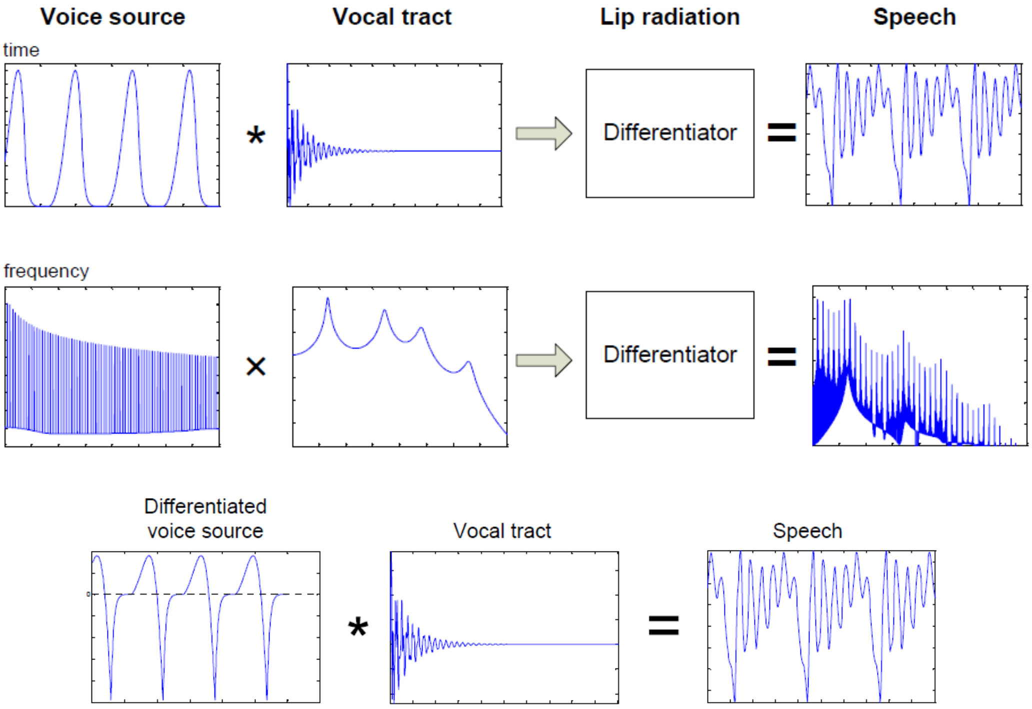


Figure 1.3: The linear source-filter model of speech production [Fan70]. Top panels: the model in the time domain. Middle panels: the model in the frequency domain. Bottom panels: the model with lip radiation effect integrated into the source (from [Shu10a]).

1.2.1 The voice source

In order to accurately represent the voice source signal, many models have been proposed. Voice source models can be generally grouped into two categories: interactive models which explicitly define the effects of interaction between the glottal source and the vocal tract, and non-interactive models which are based on the linear source-filter theory, assuming no interaction between the glottal source and vocal tract. Interactive models usually involve aerodynamic and mechanical theories such as coupling effects of the voice production system, aiming at explaining how the voice source signal is generated from a physiological point of view. Non-interactive models are commonly used in speech processing applications in the form of parametric signal models, in which source signal variation can be characterized by a few parameters. Even with simplified assumptions, non-interactive models have been shown to be effective in various applications such as speech coding and speech synthesis. Since the aerodynamic coupling effects in the interactive models are far from well understood, the discussion is limited to non-interactive models in this dissertation.

1.2.1.1 Traditional voice source models

Many non-interactive models have been proposed with varying level of complexity, such as the Liljencrants-Fant (denoted as LF) [FLL85], Rosenberg (denoted as Ros) [Ros71], Rosenberg++ (denoted as R++) [Vel98], Fujisaki-Ljungqvist (denoted as FL) [FL86] models, and the model proposed by Shue and Alwan (denoted as SA) [SA10] (see [CC95, DAA14] for review). With three parameters, the Ros model has two separate trigonometric functions for the opening and closing phases to represent the glottal flow volume velocity. Because the effects of lip radiation can be modeled as a first-derivative filter, modeling the glottal source can also incorporate the radiation effect. This results in a model of the first derivative of glottal flow pulse rather than the glottal volume velocity pulse. Flow derivative models include the LF, FL, and R++ models. The four-parameter LF model [FLL85] uses a combination of sinusoidal and exponential functions, and is commonly used in speech synthesis. The definition of the LF model is shown in Equation 1.3. An example of the LF model is shown in Figure 1.4. With six parameters and polynomial functions, the FL model provides greater detail in modeling the glottal pulse shape, but the increased number of parameters also makes it more difficult to use in practice. The R++ model in [Vel98] is computationally more efficient but perceptually equivalent when compared to the LF model. The SA model uses a combination of sinusoidal and exponential functions similar to the LF model, but with the ability to adjust the slopes of the opening and closing phases separately (see Section 1.2.1.2 for more discussion).

$$u(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t), & 0 \leq t \leq t_e \\ (\frac{-E_e}{\epsilon T_a})[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}] & t_e < t \leq t_c \end{cases} \tag{1.3}$$

These non-interactive voice source models differ in waveform factors such as onset, offset, and slope, but the gross shapes are generally similar to the waveform in Figure 1.4. All of the models reviewed in this section are suitable for applications such as speech coding, speech synthesis, and speech analysis in which the voice source and the vocal tract are assumed to be linearly separable under the linear speech production model.

1.2.1.2 The SA model

Based on the HSV recording of the vocal folds in [Shu10a], it was reported that a modification of the LF model may be necessary for accurately modeling the observed vibration of the vocal folds. For example, it was noticed that in many cases the opening phase duration was shorter than the closing phase duration, which

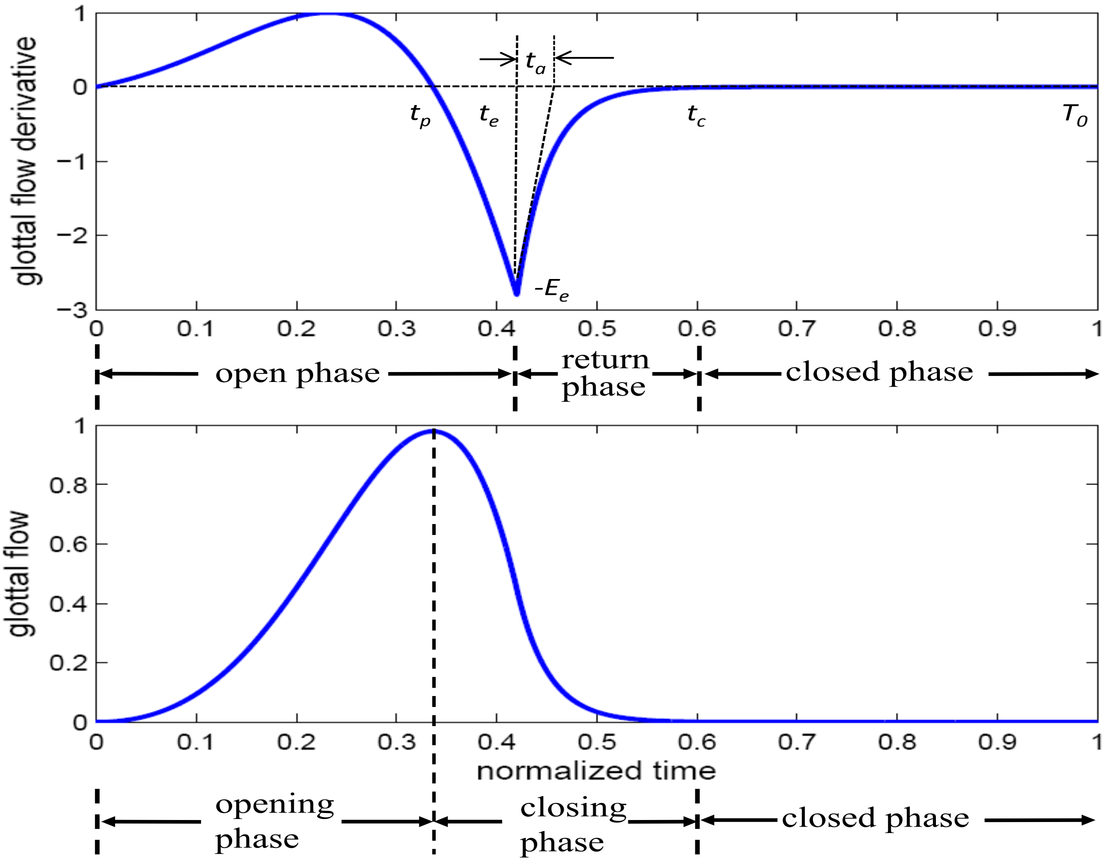


Figure 1.4: The LF model. Top panel illustrates the glottal flow derivative: instant of maximum airflow (t_p), instant of maximum airflow derivative (t_e), effective duration of return phase (t_a), beginning of closed phase (t_c), fundamental period T_0, and amplitude of maximum excitation of glottal flow derivative (E_e). Bottom panel illustrates the glottal flow model.

is not accounted for in the LF model. In addition, both the opening and closing phases can occur very quickly for some phonations (e.g., pressed voices), and this flexibility is beyond what can be specified by the LF model. In order to account for the source variabilities observed from HSV data described in [Shu10a], the SA model was proposed based on a combination of an exponential function with a sine function [SA10], which is similar to the first equation in the LF model definition in Equation 1.3. An example of the SA model is shown in Figure 1.5, represent-

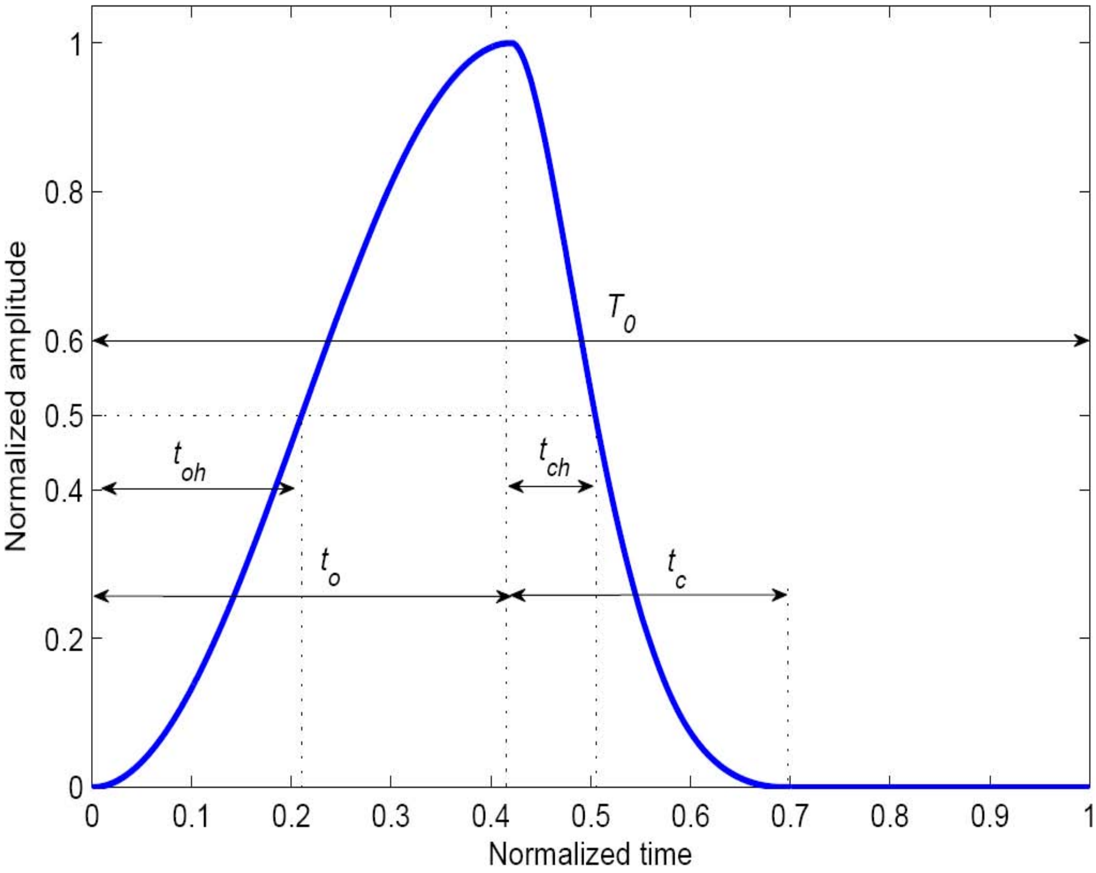


Figure 1.5: The SA model with $OQ = 0.7$, $\alpha = 0.6$, $S_{op} = 0.5$, and $S_{cp} = 0.7$. Note that the SA model defines the glottal flow waveform.

ing one cycle of the **glottal flow**, not the **flow derivative**. The definition of the SA model is shown in Equation 1.4. The SA model consists of 4 parameters: open quotient (OQ), asymmetry coefficient (α), speed of opening phase (S_{op}), and speed of closing phase (S_{cp}). Using the notation from Figure 1.5, T_0 denotes the fundamental period, $OQ = \frac{t_o+t_c}{T_0}$, $\alpha = \frac{t_o}{t_o+t_c}$, $S_{op} = \frac{t_{oh}}{t_o}$, and $S_{cp} = 1 - \frac{t_c h}{t_c}$, where t_{ch} and t_{oh} are at 50% of the maximum amplitude for the opening and the closing phases. The four parameters all range from 0 to 1.

$$u(t) = \begin{cases} f(\beta_o t, \lambda_{S_{op}}), & 0 \leq t \leq t_o \\ f(\beta_c(t_o + t_c - t), \lambda_{S_{cp}}), & t_o < t \leq t_o + t_c \\ 0, & t_o + t_c < t \leq T_0 \end{cases} \tag{1.4}$$

where $A(\lambda^*) = \frac{1}{\pi(e^{\lambda^*}+1)}$ and

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}} \left| \frac{e^{\lambda_s}(\lambda \sin(\pi s) - \pi \cos(\pi s))}{\pi(e^{\lambda} + 1)} + \frac{1}{e^{\lambda} + 1} - \frac{1}{2} \right| \tag{1.5}$$

The SA model provides more flexibility in modeling pulse shape variation than does the LF model. Equation 1.4 allows for quicker transitions from the pulse onset to the pulse peak and also from the pulse peak to the pulse offset, as observed in the HSV recordings. However, as shown in Equation 1.5, the generation of the SA model involves the slope parameter λ, which needs to be calculated through an optimization step. Although simple optimization techniques such as the gradient descent algorithm can be used, it is still a time-consuming step in analysis applications such as model fitting, where the model has to be generated with numerous iterations.

1.2.2 The vocal tract

The vocal tract consists of many articulators such as the tongue, palate, jaw, and lips. The configuration of these articulators forms an acoustic tube where the voice source signal is further modulated. According to the linear source-filter theory, the vocal tract is a linear system represented by the vocal tract transfer function (VTTF). Because the vocal tract changes shape relatively slowly, the VTTF can be assumed to be time-invariant over time intervals on the order of 10 ms [RS07].

The VTTF usually contains poles and zeros and can be expressed as:

$$V(z) = \frac{b_0 \prod_{k=1}^{M} (1 - d_k z^{-1)}}{\prod_{k=1}^{N} (1 - c_k z^{-1)}} \tag{1.6}$$

where b_0 is the gain factor, c_k are the poles of $V(z)$, and d_k are the zeros of $V(z)$. For voiced sounds, the VTTF consists of resonances and anti-resonances. The poles characterize several peaks corresponding to resonances of the acoustic cavities that form the vocal tract. These resonances are measured by *formants.* Each formant is described by its formant frequency (resonance frequency) and its formant bandwidth (resonance bandwidth). For example, the first (lowest) formant frequency is referred to as F_1, and the first formant bandwidth, B_1. Changing the vocal tract configuration causes a shift in the formants to different frequencies. Generally, shifts in the frequencies of the lowest three or four formants are associated with changes in the vowel being spoken [KS11].

The zeros (anti-resonances) of the VTTF represent energy loss and are useful when modeling consonants. For modeling vowels, it is a common approach to include only poles in the model because the zeros are few and located at very high frequencies.

1.3 Voice quality

The perceptual characteristic of a voice is its quality. As stated in [KS11]: "Voice quality is one of the primary means by which speakers project their identity– their 'physical, psychological, and social characteristics' [Lav80] or their 'auditory face' [BFB04]–to the world." Changes in vocal fold vibration manner may result in perceptible changes in the voice. Modal voice often refers to the kind of phonation humans normally produce. For example, modal phonations include the range of fundamental frequencies normally used for speaking or singing—the mode of the fundamental frequency distribution for an individual (e.g., [Hol74]).

In contrast, “nonmodal” is used to describe phonations that differ from the most “usual” voices. Many kinds of phonations may contrast with modal voices. Breathy voices are produced when the vocal folds close gradually and less high-frequency energy is generated. Sometimes the glottis may not fully close at the end of a glottal cycle during vibrations, allowing unmodulated airflow to pass through the glottis and generating turbulence noise [KS11]. Pressed phonations are usually characterized by tense vocal folds and smaller glottal openings [KK90]. Whispered voices are produced when the vocal folds only partially vibrate (or do not vibrate at all), and acoustic energy is solely produced by the turbulence airflow that passes the partially-closed glottis [KS11].

Speakers can differ from one another in voice quality (*inter-speaker variability*), and an individual speaker’s voices can vary from occasion to occasion (*intra-speaker variability*). Although in a broad sense the physiology underlying voice quality involves respiratory, laryngeal, and vocal tract configurations, this dissertation focuses on the voice source contributions to quality.

1.3.1 Acoustic measures of the speech signal

According the linear speech production model in Equation 1.2, the speech spectrum $S(w)$ is the product of its source spectrum $U(w)$ and its vocal tract transfer function $V(w)$. Figure 1.6 shows the spectrum of a vowel /a/ and some spectral measures used in this dissertation. The periodicity of the voice source signal results in the harmonic peaks at integer multiples of F_0. H_1, H_2, and H_4 are the first, second, and fourth harmonic magnitudes, respectively. The vocal tract resonances result in peaks of the spectral envelope, such as F_1, F_2, and F_3 (the first three formant frequencies, in Hz). Spectral magnitudes at the formant frequencies are denoted as A_1, A_2, and A_3 (in dB). Corrections can be made to remove the VTTF influence so that the measures capture the characteristics of the voice source signal (e.g., [ISA07]). Corrected measures are usually denoted by an asterisk (*), such as $H_1^* - H_2^*$. Table 1.1 lists some of the voice source related measures used in this dissertation.

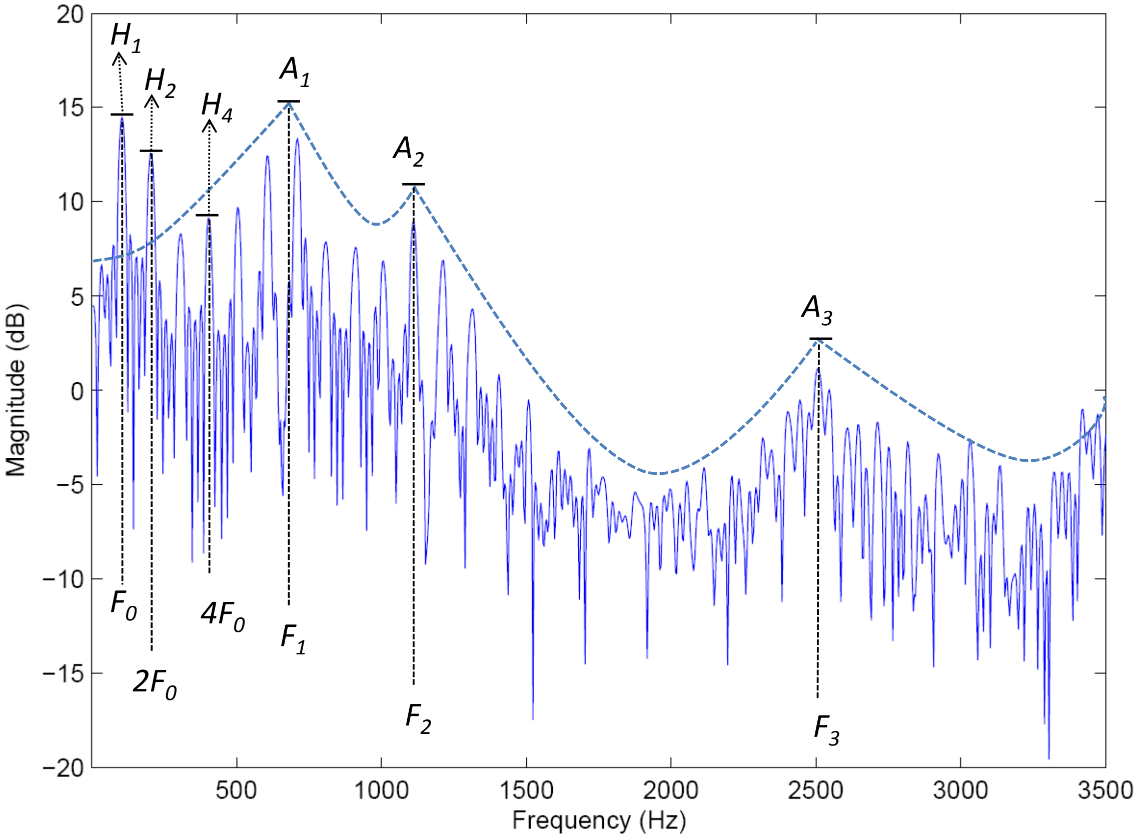


Figure 1.6: Magnitude spectrum of a male speaker's vowel /a/ showing the measures: fundamental frequency (F_0), harmonic magnitudes (H_1, H_2, and H_4), formant frequencies (F_1, F_2, and F_3), and the spectral magnitudes at the formant frequencies (A_1, A_2, and A_3).

$H_1^* - H_2^*$, the difference between the first two harmonic magnitudes corrected for the effects of the VTTF, has been widely assumed to be related to the open quotient (OQ, defined as the proportion of time the vocal folds are open during a

Table 1.1: Description of voice source measures used in this dissertation.

Measure	Description
$H_1^* - H_2^*$	The difference between the first two source spectral harmonic magnitudes [Han97].
$H_1^* - A_3^*$	The difference between the first source spectral harmonic magnitude and the source spectral magnitude at the frequency of the third formant [Han97].
$H_2^* - H_4^*$	The difference between the second and fourth source spectral harmonic magnitudes [KGB07].
CPP	Cepstral Peak Prominence; a measure of the amplitude of the cepstral peak corresponding to the fundamental period, normalized for overall signal amplitude [HCE94].
HNR	Harmonic-to-Noise Ratio; a measure of harmonic energy normalized by the spectral noise level [Kro93].

phonation cycle). As the OQ increases, the glottal open phase becomes closer to the fundamental period, leading to a stronger fundamental component (H_1^*) in the speech spectrum. The increase of $H_1^* - H_2^*$ presumably contributes to increased "breathiness" in perceived voice quality (e.g., [KK90]).

Cepstral Peak Prominence (CPP) is defined in [HCE94] as "a measure of the amplitude of the cepstral peak corresponding to the fundamental period, normalized for overall signal amplitude". A signal with well defined periodic structure is expected to show a very prominent cepstral peak. The turbulent airflow pattern that is associated with breathy voice results in an acoustic signal that tends to be less periodic than a nonbreathy voice. Hence, CPP has been used to differentiate between breathy signals (low CPP values) and nonbreathy signals (high CPP values). In [HCE94], the effectiveness of several acoustic measures in predicting breathiness was evaluated. Perceptual tests were conducted to obtain breathiness ratings from a sustained vowel and a 12-word sentence spoken by 20 speakers with voice pathologies and 5 speakers with no voice pathologies. Results showed that CPP is highly correlated with breathiness ratings ($|r| = 0.89$).

The harmonic-to-noise ratio (HNR) is a measure of harmonic energy normalized by the spectral noise level [Kro93]. In [Kro93], harmonics are defined as band-limited peaks at integer multiples of F_0, while spectral noise is defined as frequency components that are not integer multiples of F_0. The level of spectral noise is related to the perceptual characteristics of the voice. At the physiological level, the incomplete closure of vocal folds results in turbulent airflow through the glottis. The noise generated by the turbulent airflow results in a higher noise level in the speech spectrum. It was reported that listeners were more likely to rate a signal as being breathy, if random noise is added to the signal along with an increase in H_1 [KK90]. Because HNR reflects the noise level, HNR could presumably be an indicator of breathiness.

$H_1^* - A_3^*$ (the difference between the first source spectral harmonic magnitude and the source spectral magnitude at the frequency of the third formant) was shown in [Han97] to be related to the source spectral tilt. Source spectral tilt measures the amount of high frequency components relative to low frequency components, and was found to be associated with stress and intonation [SV96, SH96]. $H_2^* - H_4^*$ (the difference between the second and fourth source spectral harmonic magnitudes) is related to mid-frequency spectral tilt [KGB07]).

1.4 Data acquisition methods and challenges

Due to the hidden position of the vocal folds, various methods have been proposed to observe vocal fold vibrations. Modeling the voice source relies on the data collected. The accuracy of data collection is critical to the establishment of an accurate model. The wide range of models reviewed in Section 1.2.1 also reflects the different types of data and observations upon which the models are built.

1.4.1 A review of data acquisition methods

One way of recovering the voice source signal from the acoustic sound pressure or oral airflow signal is via inverse filtering, which attempts to remove vocal-tract filtering effects. Inverse filtering is sensitive to recording conditions and experimental setup, and several methods have been proposed (e.g., [Rot73, JBM87, Alk92, AAB06, AMY09]).

Another non-invasive method is electroglottography (EGG) [HdD04], which measures the changes in contact area between the vocal folds during phonations. A high-frequency modulated current (about 1 MHz) is sent through two electrodes, placed at each side of the thyroid cartilage during a voiced phonation. The electrical admittance increases as the vocal folds contact increases, and provides a relative measure of vocal folds contact area. EGG signal and its derivative provide useful information such as glottal opening and closing instants [HdD04]. However, many pitfalls have been noted when applying EGG, such as variation across speakers due to physiological characteristics (e.g., males, females, children, and patients) [CC90]. Most importantly, EGG reflects only the degree of vocal fold contact, not area of opening or glottal airflow. In this sense, the EGG signal is an indirect measure of the voice source and is far from sufficient for modeling studies.

A more direct way of observing vocal fold vibration is through laryngeal high-speed videoendoscopy (HSV) recording, which captures the true intracycle vibratory behavior through a full image of the vocal folds. Although high-speed motion films have been used for studying the motion of the vocal folds as early as the 1940s [Far40], commercial HSV systems were introduced relatively recently in the 1990s [DPB07]. Images of vocal folds vibration are captured by a camera equipped with a endoscope, typically at thousands of image frames per second. The glottal area waveform can be extracted from high-speed images to represent the voice source signal (e.g., [SA10])[1].

Other methods, which are more popular among clinicians, include laryngeal stroboscopy [BHF87] and videokymography [SS96].

1.4.2 Challenges in analyzing HSV data

Although laryngeal HSV has emerged as the state of the art in laryngeal imaging, the study of HSV remains limited due to a few challenges. Firstly, the large amount of data produced in HSV recordings limit its applicability in scientific research and clinical applications. For example, 5 s of HSV recording at a speed of 4,000 fps generates 5 gigabytes of data, which would require about 30 min to view the whole recording at a playback speed of 10 fps. It is challenging to interpret the images visually from HSV playback and it usually requires subjective assessment. A substantial amount of effort is devoted to processing the HSV data for subsequent analysis. A common approach is to extract glottal area waveforms from HSV data by applying glottal area segmentation. The extracted glottal area waveforms can be used for clinical diagnosis and for analyzing vibratory patterns of the vocal folds. The segmentation of glottal area is in itself a challenging task and typically requires manual interactions (see Section 2.1 for more discussion). There is a need for an automatic or semi-automatic analysis technique to provide a complementary way to assist diagnosis. Chapter 2 presents a new computationally-efficient method—the glottaltopogram—to compactly summarize the overall spatial synchronization pattern of vocal fold vibration for the entire glottal area, in a manner that can be intuitively interpreted. Such a method may produce plots that are spatially similar to the original images, and which can be easily interpreted by physicians and clinicians during diagnosis.

Additionally, many measures have been used to parameterize the voice source and to study acoustic and perceptual consequences of changes in glottal pulse

[1] See Section 1.5.1 for more discussion

shapes. However, these conventional measures are typically used for **glottal flow** signals and are not specifically designed for **glottal area** signals. There is a need to derive a measure to capture variations in glottal area pulse shape that have perceptual importance. Chapter 3 investigates the aspects of the glottal area pulse shape that vary with voice quality, by using HSV recordings of the vocal folds. A new measure of the **glottal area** is proposed to adequately capture variations in pulse shapes.

1.5 Research questions related to voice source modeling

The voice source provides important information to many speech research disciplines, such as speech recognition, speaker recognition, voice quality analysis, and emotion recognition. For a vast majority of these applications, voice source information has to be estimated from the speech signal recorded by a microphone, rather than from medical/clinical devices. Thus, whether or not voice source signals can be reliably estimated from the speech signal in an everyday (possibly noisy) environment is the first challenge. For instance, in [PQR99], voice source waveforms were estimated from speech signals using an automatic approach, and source parameters were applied to a speaker identification task. In this scenario, an adequate and accurate voice source model will help improve the performance of voice source estimation. In Chapter 5, a new voice source model and a noise-robust automatic source estimation algorithm are proposed.

Many voice source models have been proposed, but few studies have attempted to systematically validate glottal source models perceptually. Is is not clear which model is better in terms of fitting to the observed data. Additionally, whether deviations from perfect fit between models and data have any perceptual importance remains a question. An important application of voice source modeling is speech synthesis. A perceptually-adequate source model should capture perceptually-important aspects of the source signal, thus generating natural-sounding synthetic voices. In Chapter 6, a new voice source model, motivated by data from laryngeal HSV, is proposed to capture perceptually-important source shape aspects. Perceptual experiments show that the proposed model provides significantly better synthetic voices in comparison to four existing source models, in terms of perceived naturalness.

1.5.1 What to model: glottal flow or glottal area?

The glottal flow measures the volume velocity of the air produced at the glottis. The glottal area is the area of separation between the vocal folds as projected by the image of the glottis. These two entities are closely related because the size of glottal area directly affects the amount of airflow that passes through the glottis. The glottal area can be quantitatively measured from data such as HSV recordings of the vocal folds, while the glottal flow can not currently be directly measured.

Previous studies have supported the argument that the glottal flow and glottal area are somewhat similar in gross waveform shape. It was reported in a computational simulation study [HM07] that, while the acoustic source pulse shapes differed from the glottal area waveforms, the differences were small relative to the larger differences across the waveforms. Therefore, the glottal area data extracted from HSV recordings of the vocal folds were assumed to represent the glottal flow in [Shu10a, SA10].

However, because the production of glottal flow involves the interaction between lung pressure and the glottal area function [Fan82] as well as the interaction between the glottal area and the vocal tract system [TS97, Tit08a], the glottal area function is not identical to the glottal flow. The differences between these two waveforms have been documented in many studies (e.g., the glottal flow pulse has a notable skewing rightward in time [Ste98, Rot81]; see Figure 5.1 for an example). Although it is not possible to validate experimentally, the relationship between the glottal area and the glottal flow signals was quantitatively modeled using the three-mass vocal fold model in theoretical modeling studies [ST95, TS02]. Therefore, it is reasonable and necessary to clearly distinguish between the glottal area and the glottal flow for model evaluation purposes, which will be discussed later in Chapter 5.

Although HSV recordings of the vocal folds are used in this dissertation to aid analysis and evaluation, it is worth noting that the models proposed in Chapters 5 and 6 are glottal flow models rather than glottal area models.

1.6 Dissertation outline

The remainder of this dissertation is organized as follows:

Chapter 2 provides a survey of methods used to analyze HSV data. A new computationally-efficient method—the glottaltopogram—is presented to reveal the overall synchronization of the vibrational patterns of the vocal folds over the entire laryngeal area.

Chapter 3 investigates the aspects of the glottal area pulse shape that vary with voice quality, by using HSV recordings of the vocal folds. A new measure of the **glottal area** is proposed to adequately capture variations in pulse shapes.

In Chapter 4, voice source related acoustic measures are analyzed in the context of a physiological vocal-fold vibration pattern—the glottal gap. These acoustic measures are then applied to an automatic gender classification task of children's voices.

In Chapter 5, a new source model and a noise-robust automatic source estimation algorithm are proposed to estimate the voice source from speech signals.

Results in both clean and noisy conditions show that the proposed model and algorithm are robust in accurately estimating the voice source signal.

In Chapter 6, a new voice source model, motivated by HSV recordings of the vocal folds, is proposed to capture **perceptually-important** source shape aspects. Perceptual experiments show that the proposed model provides significantly better synthetic voices in comparison to four existing source models, in terms of perceived naturalness.

Finally, Chapter 7 summarizes this dissertation and discusses future research directions.

CHAPTER 2

The Glottaltopogram: A method of analyzing high-speed images of the vocal folds

As described in Chapter 1, clinicians and speech scientists have developed a number of techniques to observe vocal fold vibrations, including electroglottography [Bak92], photoglottography [Son59], stroboscopy [Kit85], and videokymography [SS96]. Recently, high-speed video (HSV) of the larynx has emerged as the state of the art in laryngeal imaging, due to increased recording frame rates, improved image resolution, and the decreasing cost of high-speed recording devices. HSV data have provided valuable information to the study of the voice source.

The study of HSV remains limited, however, by the large amount of 3-dimensional data produced (Figure 2.1), so that images are inherently difficult to interpret visually and usually require subjective assessment. Because humans are better at discriminating characteristics of static than dynamic images (which impose a memory load), many methods have been proposed to reduce the dimensionality of spatial-temporal HSV data and condense the time-varying video into a few static images that preserve the most important characteristics of the vibratory patterns. This chapter proposes a new computationally-efficient method—the glottaltopogram—to compactly summarize the overall spatial synchronization pattern of vocal fold vibration for the entire glottal area.

This chapter is based on the following publication:

- Gang Chen, Jody Kreiman, Abeer Alwan, "The glottaltopogram: a method of analyzing high-speed images of the vocal folds," Computer Speech and Language, 2013, in press, http://dx.doi.org/10.1016/j.csl.2013.11.006.

2.1 Background

Many previously-described methods for analyzing HSV data depend on glottal area segmentation [LET08, KHd12, DLS11, YAK05]. Automatic segmentation of the glottal area from HSV is in itself a challenging task, and several methods have been proposed. The most straightforward technique is thresholding, in which pixels with brightness lower than a certain threshold are treated as part of the glottis (e.g., [MDZ10, MDQ11]). The threshold is typically specified based on a histogram of the image, where several peaks are assumed to exist due to clustering of glottal and non-glottal regions. However, this method is unsatisfactory when contrast is low, because segmentation performance is sensitive to threshold selection. In addition, this method is not fully automatic because it typically requires manual adjustment of thresholds over time. Other approaches to glottal area segmentation apply seeded region-growing algorithms. After manually selecting seeds from the image, neighboring pixels are examined to decide whether they should be added to the region, subject to a criterion that varies from implementation to implementation [AB94, YCB06, LTR07]. This method typically requires clear glottal edges to produce a correct result.

The segmented glottal area can subsequently be analyzed to reveal spatial and/or temporal variations in glottal vibratory patterns. For example, in phonovibrography (PVG; [LET08]), the segmented glottal area is transformed into a geometric pattern representing the distance from the glottal edges to the glottal center line axis. In terms of the representation in Figure 2.1, PVG condenses the x and y axes into one axis by mapping along the glottal edge trajectory, so that temporal resolution is perfectly maintained but spatial resolution is limited to the glottal edge trajectory. This method is sensitive to the detection of the glottal center line axis, which strongly depends on the geometry of the detected glottal area [KHd12] and can be difficult to identify accurately in the presence of a posterior glottal chink (glottal gap). A visual representation termed the "glottovibrogram" extends the PVG method [KHd12, DLS11]. Glottovibrograms measure the distance between vocal fold contours instead of the distance to the glottal center-line axis, but visualization and interpretation of alterations in subsequent cycles remain unintuitive. Recently, Unger *et al.* proposed a PVG-wavegram to reveal inter-cycle characteristics of vocal fold vibrations across long sequences, where individual cycles of a PVG are segmented, normalized for cycle duration, and concatenated over time [UMH13]. Yan *et al.* applied a Hilbert transform to glottal area waveforms to analyze perturbation and periodicity [YAK05]. However, analyses of the glottal area waveforms do not preserve spatial information about vocal fold vibration, limiting applicability for interpreting spatial vibratory features such as asymmetry.

Despite these efforts, segmentation of the glottal area remains a non-trivial task. Results depend on the quality of the HSV data, including image contrast and the clarity of the glottal edge. Manual interactions are typically needed, such as initial seed assignment or threshold selection, and the segmented glottal area sequence requires inspection. In addition, segmentation of the glottal area typically requires processing the HSV data on a frame-by-frame basis, and the long computational time required for image processing limits the applicability of glottal-area based approaches under clinical conditions, where prompt results are preferred.

Other HSV analysis tools do not rely on glottal area segmentation. The most common of these, kymography [TWM99, LHL00], reduces data dimensionality by selecting pixels with a given value on the y axis (anterior-posterior dimension; Figure 2.1)—or several values in multiplane kymography—usually chosen near the

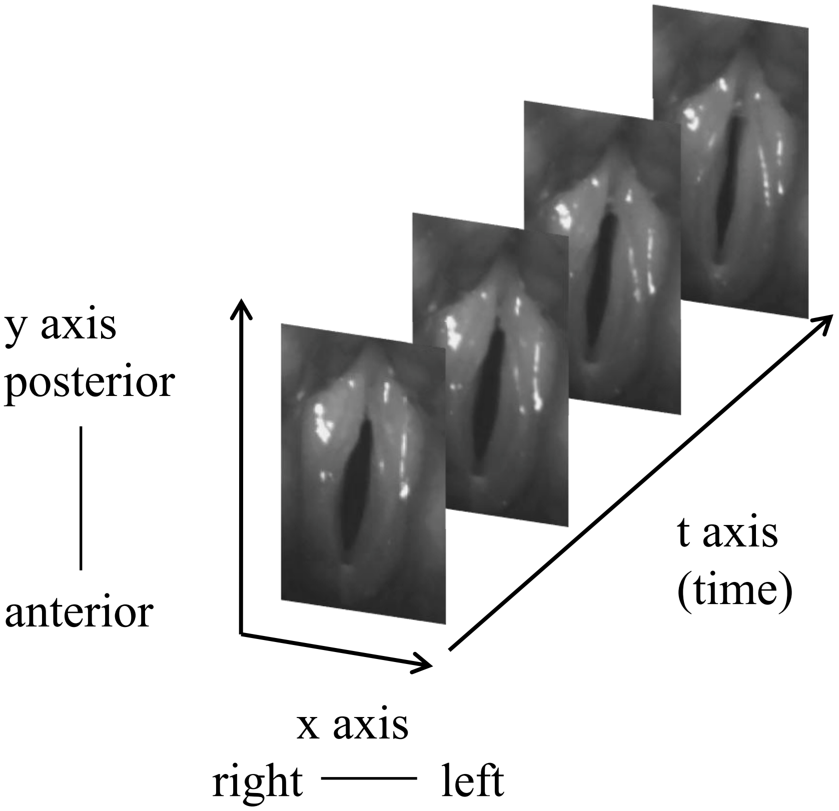


Figure 2.1: The 3 dimensions of variability in high-speed video data: left-right (x), posterior-anterior (y), and time (t).

glottal midpoint. By limiting resolution along the y axis, kymography essentially collapses image analysis along the anterior-posterior dimension, so that temporal resolution is lossless but spatial resolution is limited to at most a few points. In a second method, temporal oscillation patterns across the entire laryngeal area are visualized by applying a Fourier transform to the light intensity time sequences from sequential high-speed images [GL01]. The resulting signal contains amplitude and phase information as a function of frequency, and is displayed as color saturation on top of a single image selected from the original sequence, to characterize vibrational characteristics of the entire laryngeal area. On the basis of this work, Sakakibara *et al.* proposed a third method they called "laryngotopography" to visualize spatial characteristics of the Fourier spectra of the pixel-wise brightness curves (e.g., the frequency component that has the maximum amplitude in the Fourier spectra), which they claimed was effective in visualizing various vibrational modes of the vocal folds of patients with paralysis and cysts [SIK10]. Laryngotopography compresses the time axis by mapping the pixel-wise brightness scale time course into several transformed coefficients, where temporal information is condensed but spatial resolution is fully preserved. In other words, while kymography has limited spatial resolution, laryngotopography maintains the spatial characteristics of the entire image but focuses only on a single frequency component of the spectrum of the vibrational pattern.

This chapter proposes the "glottaltopogram" to visualize HSV data. In this method, principal component analysis (PCA) is applied to light intensity time sequences from consecutive high-speed images and PCA coefficients are visualized.

2.2 Data and methods

2.2.1 Subjects and equipment

High-speed images were recorded at 4000 frames/s using a 70° rigid laryngoscope (KayPentax, Lincoln Park, New Jersey) with a 300W Xenon light source (KayPentax, Lincoln Park, New Jersey) and a Color High-Speed Video System, Model 9710 (KayPentax, Lincoln Park, New Jersey). The image resolution was 512 × 256 pixels and the color mode was 8 bit RGB. Audio signals were synchronously recorded with a Brüel & Kjær microphone (1.27 cm diameter; type 4193-L-004) and directly digitized at a sampling rate of 40 kHz, with a conditioning amplifier (NEXUS 2690, Brüel & Kjær, Denmark). Four subjects (3 males, denoted by M1-M3, and 1 female, denoted by F1) without voice disorders were recorded saying the vowel /i/ with breathy, modal, and pressed voice qualities (although for the male speakers only the modal voice samples were examined in this chapter). Similar to [CKG13], healthy subjects were phonetically knowledgeable and voice quality was demonstrated by a phonetician prior to each recording. Four additional male subjects with voice disorders (denoted by PM1-PM4) were also recorded while saying /i/ using their habitual pitch and loudness. All subjects were asked to sustain the phonation for at least one second during rigid endoscopy.

2.2.2 Image preprocessing

High-speed images were first converted from RGB to brightness scale. Due to illumination conditions, brightness of some glare spots needed to be adjusted before subsequent pixel brightness scale analysis (Figure 2.2), because their brightness did not reflect actual vocal fold movement. Histogram equalization was performed manually (through an interactive graphical user interface) to enhance edge contrast of the vocal folds and remove the glare spots as much as possible. Compared to the original image in panel (a), the glare spots in the posterior glottis have been removed after the brightness adjustment in panel (b). Note that although the overall brightness increased after the adjustment, the contrast between glottal and non-glottal areas in the image was enhanced. The brightness of vocal folds approaches its maximum value and the brightness of the glottal open area approaches 0 (a non-linear transformation from physical position to light intensity), so that brightness curves better represent movements of the vocal folds.

2.2.3 PCA implementation

One PCA was performed for each HSV recording. A rectangular window was manually selected to isolate the image region containing the vocal folds (Figure 2.3). To ensure the representativeness of each function, the brightness scale time course was extracted across 300 consecutive frames (roughly 8 to 15 glottal cycles depending on the speaker's fundamental frequency) for each pixel inside the rectangular window. The number of pixels included in each analysis differed

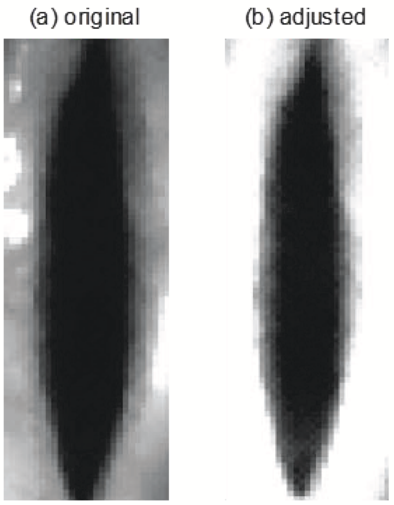


Figure 2.2: (a) The original image of the glottis. (b) The image after brightness adjustment. The posterior glottis is shown at the top of the images, and the anterior glottis is at the bottom.

across recordings, ranging roughly from 5,000 to 10,000, depending on the distance of the laryngoscope from the glottis.

The amplitude values for the brightness scale time course for each pixel served as input to the PCA, which was implemented using the Matlab Toolbox for Dimensionality Reduction [Maa11]. Specifically, for a given HSV, if $g_{(i,j)}(t)$ is a 1-by-N vector and contains the glottal vibration information at pixel location (i, j), then:

$$g_{(i,j)}(t) = [b_{(i,j)}(1), b_{(i,j)}(2), ..., b_{(i,j)}(N)] \tag{2.1}$$

denotes the brightness time sequence (from frame 1 to frame N) at pixel location (i, j), where W is the image width, H is the image height, N is the total number of image frames, t is the frame index, and $b_{(i,j)}(k)$ denotes the brightness value of pixel (i, j) at frame index k. Examples of $g_{(i,j)}(t)$ are shown in the panels surrounding the central image in Figure 2.3.

After performing a mean subtraction (for each frame) to ensure each frame

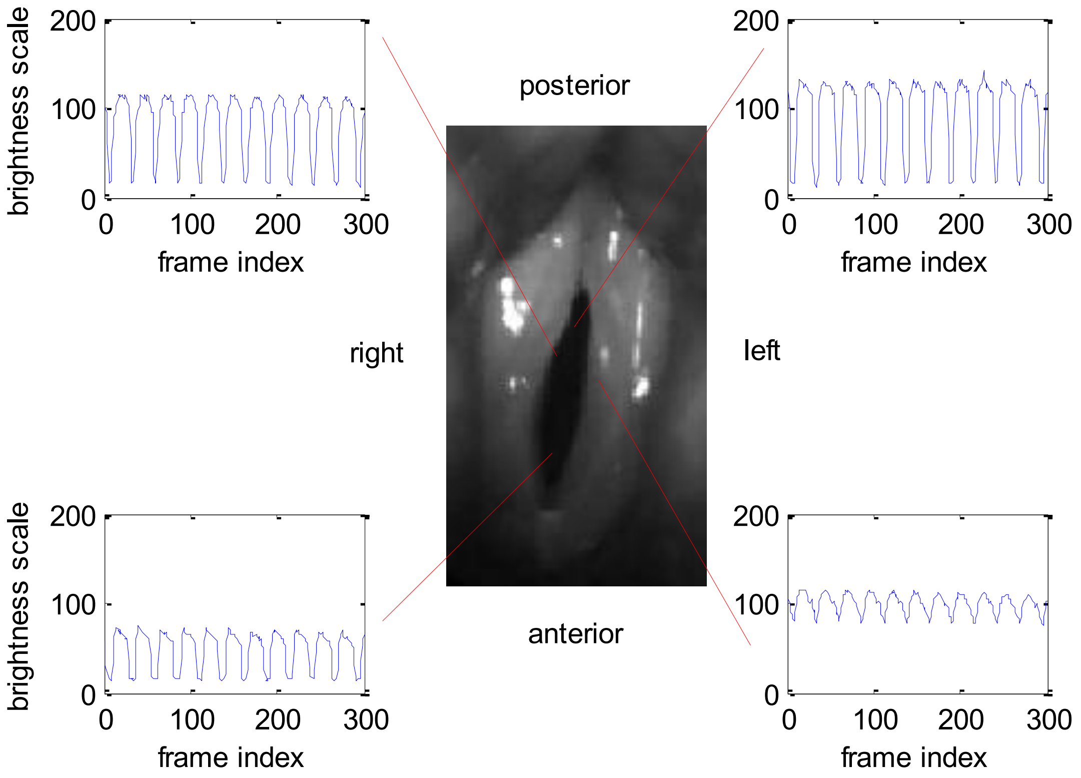


Figure 2.3: Center: image selected for analyses. Surrounding panels: Brightness scale time functions of pixels at different locations in and around the glottis.

has a zero mean for the brightness scale, a PCA was conducted. PCA models the brightness time sequence $g(t)$, treating each spatial pixel as a "repetition" of the experiment and each frame as a "feature". The matrix G in Equation 2.2 was thus built and used as the input to PCA. This $W \times H$-by-N matrix G was constructed by concatenating all the brightness scale time sequences across all pixels in the video:

$$G_{W*H,N} = \begin{bmatrix} g_{(1,1)}(t) \\ g_{(1,2)}(t) \\ \vdots \\ g_{(1,H)}(t) \\ g_{(2,1)}(t) \\ g_{(2,2)}(t) \\ \vdots \\ g_{(2,H)}(t) \\ \vdots \\ g_{(W,H)}(t) \end{bmatrix} \tag{2.2}$$

This matrix G losslessly contains all the glottal vibration information from the video under study. Each brightness time sequence $g_{(i,j)}(t)$ can be decomposed as:

$$g_{(i,j)}(t) = \alpha_{i,j} \cdot PC1(t) + \beta_{i,j} \cdot PC2(t) + e_{i,j}(t) \tag{2.3}$$

where $PC1(t)$ and $PC1(t)$ are the first two principal components (orthogonal bases), $\alpha_{i,j}$ and $\beta_{i,j}$ are projections on the principal components, and $e_{i,j}(t)$ is the error term. Unlike conventional PCAs which are applied to model multiple images in other studies (e.g., face recognition), the PCA used in this chapter was applied to model the brightness time sequence, treating a spatial pixel's sequence as a "repetition". One PCA was conducted to model the brightness scale time sequences from all spatial points within a single recording. Thus, the basis of the PCA (principal component) was the same for all spatial points within that recording. That is, a single matrix G was derived for each individual video, so that $PC1(t)$ and $PC2(t)$ did not depend on pixel locations (i, j).

2.2.4 Analysis and visualization

For each brightness scale time sequence $g_{(i,j)}(t)$, the first two PCA coefficients $\alpha_{i,j}$ and $\beta_{i,j}$ (projections on the first two principal components, PC1 and PC2) were calculated. The coefficients were normalized to an 8 bit (0-255) scale and visualized at the original pixel location (i, j) in terms of color saturation to facilitate interpretation. The brightness scale curve was then reconstructed using the first two coefficients and principal components. Mean square reconstruction errors (mean square of $e_{i,j}(t)$) were calculated and visualized in the same way. In the final stage, the percentage of variance explained by the first two principal components (eigenvalues, or energy, corresponding to the orthogonal bases) was calculated, which partially reflects the energy compactness of PCA (synchronization of the glottal vibration).

By performing PCA, the glottal vibratory pattern represented by the brightness scale time courses is presumably "mapped" to a two-dimensional space captured by PC1 and PC2, given that PC1 and PC2 can account for the majority of the variance in the time-varying data. That is, glottaltopography compresses the time axis by mapping the pixel-wise brightness scale time course into the PCA coefficients, where temporal information is condensed into a single static image but spatial resolution is fully preserved. Pixels with similar brightness scale time courses should have similar PCA coefficients, which are represented in the glottaltopogram as similar colors. Recall that the PCA for each HSV recording was based on brightness scale time sequences from all spatial points within this video, which ensures homogeneity across the spatial points within one HSV recording. Thus, if the left and right vocal folds are vibrating symmetrically, the pixels on the two folds should also exhibit similar brightness scale time sequences. This similarity should be captured by the first two PCA coefficients and the derived images should exhibit symmetric color patterns. If the left and right vocal folds are vibrating asymmetrically, as might occur in a vocal fold paralysis, this asymmetry should result in a glottaltopogram with asymmetric color patterns. Similarly, a glottal region with highly aperiodic vibrations will appear with a distinct color pattern with respect to the remaining steady-vibrating region. When vibration of the two vocal folds is synchronized, the variance accounted for by the principal components should be higher (more compact energy concentration) than when vibrations are unsynchronized, because synchronization results in similar pixel-wise brightness scale time sequences. Similarly, the pixel-wise mean square reconstruction error should be generally low and (roughly) evenly distributed across pixels when glottal vibration is synchronized, while higher reconstruction errors should be observed in laryngeal regions exhibiting unsynchronized glottal vibrations.

2.3 Results

In this section, results of the glottaltopographic visualization approach are presented for both healthy speakers and subjects with voice disorders. Each HSV recording was visualized using a glottaltopogram to determine the underlying glottal vibratory pattern. In some cases, kymograms are also presented, to highlight the complementary information available from each type of display.

2.3.1 Variations in voice quality within and across healthy subjects

This subsection applies glottaltopography first to modal voices of three healthy male subjects (speakers M1, M2, and M3) and secondly to modal, breathy, and pressed voices of a healthy female subject (speaker F1). These simple cases demonstrate the manner in which glottaltopograms can be interpreted, and how these analyses can augment information available from existing analysis approaches.

2.3.1.1 Variations in modal quality among healthy subjects

Modal voice as produced by 3 male speakers (M1, M2, and M3) without voice disorders is first examined. [1] Figure 2.4 shows the glottaltopograms from each speaker, and variance accounted for by each analysis is given in Table 2.1. The first principal coefficient distributions ((a) panels) display symmetric patterns, roughly representing the means of the pixels' light-intensity time courses, which are predominantly determined by the average shape of the time-evolving glottal area (the glottal area generally has lower brightness than the non-glottal area). Recall that a mean subtraction was conducted (for each frame) before performing PCA to ensure each frame has a zero mean brightness, whereas each pixel's brightness scale time sequence was not normalized.

The color differences between the left and right folds in the second principal coefficient are shown in Figure 2.4 (b), and reflect the difference between folds in vibratory pattern. Figure 2.5 shows the corresponding kymogram from the first speaker (M1), and reveals a phase difference between the left and right vocal folds but no obvious differences between folds in the amplitude or frequency of vibration. In this case, the asymmetric vibrational patterns are captured in both Figures 2.4 (first row (b)) and 2.5. Speakers M2 and M3 have symmetric vocal fold vibratory patterns, which are visualized in Figure 2.4 (second and third rows).

2.3.1.2 Comparing phonation types within a single subject

Figure 2.6 shows three glottaltopograms for subject F1, representing modal, breathy, and pressed phonation, respectively. Variance accounted for by each analysis is included in Table 2.1. As shown in the first column of this figure, the first principal coefficient distributions for modal (first row) and pressed (third row) phonation display highly symmetric patterns, although more movement is apparent in the

[1] HSV and audio recordings are available at `http://www.sciencedirect.com/science/article/pii/S0885230813001137`

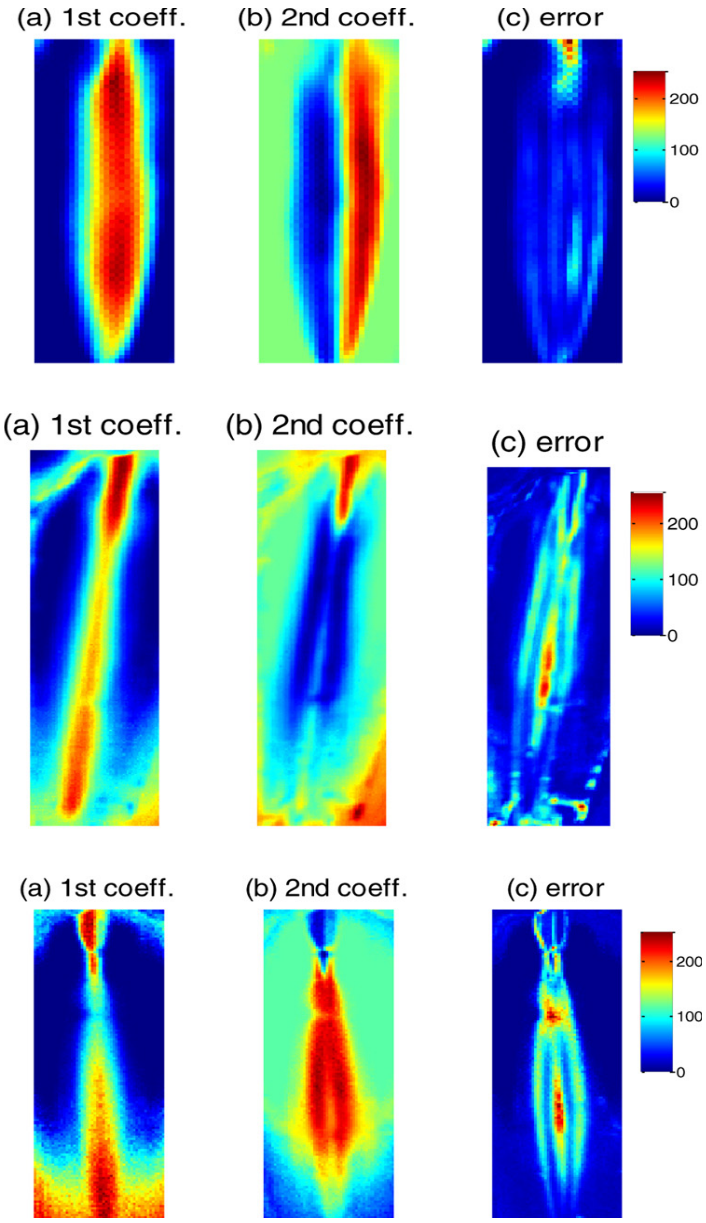


Figure 2.4: Glottaltopograms of modal voice produced by three males without voice disorders. (a) and (b): the first and second principal coefficients, displayed in terms of color saturation. (c): reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The first row represents speaker M1; the second row represents speaker M2; and the third row represents speaker M3. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

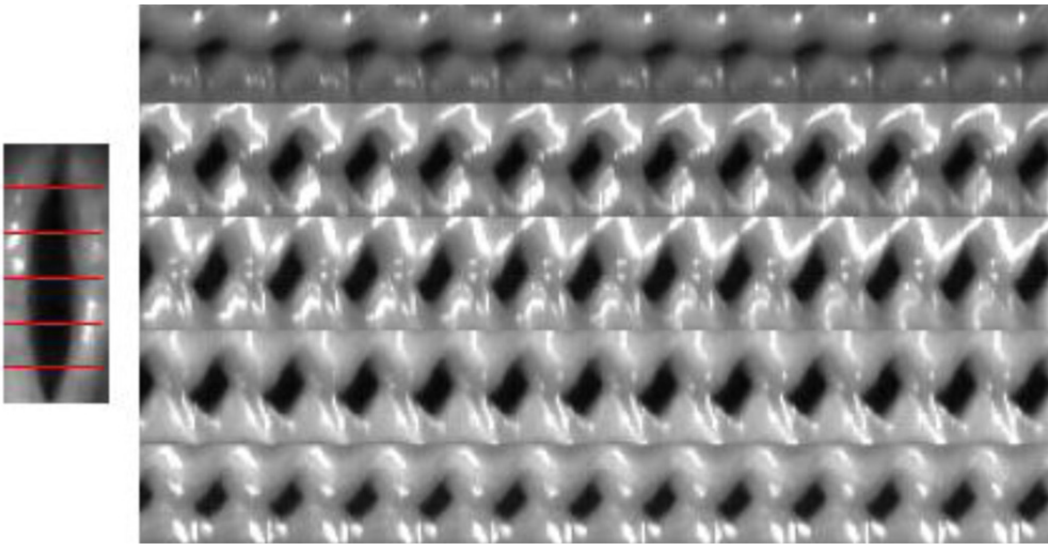


Figure 2.5: Multi-line kymogram of a modal voice from a healthy subject (speaker M1). The x axis represents time, and the y axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom.

anterior glottis than in the posterior (see panel (b) in the third row of Figure 2.6) when phonation is pressed, possibly due to a recording artifact. [2] In contrast, breathy phonation in this speaker (middle row) is characterized by some irregularity in the posterior glottis (presumably representing a glottal gap), which is symmetric for both modal and pressed phonation. Similar roughly symmetric patterns are also observed in the second principal coefficient distributions, with slight asymmetries at the posterior end for modal and breathy phonation. Reconstruction error distributions are visualized in the third column of the figure. These show that reconstruction error is consistently highest in the posterior glot-

[2] Examination of the HSV for the pressed example suggests that this apparent movement may be due in part to a recording artifact, which resulted in a poor view of the most anterior part of the glottis. The high amount of variance accounted for by the second PC may also be due to this effect. Note, however, that the difference between healthy and pathological speakers in variance accounted for by the glottaltopograms remains significant when values from the pressed case were excluded from analysis $[F(1, 7) = 10.68, p = .01, R^2 = 0.60]$.

Table 2.1: Variance accounted for by the first and second principal components for each speaker

	Percent Variance Accounted For			
Speaker	PC1	PC2	Total Variance	Comment
M1	72	19	91	Asymmetric vibrations
M2	74	17	91	–
M3	79	14	93	–
F1	77	11	88	Modal phonation
F1	71	13	84	Breathy phonation
F1	70	21	91	Pressed phonation
PM1	66	14	80	Complex and asymmetric vibrations; creaky
PM2	72	15	87	Phase difference, anterior glottis; breathy
PM3	76	12	88	Phase difference, whole glottis; breathy
PM4	78	7	85	Hyperfunctional quality

tis, presumably due to variability in glottal gap configurations and to the region's small vibration amplitude and slight phase lag compared to the middle portion of the vocal folds.

2.3.2 Patients with voice disorders

In this subsection, glottaltopography is applied to visualize more complex phonatory patterns in four patients with voice disorders. Three patients (PM1-PM3) had asymmetric vocal fold vibrations to different degrees, while one patient (PM4) had symmetric vibrations. As expected, analyses for these speakers accounted for significantly less variance in the underlying HSV data than did analyses for healthy speakers (one-way ANOVA; $F(1, 8) = 13.95, p = .006, R^2 = 0.64$), due to greater irregularity in vibratory patterns.

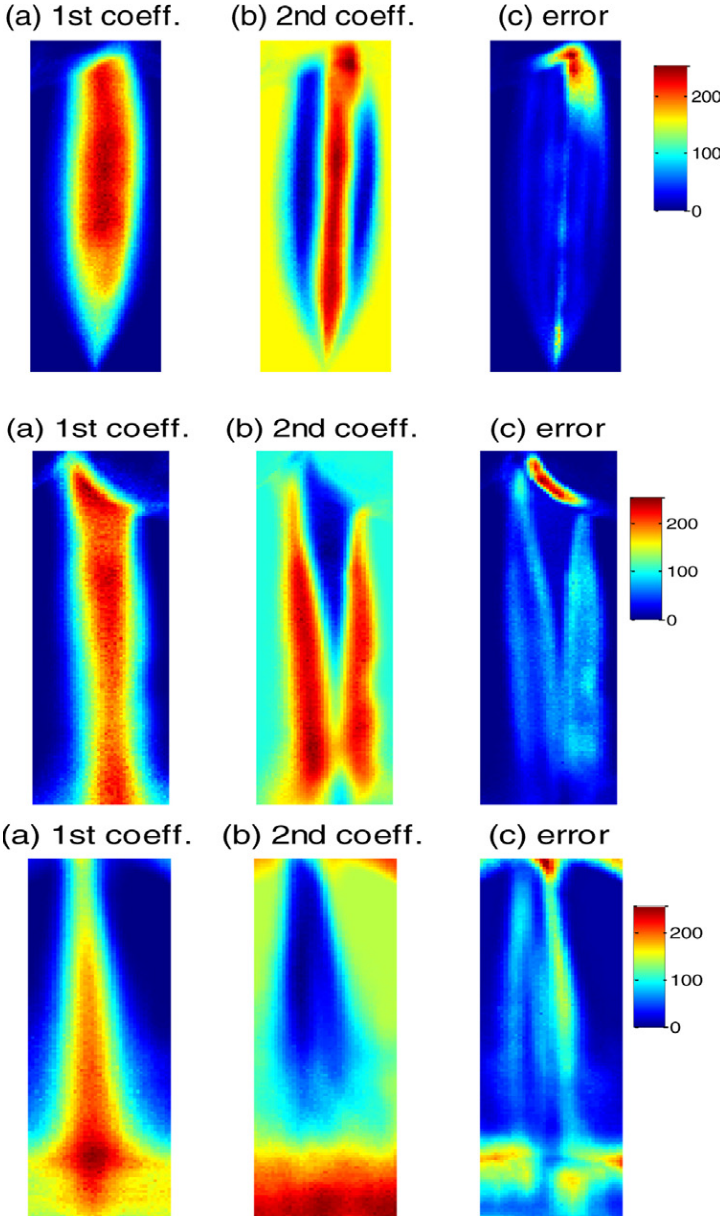


Figure 2.6: Glottaltopograms for modal, breathy, and pressed phonation produced by a healthy subject (speaker F1). (a) and (b): the first and second principal coefficients, displayed in terms of color saturation. (c): reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The first row represents modal phonation; the second row represents breathy phonation; and the third row represents pressed phonation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

2.3.2.1 A patient with a creaky voice and asymmetric glottal vibration

Figures 2.7 and 2.8 show a kymogram and a glottaltopogram, respectively, for a male patient (speaker PM1) exhibiting complex asymmetry in vibrational amplitude between the left and right vocal folds. The percent of variance accounted for by each principal component is given in Table 2.1. As Figure 2.7 shows, the right fold vibrated with larger amplitude than the left, and the vibrating amplitude of the left fold alternated cycle by cycle. Both left and right vocal folds have approximately the same vibratory frequency, although the frequency of phonation appears to alternate in an A-B-A-B pattern. The corresponding acoustic signal sounds creaky.

Figure 2.8 shows a glottaltopogram corresponding to this kymogram. While the glottaltopogram does not reveal the alternations in amplitude and period that are apparent in the kymogram, it does show that the vibrational patterns are distinct between the left and right vocal folds. Note that PC1 accounts for relatively little variance compared to the other cases listed in Table 2.1, possibly reflecting the complex synchronization of this example. Compared to the kymogram, the glottaltopogram provides better spatial resolution in visualizing the different vocal fold vibratory patterns, in a display that includes only 3 images (PC1, PC2, and reconstruction error) rather than the 60 frames included in the kymogram.

2.3.2.2 A patient with a breathy voice and unsynchronized glottal vibration

The glottaltopogram of a second male patient (speaker PM2) with a breathy voice is shown in Figure 2.9. Variance accounted for is included in Table 2.1. The first principal coefficient distribution (panel (a)) displays a symmetric pattern, representing the means of the pixels' brightness scale time sequences (roughly dependent on the average shape of the glottal area). Frame-by-frame visual inspection

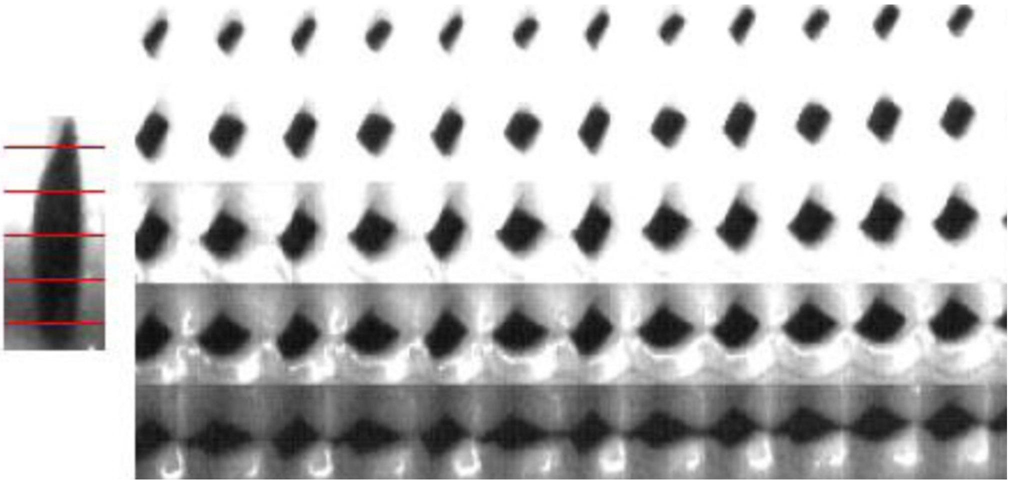


Figure 2.7: Multi-line kymogram of a patient (speaker PM1) with creaky voice. The x axis represents time, and the y axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of each frame, and those of the left vocal fold are shown at the bottom.

of the video recording shows that the left anterior portion of the vocal folds is in opposite phase with respect to the rest of the vocal folds. This is manifested in panel (b) as two distinct portions in the second PCA coefficient distribution: the left anterior portion (lower right of the image) versus the rest of the vocal folds. In (c), the left middle portion of the vocal folds has the largest reconstruction error. This is due to the fact that the first two PCA coefficients poorly model this part. Thus, the vibration pattern is more complex than a synchronous pattern or a pattern with perfectly opposite phase. The left middle portion is the border where normal phase and opposite phase meet, which produces an irregular vibratory pattern.

Figure 2.10 shows a multi-line kymogram of speaker PM2, where the anterior portion shows the phase difference between the left and right vocal folds. However, vocal fold activity in the anterior-posterior direction is not well captured in the

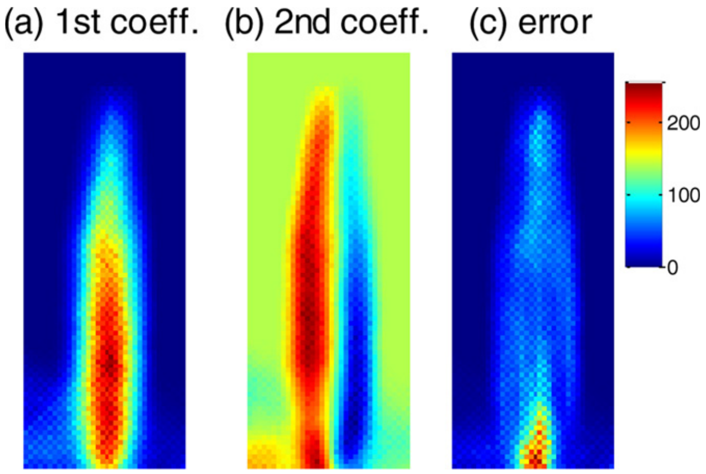


Figure 2.8: The glottaltopogram of a patient (speaker PM1) with creaky voice. (a) and (b): the first and second principal coefficients, displayed in terms of color saturation. (c): reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

kymogram. The glottaltopogram in Figure 2.9 (b) clearly shows that the "phase-unsynchronized" region is the anterior portion of the left vocal fold. The size and position of this problematic region are also visualized, but the actual degree of phase-difference can only be accessed from the kymogram.

2.3.2.3 A patient with a breathy voice and unsynchronized glottal vibration

The glottaltopogram of a third male patient (speaker PM3) with a breathy voice is shown in Figure 2.11. Variance accounted for is included in Table 2.1. Frame-by-frame visual inspection of the HSV recording shows that most of the left vocal fold has a phase lag of about 90° relative to the right fold. This manifests in (b) as two distinct portions in the second PCA coefficient distribution: the left fold (with the exception of the posterior-most segment, near the arytenoids) versus the rest of the vocal folds. The symmetric pattern of the first principal coefficient distribution (panel (a)) illustrates the means of the pixels' brightness scale time

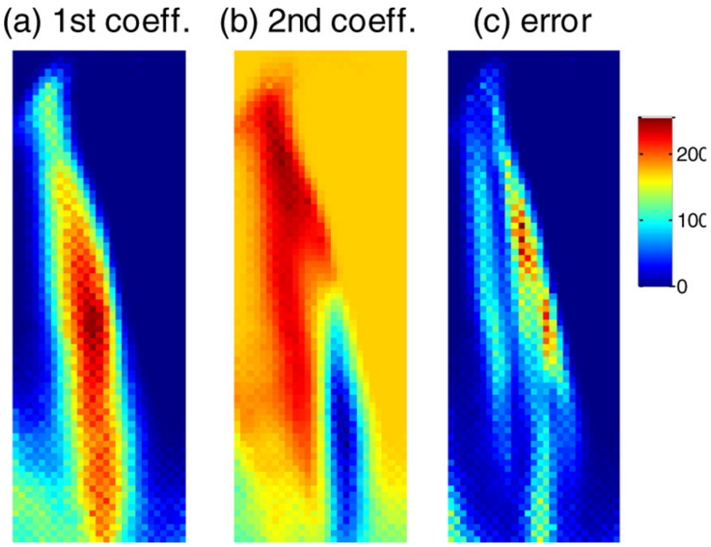


Figure 2.9: The glottaltopogram of a patient (speaker PM2) with breathy voice. (a) and (b): the first and second principal coefficients, displayed in terms of color saturation. (c): reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

sequences, roughly showing the average shape of the time-evolving glottal area.

Figure 2.12 shows a multi-line kymogram of speaker PM3, where the anterior (but not the posterior) glottis shows the phase difference between the left and right vocal folds. Similar to Section 2.3.2.2, the glottaltopogram in Figure 2.11 clearly shows the "phase-unsynchronized" region, providing better spatial information about the overall vocal fold vibrational pattern than does kymography.

2.3.2.4 A patient with pressed voice and synchronized glottal vibration

Figure 2.13 shows the glottaltopogram of a fourth male patient (speaker PM4) with vocal hyperfunction. Variance accounted for is included in Table 2.1. A multi-line kymogram for this speaker is shown in Figure 2.14, where synchronized vibrations can be observed for the left and right vocal folds. This symmetric vibra-

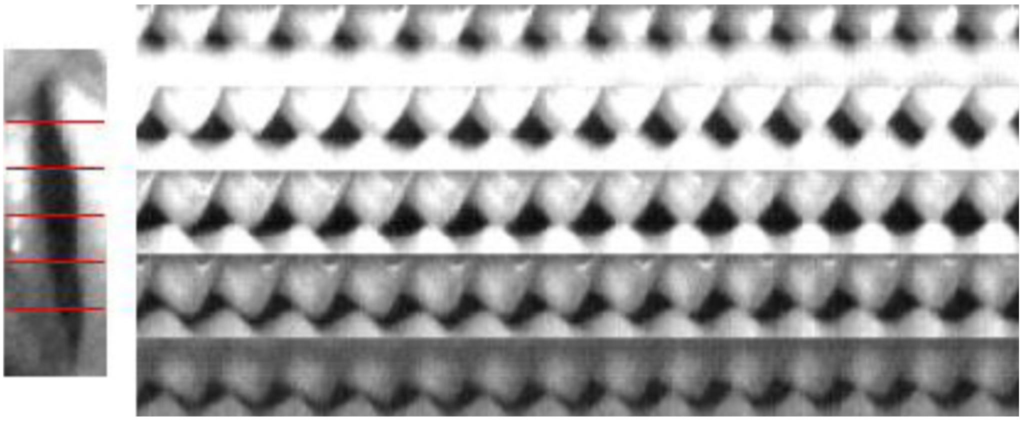


Figure 2.10: Multi-line kymogram of a patient (speaker PM2) with breathy voice. The x axis represents time, and the y axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom.

tional pattern is also captured in Figure 2.13 as a left-right symmetric color distribution. In this case of highly symmetrical vibration, glottaltopography illustrate the spatial synchronization pattern, while kymography visualizes the temporal synchronized evolution within pre-selected lines.

2.4 Discussion

Data reduction methods like glottaltopography reduce HSV data from 3 dimensions to 2, which inevitably leads to loss of information, either temporal or spatial. In this sense, glottaltopography, kymography, and laryngotopography visualize different aspects of HSV data, by maintaining information from different dimensions, but no one method “outperforms” the others. However, the results presented here show how methods can be combined to analyze and interpret HSV data while overcoming the limitations inherent in each individual visualization approach.

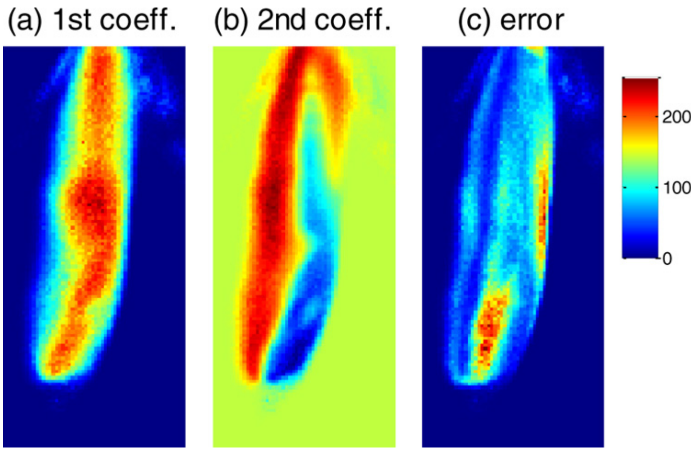


Figure 2.11: The glottaltopogram of a patient (speaker PM3) with breathy voice. (a) and (b): the first and second principal coefficients, displayed in terms of color saturation. (c): reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

Two attributes of glottaltopography make it a particularly useful addition to the set of methods available for working with HSV data. First, glottaltopography is robust (especially when compared to methods like PVG requiring glottal area segmentation) when used with HSV data with variations in contrast levels, random noise during recordings, and multiple glottal gaps, where detection of the glottal edges is inherently difficult. Because some subjects have difficulty tolerating a rigid endoscope, it can be impractical to create multiple high-speed recordings of the same subject in clinical application, and the ability to adjust focus and illumination levels during recording may be limited by the need to complete an exam quickly. As a result, recorded HSV data are often suboptimal in quality [LTR07], so that robustness is an important advantage of the method described here. Secondly, the computational complexity of the glottaltopogram is much lower than that of methods based on glottal area segmentation (e.g., PVG), where the detection of glottal area has to be implemented for each image on a frame-by-frame basis. A glottaltopogram can be generated from 300 video frames in under 5 seconds, while calculating a PVG typically takes a few minutes and involves

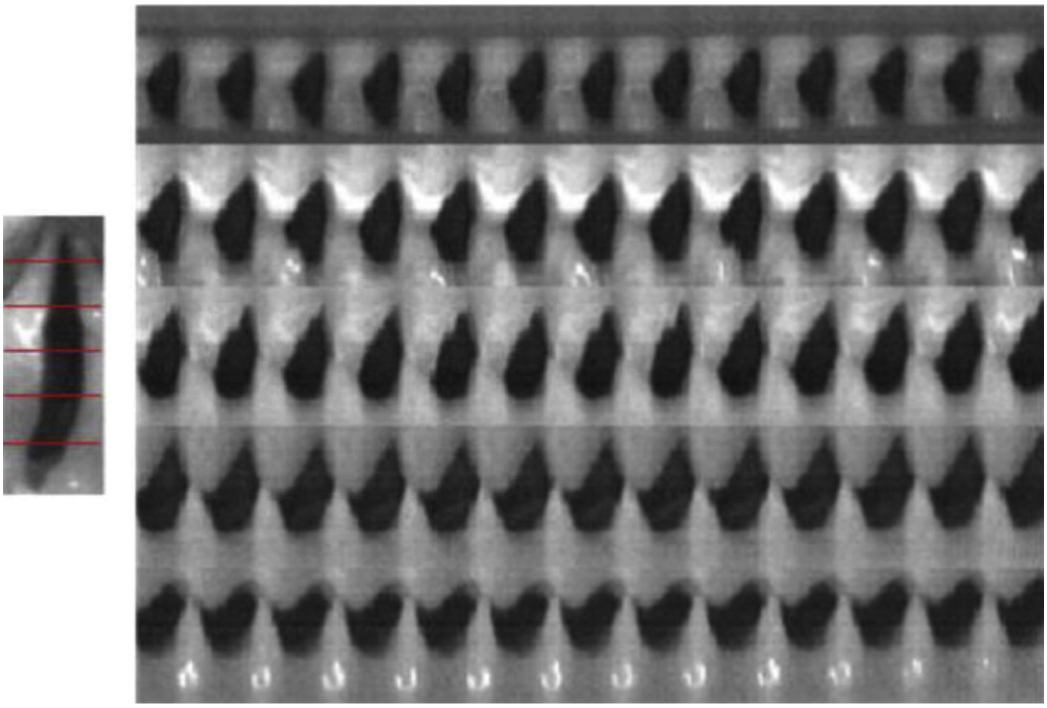


Figure 2.12: Multi-line kymogram of a patient (speaker PM3) with breathy voice. The x axis represents time, and the y axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom.

visual inspection of (at least a few) key frames to ensure the accuracy of glottal area detection.

The first PCA coefficient describes the projection on the dimension that represents the maximum variance in the underlying HSV data. In the present data, this first coefficient always roughly represents the mean of the pixel's brightness scale time sequence, which predominantly depends on the average shape of the glottal area. The second PCA coefficient shows more variability in vibrational pattern across pixel locations, and thus differed more from speaker to speaker. For both synchronized and unsynchronized vocal fold vibrations, the first two PCA coefficients accounted for an average of almost 88% of the variance, largely

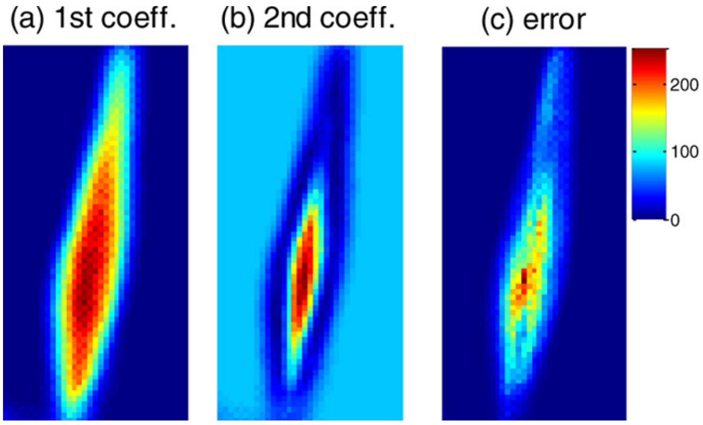


Figure 2.13: The glottaltopogram of a patient (speaker PM4) with vocal hyperfunction. (a) and (b): the first and second principal coefficients, displayed in terms of color saturation. (c): reconstruction error using the first two principal coefficients, displayed in terms of color saturation. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

due to the prevalent quasi-periodic shapes of the brightness scale time sequences among pixels that resulted from quasi-periodic vocal fold vibrations (Table 2.1). This also indicates that the mapping into PCA coefficients substantially maintains the characteristics of vocal fold vibration, as represented by brightness scale time sequences.

It is often claimed that healthy voices are characterized by symmetric, periodic vocal fold vibrations [HLW03, DBL03], and previous studies have sometimes found links between the presence of asymmetric vocal fold vibration and degradations in perceived voice quality in patients with voice disorders [NM00, LFM01]. The present data are not entirely consistent with this scenario. Although the manner in which vibratory asymmetries or phase lags affect perceived voice quality is far from well understood, virtually all of the glottaltopograms of phonation from healthy speakers revealed at least minor asymmetries (and in some cases very large asymmetries) in vibratory patterns. It is noted that a recent study based on physical vocal fold models showed that left-right asymmetry in vocal fold vibration does not produce a perceivable perceptual effect unless the asymmetry is so

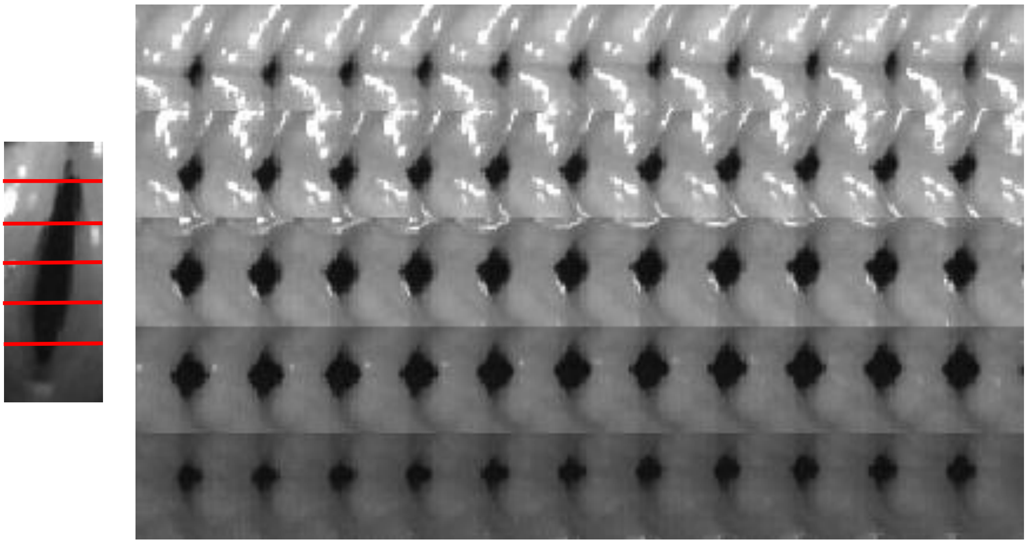


Figure 2.14: Multi-line kymogram of a patient (speaker PM4) with vocal hyperfunction. The x axis represents time, and the y axis represents the amplitude of vocal fold vibration. Each row of images corresponds to movement of the folds at one glottal location (indicated by the red lines through the frame at the left of the figure). Movements of the right vocal fold are shown at the top of the kymogram, and those of the left vocal fold are shown at the bottom.

large that it causes a change in the vibratory mode [ZKG13]. The potential applicability of detecting unsynchronized vocal fold vibration via glottaltopography in clinical settings may provide the data needed to explicate which asymmetries are clinically significant, and which have little or no impact on voice quality. In this way, the proposed method constitutes a promising aid in studying the perceptual consequences of irregular vocal fold vibrations among healthy subjects and among patients with voice disorders.

2.5 Summary

This chapter proposes a new computationally-efficient method—the glottaltopogram—to visualize HSV data. In this method, PCA is applied to light intensity time sequences from consecutive high-speed images and PCA coefficients are visualized. The proposed method reveals the overall spatial synchronization pattern of the vocal fold vibrations for the entire laryngeal area, rather than focusing on a specific location or frequency. Full spatial resolution is maintained, although the time axis is not preserved. Further, the proposed method does not rely on segmentation of the glottal area, and is robust to perturbations of video quality that might result in artifacts during glottal area detection. With minimal user interaction and fast processing time, glottaltopography provides an automatic way of finding the region of interest from the entire image and is suitable for clinical applications. Comparisons between analyses of pathological and healthy data show that the proposed method is effective in visualizing a wide variety of vocal fold vibrational patterns. Additional comparisons between glottaltopograms and kymograms show the manner in which these two analysis techniques (one that compresses the time axis, and one that compresses area) can complement each other in understanding glottal vibration. A Matlab Graphical User Interface—`GTG analyze tool`—is implemented for the glottaltopogram algorithm. A brief description of this tool can be found in Appendix A.

CHAPTER 3

Development of a glottal area index that integrates glottal gap size and open quotient

Many measures have been used to parameterize the voice source and to study acoustic and perceptual consequences of changes in glottal pulse shapes, including open quotient (OQ, the relative duration of the open part of the glottal vibratory cycle), speed quotient (SQ; [TLM58]), closing quotient (ClQ), alternating-current to direct-current ratio (AC-DC ratio; [HHP88, HHP89]), and normalized amplitude quotient [ABV02]. However, the ability of these conventional source measures (commonly used for glottal flow) to relate area waveform variations to spectral changes is limited by the difficulty of modeling both complete and incomplete glottal closure appropriately [KSC12]. This chapter investigates the aspects of the glottal area pulse shape that vary with voice quality, by using high-speed videoendoscopy (HSV) of the vocal folds. A new measure of the glottal area is proposed to adequately capture variations in pulse shapes. These variations are related to corresponding acoustic changes, across glottal configurations both with and without complete closure of the cartilaginous and/or membranous glottis.

This chapter is based on the following publication:

- Gang Chen, Jody Kreiman, Bruce Gerratt, Juergen Neubauer, Yen-Liang Shue, and Abeer Alwan, "Development of a glottal area index that integrates glottal gap size and open quotient," Journal of the Acoustical Society of America, Vol. 133, Issue 3, March 2013, pp. 1656-1666.

3.1 Introduction

3.1.1 Voice source measures

The voice source can be parameterized by fitting the source data with a pre-defined mathematical model subject to certain optimization criteria, so that the source can be represented by a set of model parameters. As reviewed in Chapter 1, the LF model of the glottal flow derivative is the most commonly used, but analyses of singing (and other) voices showed that it provides a suboptimal fit to some source spectra, suggesting that it is not able to accommodate all observed variability in vocal production [HdD01]. Similar results were reported by Shue *et al.*, who estimated the open quotient (OQ) using a codebook of the LF model from voices of four subjects [SKA09]. The estimated OQ and physiological measurements from high-speed imaging data were well-correlated for only two of four speakers, suggesting again that the LF model may be suboptimal for representing some source signals [SKA09].

Research efforts have also been devoted to studying the spectral and perceptual consequences of changes in source waveform shape, as represented by source model parameters. For example, Mehta *et al.* parameterized the glottal area waveform from high-speed videoendoscopy to obtain OQ, plateau quotient (PQ), SQ, and ClQ [MZQ11]. PQ did not correlate significantly with any spectral tilt measures, while OQ and ClQ exhibited statistically significant but small correlations ($|r| = 0.27$ to $|r| = 0.48$) with spectral tilt measures. As the OQ increases, energy in the first source harmonic relative to the second (denoted as $H_1^* - H_2^*$) is assumed to increase, which presumably contributes to increased "breathiness" in perceived voice quality (e.g., [KK90]). However, a recent study based on high-speed imaging of the vocal folds during a "glide" phonation, where quality changed continuously from breathy to pressed, showed that two different relationships hold between $H_1^* - H_2^*$ and OQ, depending on whether glottal closure is complete or not [KSC12]. In the presence of a glottal gap, $H_1^* - H_2^*$ was best predicted by glottal pulse skewness (also called the asymmetry coefficient; [HdD01]), [1] with no significant contribution of OQ; but in the absence of a posterior gap, $H_1^* - H_2^*$ was best predicted by OQ, with pulse skewness making no significant contribution to prediction. An additional study of the same data showed that the size of the glottal gap was strongly correlated with $H_1^* - H_2^*$ when glottal closure was incomplete [CKS11]. Thus, although quality changed continuously in this utterance, it appears that the relationship between glottal configuration and quality is discontinuous when described in terms of existing measures of the voice source like OQ, which do not reflect the presence or absence of a glottal gap. A measure that reflects both the timing of glottal opening and closing, and the presence and size of a posterior glottal gap could overcome this difficulty, giving insight into the physical precursors of changes in perceived quality and providing a linkage between changes in glottal vibratory patterns and perceptual consequences. Such a measure is of particular importance because glottal gaps commonly occur during phonation in both normal and clinical subjects, especially in women [KH73, MRB83, SL90].

3.1.2 The glottal gap effect

Current time-domain source models lack an effective way of modeling incomplete glottal closure, which has been shown to be an important physiological parameter in voice production [CS95, HC99]. Computer simulation compared gaps extending to the membranous glottis ("linked leaks," corresponding to variations in AC flow) to gaps forming an orifice in the cartilaginous glottis separated from the vibrating part of the glottis (a DC component, or "parallel chink") [CS95]. Modeling results showed that gaps in the cartilaginous glottis and corresponding DC flow components had little or no effect on spectral slope relative to cases with

[1] Defined as $t_o/(t_o + t_c)$, where t_o is the duration of opening phase and t_c is the duration of closing phase.

no cartilaginous gap, while persistent gaps in the membranous glottis and corresponding AC modulations in flow resulted in much steeper spectral rolloffs than those of no-gap cases. Omori *et al.* measured glottal gap area at the most closed point of vibration from video-stroboscopic images of speakers with varying vocal pathologies [OSK98]. Glottal gap area affected pitch perturbation, HNR, high-frequency power ratio, mean flow rate, and maximum phonation time. Acoustic and aerodynamic measures were similar when glottal gap sizes (and presumably DC flow levels) were similar, regardless of the underlying vocal pathologies.

In [SL90], glottal closure and perceived breathiness were evaluated in 9 female and 9 male subjects with no known speech pathology. Video-fiberstroboscopic recordings and audio recordings were judged by speech clinicians to evaluate the degree of glottal closure and the degree of perceived breathiness, respectively. Results showed that the degree of incomplete closure and the degree of perceived breathiness were significantly higher for females than for males; the degree of incomplete closure was not significantly affected by F_0 levels. It was hypothesized in [Han97] that speakers with larger posterior glottal openings would have larger spectral tilt (measured by $H_1^* - A_1$ and $H_1^* - A_3^*$). The fiberscopic observation in that study confirmed that subjects with larger spectral tilt measures did exhibit larger posterior openings.

In [SCA10], glottal area waveforms were extracted from HSV of the vocal folds. The effects of glottal gaps on voice source model parameters and acoustic measures were examined. Results showed that OQ, CPP, and spectral tilt measures ($H_1^* - A_2^*$, $H_1^* - A_3^*$, and $H_1^* - H_2^*$) were significantly affected by the presence/absence of a glottal gap. Phonations with glottal gaps had significantly higher $H_1^* - A_2^*$ and $H_1^* - A_3^*$ values than those without glottal gaps. Note that in [SCA10], only the effect of the presence/absence of a glottal gap was analyzed, without quantitative measures of glottal gap size. In [CKS11], glottal gap sizes were quantitatively measured from HSV of the vocal folds. The glottal gap size was shown to affect

CPP and HNR, indicating the presence of relatively more spectral noise with increasing glottal gap size. Simulation using a computational, kinematic model of the vocal folds showed that the acoustic measure CPP decreased with increased separation of the vocal processes, which was partially manifested as the size of the glottal gap during the maximum glottal closure [SS11].

3.1.3 Motivation of the proposed measure

Studies of the acoustic consequences of changing glottal configurations are limited by the lack of a measure of glottal configuration that varies continuously with quality, as described above. To the best of our knowledge, no measure has been proposed that reflects both overall pulse shape and the presence and size of a glottal gap. Studies of the perceptual consequences of changes in the voice source can also benefit from a source measure that adequately relates variations in glottal area waveforms to spectral variations, across a wide range of glottal configurations. For example, the analyses described above [KSC12] did not reveal any abrupt quality change at the instant when the glottal gap disappeared. The continuous, smooth transition in voice quality suggests that a single physiologically-based glottal measure might successfully map the continuum in waveform variation to corresponding changes in voice quality, particularly if that measure reflects the changing relationship between quality, OQ, and glottal gap described above. This chapter describes such a measure, AC/OQ (the ratio of AC to OQ; see Section 3.2.3 for the definition of AC), which was developed based on analyses of high-speed videoendoscopy of vocal fold vibrations during productions of steady-state vowels that varied statically in voice quality. This chapter then tests the ability of this measure to capture continuous variations in voice quality across a range of glottal configurations by analyzing additional high-speed videoendoscopy of phonation during which quality varied continuously along a continuum from breathy to pressed.

3.2 Data and methods

3.2.1 High-speed videoendoscopy data and audio recording

Two sets of synchronous audio recordings and high-speed videoendoscopic images of the vocal folds were collected. The first set included recordings from six phonetically-knowledgeable subjects, three females (denoted by F1-F3) and three males (denoted by M1-M3) [KSC12]. None of the subjects had a history of a voice disorder. Speakers were asked to sustain the vowel /i/ for approximately 10 seconds while holding voice quality, fundamental frequency (F0), and loudness as steady as possible. Across tokens, speakers varied their F0 (low, normal, and high) and voice quality (pressed, modal, and breathy) quasi-orthogonally, resulting in nine steady-state recordings from each speaker. The vowel /i/ was selected to optimize the view of the vocal folds [DBP07]; across tokens vowel quality ranged from /I/ to approximately cardinal vowel /ɛ/. Voice quality was modeled by a phonetician prior to each recording. Because the purpose of the quality and pitch variations was simply to generate a variety of glottal configurations, no effort was made to ensure that voice quality types produced were comparable across speakers. For example, one person's modal phonation might have resembled another speaker's breathy or pressed. Images were recorded at 3000 frames/second at a resolution of 512×512 pixels using a FASTCAM-ultima APX camera (Photron Ltd., San Diego). Microphone signals were bandpass filtered between 20 Hz and 22.4 kHz. The A/D converter (PCI-DAS64/M1/16, Measurement Computing, Norton, MA) had a voltage resolution of 16 bits with input range +/- 5 volts. The audio recordings were later downsampled to 16 kHz for analysis. The other recording settings were identical to those described in Chapter 2.

The second set of recordings was gathered from four speakers, two of whom (speaker F1, M1) participated in the previous recording session, and two additional male speakers (denoted as M4 and M5). These speakers gradually changed their phonations from breathy to pressed while holding F0 and vowel quality as constant as possible. High-speed images of the vocal folds were recorded using a Phantom V210 camera (Vision Research, Wayne, NJ) at a sampling rate of 10,000 frames/second, with a resolution of 208×352 pixels. The camera was mounted on a Glidecam Camcrane 200 (Glidecam Industries, Kingston, MA). The A/D converter (Module 9223, National Instruments, Austin, TX) had a voltage resolution of 16 bits with input range +/- 10 volts. Synchronized audio and high-speed images were recorded for 6 seconds. The other recording settings were identical to those described in the previous paragraph.

In both sets of recordings, most tokens provided satisfactory views of the posterior glottis, but additional tokens were recorded when necessary. The recording that provided the best view of the complete glottis (as judged by a speech-language pathologist) was selected for subsequent analysis.

3.2.2 Glottal area waveform extraction

For the first set of data, a 1-sec sample of auditorily-stable phonation was excerpted from each high-speed videoendoscopic recording. This sample excluded the beginning of the recording in order to avoid possible transient information from initiation of vibration. The glottal area waveform was calculated from the first 150 frames (50 ms) of each sample using a series of edge-detection and region-growing algorithms, described in detail in [SA10]. Factors such as shadows, random noises, over-exposures, and variations in contrast levels affected visualization of the glottis and the accuracy of glottal area extraction. Hence, these analyses were limited to 150 frames (50 ms) instead of the entire token (1 second), allowing visual examination on a frame-by-frame basis and manual adjustment if necessary for accuracy. For several tokens from speaker F1, glottal area waveforms were extracted for the entire 1 second (3000 frames) period and compared with data from the first 150 frames. Comparison showed that the glottal area waveform of the first 150 frames was representative of the entire token. The number of glottal cycles contributed by each speaker depended on the speaker's F0. A total of 442 glottal cycles were included for analysis, among which speakers F1, F2, F3, M1, M2, and M3 contributed 98, 89, 83, 47, 53, and 72 cycles, respectively. Subsequent analyses were also performed on the glottal waveforms from each speaker separately, which minimized the effect of different number of cycles contributed by different speakers.

For the second set of video recordings, glottal area waveforms of the complete utterances were extracted using "GlotAnTools," a software toolkit that automatically segments the glottal area from high-speed images (supplied by the Department for Phoniatrics and Pedaudiology of the University Hospital, Erlangen, Germany). Note that in both sets of recordings, each glottal area cycle was kept rather than averaging across cycles within each recording.

3.2.3 Calculation of glottal measures

Based on analyses showing a trading relationship between changing OQ and glottal gap size as quality varied continuously [KSC12], it was hypothesized that a measure capturing these two aspects of vocal function would correspond reasonably well to changing quality in a larger set of voice samples. As part of the process of developing this measure, values of OQ, DC, and AC were calculated for each glottal cycle in the glottal area waveforms. Figure 3.1 shows how these measures were determined from sample waveforms. Each cycle of glottal vibration was tracked from the extracted glottal area waveforms by marking the first instants of glottal opening when glottal closure was complete. When no complete glottal closure occurred, the moments of minimal glottal area were tracked. DC offsets of the glottal area waveforms were maintained, so that when closure was incomplete, the minimum glottal area was non-zero. The glottal area waveform amplitude was measured in numbers of pixels, and therefore did not represent the actual glottal area. Further, due to variable positioning of the laryngoscope relative to the vocal folds across recordings, glottal area waveform amplitudes were not directly comparable across recordings. Thus, for each glottal cycle, the waveform was normalized by the maximum glottal area within each cycle, so that the maximum amplitude was always 1. DC was defined as the minimum normalized glottal area in each glottal cycle. This process results in a smaller AC waveform when a DC component is present, relative to cases with full glottal closure (see Figure 3.1 for an example). Because this normalization factors the DC component into the AC value, AC was then defined as the root-mean-square (rms) of the AC portion around its mean [HHP88]. [2] Finally, following [KSC12], when the glottis did not close completely, the moment when glottal area began to increase and the onset of maximum closure was treated as opening and closing instants, respectively. For each individual cycle of phonation, OQ was calculated as the time from the first opening instant to the onset of maximum closure (or minimum area), divided by cycle duration (the time from the opening instant to the opening instant of the following cycle). Note that these measurements, although commonly used for glottal flow waveforms (e.g., [HHP88]), were calculated from glottal area waveforms in this chapter.

[2] The actual measure used in [HHP88] was an AC-DC ratio, defined as the rms of the AC portion around its mean divided by the mean of the AC portion. In that study, glottal flow was calculated by inverse-filtering the oral flow measured using a flow mask, which quantified the absolute amount of glottal flow. The division by the mean of the AC portion compensated for the dynamic range of actual glottal flow. In studies using laryngeal high-speed videoendoscopy, the absolute glottal area is not available due to the varying distance between the laryngoscopy and the glottis across recordings. Therefore, in this chapter the extracted glottal area waveforms were normalized to have a maximum value of 1 (divided by the maximum glottal area in each glottal period) so that waveforms were comparable across recordings. This normalization process compensated for the dynamic range of the glottal area. Therefore, the AC component (calculated as rms of the AC portion) was directly used in the chapter without being divided by the mean of the AC component.

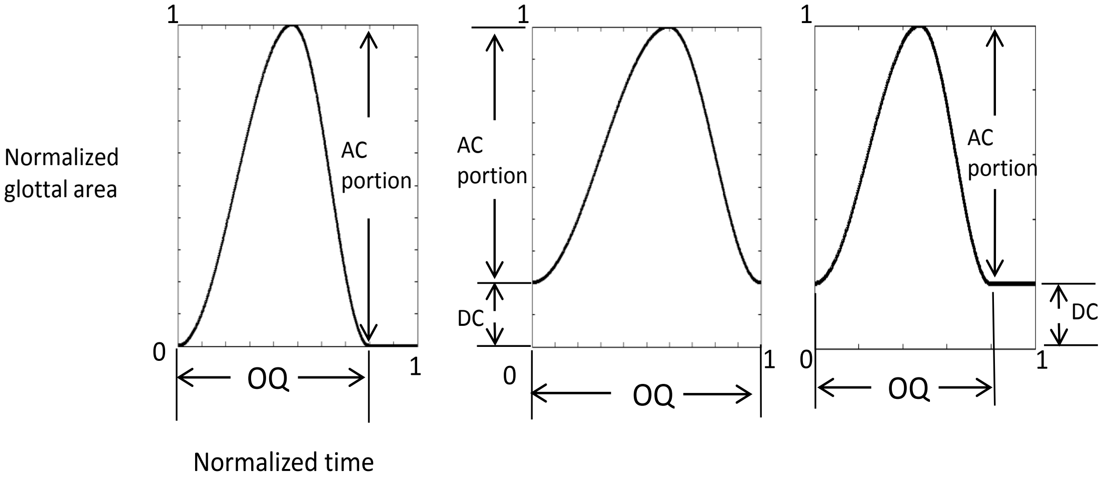


Figure 3.1: Examples showing how glottal measures (AC, DC, and OQ) were determined from glottal area waveforms. (a): complete glottal closure. (b) and (c): incomplete glottal closures

3.2.4 Acoustic measures

The cepstral peak prominence (CPP; [HCE94]), which robustly measures the relative energy in the harmonic and inharmonic aspects of a voice signal, was measured pitch-synchronously from the audio signals with VoiceSauce software [Shu10b] using an analysis window of four periods with a 1 ms shift. F0 values were obtained from the STRAIGHT algorithm [KMC99] to determine the period of a glottal cycle. Values were aligned with glottal area waveforms extracted from the imaging signal for subsequent analysis.

3.2.5 Principal component analysis

On the first set of data, principal component analysis (PCA) was applied to investigate factors that describe variations in the glottal pulse shape. The first PCA was conducted using glottal waveforms from all speakers. Each cycle of each waveform from every speaker was resampled to 1000 points to normalize for differences in F0. Resampled waveforms were visually examined to ensure the pulse shapes of the original waveforms were preserved after the resampling procedure. The amplitude values for each glottal pulse at each sampling instant served as input to the PCA. A second PCA was performed on glottal area waveforms from tokens that exhibited glottal gaps, and a third was performed on glottal area waveforms from tokens where no glottal gap presented. An additional set of PCAs was performed on the glottal waveforms from each speaker separately. The waveforms with maximum and minimum projections on the first two principal components (PCs) were plotted to visualize the variation in waveforms for each speaker. Finally, regression analyses relating source measures to PCs were conducted.

3.3 Analysis and results

3.3.1 Principal component analyses (PCA)

Results of the PCAs and multiple regression analyses relating PCs to measures of pulse shape are shown in Table 3.1. In the first PCA (which included glottal area waveforms from all speakers), PC1 and PC2 accounted for 66.4% and 19.4% of the variance in pulse shapes, respectively. PC1 was most strongly related to OQ and PC2 was most strongly related to AC. The measures of AC and DC were highly correlated ($r = -0.96, p < 0.001$). As noted in Section 3.2.3, the varying glottal gap size directly affects AC, which decreases with increasing glottal gap size, indicating more inharmonic noise relative to the harmonic energy. In this sense, the measure AC incorporates the glottal gap effect and provides a basis for capturing this aspect of glottal area waveform variation. In the second PCA (which included only tokens with glottal gaps), PC1 was best predicted by AC, with no significant contribution of OQ. In the third PCA where tokens without glottal gaps were included, PC1 was best predicted by OQ, with AC making no significant contribution to prediction.

Table 3.1: Standardized regression coefficients for multiple linear regression analyses relating source measures to the first two PCs. Percentage of variance accounted for by each PC is shown in parentheses. "All" denotes PCA using glottal area waveforms from all speakers. "Gap" denotes PCA using only tokens that exhibited glottal gaps. "No gap" denotes PCA using only tokens with no glottal gap. All values except those with an asterisk ($*$) are significant at $p < 0.01$.

	All		Gap		No Gap	
	PC1 (66.4%)	PC2 (19.4%)	PC1 (55.2%)	PC2 (29.8%)	PC1 (74.9%)	PC2 (10.7%)
OQ	-0.88	0.51	0.07*	-0.41	-0.87	-0.24
AC	0.06*	0.97	0.87	0.07*	0.08*	0.06*
R^2	0.86	0.60	0.75	0.19	0.72	0.06

For the PCAs performed on the glottal waveforms from individual speakers, results of multiple regression analyses relating PCs to measures of pulse shape are shown in Table 3.2. The time-based measure OQ and the amplitude-based measure AC showed significant effects on PC1 and PC2 for all speakers except M3. For speaker M3, PC1 accounted for 82% of the variance and was best predicted by OQ only, with AC making no significant contribution to prediction.

Figure 3.2 shows the waveforms representing minimum and maximum values of the first two PCA factors, for each speaker. For PC1, the minimum and the maximum cases differ greatly in OQ for all speakers. Changes in AC (easiest to see in Figure 3.2 as changes in DC offset) between the minimum and the maximum cases also exist for PC1 for speakers F2, F3, M1, and M2. For PC2, speakers F1 and M2 exhibit differences in OQ and AC; speakers F2 and M1 show differences mainly in OQ; speaker F3 exhibits differences in AC. These analyses show that, across speakers and voice qualities, variations in glottal area waveforms (including the effects of glottal gaps on normalized pulse amplitude) are well-summarized by the combination of AC and OQ.

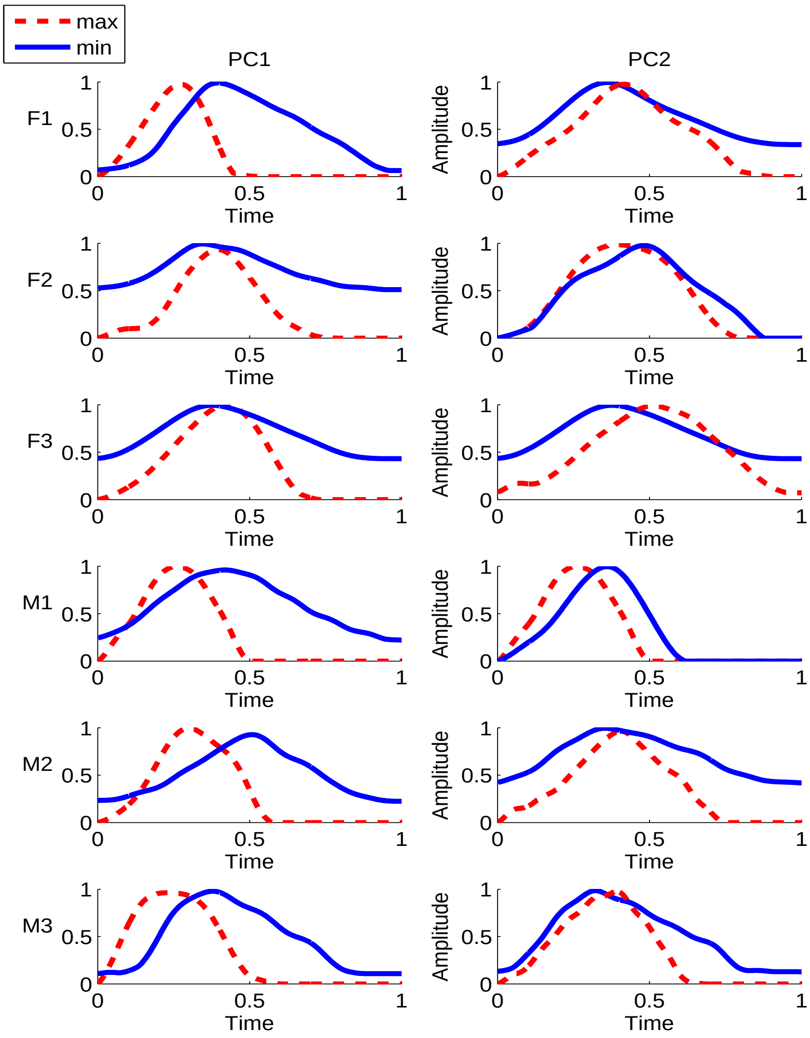


Figure 3.2: Waveforms representing minimum and maximum values of the first two PCA factors for each speaker (F1, F2, F3, M1, M2, and M3).

3.3.2 Data distribution in the PCA space

Projections (scores) on the first two PCs were calculated for each of the glottal area waveforms on the first set of data. Waveforms were reconstructed using the first two PC scores and visually examined to ensure that they captured the shape of the original waveforms. Figure 3.3 shows 3 examples (breathy, modal, and pressed) of reconstructed and original waveforms from speaker F1. Although detailed differences exist, the reconstructed waveform represents the gross shape

Table 3.2: Standardized regression coefficients and R^2 values for multiple linear regression analyses relating source measures to the first two PCs for each speaker. "-" denotes not significant. All other values are significant at $p < 0.01$. Percentage of variance accounted for by each PC is shown in parentheses.

Speaker		OQ	AC	R^2
F1	PC1 (80%)	-0.92	0.08	0.95
	PC2 (12%)	-0.77	-1.17	0.82
F2	PC1 (54%)	-0.29	0.66	0.75
	PC2 (23%)	-0.97	-0.91	0.73
F3	PC1 (74%)	-0.68	0.38	0.94
	PC2 (20%)	-0.93	-1.14	0.78
M1	PC1 (78%)	-0.57	0.42	0.93
	PC2 (16%)	-0.99	-1.14	0.23
M2	PC1 (76%)	-0.69	-0.34	0.96
	PC2 (16%)	1.13	1.25	0.58
M3	PC1 (82%)	-0.92	-	0.86
	PC2 (8%)	-	-	0.05

of the original waveform, because the first two PCs accounted for 85.8% of the variance. The distribution of nominally breathy, modal, and pressed cases across speakers is shown in Figure 3.4, and the distribution of data in the PCA space for each individual speaker is shown in Figure 3.5. Although the speakers were asked to produce sounds in three "categories" of voice qualities, the highly overlapped data distribution between categories indicates the existence of a continuous axis to which the voice source variation continuum can be mapped. This axis should approximately capture the glottal area pulse shape variation along a breathy-to-pressed dimension, from bottom left to bottom right clockwise as shown in Figure 3.4. For speaker M1, the modal and pressed cases overlap substantially, while the breathy cases are well separated from the other two types. For the other speakers, modal cases overlap partially with breathy cases and partially with pressed cases. Neither PC1 nor PC2 alone quantified the three voice qualities sequentially, as expected given the large interspeaker differences in how the stimuli

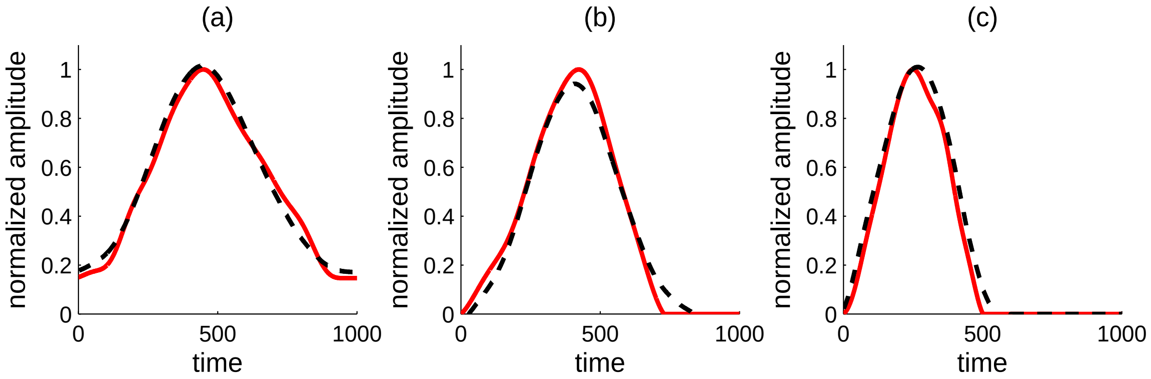


Figure 3.3: Examples of reconstructed waveforms using the first two PC scores (dashed line) and original waveforms (solid line) from speaker F1. (a) breathy. (b) modal. (c) pressed.

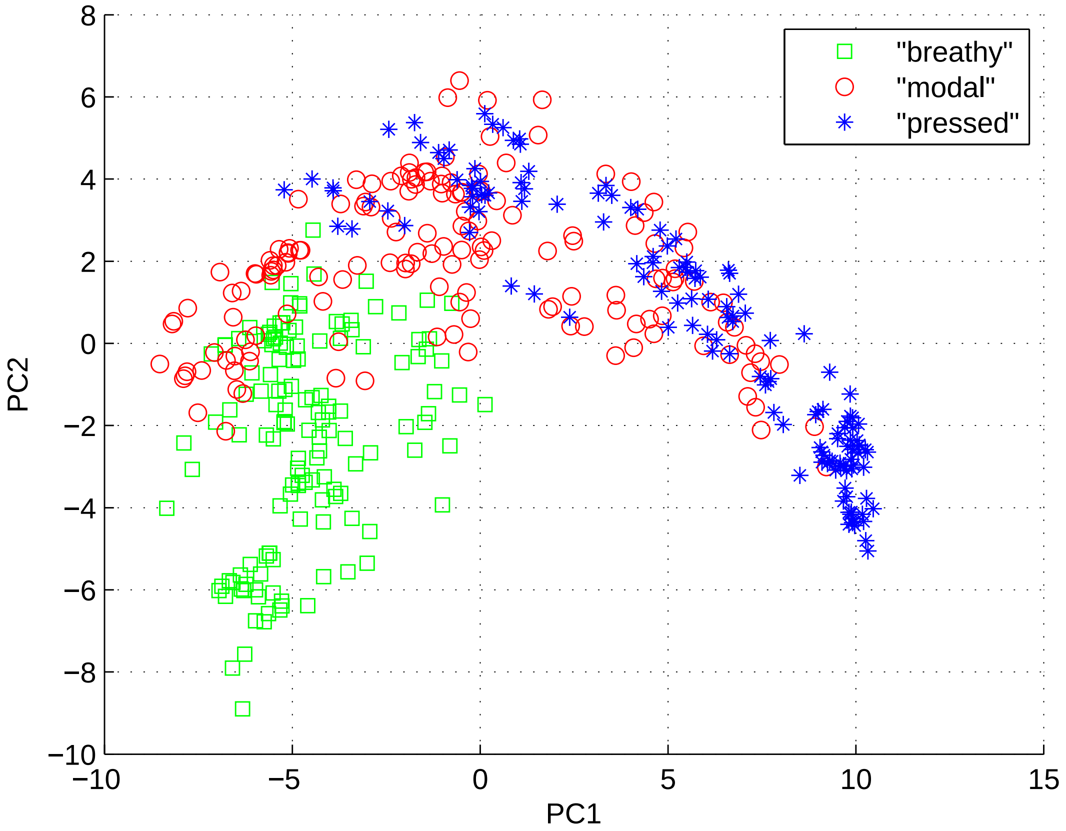


Figure 3.4: Data distribution, for all speakers, in the PCA space labeled by nominal voice qualities.

were produced.

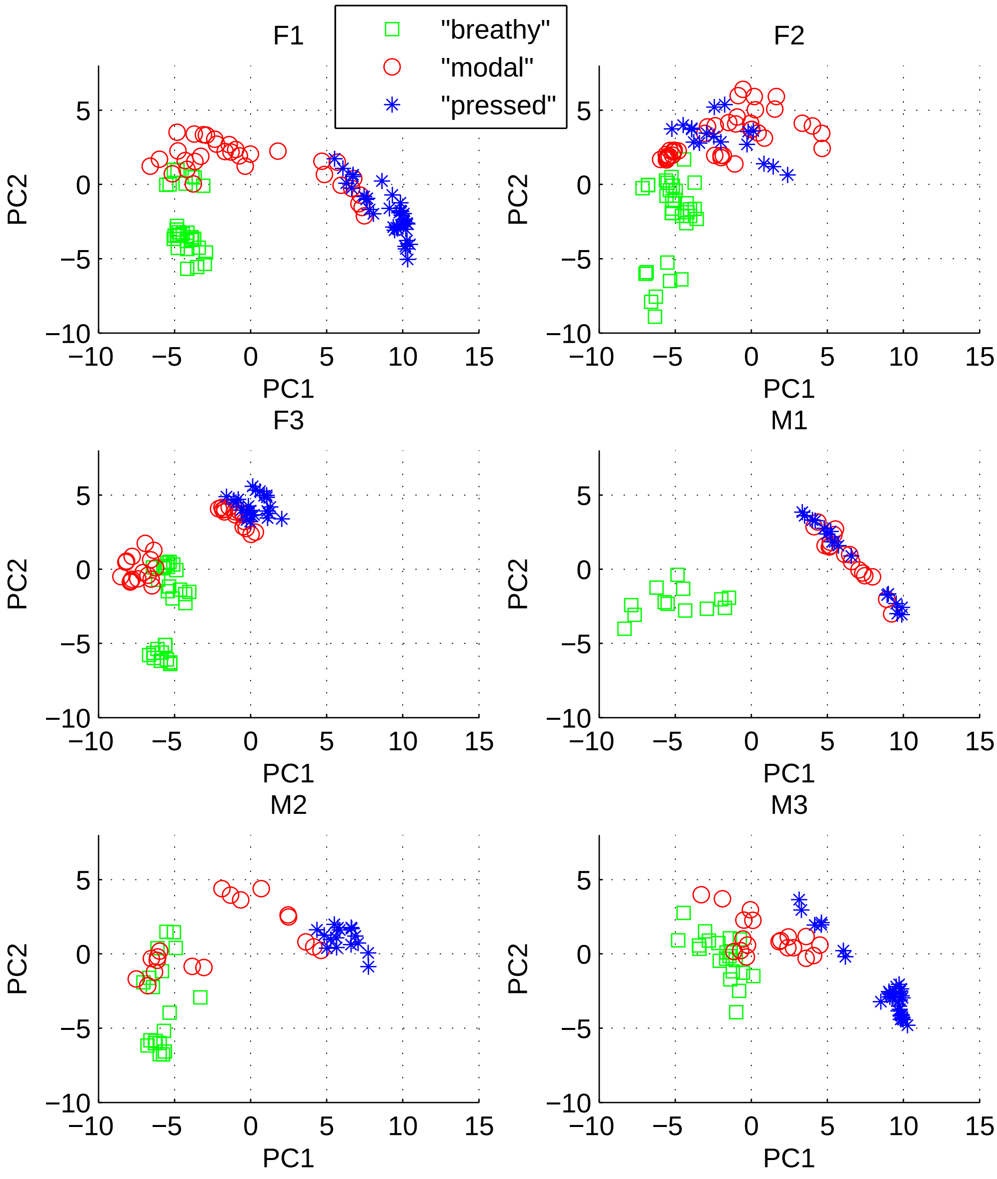


Figure 3.5: Data distribution in the PCA space labeled by voice qualities for each speaker.

3.3.3 The proposed measure: AC/OQ

Measures of the physical voice source ideally should quantify the most prominent factors characterizing glottal pulse shapes, and should also reflect physical precursors of voice quality variation, including the overall glottal pulse shape variations and glottal gap configurations. PCA results showed that pulse shape variations can be efficiently characterized by the time-based measure OQ and the pulse-amplitude-based measure AC. Further, as noted above, the relationship between acoustic measures, quality, and OQ varies depending on the extent of glottal closure [KSC12]. A measure AC/OQ, defined as the ratio of AC to OQ, is proposed to combine both amplitude and temporal characteristics of the glottal area waveform in cases of both complete and incomplete glottal closures. In the numerator, the AC component (reflecting glottal gap presence and size) quantifies the oscillating energy elicited during the glottal open phase. In the denominator, OQ measures the relative duration of the open phase during a glottal period. In this sense, the AC/OQ measure quantifies the oscillating energy produced within a unit time slot. AC/OQ reaches its minimum value of 0 when the glottal area waveform is a constant (i.e., vocal folds are open and no sound is being produced). Theoretically, AC/OQ can reach infinity when the glottal pulse is an impulse (delta function), but this does not occur in human phonation (although following this logic values should be highest for vocal fry, in which the laryngeal excitations are a discrete train of pulses; e.g., [HMW66]).

Figure 3.6 shows AC/OQ values for 4 examples of glottal area waveforms. The first panel (a) shows an area waveform with no DC offset, normalized to peak amplitude; (b) shows the same area waveform with the addition of a DC offset. The presence of a DC offset in (b) has the effect of "compressing" the area waveform, as described in Section 3.2.3, and, hence, AC/OQ decreases. In (c), OQ decreases from 1 to 0.8 for the waveform in (a); and in (d) a DC offset is added to the waveform in (c). As illustrated in this figure, the decrease in OQ results in an increase in AC/OQ (compare (c) to (a)). Adding the DC offset reduces the AC/OQ value (compare (b) to (a) and (d) to (c)). Similar to kurtosis in probability theory, AC/OQ measures the "peakedness" of the glottal area waveform. A higher AC/OQ value indicates a sharper peak in the glottal area pulse and stronger periodic oscillating energy in the spectral domain. On the other hand, a lower AC/OQ value corresponds to a flatter glottal pulse and weaker periodic oscillating energy. Values in Figure 3.6 show how the AC/OQ measure captures the tradeoff evident in the glide phonation described in Section 3.3.6 between effects of changing OQ and changing DC levels on voice quality. As quality moved from breathy towards pressed, glottal configuration initially resembled (b) (with the lowest AC/OQ value), then (a), with an intermediate value, and finally (c), with the highest value. The prediction implied by the comparison between panels (a) and (d) is that voices with a smaller OQ plus a DC offset should fall perceptually in roughly the same range along a breathy-to-pressed continuum as those with a large OQ but no DC offset. This prediction remains to be tested. Such a comparison requires a comparatively large glottal gap only in the cartilaginous region that is separated from the vocal fold vibration (the membranous glottis vibrating with a relatively small OQ). This scenario was not available in the current data.

3.3.4 Evaluating AC/OQ in parameterizing differences in glottal area waveforms across voice qualities

Assuming a quality continuum from breathy to pressed, Table 3.3 shows regression analyses relating AC/OQ to the nominal voice quality continuum for each speaker, and Table 3.4 shows the means and standard deviation of AC/OQ for the three productive categories for the 6 speakers in the first set of data. AC/OQ was significantly correlated with the voice quality continuum for all 6 speakers ($p < 0.001$). Except for modal vs. pressed for speaker M1, AC/OQ values also differed significantly between categories ($p < 0.001$) for each speaker. For speaker M1, whose modal phonation was quite pressed-sounding, neither OQ nor $H_1^* - H_2^*$ differed significantly between pressed and modal phonations ($p > 0.05$). Previous studies have argued that pressed phonation has lower OQ and $H_1^* - H_2^*$ values than modal phonation [KK90, Han97], suggesting that speaker M1's productions of the

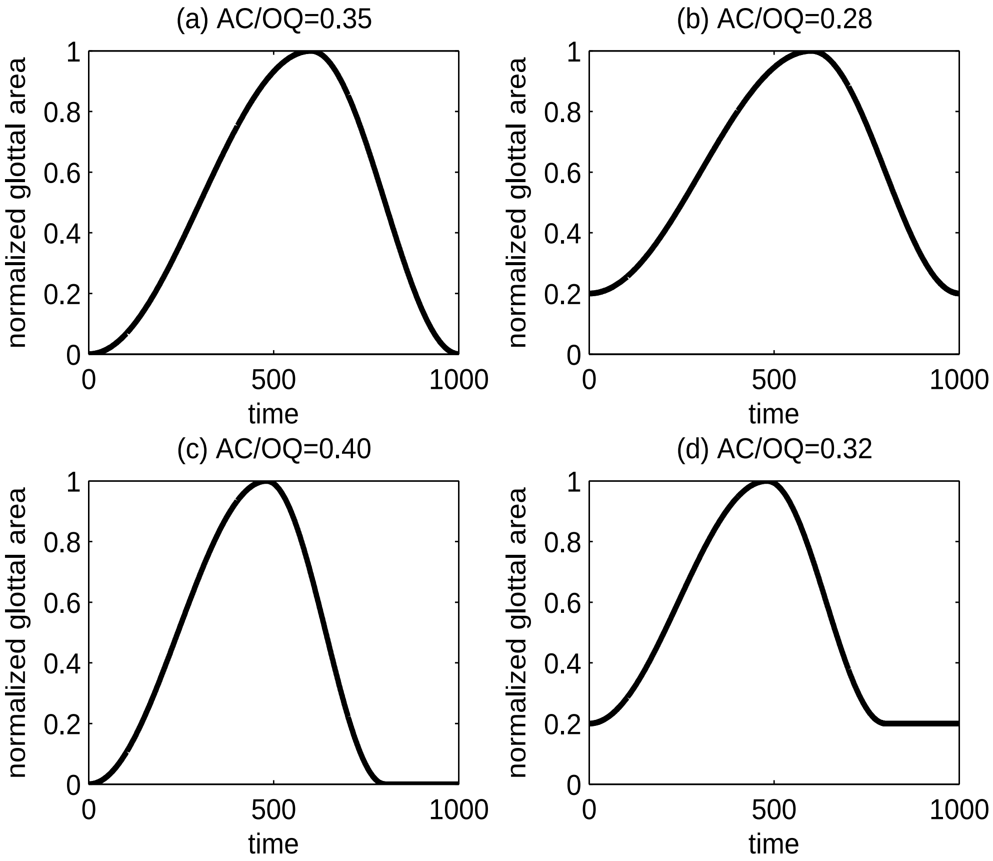


Figure 3.6: Four synthetic glottal area waveforms showing how the changes in OQ and DC offset affect AC/OQ values. Note that OQ for panels (a) and (b) are equal to one.

designated voice qualities were inconsistent with the most usual understanding of quality labels. Despite these anomalies, the correlation between AC/OQ and the productive continuum was still significant for this speaker ($r^2 = 0.63$).

3.3.5 Relating the physical measure AC/OQ to the acoustic measure CPP

Previous studies [Fis67, KK90, SL90, CKS11] showed that an increase in glottal gap size results in a higher spectral noise level; and changes in OQ are related to changes in the shape of the harmonic source spectral shape (e.g., [Fan95]). Thus,

Table 3.3: Regression coefficients and r^2 values for linear regression analyses relating AC/OQ to the nominal voice quality continuum for each individual speaker. All values are significant at $p < 0.001$.

Speaker	Regression coefficients	r^2
F1	0.25	0.85
F2	0.27	0.80
F3	0.30	0.76
M1	0.25	0.63
M2	0.23	0.71
M3	0.37	0.80

Table 3.4: Mean and standard deviation (in parentheses) of AC/OQ with changes in the target voice quality for the 6 individual speakers.

Speaker	Breathy	Modal	Pressed
F1	0.25(0.03)	0.35(0.03)	0.46(0.04)
F2	0.24(0.04)	0.36(0.03)	0.42(0.02)
F3	0.25(0.04)	0.35(0.04)	0.41(0.01)
M1	0.28(0.02)	0.46(0.01)	0.45(0.02)
M2	0.24(0.05)	0.31(0.06)	0.44(0.04)
M3	0.32(0.02)	0.36(0.01)	0.46(0.03)

the changes in glottal configuration measured by AC/OQ should be manifest in the cepstral domain as the prominence of the cepstral peak, which can be quantified by the acoustic measure CPP ([HCE94]), reflecting the relative amounts of periodic and aperiodic energy in a voice signal. AC/OQ and CPP values for the 6 speakers on the first set of data are shown in Figure 3.7 ($r = 0.68, p < 0.001$). Note that the function relating these measures flattens out for large values of AC/OQ, for which OQ is small and glottal gaps are small or absent. When these conditions pertain, noise levels are relatively constant across stimuli, reducing variability in the CPP and consequently reducing the overall correlation between these measures.

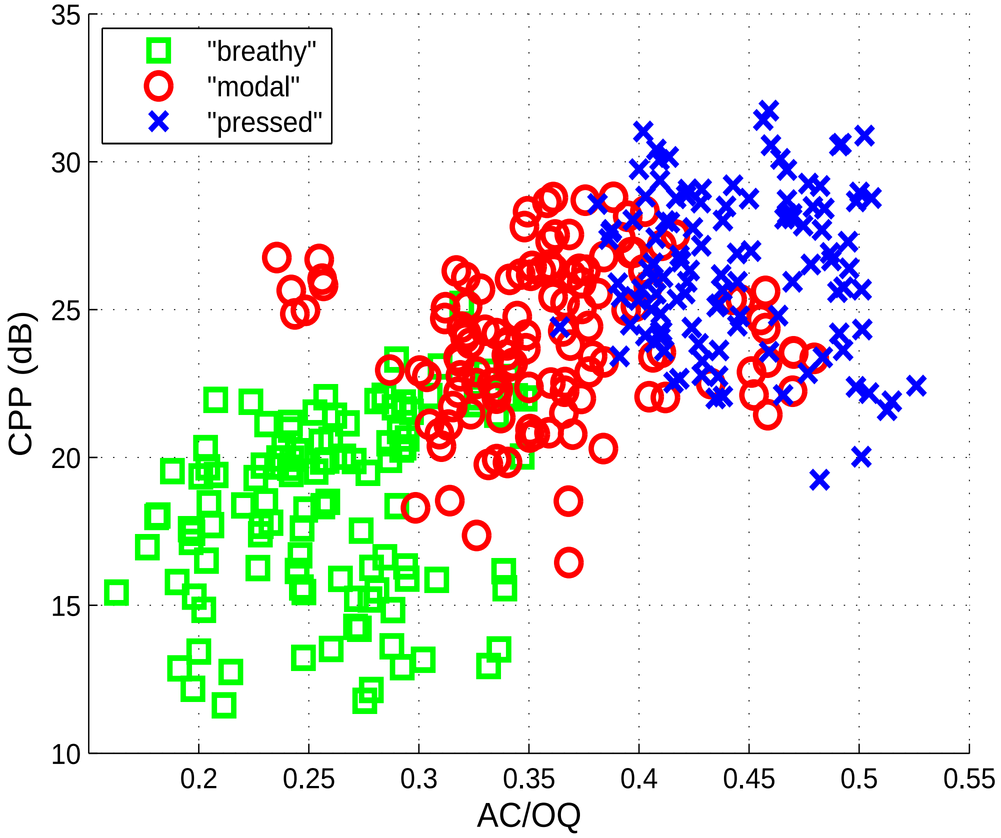


Figure 3.7: AC/OQ and CPP values with changes in the target voice quality (breathy, modal, and pressed) for the 6 speakers (F1, F2, F3, M1, M2, and M3).

3.3.6 Testing AC/OQ on glide phonations from breathy to pressed

The large interspeaker variability in how stimuli in different voice quality "categories" were produced limited our ability to validate the relationship between AC/OQ and changes in voice quality. To address this limitation, AC/OQ was further measured for the second set of high-speed recordings, in which 4 speakers (F1, M1, M4, and M5) produced voice quality glide phonations during which voice quality changed continuously from breathy to pressed within a single utterance. The glottal area waveforms for a glide phonation from speaker F1 are shown in Figure 3.8. As observed previously [KSC12], the DC component (i.e., the glottal gap size) gradually decreases during the first half of the recording (from cycle index 0 to 300 with incomplete glottal closures). During the second half of the recording (from cycle index 300 to 700, when glottal closure is complete), the OQ continuously decreased as quality became increasingly pressed.

The measures OQ, DC, AC, AC/OQ, and CPP during glide phonations from all 4 speakers are plotted in Figure 3.9. The glottal configuration is mainly characterized by a decrease in DC over approximately the first half of the recording, and by decreasing OQ during the phonatory phase with complete glottal closure (approximately the second half of the recording). This two-part physical process is captured by AC/OQ as illustrated in Figure 3.9. Recall that AC reflects the effect of DC in the presence of a glottal gap. Despite this effect of glottal gap, for all speakers, AC/OQ increased approximately steadily as quality changed from breathy to pressed, apparently capturing the waveform variation along the voice quality axis of breathy-to-pressed. Linear regression analyses modeling AC/OQ as a function of time show $r^2 = 0.96, 0.96, 0.80$, and 0.89 for speakers F1, M1, M4, and M5, respectively. AC/OQ is also strongly correlated with the CPP for these 4 speakers ($r = 0.86, 0.95, 0.77$, and 0.92, $p < 0.001$, for speakers F1, M1, M4, and M5, respectively). In this sense, AC/OQ appears to map between glottal vibratory patterns, acoustic consequences, and quality.

3.3.7 Discussion

PCA was applied to glottal area waveforms to investigate the factors that vary with voice quality, based on the assumption that perceptually-meaningful vibratory measures should quantify those aspects of vibration that correspond to differences in voice quality. Because the PCAs on which this measure is based weigh each sample of the glottal pulse equally and capture the dimensions with the largest variation, AC/OQ (which is based primarily on the first two principal components) mostly reflects the gross shape of the glottal pulse, which corresponds largely to the low-frequency part of the source spectrum [Ste98]. Listeners are

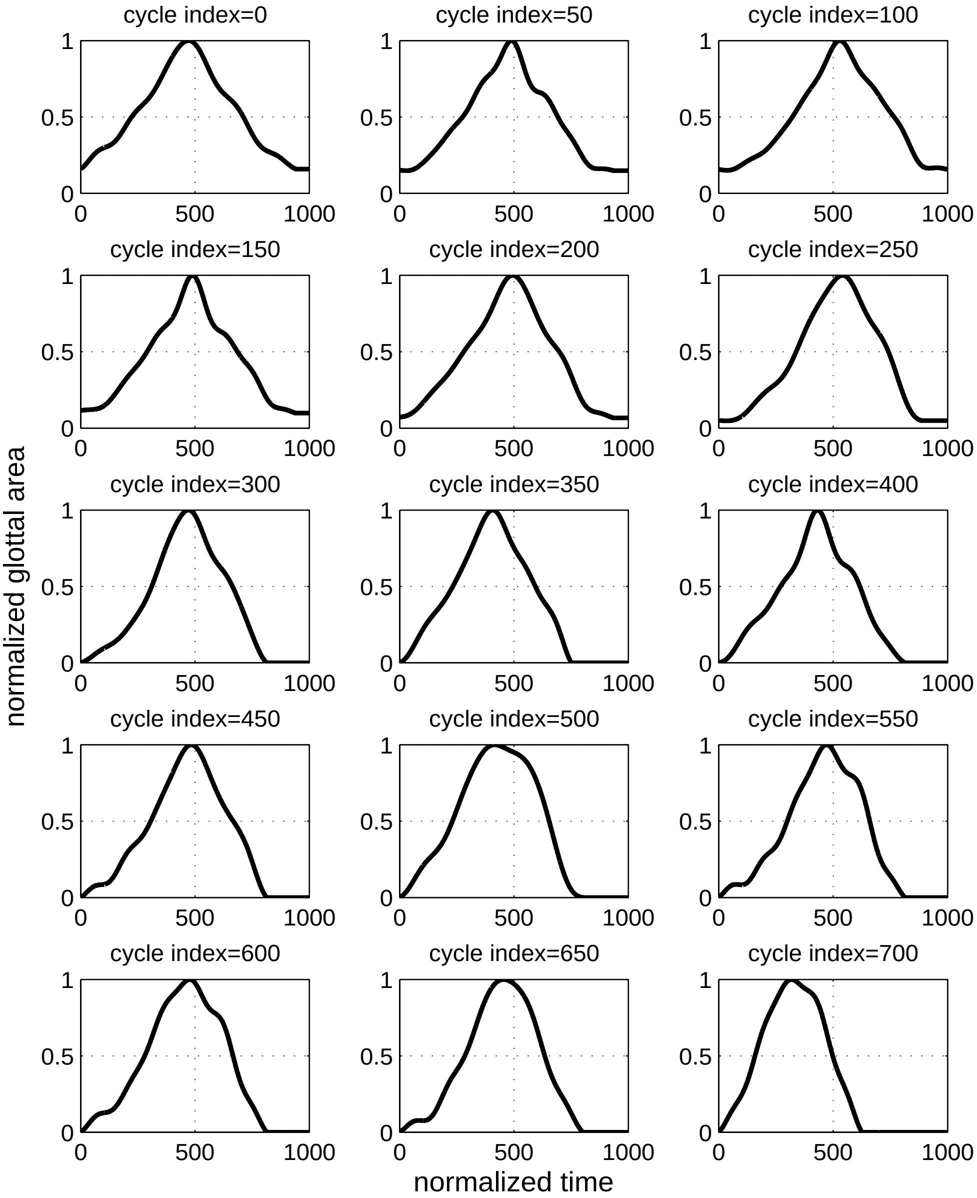


Figure 3.8: The glottal area waveforms during a voice quality "glide" phonation (from breathy to pressed) from speaker F1. The plots are sequential from left to right and top to bottom, according to cycle index numbers.

highly sensitive to the relative amplitudes of the lower harmonics [KGD10], which convey both paralinguistic information about a variety of personal and interper-

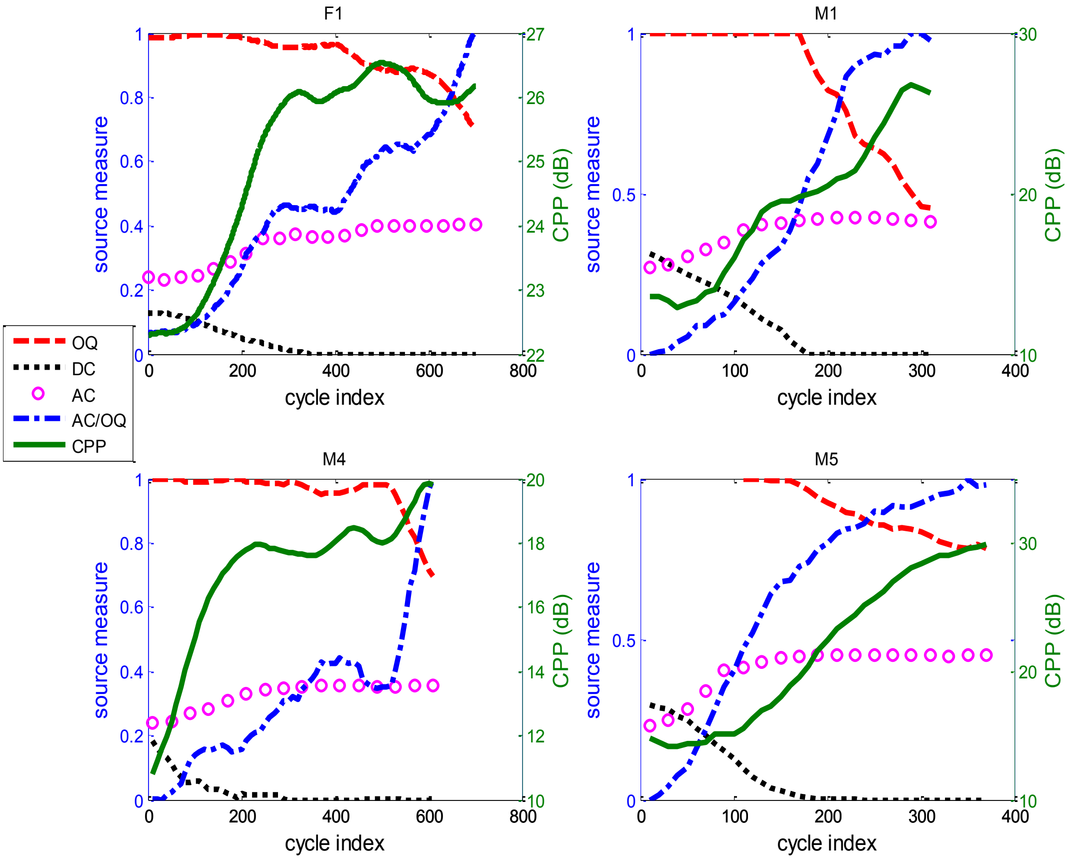


Figure 3.9: The voice source measures (OQ, DC, AC, and AC/OQ) and the acoustic measure CPP for voice quality "glide" phonations from breathy to pressed for 4 speakers (F1, M1, M4, and M5). For clarity, AC/OQ has been normalized to a maximum value of 1 and a minimum value of 0.

sonal attributes (see [KS11], for review) and linguistic information in languages like Gujarati [Fis67] and White Hmong [Huf87]. In this sense, AC/OQ potentially provides insight into how changes in glottal vibration patterns result in acoustic patterns that are perceptually salient.

The primary advantage of AC/OQ relative to existing source measures (and OQ in particular) is that it provides a unified framework for measuring the glottal area waveform along a voice quality axis of breathy-to-pressed, regardless of whether glottal closure is complete or not. Examination of changes in glottal area functions with changes in quality showed that the breathiest voice qualities were accompanied by glottal gaps which decreased in size with increasing pressedness. Only after the membranous glottal gap had completely disappeared did OQ begin to decrease with ongoing changes in voice quality. Despite this two-part physical process, AC/OQ is linearly related to continuous changes in quality and to the acoustic measure CPP, linking these three domains in a straightforward manner. These findings are consistent with results of Samlan and Story [SS11], whose computer simulation showed that increasing the separation between the vocal processes at maximum closure (controlled by a vocal fold adduction parameter) generally led to decreased harmonic energy and increased random (noise) energy, which resulted in decreased CPP. The two-way physical process (captured by AC/OQ) and CPP values observed in this chapter lend experimental support to the simulated relationship between kinematic (anatomical) parameters and acoustics measures in [SS11].

Additionally, glottal measures derived from the glottal flow (or the flow derivative) usually involve measuring the characteristics of the glottal closing phase (e.g., the negative peak of the flow derivative) because of its association with the main acoustic excitation of the vocal tract [Fan93]. AC/OQ captures the gross shape of the glottal area pulse, and thus is able to quantify the variation in glottal area waveforms. Finally, the calculation of AC/OQ does not rely on the sample value of the waveform at a single time instant (i.e., the negative peak of the waveform derivative). Thus, distortion of glottal area functions due to recording conditions or the area calculation algorithm does not significantly affect the accuracy of AC/OQ.

Measures of the glottal area waveform pulse skewness (speed quotient/closing quotient/asymmetry coefficient) have been linked with acoustic measures. Previous studies on this topic commonly made measures on glottal flow signals (e.g., [HdD01, HHP88]). However, recent studies using laryngeal high-speed videoendoscopy report varying levels of correlation between glottal area waveform skewness and acoustic measures. For example, Mehta *et al.* reported that the speed quotient of the glottal area function showed only weak correlation with spectral tilt measures [MZQ11]. Kreiman *et al.* reported that the relationship between the glottal area pulse skewness and $H_1^* - H_2^*$ depended on whether a glottal gap existed (see [KSC12], Table IV) and regression model parameters were also speaker dependent [KSC12]. Simulations using a computational model of the kinematics of vocal fold showed that speed quotient of the glottal area function was not a direct measure of the maximum area declination rate and had significant variability due to adduction, vertical difference, and glottal convergence [Tit06]. Measures of glottal area waveform skewness were initially tested in the current study but did not vary consistently with variations in voice quality and/or acoustic measures. Because the goal of this study was to investigate the aspects of the glottal area waveform that vary consistently with acoustic measures and voice qualities, the skewness measures were not included in the results.

3.4 Summary

This chapter investigated the aspects of the glottal area pulse shape that vary with voice quality, by using high-speed videoendoscopy of the vocal folds. A new measure, AC/OQ, was proposed to capture variations in glottal area pulse shapes in a manner that reflects both acoustic and perceptual consequences of those variations. This measure is defined as the AC component divided by OQ, so that an increase in glottal gap size (DC) or an increase in OQ results in lower AC/OQ values. Analyses of phonations differing both discretely and continuously in voice quality showed that across speakers AC/OQ values also increased monotonically along a breathy-to-pressed continuum. Thus, AC/OQ is capable of characterizing the continuum of glottal area waveform variation corresponding to a range of voice qualities, regardless of the existence or absence of glottal gaps.

CHAPTER 4

Analyzing the glottal gap effects with application to automatic gender classification of children's voices

Because voice signals result from vocal fold vibration, perceptually meaningful physiological vocal fold vibration patterns are associated with acoustic characteristics that humans can distinguish, such as the glottal gap (GG) discussed in Chapter 3. In this chapter, high-speed imaging data are analyzed to investigate the relationship between the GG area and acoustic measures for 6 subjects. Then, these acoustic measures are applied to a gender classification task of children's voices. Gender classification is relatively easy for adult voices, but is more challenging for children's speech. This chapter shows how understanding the role of voice source in speech production could benefit a practical application.

This chapter is based on the following publications:

- Gang Chen, Jody Kreiman, Yen-Liang Shue, and Abeer Alwan, "Acoustic Correlates of Glottal Gaps," Interspeech 2011, pp. 2673-2676.
- Gang Chen, Xue Feng, Yen-Liang Shue, and Abeer Alwan, "On Using Voice Source Measures in Automatic Gender Classification of Children's Speech," Interspeech 2010, pp. 673-676.

4.1 Analyzing the glottal gap effects in voice source

4.1.1 Data and Methods

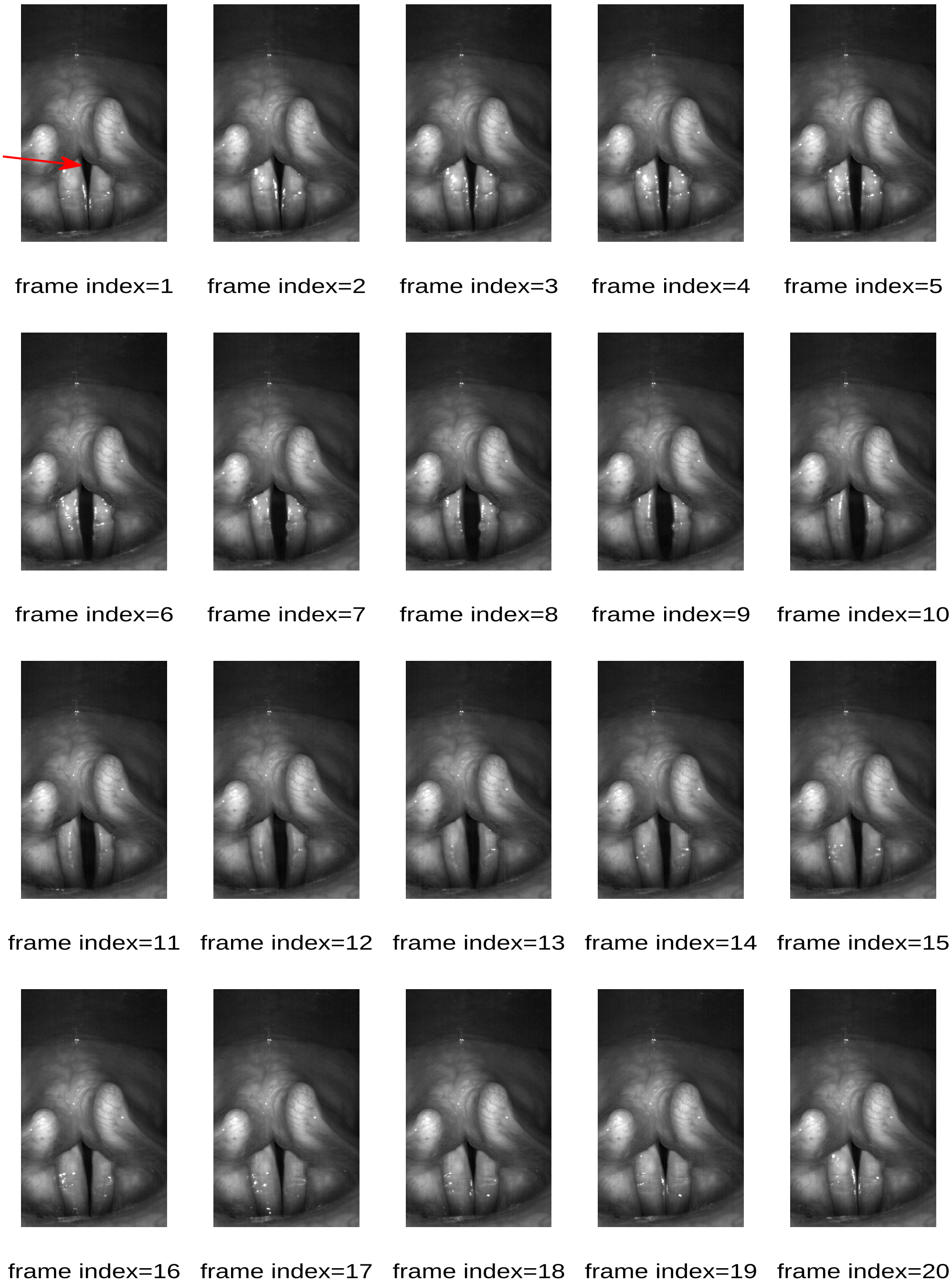
4.1.1.1 Data

The data used are the first set of recordings described in Chapter 3. Figure 4.1 shows consecutive high-speed images of a complete glottal cycle with a glottal gap. In this section, only phonations which exhibited GGs were selected for analyses. These phonations include: 16 out of 17 breathy phonations, 7 out of 33 non-breathy phonations, 8 out of 18 high-F0 phonations, 15 out of 32 normal and low-F0 phonations, 14 out of 26 phonations from female speakers, and 9 out of 24 phonations from male speakers.

4.1.1.2 Calculation of glottal measures

The glottal area waveform was calculated from the first 150 frames (50 ms) of each HSV recording. Values of OQ and DC were calculated for each glottal cycle in the glottal area waveforms using the methods described in Chapter 3. Recall that DC was defined as the minimum normalized glottal area in each glottal cycle, representing the GG size. Additionally, the asymmetry coefficient (AQ; [HdD01]) was calculated from the same glottal area waveform data by locating the first instants of glottal opening, the instants of maximum opening, and the onsets of maximum closure. AQ is defined as $t_o/(t_o+t_c)$, where t_o is the duration of opening phase and t_c is the duration of closing phase [HdD01].

Note that in [SCA10], the glottal area waveform of each individual phonation was averaged to obtain a single pulse which was representative of that particular phonation. In this section, each individual cycle of glottal area waveform generated one set of DC and AQ measures without averaging, allowing sufficient data for analyses of individual speakers.



frame index=1 frame index=2 frame index=3 frame index=4 frame index=5

frame index=6 frame index=7 frame index=8 frame index=9 frame index=10

frame index=11 frame index=12 frame index=13 frame index=14 frame index=15

frame index=16 frame index=17 frame index=18 frame index=19 frame index=20

Figure 4.1: High-speed images of a female (F1) speaker's breathy voice showing a complete glottal cycle with a glottal gap (from maximum closure to open to maximum closure). The glottal gap is denoted by an arrow in the first image. The plots are sequential from left to right and top to bottom, according to cycle index numbers. The posterior glottis is shown at the top of each image, with the anterior glottis at the bottom.

4.1.1.3 Acoustic measurements

Acoustic measures were calculated for each phonation and include $H_1^* - H_2^*$ and $H_1^* - A_3^*$ (related to the spectral tilt), CPP (related to the periodic structure of the source), and HNR between the frequencies 0–3.5 kHz (measuring the spectral noise level [Kro93]).

These measures were calculated using VoiceSauce software [Shu10b] at a resolution of 1 ms. In each glottal cycle, acoustic measures were averaged to match the DC and AQ of that cycle. Statistical analyses were performed using SPSS (v16.0). Tests where the null hypothesis had a probability of $p < 0.001$ were considered to be statistically significant.

4.1.2 Results

4.1.2.1 Experiment 1

In Experiment 1, DC was measured for all phonations with GGs.

Voice quality and F0 effects Table 4.1 lists means and standard deviations of DC for the three voice quality types and three F0 levels. Statistical analysis showed that breathy phonations had significantly larger DC than modal and than pressed ones. There was no significant difference in the DC value between modal and pressed phonations. High-F0 cases had significantly larger DC than normal-F0 and low-F0 cases. No significant difference was found in DC between normal and low-F0 cases.

Correlation analysis Table 4.2 lists the correlations between DC and various acoustic measures for the whole dataset. Correlation results for individual speakers are shown in Table 4.3.

In the presence of GGs, variation of OQ is small ($mean = 0.955$, $sd = 0.071$),

Table 4.1: Means and standard deviations (in parentheses) of DC for the three voice quality phonations and the three F0 levels averaged across all speakers.

Voice quality	Breathy	Modal	Pressed
DC	0.244 (0.135)	0.108 (0.077)	0.061 (0.005)

F0 level	Low	Normal	High
DC	0.121 (0.082)	0.139 (0.077)	0.348 (0.189)

suggesting that OQ may not be a good predictor of acoustic measures in the presence of GGs; hence, it is not surprising to see that OQ did not show a significant correlation with DC. AQ did not show a significant correlation with DC for the whole dataset, but showed a significant negative correlation in the individual analyses for all speakers except M3, suggesting that the way and/or degree of varying AQ could be speaker dependent. The earlier study [SCA10] on the same data showed a significant negative correlation between AQ and OQ ($r = -0.5546$, $p < 0.001$). Similar to OQ, the increase in GG size characterizes a more "open" glottic configuration and associates with the trend of decreasing AQ.

The noise measures CPP and HNR were negatively correlated with DC for the whole dataset, which was confirmed in individual analyses. Increased GG size allows more airflow, producing more aspiration noise. This result is consistent with earlier studies [Fis67, KK90], which suggested that an increase in GG size results in an increase in the noise floor of the speech spectrum.

Table 4.2: Correlation coefficients between DC and various acoustic measures (all speakers). All values are significant at $p < 0.01$.

Measures	F0	CPP	HNR	$H_1 - A_3$
Correlation with DC	0.491	-0.646	-0.330	-0.522

Interestingly, F0 showed a significant positive correlation with DC on the whole dataset; moderate correlations were also observed for all 3 female speakers. One explanation could be the increased stiffness of the vocal folds when increasing F0. A common approach to increase F0 is to increase activity in the cricothyroid muscle to stiffen the vocal folds. The increased tension could prevent the vocal folds from complete closure at high F0 [HKK88]. It has been reported [Lin92] that some individual elderly speakers tended to vary glottic configuration consistently with F0 level changes. In [SL90], the tendency of increasing degree of GG from habitual to high F0 for several female subjects was reported, but the effect of F0 was not significant for female subjects and no effect of F0 on degree of closure was found for male subjects. In that study, the lack of a significant effect of F0 may be attributed to the fact that the highest F0 was 262 Hz, while in our data F0 was as high as 330 Hz. In our study, a correlation analysis of female voices showed that DC is modestly correlated with F0 ($r = 0.356$, $p < 0.001$). Male subjects did not show a significant effect of F0 on DC. For the largest GGs, the GG extended to the membranous portion, resulting in an anterior-posterior gap configuration. For each speaker, the largest GG size is from the breathy phonation at a high F0. The increasing longitudinal tension upon the vocal ligaments would reduce the vibration amplitude, thus producing a larger GG. This statement is supported by similar findings among elderly [Lin92] and female speakers [SL90], since those speakers are generally assumed to be more breathy.

Surprisingly, moderate negative correlations between the spectral tilt measure $H_1^* - A_3^*$ and DC were observed for the whole dataset; strong negative correlations were also observed for speakers F1, F2, F3, and M1 in individual analyses. Thus, regression analyses were used to relate F0, CPP, and $H_1^* - A_3^*$ to DC. Results are shown in Table 4.4. The multiple linear regression models capture the acoustic consequences of DC well, in terms of R^2 values. Speaker F1, F2, F3, and M1 showed strong negative effects of DC on $H_1^* - A_3^*$. This is somewhat inconsistent with the hypothesis in [Han97] that larger GG would result in larger spectral tilt measures. This could be explained by the contribution of the breathy phonation at a high F0. During this falsetto-like phonation, the small-amplitude vibration of the vocal folds is associated with strong aspiration noise caused by the ultra-large GG which extends to the membranous portion of the glottis. The relatively weak harmonic structure is overwhelmed by the high noise floor in the spectrum, leading to a small spectral tilt. Thus a large GG produces a small $H_1^* - A_3^*$ value under this mechanism.

Table 4.3: Correlation coefficients relating DC to source parameters and acoustic measures for each speaker. “ns” denotes not significant. All other values are significant at $p < 0.01$. Measures which correlate with DC significantly for 4 or more speakers are shown.

Speaker	F1	F2	F3	M1	M2	M3
AQ	-0.551	-0.681	-0.933	-0.903	-0.701	ns
F0	0.725	0.807	0.566	ns	ns	0.697
CPP	-0.432	-0.770	-0.648	-0.947	-0.974	-0.631
HNR	ns	-0.608	-0.947	-0.827	-0.744	ns
$H_1^* - A_3^*$	-0.858	-0.969	-0.969	-0.958	ns	ns

Table 4.4: Standardized regression coefficients and R^2 values for multiple linear regression analyses relating DC to F0, CPP, and $H_1^* - A_3^*$ for each speaker. “ns” denotes not significant. All other values are significant at $p < 0.01$.

Speaker	F1	F2	F3	M1	M2	M3
F0	ns	0.296	0.164	ns	-0.262	0.818
CPP	-0.276	-0.135	ns	ns	-0.929	-0.670
$H_1^* - A_3^*$	-0.944	-0.665	-0.869	-0.886	ns	ns
R^2	0.828	0.962	0.956	0.972	0.975	0.949

4.1.2.2 Experiment 2: A glide phonation

Speaker F1 from Experiment 1 (a female phonetician) participated in this experiment. She gradually changed the phonation type from “breathy” to “pressed” while holding F0 and vowel quality steady. The duration of this utterance was 8 seconds. High-speed image and audio signals were recorded synchronously and analyzed. Glottal area waveform for the whole utterance was calculated from high-speed images. DC, OQ, and AQ were calculated from the glottal area waveform using the same method described in Experiment 1. Since our focus is the role of GG size, measures from the portion of the utterance with a GG were used for correlation and regression analyses. Direct observations showed that the GGs are situated in the membranous portion of the glottis and kept decreasing in size during the phonation.

Correlation analysis Table 4.5 lists correlation coefficients relating DC to source parameters and acoustic measures. As expected, DC correlated with AQ and the noise measures (CPP and HNR) negatively, confirming the results in Experiment 1. $H_1^* - H_2^*$ showed a strong correlation with GG size. It was reported in [KK90] that listeners were more likely to rate a phonation as breathy if an increase in $H_1^* - H_2^*$ was accompanied by noise. In this glide phonation, the transition from breathy to modal is characterized by a gradual decrease of GG size, associated with decreasing noise measures and $H_1^* - H_2^*$. The close correlation with both noise measures and $H_1^* - H_2^*$ suggests that GG size could be a physiological indicator of breathiness. The measures $H_1^* - H_2^*$ and DC for the glide phonation are shown in Figure 4.2. Both measures displayed similar falling trend.

The spectral tilt measure $H_1^* - A_3^*$ showed a strong positive correlation with GG size. This is consistent with the hypothesis in [Han97]. Note that F0 is within habitual range and kept steady, which is different from Experiment 1 where varying F0 level was predominantly affecting the physiological mechanism.

Table 4.5: Correlation coefficients relating DC to source parameters and acoustic measures. All values are significant at $p < 0.01$.

Parameters/Measures	AQ	$H_1^* - H_2^*$	CPP	HNR	$H_1^* - A_3^*$
Correlation with DC	-0.821	0.904	-0.788	-0.592	0.801

Relationship between $H_1^* - H_2^*$ and source parameters A high correlation between DC and $H_1^* - H_2^*$ ($r = 0.904$) was observed. Regression analyses relating DC, AQ, and OQ to $H_1^* - H_2^*$ were conducted. Standardized regression coefficients and the proportion of explained variance (R^2) are listed in Table 4.6. $H_1^* - H_2^*$ could be predicted well using only DC, with an R^2 of 0.817. $H_1^* - H_2^*$ could also be well predicted by AQ with an R^2 of 0.765, while the model is further improved by combining DC and AQ. The regression model using only OQ to predict $H_1^* - H_2^*$ had a low R^2 of 0.107. In the presence of GG, DC and AQ are better predictors of $H_1^* - H_2^*$ than OQ, suggesting the emergence of a new degree of freedom when GG persists.

Table 4.6: Standardized regression coefficients and R^2 relating DC, AQ, and OQ to $H_1^* - H_2^*$ for the glide phonation (speaker F1).

DC	AQ	OQ	R^2
–	–	0.329	0.107
0.904	–	–	0.817
–	-0.875	–	0.765
0.569	-0.408	–	0.871

4.1.3 Discussion

From the individual analyses in Experiment 1 and the glide phonation in Experiment 2, AQ showed a significant negative correlation with DC. An early study on the same data [SCA10] showed a similar trend between AQ and OQ: AQ is

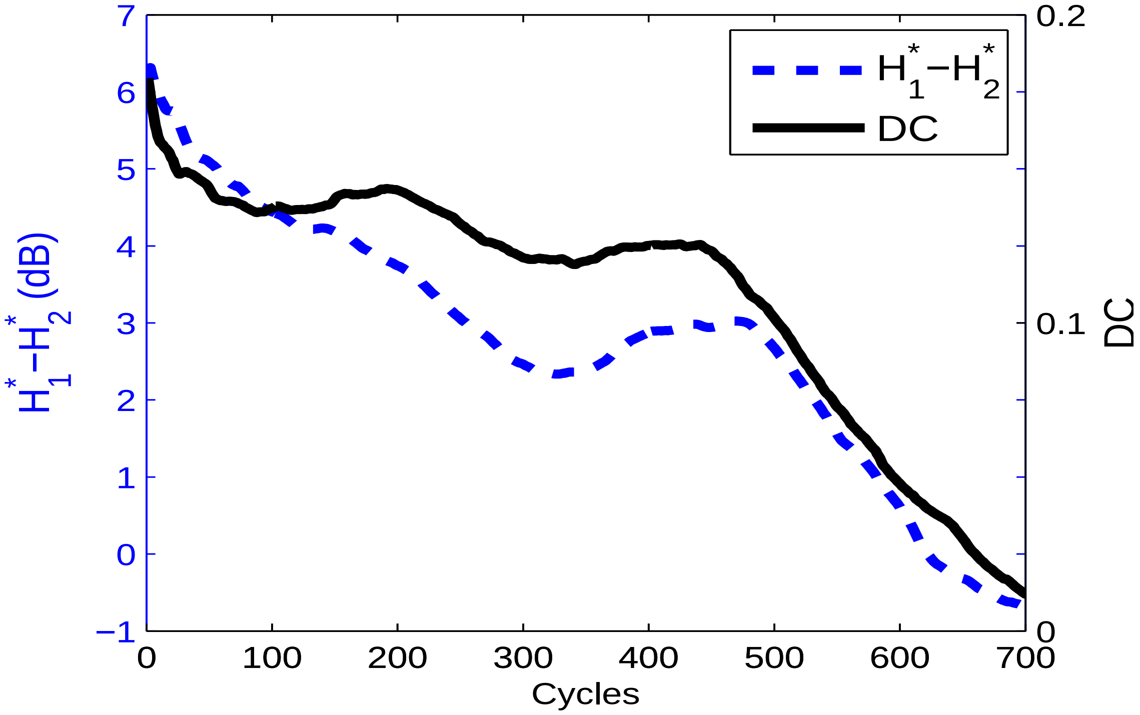


Figure 4.2: $H_1^* - H_2^*$ and DC for a glide phonation (breathy to pressed, female speaker F1). DC was normalized relative to the maximum glottal area in each cycle.

negatively correlated with OQ ($r = -0.5546$, $p < 0.001$). Similar to OQ, the increasing GG size characterizes a more "open" glottic configuration, associating with the trend of decreasing AQ. The regression model in Experiment 2 showed that DC and AQ are better predictors of $H_1^* - H_2^*$ than OQ in the presence of GG. When the correlation between OQ and $H_1^* - H_2^*$ decreases, DC emerges as a new degree of freedom in predicting acoustic measures.

The existence of GGs among female speakers acts as an additional degree of freedom, resulting in more variation in voicing acoustics. As stated in [HC99]: "the size of a posterior glottal opening can be considered to provide an additional degree of variability in the acoustic parameters considered."

The noise measures (CPP and HNR) are negatively correlated with DC, indicating the presence of more spectral noise with increasing GG area. F0 showed a positive correlation with GG size, especially among female speakers.

The positive correlation between GG size and spectral tilt measure $H_1^* - A_3^*$ in Experiment 2 supported the hypothesis in [Han97]; while a negative correlation was found in Experiment 1 under varying F0 level. This suggests that Hanson's hypothesis that larger GG would result in larger spectral tilt measures is only valid under a steady F0 constraint. In some phonatory modes, increasing F0 may reduce the amplitude of vocal folds vibration, increase GG size, and produce a lower spectral tilt due to significant aspiration noise, leading to a negative correlation between DC and the spectral tilt.

4.2 Automatic gender classification of children's voices

4.2.1 Introduction

The previous section investigated the acoustic correlates of the glottal gap effects. The noise measures (CPP and HNR) showed significant correlations with the glottal gap size. Because female speakers are more likely to exhibit glottal gaps than male speakers, these acoustic characteristics associated with glottal gaps could presumably be useful in distinguishing gender from speakers' voices. In this section, these voice source measures are explored in terms of their roles in gender classification of children's voices.

4.2.1.1 Differences between female and male speech

As stated in [WC91],"Generally, there exist three types of parameters: physiological and acoustical, which can be measured objectively, and perceptual, which is subjective but can be assessed psychophysically". Differences in acoustic properties between female and male speech signals are well known. The differences have been observed across different languages. Acoustic characteristics of speech signals differ between male and female voices due to physiological differences of the glottis and the vocal tract.

Typically, adult females have shorter vocal tract lengths (VTL) and smaller vocal folds than adult males. Fant showed that the ratio of total length of the female vocal tract to that of a male is about 0.87 [Fan76]. In [HKN83], the ratio of the length of the female vocal fold to that of the male is about 0.8. In [Tit87, Tit89], anatomical results showed that the female larynx differs from the male larynx in thickness, angle of the thyroid laminae, resting angle of the glottis, vertical convergence angle in the glottis, etc.

The gender-dependent physiological differences result in acoustical differences. This implies, according to the linear speech production theory, that females' shorter vocal tract length and smaller vocal folds would lead to higher formant frequencies and fundamental frequency (F_0), respectively. The differences in voice-source and vocal-tract related measures, such as fundamental frequency and formant frequencies are well documented. In [Lin89], results showed that the F_0 level of female speakers is approximately one octave higher than that of males. Hollien and Shipp found that the F_0 range of male speakers is 112-146 Hz [HS72], while the range for female speakers is 170-275 Hz [Sto81]. The nonlinear relationship between female and male formant frequencies has been well established, with female formants being, on average, 20% higher than those of males [PB52].

Previous studies have shown that glottal excitation provided gender discriminative information. Holmberg *et al.* [HHP88] studied voice-source related measures in normal, soft and loud voices produced by male and female subjects. Statistically significant differences between male and female voice-source related measures were found. In normal and loud voices, female glottal waveforms exhibited lower vocal fold closing velocity, lower AC flow, and a proportionally shorter closed phase of the cycle (indicating a steeper spectral slope). In soft voices, the spectral slopes of female and male speakers are similar [HHP88].

Differences also exist in perceptual factors, such as voice quality. In [KK90], two sentences by ten female and six male talkers were analyzed and results showed that, on average, females are more breathy than males among English speakers.

4.2.1.2 Gender classification

Automatic gender classification has applications in several areas. In automatic speech recognition and speaker identification, gender dependent models could be established if gender information is available; gender dependent models provide higher recognition/identification accuracies than gender independent models. In speech synthesis systems, an automatic gender recognition technique could assist the identifying which features are important for synthesizing male and female speech.

Previous studies on automatic gender classification from speech signals of adult speakers achieved high accuracy by using only features related to the fundamental frequency (F_0) and the first four formant frequencies [WC91]. This is mainly due to the well-known physiological differences between adult male and female speakers.

4.2.1.3 Characteristics of children's speech

Automatic gender classification from speech signals for children and adolescents remains a challenge because F_0 and formant frequencies are not easily distinguishable between boys and girls.

Existing studies of children's voices have mainly focused on the formant properties. In [BP95], the voices of children between the ages of 5 and 11 years old were studied. By using target words, which represented non-diphthong vowels in Australian English, the study was able to show that the value of the first three formant frequencies for girls were higher than those for boys, while boys have higher formant amplitudes than girls. In [LPN99], F_0, formant frequencies and measures related to the spectral envelope were studied as a function of age for speakers between ages 5 and 50. That study showed that the F_0 value dropped between ages 12 and 15 for males, and formant frequencies decreased between ages 10 and 15. With increasing age, male speakers also showed a faster decrease in formant frequencies than female speakers, and the formant frequencies after the decrease were lower for male than for female speakers. In [POA01], speech from children of ages 4, 8, 12 and 16 were studied; each age group consisted of 10 boys and 10 girls. An analysis of seven non-diphthong vowels of American English showed that the formant frequencies differentiated gender before 12 years of age, while formant frequencies along with F_0 differentiated gender after 12 years of age. These studies suggest the usefulness of F_0 as a distinguishing feature diminishes as the differences between F_0 for the two genders decrease.

Although vocal-tract related features, including formant frequencies and their amplitudes, have been studied to differentiate gender, the role of voice-source related measures (except for F_0) in gender classification have not be systematically investigated. The effects of age, gender and vowel dependencies on some measures related to the voice source were analyzed in [ISA07]. The measures were analyzed from the speech data of speakers between the ages of 8 and 39, and included: F_0, $H_1^* - H_2^*$ (related to the open quotient [HHP95]), and $H_1^* - A_3^*$ (related to spectral tilt [HHP95]). The asterisk indicates a correction for the influence of vocal tract resonances using the formula given in [IA04]. Results showed that $H_1^* - A_3^*$ continuously decreases between ages 8 and 39 by about 10 dB for males and decreases slightly by about 4 dB for females. It also suggested that $H_1^* - H_2^*$ drops about 5 dB at around age 15 for males but remains relatively unchanged for females. These differences motivated the study in [SI08] where acoustic measures from both the voice source and the vocal tract were used for automatic gender classification for speakers aged 8 to 39. It was found that the addition of two measures, $H_1^* - H_2^*$ and $H_1^* - A_3^*$, yielded the most consistent classification accuracy improvement when compared to the baseline (F_0 and formant frequencies). The results suggested that voice source measures could contain discriminative information for gender classification based on children's speech.

4.2.2 Data

Speech data are from the CID database [MLU96], produced by five age groups: ages 8–9, 10–11, 12–13, 14–15 and 16–17. Each recording was of the form: "I say uh, bVt again", where the vowel 'V' was /ih/, /eh/, /ae/ or /uw/. The vowel /iy/ in 'bead' was also used. Each speaker had, on average, 20 utterances of this form with different vowels. Table 4.7 shows the distribution of the utterances in terms of gender and age groups. The total number of male/female speakers is 174/140, and the total number of utterances is 3418. The steady state part of each vowel was extracted manually for analysis.

Table 4.7: Distribution of utterances in terms of gender and age (CID database).

Age group	No. of males/females	No. of utterances
8–9	48/36	810
10–11	48/33	807
12–13	38/34	708
14–15	22/21	413
16–17	18/16	680

4.2.3 Methods

The vocal tract parameters used in this section were the first three formant frequencies (F_1, F_2 and F_3) and the formant bandwidths (B_1, B_2). Also used were measures related to the voice source: F_0, CPP (related to breathiness [HCE94]), HNR (the harmonic-to-noise ratio [Kro93]) , $H_1^* - H_2^*$, $H_1^* - A_3^*$, and $H_2^* - H_4^*$ (the difference between the second and fourth source spectral harmonic magnitudes;

related to mid-frequency tilt [KGB07]). Additional measures used are amplitudes of the first three formant frequencies (A_1, A_2 and A_3).

HNR was calculated in the frequency band of 0-3500 Hz. The formant frequencies were estimated using the "Snack Sound Toolkit" software [Sj04] using the following settings: window length 25 ms, window shift 1 ms and preemphasis factor of 0.96. F_0 values were estimated using the STRAIGHT algorithm [KCP98]. The harmonic magnitudes, H_1^*, H_2^* and H_4^*, were calculated from the speech spectrum using the F_0 values obtained from STRAIGHT, and were corrected for the effects of the first two formant frequencies using the formula in [IA04]. Similarly, A_3^* were calculated from the speech spectrum using the formant frequencies obtained from Snack and were also corrected for the effects of the first two formants. A_3^* was additionally corrected for the effects of F_3. All measures were calculated using the "VoiceSauce" software [Shu10b] (`http://www.seas.ucla.edu/spapl/voicesauce/`).

For each classification experiment, 70% of the utterances were selected randomly for training and the remaining 30% of utterances were used for testing. Utterances from a particular speaker were used either for training or testing. Five experiments were conducted for each combination of acoustic measures and average accuracies were recorded. Note that for each utterance, acoustic measures were calculated frame by frame and then averaged over the utterance.

In this section, classification was done using an SVM classifier with a Radial Basis Function kernel. The LIBSVM toolkit [CL01] was used for training and testing. The results of the SVM classifier were compared with traditional MFCC features, using the first 12 MFCCs extracted from the vowel segment in each utterance. Due to the small number of vowels, training was implemented using 2 GMMs, each with 6 mixtures.

4.2.4 Analysis

F_0 and formant frequency values, averaged across all subjects for each age group, are provided in Table 4.8, and the means and standard deviations of F_0 for each group are shown in Figure 4.3. Results are similar to [LPN99, ISA07]. It is observed that F_0 for male and female speakers is not distinguishable for the age groups 8–9 and 10–11. The F_0 difference between male and female speakers becomes significant beginning from age 12, partly due to the drop in F_0 for male speakers between age 12 and 15 [LPN99].

Table 4.8: Mean and standard deviation (in parentheses) of fundamental frequency and formant frequency values for male and female speakers for each age group (in Hz)

Age group	Gender	F_0	F_1	F_2	F_3
8–9	female	267(40)	609(257)	2154(618)	3196(419)
	male	257(41)	578(236)	2109(620)	3170(428)
10–11	female	255(43)	629(242)	2242(532)	3170(411)
	male	253(41)	578(226)	2145(575)	3143(417)
12–13	female	239(33)	590(223)	2233(493)	3198(339)
	male	212(47)	546(207)	2093(469)	3053(383)
14–15	female	227(27)	594(198)	2113(409)	3002(334)
	male	150(45)	527(191)	2013(396)	2883(339)
16–17	female	223(25)	581(199)	2112(446)	3007(299)
	male	128(31)	490(193)	1952(361)	2804(347)

Values of CPP, HNR and $H_2^* - H_4^*$, averaged across all subjects for each age group, are provided in Table 4.9. The means and standard deviations of CPP are shown in Figure 4.4. It can be seen from the figure that the difference in CPP between male and female speakers is not significant for age group 8–9, which is relatively of the same scale as the difference in F_0. For age groups 10–11 and 12–13, however, the difference between male and female speakers in CPP increases, which is relatively larger than the difference in F_0. With increasing age, HNR and $H_2^* - H_4^*$ begin to differentiate male and female speakers from age 12

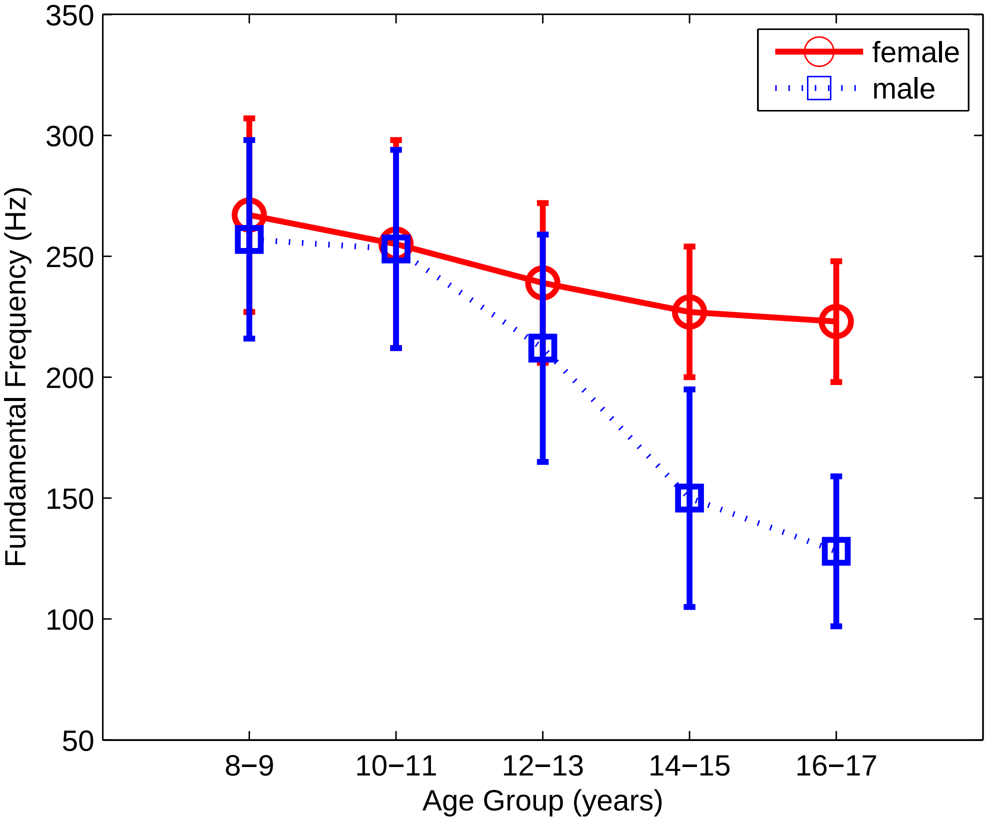


Figure 4.3: F_0 averaged across all subjects is shown for each age group

and 14, respectively; but the differences are overshadowed by the large standard deviations. This suggests that, the involvement of voice source measures, such as CPP, could improve gender classification accuracy for pre-adolescents, whereas F_0 values do not help differentiate between male and female children's speech.

4.2.5 Classification results

In this section, M0 is used to denote a feature set representing formant information (F_1, F_2, F_3, B_1 and B_2) and F_0, as in [SI08], and it is used as the baseline feature set. M1 is used to denote the feature set M0 with $H_1^* - H_2^*$ and $H_1^* - A_3^*$ from [SI08] which gave the best performance in that study. M2 denotes the feature set M0 with CPP. M3 denotes the feature set M0 with CPP and HNR. M4 denotes

Table 4.9: Mean and standard deviation (in parentheses) of CPP, HNR and $H_2^* - H_4^*$ values for male and female speakers for each age group (in dB)

Age group	Gender	CPP	HNR	$H_2^* - H_4^*$
8–9	female	22.42(3.00)	31.30(8.71)	2.61(7.80)
	male	22.59(3.04)	31.24(8.04)	3.09(7.52)
10–11	female	22.62(2.51)	31.25(7.84)	3.18(7.18)
	male	23.25(2.75)	30.14(7.39)	3.32(7.05)
12–13	female	22.97(2.28)	30.79(7.10)	2.96(6.59)
	male	23.88(2.93)	28.46(7.35)	3.90(6.47)
14–15	female	23.75(1.84)	31.57(7.47)	3.07(5.09)
	male	24.27(3.38)	25.06(8.05)	6.22(7.03)
16–17	female	23.21(2.25)	32.68(8.49)	2.41(5.49)
	male	24.68(2.40)	25.32(8.86)	6.21(5.86)

the feature set M0 with CPP, HNR and $H_2^* - H_4^*$. Table 4.10 summarizes these sets.

Table 4.10: Measure sets (M0-M4) used in the gender classification tests.

Set	M0	$H_1^* - H_2^*$	$H_1^* - A_3^*$	CPP	HNR	$H_2^* - H_4^*$
M0	✓					
M1	✓	✓	✓			
M2	✓			✓		
M3	✓			✓	✓	
M4	✓			✓	✓	✓

4.2.5.1 Results using additional voice-source related measures for each age group

Figure 4.5 compares gender classification accuracies from different measure sets. It can be seen from the figure that the classification accuracies of M4 are higher than the baseline and M1. Table 4.11 shows classification accuracies for each age group compared with the results for M0, M1 and also for the MFCC/GMM method.

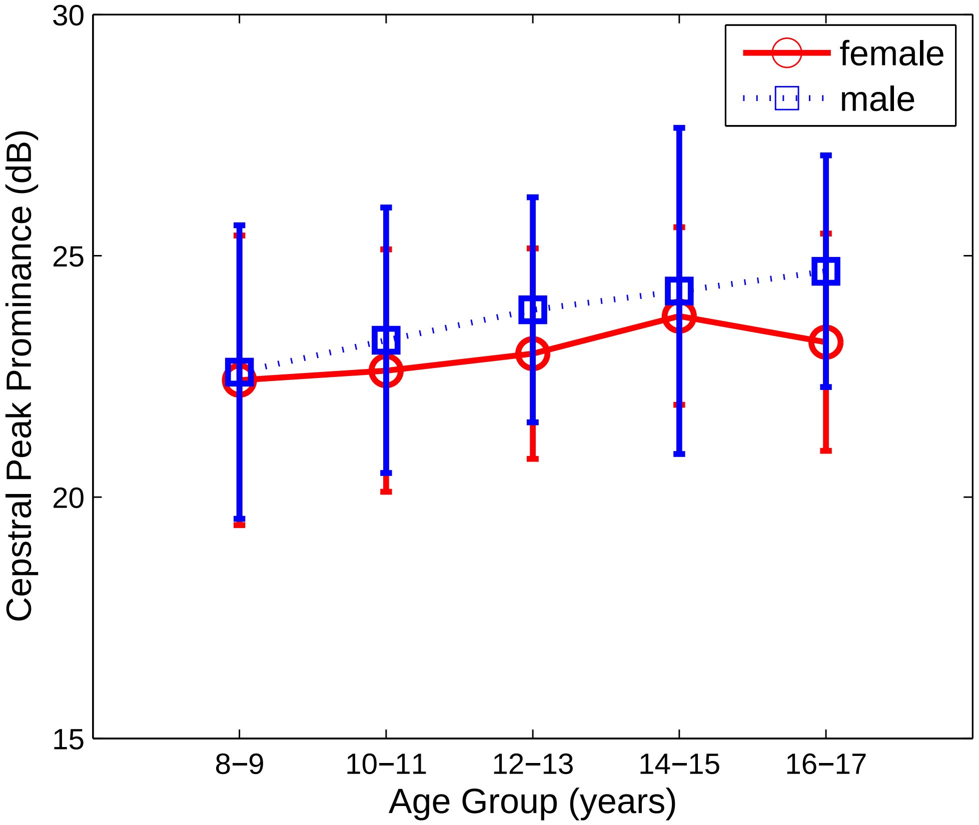


Figure 4.4: Cepstral Peak Prominence value averaged across all subjects is shown for each age group

It can be seen that the addition of voice source measures CPP, HNR, and $H_2^* - H_4^*$ constantly improved classification accuracies, compared to M0 and M1, for all age groups. An average of 3.2% improvement was achieved by adding measure CPP to the baseline set M0 (the M2 set). With the exception of age group 8–9, classification accuracies were further improved by adding the HNR. The change in classification accuracies by adding measure $H_2^* - H_4^*$ was not significant (the M4 set). The performance of voice source measures set M4 is about 4.4% higher than the result for M0 and about 3% higher than the result for M1. A large improvement of about 8% is obtained on the age group 12-13 when comparing M4 with M0.

Table 4.12 shows the classification accuracies of the M4 set for males and

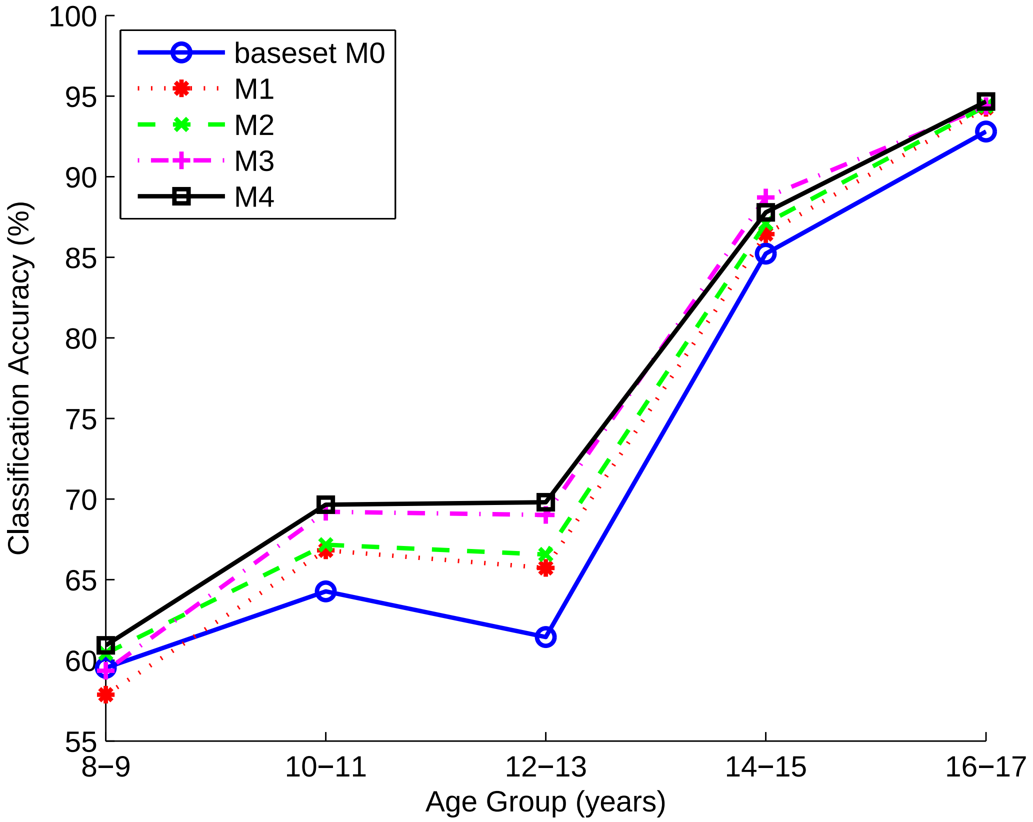


Figure 4.5: Gender classification accuracy for each age group using the measures sets M0, M1, M2, M3, M4.

females. Interestingly, the accuracy is higher for females than that of males for all age groups. Similar results were reported in [SI08].

4.2.5.2 Discussion

The results in Table 4.11 show that using CPP and HNR is useful in improving gender classification accuracy for children's speech. This suggests that the voice source measures CPP and HNR contain characteristics which are unique for young male and female speakers. Since CPP is highly correlated with breathiness [HCE94], the results confirm that, in general, females are breathier than males [KK90]. Interestingly, HNR is higher for females than males for all age groups.

Table 4.11: Gender classification accuracy for the different measurement sets (M0-M4) on each age group (in %). Boldface represents the highest accuracy for each age group

Age group	M0	M1	M2	M3	M4	MFCC/GMM
8–9	59.54	57.87	60.43	59.35	**60.93**	59.30
10–11	64.27	66.82	67.17	69.21	**69.66**	60.62
12–13	61.45	65.73	66.56	69.02	**69.81**	68.08
14–15	85.23	86.43	87.10	**88.71**	87.78	82.30
16–17	92.80	94.26	94.37	94.38	**94.66**	90.79

Table 4.12: Gender classification accuracy for the M4 feature set, using SVMs on each age group, distinguishing between males and females (in %).

Age group	8-9	10-11	12-13	14-15	16-17
M	56.66	69.37	67.47	87.30	93.38
F	65.32	69.88	72.13	88.19	95.95

The difference in HNR between females and males increases with increasing age. This is inconsistent with the expectation that females should have lower HNR values than males, since in general females are more breathy than males [KK90]. This result requires further exploration on what signal property contributed to the high HNR of females. As stated in [Kro93]: "All kinds of signal properties may result in a noise-like appearance of the spectrum, such as a perturbation of the excitation signal (jitter and shimmer), rapid directional changes in fundamental frequency, formant transitions, and so forth." A possible explanation for these results could be the interaction effects of the noise level perception. A recent study [GK10] showed that listener's perception of noise levels in voice depends on the shape of the harmonic spectrum; but the interaction effects of voice quality on perception are not well understood.

The measure $H_2^* - H_4^*$ also assisted in improving the classification accuracies, with the exception of age group 14–15, suggesting that the mid-frequency tilt also differentiates between the male and female speech spectra. A large improvement of about 8% is obtained for the age group 12–13 when comparing M4 with M0. This could be attributed to the fact that puberty of males and females begins at around 11 [LPN99], an age when emerging gender-dependent physiological differences result in acoustical differences of children's voices. Adding other measures, such as formant amplitudes, didn't improve the classification accuracy significantly. For age group 8–9, the classification accuracy for all measure sets are below 61%. The improvement by adding features CPP, HNR and $H_2^* - H_4^*$ is not significant.

Considering the performance of all age groups, the addition of measures CPP, HNR and $H_2^* - H_4^*$ improved classification accuracy by 4.4% compared with the baseline feature set. When compared with M1, the feature set M4 provides about 3% improvement (on average) for all age groups. While the performances of feature set M4 are similar to the MFCC/GMM results for age groups 8–9 and 12–13, the classification accuracies for M4 is about 9%, 5% and 4% higher for age groups 10–11, 14–15 and 16–17, respectively.

4.3 Summary

In the first section of this chapter, HSV data were used to investigate the relationship between glottal gap area, source parameters, acoustic measures, and voice quality for 6 subjects. Results showed that CPP and HNR were affected by glottal gap area, indicating the presence of more spectral noise with increasing glottal gap area. Analysis of a glide phonation from breathy to pressed for one female speaker showed that the measures $H_1^* - H_2^*$ and $H_1^* - A_3^*$ were positively correlated with GG area under a steady fundamental frequency (F0). In some phonatory modes, increasing F0 may reduce the amplitude of vocal folds vibration, increase GG area, and produce a lower spectral tilt due to significant aspiration noise, leading to a negative correlation between GG area and the spectral tilt measure $H_1^* - A_3^*$.

In the second section, measures related to the voice source were applied to gender classification of children's voices and the results were compared with those in [SI08]. The experiments were done using the CID database which consisted of 3418 utterances spoken by 174 male and 140 female subjects. Measures related to the voice source and vocal tract were extracted from 5 target vowels and used in gender classification tests. The feature set consisting of F_0, the first three formant frequencies (F_1, F_2 and F_3) and the first two bandwidths (B_1 and B_2) were used as baseline feature set (M0). Features were added to the baseline set to test their effect on gender classification. Results show that adding the three measures CPP, HNR and $H_2^* - H_4^*$ yielded best overall performance, suggesting that measures related to breathiness and mid-frequency tilt carry discriminative information for automatic gender classification. The accuracy improvements of adding voice source features were highest for age group 12–13. After age 13, the accuracy improvements of adding voice source features decreased as the role of F_0 becomes more prominent.

CHAPTER 5

A new voice source model with application to voice source estimation in noise

An effective and robust method to estimate the voice source from speech signals is desirable for several applications. An accurate voice source model will help improve the performance of voice source estimation. In this chapter, a new voice source model (denoted EE1) and a noise-robust automatic source estimation algorithm are proposed. The source signal is estimated using a codebook search approach. Glottal area waveforms extracted from high-speed images are converted to glottal flow waveforms to calibrate the proposed algorithm. Results in both clean and noisy conditions show that the novel approach is relatively robust in accurately estimating the glottal flow waveform.

This chapter is based on the following publication:

- Gang Chen, Yen-Liang Shue, Jody Kreiman, and Abeer Alwan, "Estimating the voice source in noise," Interspeech 2012, pp. 1600-1603.

5.1 Background and introduction

As reviewed in Section 1.2.1, many models have been proposed to represent the voice source. In a previous study [SKA09], the LF model was used for source estimation. Some inconsistencies exist between the open quotient (OQ) estimated from the acoustic signal and OQ measured from high-speed imaging of the vocal folds, suggesting that a modification of the LF model may be necessary for accurately modeling the observed vibration of the vocal folds. In [SA10], the SA source model was proposed based on high-speed imaging of the larynx in order to provide greater glottal pulse shape flexibility than the LF model. Results showed that the SA model provided more accurate source estimation than the LF model [SA10].

According to the linear speech production model [Fan70], speech signals are generated by filtering the voice source with the vocal tract transfer function (VTTF). Generally, there are two types of approaches in source estimation. The first method relies on estimating the VTTF explicitly and then uses it to inverse-filter the speech signal. The residual signal obtained from inverse filtering is then fitted by a source model [Alk03, KKG10, MC04]. Inverse filtering typically requires estimating the formant frequencies accurately. However, the widely used LPC-based formant trackers are known to be inaccurate for high-F0 phonations, and estimating formant frequencies in noisy conditions remains far from robust. The inaccuracy in VTTF estimation would lead to inevitable inaccuracy in source estimation. In the second approach to source estimation, the voice source and the VTTF are estimated jointly and iteratively [FMS01, MC04], where the source estimation error due to the inaccurate VTTF estimation is compensated for by searching a wide range of source-filter combinations. The convergence of the iterative source estimation approach is sensitive to the initial estimation of the VTTF, which is far from reliable in noise. Synthesized speech and electroglottograph (EGG) signals recorded from natural speech have been used as references to evaluate the source estimation algorithms. However, the EGG signal is related to the contact area of the vocal folds, and thus does not provide an accurate shape of the glottal source signal.

In a recent study [SA10], glottal area waveforms obtained from high-speed recordings of the vocal folds were used as a reference to evaluate a source estimation algorithm. In that study, the glottal area was assumed to represent the glottal flow. As discussed earlier in Section 1.5.1, the glottal area does not fully represent the glottal flow (e.g., the glottal flow pulse has a notable skewing rightward in time [Ste98, Rot81]). The relationship between the glottal area and the glottal flow signal was quantitatively modeled using the three-mass vocal fold model in [ST95, TS02]. In this chapter, the glottal area obtained from high-speed imaging is converted to glottal flow using a three-mass model. The resultant glottal flow signal is used as the reference source signal to evaluate the accuracy of the proposed source estimation algorithm.

The source estimation method in [SA10] required estimating formant information. LPC-based formant estimators remain far from robust in noisy conditions, while manually-derived formants are impractical. In this chapter, a noise-robust automatic source estimation algorithm is proposed. This algorithm does not rely on explicitly estimating the formant frequencies to inverse-filter the speech signal. The source signal is estimated using a codebook search approach, and the method is a modified version of [SA10].

5.2 Data

The data used in this chapter are the first set of recordings described in Chapter 3. Gaussian white noise was added to the audio signal to test the robustness of the source estimation algorithm. Three SNR levels were used: 20 dB, 10 dB, and 5 dB. Recall that in Chapter 3, each glottal cycle was included for analysis without averaging. In this chapter, the glottal area waveform was averaged across the glottal cycles to produce a single-cycle waveform which was representative of the 150 frames (50 ms) analyzed for that utterance. In order to evaluate the proposed source estimation method, the OQ was calculated from the averaged glottal area waveform as the time from the first opening instant to the beginning of maximum closure (or minimum area), divided by cycle duration.

5.3 Method

5.3.1 Area to flow conversion

Glottal areas extracted from high-speed images were converted to glottal flow by using the Matlab toolkit LeTalker [Sto12]. LeTalker is a Matlab GUI version of the three-mass vocal fold model originally published in [ST95] and updated in [TS02]. Parameters such as muscle activation level and respiratory pressure can be specified as inputs to calculate the glottal area, the glottal flow, and the resultant speech signal. Both subglottal and supraglottal (vocal tract) systems are included to simulate their interactions with the vocal folds. In this chapter, the vocal tract shape was set to that of the vowel /i/ according to vocal tract area functions reported in [STH96] and all the other parameters were set to the default values in LeTalker when converting the glottal area to glottal flow.

Figure 5.1 shows an example of the glottal area extracted from high-speed recording and the resultant glottal flow calculated from LeTalker. As expected, due to the inertia of the air column [Rot81], the glottal flow pulse is notably skewed rightward in time, as noted by previous researchers [Ste98, Rot81].

5.3.2 The proposed EE1 source model

5.3.2.1 EE1 model parameters

The proposed EE1 model is a modified version of the SA model proposed in [SA10]. The model has five parameters: the fundamental period (T_0), open quotient (OQ), asymmetry coefficient (α), speed of the opening phase (S_{op}) and speed of the closing phase (S_{cp}). An example of a model waveform is shown in Figure 5.2. t_o and t_c are the durations of the opening and the closing phase, respectively.

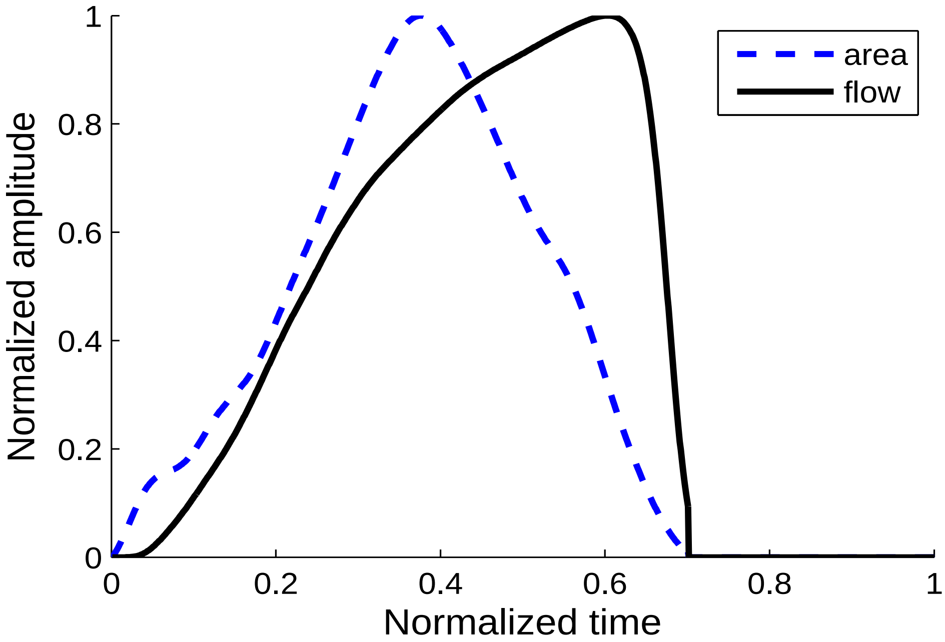


Figure 5.1: An example of a glottal area waveform extracted from high-speed images and the resultant glottal flow waveform calculated using LeTalker (speaker F1).

Using the notation from this figure, $OQ = \frac{t_o+t_c}{T_0}$, $\alpha = \frac{t_o}{t_o+t_c}$, S_{op} is the waveform amplitude at the bisect instant of the opening phase, and S_{cp} is the waveform amplitude at the bisect instant of the closing phase. With the exception of T_0, the four other parameters all range from 0 to 1.

Mathematically the proposed EE1 model $u(t)$ is defined as:

$$u(t) = \begin{cases} f(\frac{t}{t_o}, \lambda_{Sop}) & 0 \leq t \leq t_o \\ f(\frac{(t_o+t_c-t)}{t_c}, \lambda_{Scp}) & t_o < t \leq t_o + t_c \\ 0, & t_o + t_c < t \leq T_0 \end{cases} \tag{5.1}$$

where

$$\lambda = 12 \cdot (0.5 - S) \tag{5.2}$$

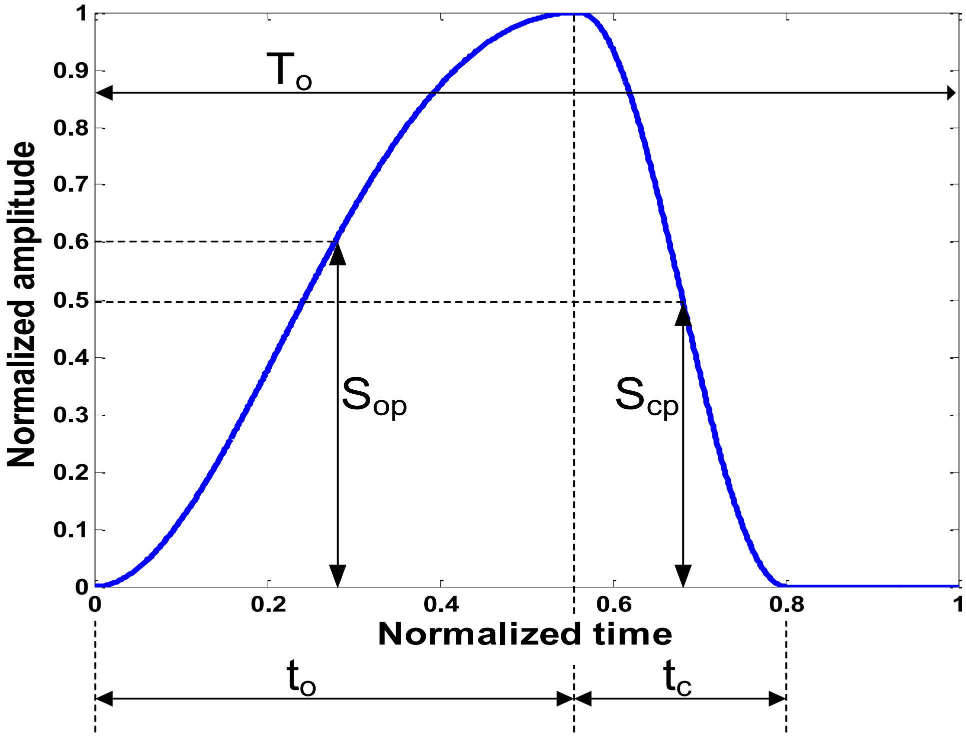


Figure 5.2: Example of the proposed EE1 model with $T_0 = 1$, $OQ = 0.8$, $\alpha = 0.7$, $S_{op} = 0.6$, and $S_{cp} = 0.5$.

$$f(x, \lambda) = \frac{1}{\pi(e^{\lambda} + 1)} \{e^{\lambda x}[\lambda sin(\pi x) - \pi cos(\pi x)] + \pi\} \tag{5.3}$$

λ is an intermediate slope parameter which controls the slopes of the waveform in the opening and the closing phase. λ_{Sop} and λ_{Scp} are the λ values when $S = S_{op}$ and $S = S_{cp}$ respectively. As shown in the equations above, given the five input model parameters (T_0, OQ, α, S_{op}, and S_{cp}), the intermediate slope parameter λ needs to be calculated in order to generate the output source waveform.

5.3.2.2 Properties of the proposed EE1 model

This modified model simplified the computational complexity of the model in [SA10] by using the amplitude measures S_{op} and S_{cp}. In [SA10], a time-consuming intermediate optimization step was required to calculate the slope parameter λ

given $S = S_{op}$ or $S = S_{cp}$ (see Equation 1.5). In EE1, an approximate closed form solution of λ exists, as shown in Equation 5.2. The output source waveform can be calculated directly without the intermediate optimization step. The computation time of calculating the model waveform given the model parameters was reduced by 90%, on average. In addition, a closed form derivative domain representation of the proposed model exist with the five parameters as inputs, which allows for a wide category of optimization methods to be applied when performing model fitting.

5.3.3 Source estimation procedures

The source estimation process is illustrated in Figure 5.3. In this method, a codebook is generated by the proposed source model (EE1). The harmonic magnitudes of the input acoustic signal are calculated and normalized to the first harmonic magnitude (the n-th normalized harmonic magnitude is denoted as S_n). The derivative of the source codebook entries are calculated to account for the radiation effect of the lips. The magnitudes of each codebook entry derivative are calculated in the same way (the n-th normalized harmonic magnitude is denoted as U_n). The vocal tract shape is obtained by subtracting the source harmonics from the acoustic signal harmonics ($S_n - U_n$). The residual signal is used for a constrained nonlinear optimization. A 3-formant VTTF is used here. In summary, for each entry in the source codebook, the following is performed:

$$E \quad = \quad \text{minimize} \sum_{n=2}^{N} (S_n - U_n - V_n)^2 \cdot W_n \tag{5.4}$$

$$\text{subject to} \quad F_1 < F_2 < F_3 \tag{5.5}$$

where V_n is the n-th harmonic magnitude of the VTTF represented by three

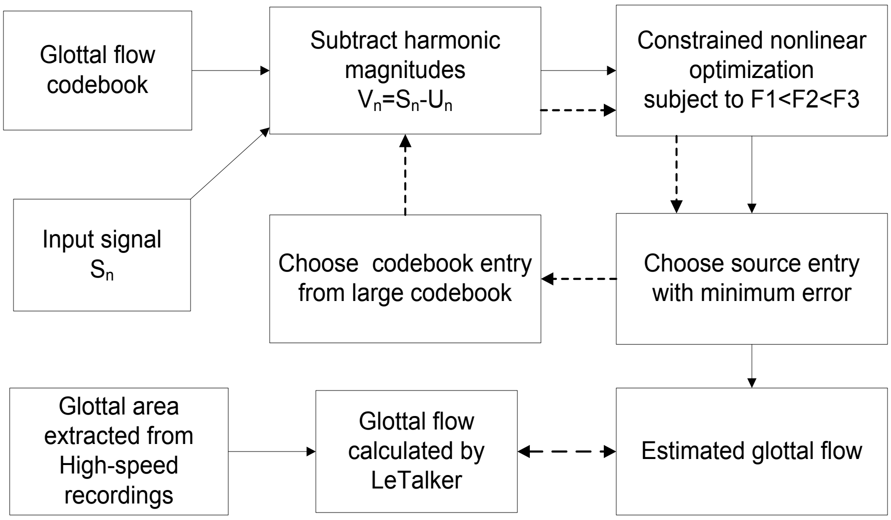

Glottal flow codebook
Subtract harmonic magnitudes $V_n=S_n-U_n$
Constrained nonlinear optimization subject to F1<F2<F3
Input signal S_n
Choose codebook entry from large codebook
Choose source entry with minimum error
Glottal area extracted from High-speed recordings
Glottal flow calculated by LeTalker
Estimated glottal flow

Figure 5.3: Flowchart of the proposed source estimation algorithm

formants F_1, F_2, and F_3. Bandwidth values are based on the formant-bandwidth mapping formula in [HM95]. W_n is the weighting function and is empirically chosen as

$$W_n = \begin{cases} 2^{12-n} & 2 \leq n \leq 12 \\ 1, & n > 12 \end{cases} \tag{5.6}$$

The value of the error term E is recorded with the source entry. After searching the entire codebook, the source entry with the minimum error E is selected.

Note that in [SA10], formant information is required for source estimation as an input, while no explicit formant information is needed in the proposed approach. It is well known that formant estimation in noisy conditions remains far from robust and LPC-based formant trackers have deficiencies for high F_0 phonations. Thus, it is desirable to develop source estimation algorithms without relying heavily on the accuracy of formant estimation. The formant frequencies and the source signal are searched and evaluated jointly, rather than determining the formant frequencies to inverse-filter the speech signal. The optimal combination of the formants and the source signal is the final output. Although the recorded data only contain the vowel /i/, the proposed algorithm is also suitable for other vowels.

As in [SA10], two iterations of the search algorithm were used to reduce the computational complexity. The first iteration used a small codebook to search for the source parameters. The small codebook was generated by varying the OQ and α in the following way: vary OQ from 0.4 to 1.0 with an increment of 0.1; α from 0.5 to 0.9 with an increment of 0.1; and S_{op} and S_{cp} were set to a constant value of 0.5. This generated a source codebook with 35 entries. The source entry selected from the first iteration was used for finding the final source entry in the second iteration from a larger codebook. The larger codebook was generated by the following setting: OQ from 0.35 to 1.00 at an increment of 0.01, α from 0.5 to 0.9 at an increment of 0.1, S_{op} from 0.4 to 0.6 at an increment of 0.1, and S_{cp} from 0.4 to 0.6 at an increment of 0.1. Once the first iteration returned the codebook entry with $OQ = OQ_s$ and $\alpha = \alpha_s$, the second iteration searched part of the larger codebook with an OQ value within $[OQ_s - 0.1, OQ_s + 0.1]$ and an α value within $[\alpha_s - 0.05, \alpha_s + 0.05]$.

For each audio recording, the first 50 ms segment (corresponding to the first 150 frames of high-speed recording) was processed and F_0 was extracted using the Straight algorithm [KCP98] with 25 ms window size and 1 ms window shift. F_0 was then averaged for the first 50 ms segment. The harmonic magnitudes were calculated based on the averaged F_0. A Hamming window consisting of 4 pitch periods was used to calculate the spectrum of the input signal. The harmonic magnitudes were calculated in the range of 0-2600 Hz. This range is associated with the number of harmonics that can be reliably estimated from the spectrum. The window step size was 10 ms and the source estimation procedure was performed for each window. The final source waveform and OQ were obtained by averaging across the estimated source waveforms and OQ values over the first 50 ms segment.

5.4 Results

In order to evaluate the proposed source estimation algorithm which incorporated the EE1 model into the codebook design, it was compared to other state-of-the-art algorithms. Firstly, the software toolkit Aparat [Air08] was used as a reference to obtain inverse-filtered source signals. Aparat is a Matlab implementation of the Iterative and Adaptive Inverse Filtering (IAIF) algorithm [Alk92]. It allows for manual parameter tuning to improve automatic inverse-filtering results. The initial inverse-filtered waveform is shown in the GUI, and parameters can be manually adjusted to minimize ripples in the inverse-filtered time waveform. However, as noted in Section 5.1, inverse filtering results are sensitive to the initial estimation of VTTF, which could degrade significantly in noise. The first 50 ms of each audio recording were inverse filtered. The cycle boundaries of the resultant glottal flow signal were marked and an average waveform was obtained by averaging across the cycles.

The source estimation approach [SA10] where formant frequencies were estimated from the Snack toolkit [Sj04] was also used for comparison. The source model and the codebook in that study are updated as described in Section 5.3.2.

Because not all aspects of the source are equally important perceptually, the error metric in evaluating source estimation algorithms has to be chosen carefully. This section first presents the source estimation results in terms of Mean Square Error (MSE) between the estimated source waveform and the reference glottal flow waveform. Although it is realized that the overall fit in MSE does not explicitly imply perceptual similarity, these MSE values serve as the preliminary evaluation of source estimation performance. The perceptual importance of various aspects of the source pulse shape will be investigated in Chapter 6. In addition, source estimation performances are evaluated in terms of estimated OQ values versus reference OQ values (from high-speed images). As noted previously in Section 5.1, OQ is an important source measure that has been associated with a quality dimension ranging from “pressed” to “breathy” [KK90]. These analyses examined utterances for which voice quality was systematically varied by experienced speakers to represent continua from breathy to pressed and from low to high F0. Assessing algorithms across a balanced set of voice quality, F0, and SNR values demonstrates the potential applicability of the proposed algorithm in real-life scenarios.

Table 5.1 shows source estimation results in terms of MSE between the estimated source waveform and the reference glottal flow waveform. Three estimation methods are shown: “Proposed” denotes the method proposed in this chapter, “Previous” denotes the estimation method in [SA10], and “Aparat” denotes manual inverse filtering using Aparat. Note that all the source waveforms are normalized both in time and amplitude for MSE calculation. Each waveform is 1000 samples in time with a maximum amplitude of 1 and a minimum amplitude of 0.

Table 5.1: Results of the source estimation in terms of waveform MSE averaged across all the recordings (in %). Best results are shown in boldface.

	Clean	20 dB	10 dB	5 dB
Proposed	**6.6**	**6.8**	**8.4**	**9.6**
Previous	6.9	7.9	10.2	11.7
Aparat	8.8	9.2	11.7	13.7

In the clean condition, the differences in MSE among the three methods are not significant (pairwise t-tests, $p > 0.01$). The performance of the proposed approach is significantly better than that of the previous approach and Aparat in noisy conditions (pairwise t-test, $p < 0.01$), and the performance improvement increases with SNR level, partially due to the inaccuracy of formant estimation under noise.

Table 5.2 shows the MSE averaged within each phonation type and F0 level

Table 5.2: Results of the source estimation in terms of waveform MSE for each phonation type and F0 level (in %). Best results are shown in boldface.

		Phonation type			F0 level		
		Breathy	Modal	Pressed	Low	Normal	High
Clean	Proposed	5.7	**6.8**	**7.5**	6.8	6.3	**6.8**
	Previous	**3.1**	8.1	10.2	**5.1**	**5.5**	9.9
	Aparat	5.8	8.5	13.1	7.8	9.9	8.5
20 dB	Proposed	5.9	**7.0**	**7.7**	7.0	**6.5**	**6.9**
	Previous	**4.1**	9.2	11.1	**6.2**	6.6	10.8
	Aparat	6.1	8.8	13.6	8.4	10.1	8.9
10 dB	Proposed	7.6	**10.2**	**7.8**	**7.5**	**7.8**	**9.7**
	Previous	**6.9**	11.6	12.8	9.2	10.2	11.8
	Aparat	11.1	11.6	12.7	11.9	11.3	12.0
5 dB	Proposed	8.8	**11.5**	**9.1**	**7.7**	**8.9**	**12.0**
	Previous	**8.4**	13.1	14.3	10.7	12.1	12.2
	Aparat	12.5	13.9	15.1	14.1	13.1	14.0

in clean and noisy conditions. In the clean condition, the proposed approach is better than the other approaches only for modal phonations, pressed phonations, and high-F0 cases. In the 20 dB SNR condition, the proposed approach is better than the other approaches for modal phonations and pressed phonations, as well as normal-F0 cases and high-F0 cases. In both 5 dB and 10 dB SNR conditions, the proposed approach is better than the other methods for each phonation type and F0 level, with the exception of breathy phonation. The MSE of the previous approach is 3.1% higher than that of the proposed approach for high-F0 cases in the clean condition (pairwise t-test, $p < 0.01$), highlighting the inaccuracies of LPC-based formant estimators for high-F0 voices.

Table 5.3 shows the OQ estimation error of the proposed method for each phonation type and F0 level, and for each gender. Recall that OQ ranges from 0 to 1. In clean and noisy conditions, the highest OQ estimation error occurs for high-F0 cases (two-sample t-tests, $p < 0.01$), and pressed phonations have the highest OQ estimation error than the other two types (two-sample t-tests,

Table 5.3: The OQ estimation error of the proposed method for each phonation type and F0 level, and for each gender.

		Phonation type			F0 level		
		Breathy	Modal	Pressed	Low	Normal	High
Clean	Male	.035	.072	.107	.025	.053	.082
	Female	.083	.049	.155	.045	.098	.148
20 dB	Male	.043	.078	.110	.028	.057	.083
	Female	.086	.069	.178	.066	.109	.152
10 dB	Male	.055	.084	.114	.031	.060	.083
	Female	.089	.089	.195	.088	.117	.156
5 dB	Male	.064	.092	.120	.035	.063	.084
	Female	.092	.108	.207	.104	.123	.161

$p < 0.01$). On average, the OQ estimation error is higher for females than males (two-sample t-tests, $p < 0.01$). One possible explanation is that women's voices have more tracheal coupling and source/tract interactions, which are not modeled in the linear speech production model.

Figures 5.4 and 5.5 show the reference and estimated glottal flow waveforms for a female subject (denoted F1), in clean and 5 dB SNR (white noise) conditions. As discussed above, these two figures roughly illustrate that the largest OQ estimation error occurs for pressed phonations, where the OQ value tends to be overestimated compared to the reference OQ.

5.5 Summary

This chapter presents a new glottal flow model (EE1) and a noise-robust source estimation method inspired by an earlier study [SA10]. The source signal was estimated using a codebook search approach. The glottal area extracted from high-speed images was converted to glottal flow to calibrate the proposed algorithm. Results in both clean and noisy conditions showed that the proposed algorithm is robust in accurately estimating the glottal flow waveform.

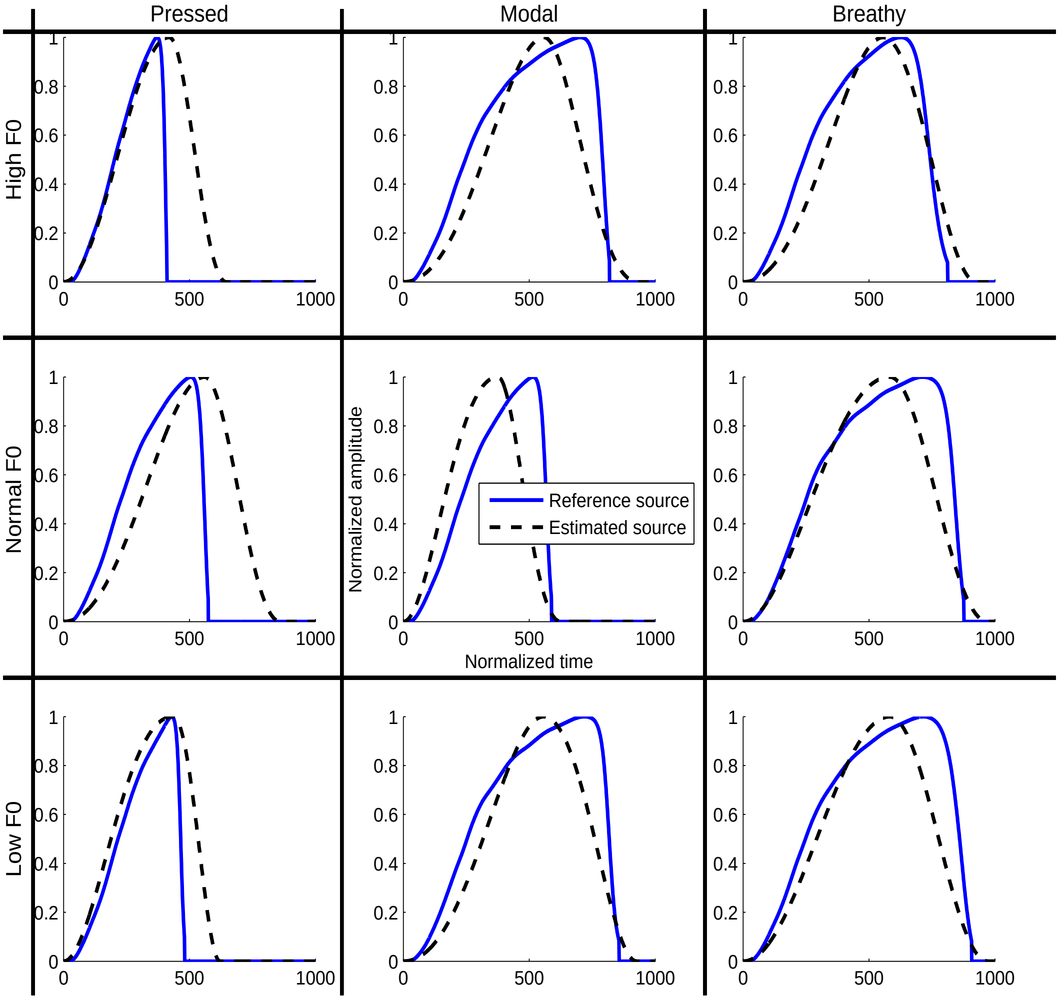


Figure 5.4: Plots of the reference (solid line) and estimated (dashed line) glottal flow waveforms for subject F1, in the clean condition. Estimation was based on the EE1 model.

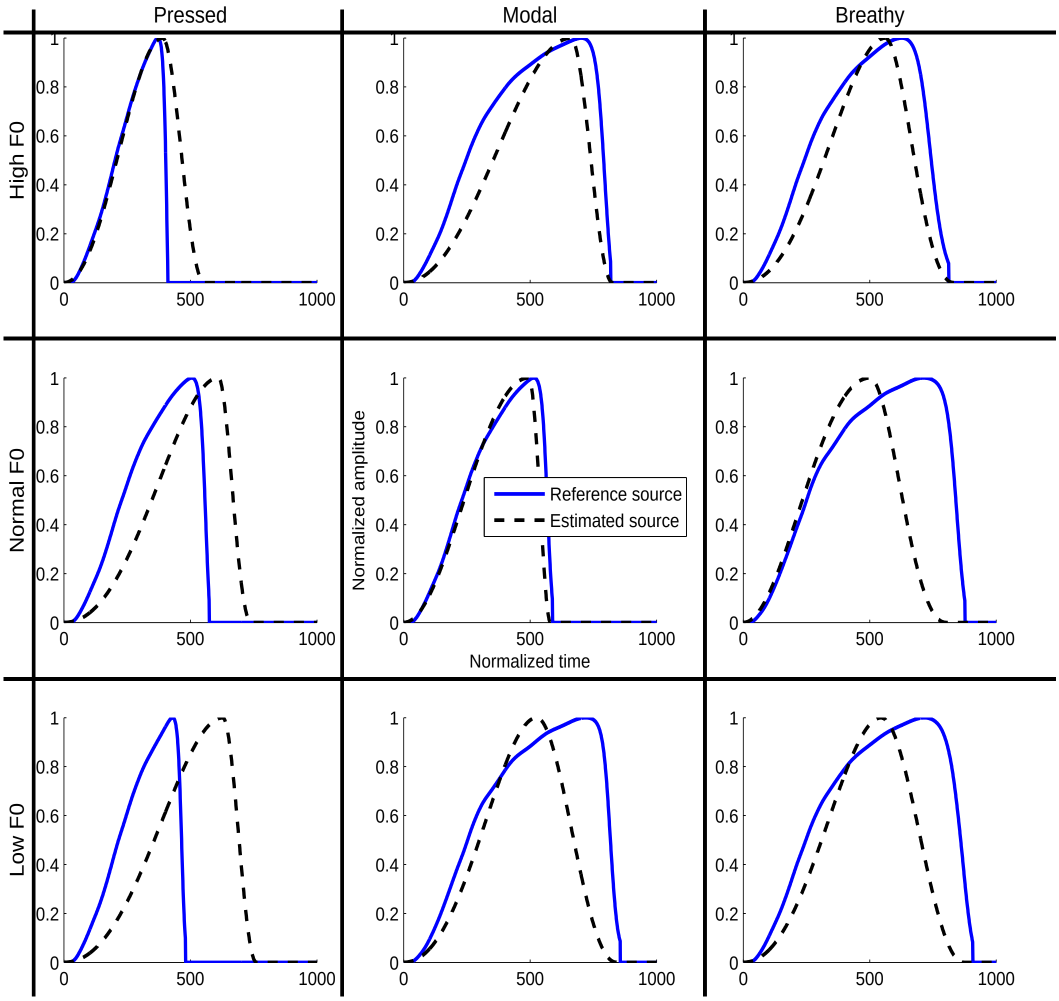


Figure 5.5: Plots of the reference (solid line) and estimated (dashed line) glottal flow waveforms for subject F1, in 5 dB SNR (white noise) condition.

CHAPTER 6

A perceptually and physiologically motivated voice source model with application to vowel synthesis

According to the linear speech production model [Fan70], speech signals are generated by filtering the voice source by the vocal tract transfer function. An important application of voice source modeling is speech synthesis, where it is important to capture perceptually-important aspects of the source signal to generate natural-sounding synthetic voices. In this chapter, a new voice source model (denoted EE2), motivated by data from high-speed laryngeal videoendoscopy, is proposed to capture perceptually-important source shape aspects. This new model, along with four other source models, is fitted to 40 voice sources (20 male and 20 female, denoted M1-M20 and F1-F20, respectively) obtained by inverse filtering and analysis-by-synthesis (AbS) of samples of natural speech. The synthetic voices are then evaluated via perceptual experiments.

This chapter is based on the following publication:

- Gang Chen, Marc Garellek, Jody Kreiman, Bruce R. Gerratt, Abeer Alwan, “A perceptually and physiologically motivated voice source model,” Interspeech 2013, pp. 2001-2005.

6.1 Introduction

As reviewed in Section 1.2.1, many source models have been proposed, and research efforts have also been devoted to studying the perceptual importance of changes in source waveform shapes. In [Ros71], listening tests using a variety of glottal excitations showed that simulated excitations with a single slope discontinuity at closure were perceived as more natural-sounding, while very small opening or closing times (or opening times approximately equal to or less than closing times) were not preferred. In [CL91], the LF model and a turbulent noise generator were used to synthesize four voice quality types (modal, vocal fry, falsetto, and breathy). Perceptual experiments showed that these four voice quality types could be characterized by four parameters: pulse width, pulse skewness, the abruptness of glottal closure, and turbulent noise. Only 3 listeners participated in the experiments. In other approaches, voice source waveforms were parameterized to capture variations in voice quality [ABV02, CKG13, KG13, AA07, IIH11, SCA10, CKS11], while those characteristics related to vocal intensity were investigated and parameterized in [BAV02, SY09, SFM05, AAB06].

Few studies have attempted to systematically validate glottal source models perceptually, and model development has focused more on replicating observed pulse shapes. As a result, it is unclear which (if any) deviations from perfect fits between models and data have perceptual importance. In [KGC12], the Ros, FL, LF, SA, and EE1 source models were fitted to 40 natural voice sources (20 male and 20 female) obtained by inverse filtering and analysis-by-synthesis (AbS), subject to the mean square error (MSE) criteria for which each point of the waveform was weighted equally. Synthetic copies of the voices were used in a perceptual experiment which showed that the match at the negative peak of the flow derivative was the most perceptually-important among source parameters. Informal listening tests using several tokens showed that a significant mismatch to the opening

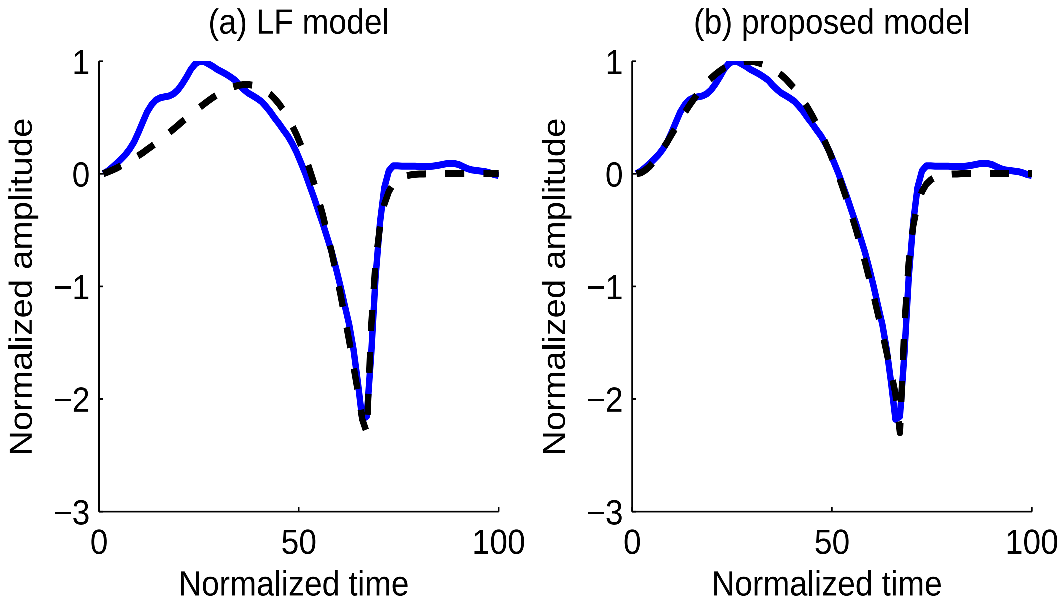


Figure 6.1: An example of fitting the LF and the proposed EE2 models to the same AbS source pulse subject to MSE criteria. Solid line: AbS source. Dashed line: model-fitted source.

phase (see Figure 6.1 (a) for an example) resulted in a noticeable perceptual difference between the target and modeled stimuli. These results indicate the need for a source model with increased flexibility to provide a close fit to all parts of the voice source signal, especially the opening phase.

In this chapter, a new voice source model, motivated by data from high-speed laryngeal videoendoscopy, is proposed to capture perceptually-important voice source shape aspects. This model is then evaluated in comparison to 4 existing source models, using both MSE and perceptual distances.

6.2 Voice source modeling

6.2.1 The proposed EE2 model

The proposed EE2 model is based on the models in [SA10, CSK12], which investigated shapes of glottal area waveforms extracted from laryngeal high-speed videoendoscopy. The model is a combination of sinusoidal and exponential functions shown to be effective in approximating a wide range of glottal flow pulse shapes. The model is then refined using AbS to eventually capture the shapes of the glottal flow derivative, as the LF model does. The model has six parameters:

- The time of the positive peak (t_i)
- The shape of the opening (S_1; amplitude of the waveform at $t_i/2$)
- The time of the peak flow (t_p; zero-crossing of the flow derivative)
- The time of the negative peak (t_e)
- The amplitude of the negative peak (E_e)
- The slope of the return phase (t_a)

The latter four parameters (t_p, t_e, E_e, and t_a) were originally defined in the four-parameter LF model [FLL85]. The first two parameters were added in the proposed EE2 model to provide an additional degree of freedom, so that the timing of the positive peak and the shape from the start to the positive peak can be manipulated directly, independent of the negative peak of the flow derivative. The parameters are perceptually-motivated, as mentioned in Section 6.1. With these parameters, the glottal opening phase could be modeled more accurately. Recall that our previous studies showed that a significant mismatch to the opening phase could lead to a noticeable perceptual difference between the target and the modeled stimuli. An example of a model waveform is shown in Figure 6.2. Given the six parameters described above, the glottal flow derivative $u(t)$ is defined as:

$$u(t) = \begin{cases} f(\frac{t}{t_i}, \lambda_1) & (0 \leq t \leq t_i) \\ [f(\frac{(2t_e - t_i - t)}{2(t_e - t_i)}, \lambda_2) - 1]\frac{12(1+E_e)}{6+\lambda_2} + 1 & (t_i < t \leq t_e) \\ \frac{-E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c - t_e)}] & (t_e < t \leq 1) \end{cases} \tag{6.1}$$

where

$$f(t,\lambda) = \frac{1}{\pi(e^{\lambda}+1)}\{e^{\lambda t}[\lambda sin(\pi t) - \pi cos(\pi t)] + \pi\} \tag{6.2}$$

$$\lambda_1 = 12 \cdot (0.5 - S_1) \tag{6.3}$$

$$\lambda_2 = \arg\min_{\lambda} \left| f(\frac{2t_e - t_p - t_i}{2(t_e - t_i)}, \lambda) - \frac{12E_e + 6 - \lambda}{12(E_e + 1)} \right| \tag{6.4}$$

$$\epsilon = \frac{1}{t_a}[1 - e^{-(t_c - t_e)/t_a}] \tag{6.5}$$

t_c is the time of closure. In practice it is convenient to set $t_c = 1$, i.e., the complete fundamental period [FLL85]. ϵ, λ_1, and λ_2 are intermediate parameters. As illustrated in Figure 6.2, the proposed parameters can be easily derived from the inverse-filtered differential glottal waveform, and directly control the shape of the glottal waveform. Unlike the LF model, which describes the open phase ($0 < t < t_e$) using one function, the proposed EE2 model uses two functions ($0 < t < t_i$ and $t_i < t < t_e$) to describe the open phase, allowing for more flexibility in modeling. Figure 6.1 (b) shows an example of constraining the proposed EE2 model to fit the negative peak of the flow derivative precisely, while still achieving satisfactory fittings in other parts.

6.2.2 Model fitting

In this chapter, each of the 40 target AbS-derived source functions (described in Section 6.3.1) was fitted with 5 source models: the Ros, LF, SA, EE1, and the proposed EE2 model. The FL model, which provided the worst fit to the target sources in our previous experiment, was excluded from further experiments. First-derivative representations were calculated mathematically for the Ros, SA, and EE1 models, which describe flow pulses in the time domain, so that all models were fitted to the target AbS source functions in the flow derivative domain. One cycle of the AbS source signal for each speaker was normalized to a maximum

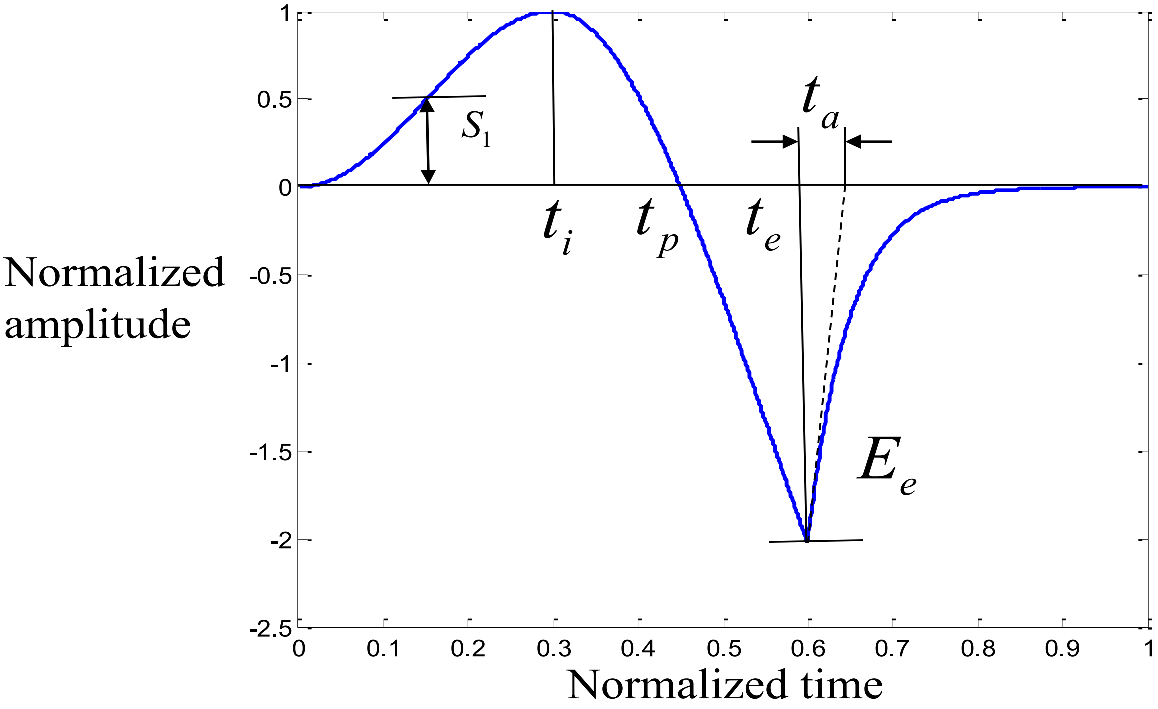


Figure 6.2: An example of the proposed EE2 model with $S_1 = 0.5$, $t_i = 0.3$, $t_p = 0.45$, $t_e = 0.6$, $E_e = 2$, and $t_a = 0.05$.

amplitude of 1. Each derivative-domain model was fitted to all of the AbS source functions using MSE criteria, for which each point of the waveform was weighted equally. Additionally, the proposed EE2 model was fitted a second time to the AbS source function with the constraint of exactly matching the "landmarks"—defined as:

- The first point $(0, 0)$
- The positive peak of the flow derivative $(t_i, 1)$
- The maximum flow (the zero-crossing of flow derivative) $(t_p, 0)$.
- The negative peak of the flow derivative $(t_e, -E_e)$

These landmarks are shown in Figure 6.3. The resulting model will be referred to as the EE2-LM model. This procedure was included in order to assess the per-

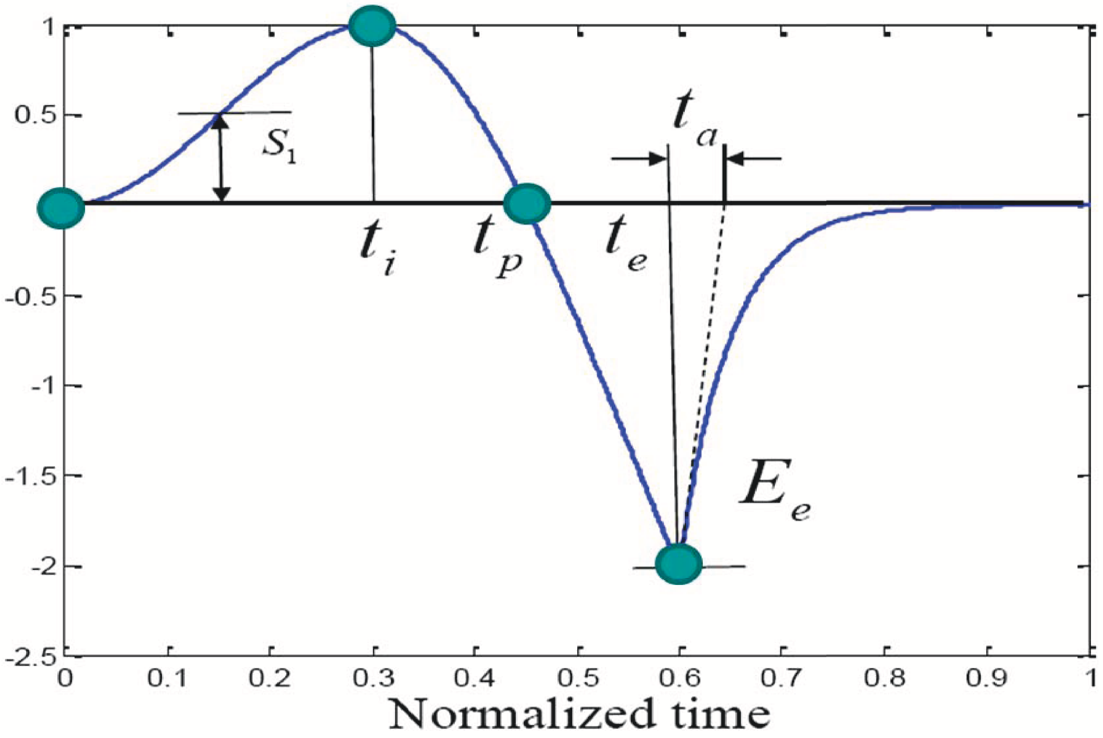


Figure 6.3: An example of the proposed EE2-LM model with landmarks shown in filled circles.

ceptual importance of the landmarks of the voice source signal. Note that it is not always possible to exactly match **ALL** landmarks for the other models, due to constraints inherent in the models and their parameters. Because of the increased flexibility, especially in modeling the opening phase, the proposed EE2-LM model is able to match all landmarks well. Target AbS source pulses and the corresponding fitted sources using the proposed EE2-LM model for six different speakers are shown in Figure 6.4 (see Appendix B for model fitting results of all 40 speakers). As this figure shows, the proposed EE2-LM model is able to approximate a wide range of pulse widths, pulse skewnesses, and abruptnesses of glottal closure. Because this model fitting is a non-linear optimization problem and suboptimal solutions might be found using standard optimization methods, model fitting was implemented using a codebook search schema (exhaustive search) similar to that in [CSK12] in order to achieve nearly optimal solutions. The codebook of each

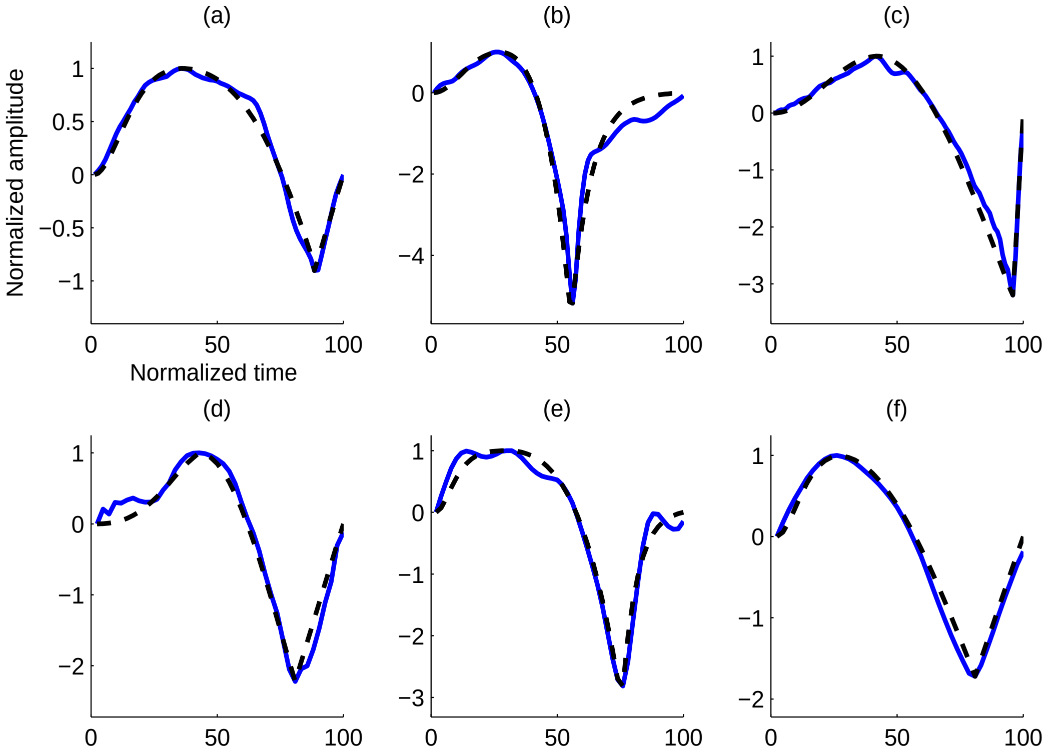


Figure 6.4: Target AbS source pulses and the corresponding fitted sources using the proposed EE2-LM model for six different speakers. Panels (a), (b), and (c): male speakers M3, M6, and M18. Panels (d), (e), and (f): female speakers F8, F9, and F20. Solid line: AbS source. Dashed line: the proposed EE2-LM model.

model has a size of 10000.

6.2.3 Model fitting results

Table 6.1 shows MSE values for fit, of each of the source models under study, to the target AbS sources. (See table caption for the meaning of model labels.) A two-way repeated measures ANOVA (model by speaker gender) showed significant main effects of model [$F(5, 190) = 12.99, p < 0.0001$] and gender [$F(1, 38) = 8.71, p < 0.01$] on mean MSE, as well as a significant model-by-gender interaction effect [$F(5, 190) = 4.27, p < 0.001$]. Tukey post-hoc t-tests (with Bonferroni adjustment for multiple comparisons) indicated that no cross-

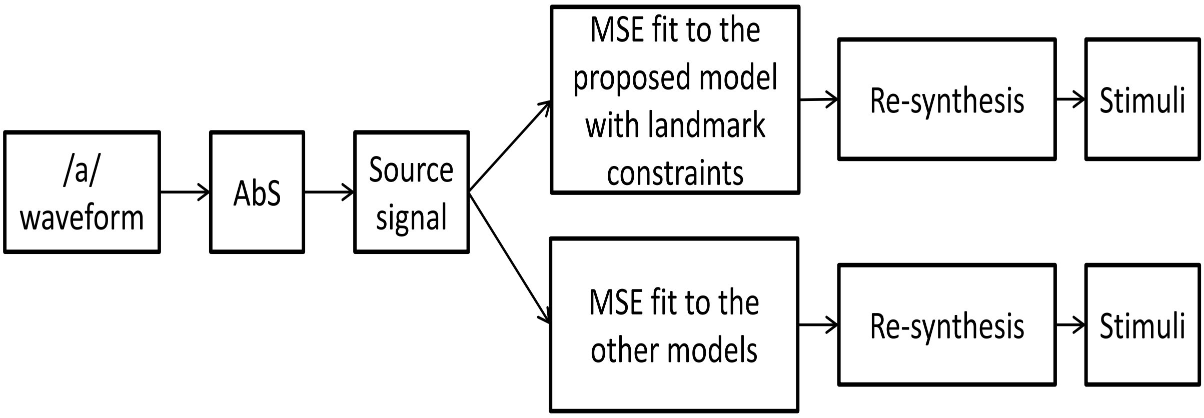

/a/ waveform
AbS
Source signal
MSE fit to the proposed model with landmark constraints
MSE fit to the other models
Re-synthesis
Re-synthesis
Stimuli
Stimuli

Figure 6.5: Flowchart showing how stimuli were generated for the perceptual experiment.

model differences were significant for female speakers. For male speakers, Tukey post-hoc t-tests (with Bonferroni adjustment for multiple comparisons) showed that the EE2 model had lower MSE values than the other models ($p < 0.05$). Note that EE2-LM has higher MSE than EE2. The following section will examine whether EE2-LM results in a better perceptual match to the target voice.

Table 6.1: MSE values (in %) of fitting models to the AbS sources. "EE2" denotes fitting the proposed EE2 model subject to MSE criteria. "EE2-LM" denotes fitting the proposed EE2 model subject to MSE criteria with the constraint of exact landmark matching.

	Ros	LF	SA	EE1	EE2	EE2-LM
Male	27.8	14.1	25.8	21.6	3.9	6.9
Female	11.3	3.6	3.8	3.5	1.2	1.6

6.3 Perceptual experiment

6.3.1 Stimuli

Source model comparisons required a target source pulse to which the models could be fitted, and the need for experimental control during perceptual evaluations mandated that this target be synthetic. That way, voice stimuli could be created that differed only in the source function. To ensure that these synthetic targets were as natural in quality as possible and that they represented a range of naturally-occurring voice qualities, target stimuli were derived via analysis-by-synthesis (AbS, [KAG10]) from 40 natural samples (20 male, 20 female) of the vowel /a/. A steady-state vowel was chosen because it is routinely used for evaluating voice quality and carries substantial information about the voice source. Further, the simpler acoustic structure of a steady-state vowel should yield responses from listeners in the perceptual studies reflecting simpler perceptual strategies that can be easily interpreted. Samples were directly digitized at 20 kHz using a Brüel & Kjær microphone (model 4193), and a 1-second-long segment was excerpted for analysis. The synthesizer sampling rate was fixed at 10 kHz. Parameters describing the harmonic part of the voice source were estimated from a representative cycle of phonation for each voice using the inverse filtering method described in [JBM87]. The harmonic and inharmonic components (the noise excitation) were identified using a comb-liftering operation in the cepstral domain [Kro93]. Spectrally-shaped noise was synthesized by passing white noise through a 100-tap finite impulse response filter fitted to that noise spectrum. F0 was estimated pulse by pulse using the time domain waveform. Formant frequencies and bandwidths were estimated using autocorrelation linear predictive coding analysis with a window of 25.6 ms. The complete synthesized source was then filtered through the vocal tract model, and all parameters were adjusted until the synthetic copy formed an acceptable match to the original natural voice sample.

A paired comparison (same/different) task ensured that the AbS tokens were indistinguishable from the natural stimuli: d prime ranged from 0 to 1.32 across voices, with a mean of 0.79 (sd=0.41). Given these results, the AbS tokens were used in place of the natural voice samples as the target stimuli in all subsequent analyses.

6.3.2 Perceptual experiment setup

To determine the perceptual importance of these results, synthetic copies of the voices were generated using a modeled source pulse for each voice, with all other synthesizer parameters held constant at the values derived during AbS, as illustrated in Figure 6.5. For the proposed EE2 model, only the model-fitted sources with exact matching at the landmark points were used in this experiment (denoted "EE2-LM"). 40 listeners (UCLA students and staff; 18-33 years of age; M=21.15 years; sd=3.03 years) assessed the similarity of all versions of each voice in a visual sort-and-rate task [Gra03, Esp10], in which listeners assessed the extent of perceived match between the original voice samples and each copy. Each listener heard 10 voice "families", where each family included an original natural voice sample, the corresponding target AbS token, and the 5 model-synthesized tokens of the same voice, such that across subjects each family was judged by 10 listeners. The stimuli were presented as distinct icons on the screen (shown in Figure 6.6). For each family (each trial), listeners were asked to play the stimuli by clicking the icons, and to place perceptually similar sounds close together on a line on the screen, while perceptually dissimilar sounds were to be placed farther away. Listeners were instructed to use as much of the line for sorting the stimuli as they wished. They could listen to the stimuli as often as they like, and the study was not timed.

Although listeners saw no numerical values associated with the endpoints of the line, the left and right endpoints were assigned values of 0 and 1000, respectively.

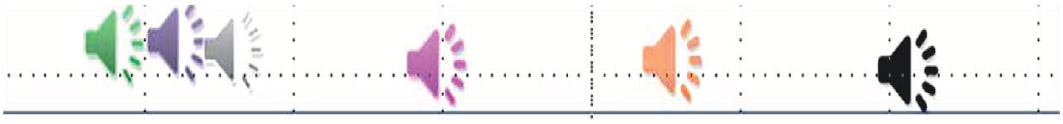


Figure 6.6: Icons representing the stimuli in a sort-and-rate experiment.

Thus, a numerical value could be assigned to the position of each token. The distance of each modeled token from the target AbS voice was then calculated, and this value was subsequently normalized within family for the range of values used on that given trial by that listener. The absolute values of these normalized distances were used in subsequent analyses, because the orientation of the line was arbitrary and varied from listener to listener.

6.3.3 Perceptual experiment results

Results of the perceptual experiment are shown in Table 6.2. Recall that 40 listeners participated in this task, but each only heard 10 of the 40 voices. Thus, every 4 subjects heard the stimuli from all 40 voices. Because a pre-test showed no significant differences in rating, the results of every 4 subjects were averaged, to make 10 "metasubjects", where each "metasubject" (consisting of 4 listeners) heard all 40 voices. This enabled us to run an ANOVA with "metasubject" as the error term. A two-way (model by gender of voice) repeated-measures ANOVA showed significant main effects of model $[F(4, 36) = 155.77, p < 0.0001]$ and gender $[F(1, 9) = 26.49, p < 0.001]$ on mean perceptual distance, as well as a significant model by gender interaction effect $[F(4, 36) = 10.62, p < 0.001]$.

Tukey post-hoc t-tests (with Bonferroni adjustment for multiple comparisons) indicated that the proposed EE2-LM model formed a significantly better match to the target AbS stimulus (lower mean perceptual distance) than the other models ($p < 0.0001$). The perceptual distance to the target token for the LF model was only lower than that of the Ros model ($p < 0.0001$), but not statistically different from those of the SA and EE1 models. The difference between male and female voices in perceptual distances between the modeled and target tokens was significant only for the Ros model, for which male voices were closer perceptual matches to the AbS voice than female voices ($p < 0.0001$). For both genders, the Ros model had a higher perceptual distance than the other models ($p < 0.0001$).

Table 6.2: Normalized perceptual distances (range from 0 to 1) between the model-fitted voices and the target AbS voice, for male and female voices. A smaller number indicates a closer perceptual distance (closer match) to the target AbS voice.

	Ros	LF	SA	EE1	EE2-LM
Male	0.57	0.46	0.38	0.40	0.26
Female	0.71	0.42	0.46	0.43	0.32

6.4 Discussion

Compared to the 4-parameter LF model [FLL85], 2 perceptually-motivated parameters were added in the proposed EE2 model to provide more flexibility in matching the glottal opening phase. With the increased number of parameters, it is not surprising that the proposed EE2-LM model provided a better model fit. Nevertheless, the significant improvement achieved by the proposed model over the LF model in perceptual experiments indicated that the source variability at the opening phase (captured by the two additional parameters) is perceptually salient. Recall that the characteristics of the glottal closing phase (e.g., the negative peak of the flow derivative) have usually been assumed to be perceptually important, because of their association with the main acoustic excitation of the vocal tract [Fan93]. However, this chapter demonstrated the perceptual importance of the glottal source shape at the opening phase, providing insights to modeling studies and synthesis applications.

6.5 Summary

This chapter presented a new voice source model with increased flexibility to capture the perceptually-important source shape aspects. Five voice source models were fitted to 40 natural voices obtained by inverse filtering and analysis-by-synthesis (AbS). Synthetic copies of the voices were generated using each modeled source pulse. Models were perceptually evaluated using a visual sort-and-rate task in which listeners assessed the extent of perceived match between the AbS copies and stimuli created with model-fitted sources. Compared to the other models, on average, the proposed model provided more accurate fittings (in terms of MSE) to the AbS-derived source. In addition, perceptual experiments showed that the proposed model provided closer perceptual matches to the target AbS voice than the other models. In order to demonstrate the potential applicability of the proposed model for improving the quality of speech synthesis, a preliminary experiment was conducted in which source models were fitted to source signals representing different voice qualities (breathy, modal, and pressed) and F0 levels. Pilot results not described in this dissertation, showed that on average, the proposed model provided a more accurate fit than did the other models. Future work will examine the effect of using this model in synthesizing continuous speech.

CHAPTER 7

Summary and future work

7.1 Summary

In this dissertation, the voice source was analyzed by using high-speed videoendoscopy (HSV) data of the vocal folds. New models of the voice source were proposed and applications were presented.

Chapter 1 introduced background information on human speech production and the linear speech production model including the voice source and vocal tract components. Existing voice source models were discussed, and the definitions and terminologies used in voice quality analysis were presented.

Chapters 2 and 3 summarized data acquisition methods. Also, in Chapter 2, a new computationally-efficient method—the glottaltopogram—to analyze HSV data was presented to reveal the overall synchronization of the vibrational patterns of the vocal folds over the entire laryngeal area. Chapter 3 investigated aspects of the glottal area pulse shape that vary with voice quality, by using HSV recordings of the vocal folds. A new measure of the **glottal area** (AC/OQ) was then proposed to capture variations in pulse shapes.

In Chapter 4, voice source related acoustic measures were analyzed in the context of a physiological vocal-fold vibration pattern—the glottal gap. These acoustic measures were applied to an automatic gender classification task of children's voices.

Chapter 5 presented a new voice source model (EE1) based on HSV data of the vocal folds. A modified codebook search technique based on the proposed EE1 model was introduced to estimate the voice source from speech signals.

In Chapter 6, a perceptually-motivated voice source model (EE2) was proposed to capture perceptually-important source shape aspects. The perceptual adequacy of the EE2 model was then evaluated in a sort-and-rate listening experiment.

The following three sections summarize the main results in this dissertation.

7.1.1 Acquisition and data analysis of HSV

In order to effectively analyze HSV data, Chapter 2 proposed the "glottaltopogram," which is based on principal component analysis of changes over time in the brightness of each pixel in consecutive video images. This method compactly summarizes the overall spatial synchronization pattern of vocal fold vibration for the entire glottal area. The proposed method may produce plots that are spatially similar to the original images, and which can be easily interpreted by physicians and clinicians during diagnosis. Experimental results showed that this method is effective in visualizing pathological and normal vocal fold vibratory patterns. A Matlab Graphical User Interface—`GTG analyze tool` was implemented for the glottaltopogram algorithm. A brief description of this tool can be found in Appendix A.

After extracting glottal area waveforms from HSV data, Chapter 3 investigated aspects of the glottal area pulse shape that vary with voice quality. A new measure, AC/OQ, was proposed to capture variations in glottal area pulse shapes in a manner that reflects both acoustic and perceptual consequences of those variations. This measure is defined as the AC component divided by OQ, so that an increase in glottal gap size or an increase in OQ results in lower AC/OQ values. Analyses of phonations differing both discretely and continuously in voice quality showed that across speakers AC/OQ values also increased monotonically along a breathy-to-pressed continuum. Thus, AC/OQ is capable of characterizing the continuum of glottal area waveform variation corresponding to a range of voice qualities, regardless of the existence or absence of glottal gaps.

7.1.2 Acoustic correlates of glottal gaps with application to gender classification

Chapter 4 used HSV data to investigate the relationship between glottal gap area, source parameters, acoustic measures, and voice quality. Results showed that CPP and HNR were affected by glottal gap area, indicating the presence of more spectral noise with increasing glottal gap area.

Three voice source related measures: CPP, HNR and $H_2^* - H_4^*$ were then used in a gender classification task from children's voices. Gender classification is relatively easy for adult voices but much more challenging for children's speech, where traditional features, such as F_0 and formant frequencies, are less useful. Results showed that using these three features improved gender classification accuracy, especially for younger (10-15 year old) speakers.

7.1.3 Modeling the voice source and applications

The voice source provides important information to many speech applications. For a vast majority of applications, voice source information has to be estimated from the speech signal recorded by a microphone. A new voice source model (EE1) and a noise-robust automatic source estimation algorithm were proposed in Chapter 5. The voice source signal was estimated, using a codebook search approach, from speech signals. The glottal area extracted from HSV was converted to glottal flow to calibrate the proposed algorithm. Results in both clean and noisy conditions showed that the novel approach is robust in accurately estimating the glottal flow waveform.

A perceptually-adequate source model should capture perceptually-important aspects of the source signal, thus generating natural-sounding synthetic voices. A refined voice source model (EE2) was proposed to capture perceptually-important source shape aspects. Two perceptually-motivated parameters were added in the proposed EE2 model to provide more flexibility in matching the glottal opening phase. The resulting model (EE2-LM), along with four other source models, was fitted to 40 voice sources (20 male and 20 female) obtained by inverse filtering and analysis-by-synthesis of samples of natural speech. Synthetic copies of the voices were generated using each modeled source pulse, with all other synthesis parameters held constant. A visual sort-and-rate task was then conducted, in which listeners assessed the extent of perceived similarity between the target voice samples and each copy. Results showed that the proposed EE2-LM model provided a more accurate fit and a better perceptual match to the target than the other models.

7.2 Unsolved issues and outlook

The proposed glottaltopogram is able to effectively visualize the glottal vibration patterns, but the manner in which the unsynchronized vibrations (e.g., vibratory asymmetries or phase lags) affect perceived voice quality is far from well understood. More data are needed to further investigate the interrelationship between vocal fold vibration pattern and perceived voice quality. Future study will also include collecting data of more voice quality types (e.g., creaky voice and falsetto voice). Exploring automatic detection of speech pathology from high-speed images will be another direction of future work.

The proposed measure AC/OQ is capable of characterizing the continuum of glottal area waveform variation corresponding to a range of voice qualities, regardless of the existence or absence of glottal gaps. However, the audio and HSV data used to examine the effect of AC/OQ were collected from speakers without voice disorders. In producing breathy phonation, these speakers usually demonstrated a gap through the cartilaginous glottis, which may extend continuously through some or all of the membranous glottis [HHP88, SL90]. However, speakers with voice disorders may have a gap that appears only in the membranous glottis, as occurs in presbylaryngis (the aged larynx) or in some patients with Parkinson disease who have breathy voices [HGW84]. Because this glottal configuration was not included in our study, it is possible that AC/OQ may not measure these voices adequately. The extent or manner in which AC/OQ may generalize to other glottal configurations will be a topic for future research.

Time-domain source models lack an effective way of modeling the incomplete glottal closure phenomenon, which has been shown to be an important physiological aspect of voice production [CS95, HC99]. The EE1 model assumes that the glottal flow signal starts at 0 and ends at 0 for each glottal cycle (i.e., complete glottal closure). This may not be true for some phonations, e.g., breathy voices for which the glottis may not fully close at the end of a glottal cycle. Future work will include incorporating the incomplete glottal closure effect into the source modeling to further improve source estimation accuracy. Estimating the voice source in other noise types, for example babble noise, will also be examined. In addition, more theoretical and experimental studies are needed to better model the relationship between glottal area and glottal flow.

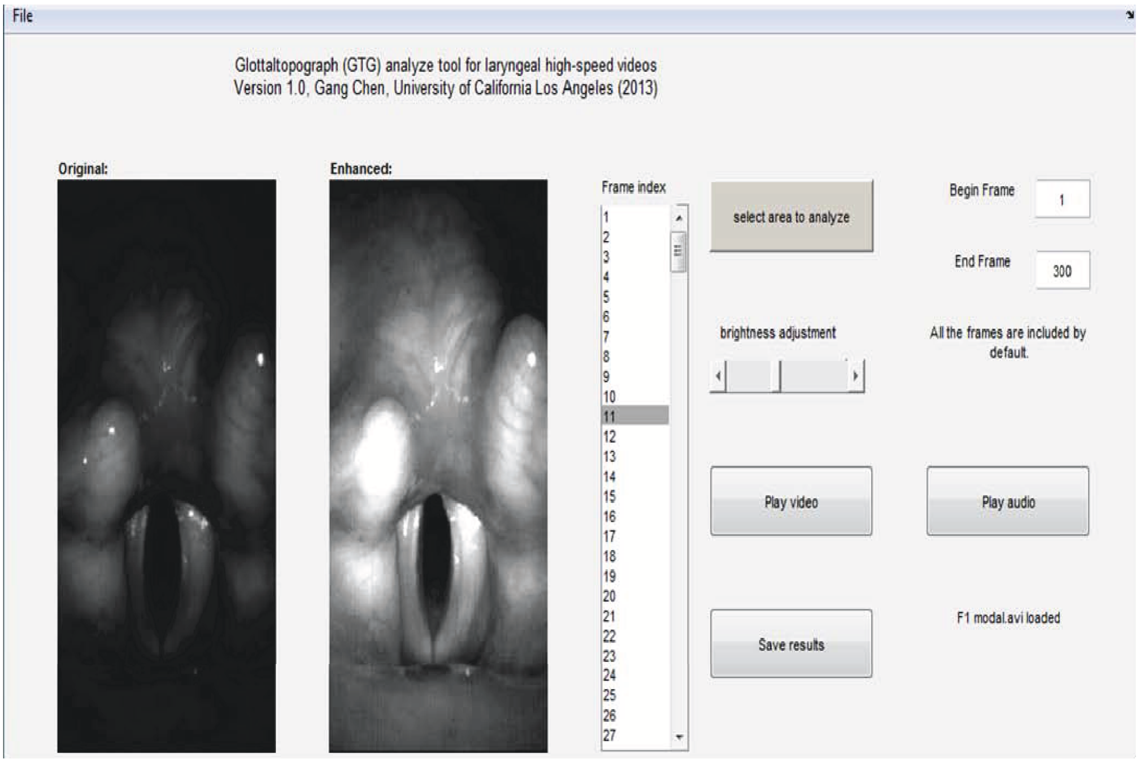
Perceptual experiments on steady-state vowels showed that the proposed EE2-LM model provided significantly better synthetic voices in comparison to four existing source models, in terms of perceived naturalness. In order to demonstrate the potential applicability of the proposed EE2-LM model for improving the quality of speech synthesis, future work will examine the effect of using it in synthesizing continuous speech. Using the model in the codebook-based source estimation approach will also be explored.

Solving the issues described above would lead to a better understanding, and eventually a better model of the voice source. This knowledge would benefit various speech applications, such as speech recognition, speech synthesis, speaker identification, age/gender classification, as well as clinical assessments.

APPENDIX A

GTG analyze tool: A Matlab Graphical User Interface for Glottaltopogram

GTG analyze tool is a Matlab Graphical User Interface (GUI) implemented for the glottaltopogram algorithm described in Chapter 2. A screenshot of the GUI is shown in Figure A.1. The GUI can be obtained freely at: http://www.seas.ucla.edu/spapl/shareware/GTG/GTG_analyze_tool.htm


Glottaltopograph (GTG) analyze tool for laryngeal high-speed videos
Version 1.0, Gang Chen, University of California Los Angeles (2013)

Begin Frame 1

End Frame 300

All the frames are included by default.

F1 modal.avi loaded

Figure A.1: Matlab Graphical User Interface of GTG analyze tool

The GUI is designed for easy access to the parameters used in the algorithm in an interactive way. Results in each intermediate processing stage can be viewed, allowing the user to adjust the parameters accordingly.

The GUI supports high-speed video recordings in “.avi” format. An input video can be loaded by clicking “Load” in the “File” tab. Once the video is loaded, the first frame of the video will be displayed on the image panels on the left of the GUI. A specific frame can be displayed by selecting the corresponding frame number in the “Frame Index” box. There are two image panels on the left of the GUI. The “Original” panel shows the original image frame of the video, while the “Enhanced” panel shows the image frame after brightness adjustment. The brightness adjustment parameter can be manipulated through the slide bar with the adjusted image shown in real time, allowing the user to perform this preprocessing step in an interactive way. The “Begin frame” and “End frame” boxes specify the range of frames that is included for analysis. All the frames are included by default, but user can also set any specific range to include only frames of interest. Once the settings above are finalized, user can click on the “select area to analyze” button. A box will pop up which allows the user to select a triangular area in the “Enhanced” frame panel. The pixels within the triangular area of the enhanced video will be processed and resultant glottaltopogram plots will be shown.

For visualization purposes, it is recommended that the triangular box selected should include only the vocal folds to avoid the interference of artifacts such as glare spots. The “Play video” and “Play audio” buttons can be pressed anytime for easy access to the data under processing. Videos are played back at a speed of 25 frames/sec for visual inspection. “Save results” button will save the analysis results in a Matlab “.mat” file for easy access and portability. Saved results include vectors of the fist two principal components, variance accounted for by each principal component, reconstruction error, and the enhanced video data.

APPENDIX B

Glottal flow model fitting performance of the proposed EE2-LM source model

Figures B.1 and B.2 show target AbS source pulses and the corresponding fitted sources using the proposed EE2-LM model, as described in Section 6.2.2.

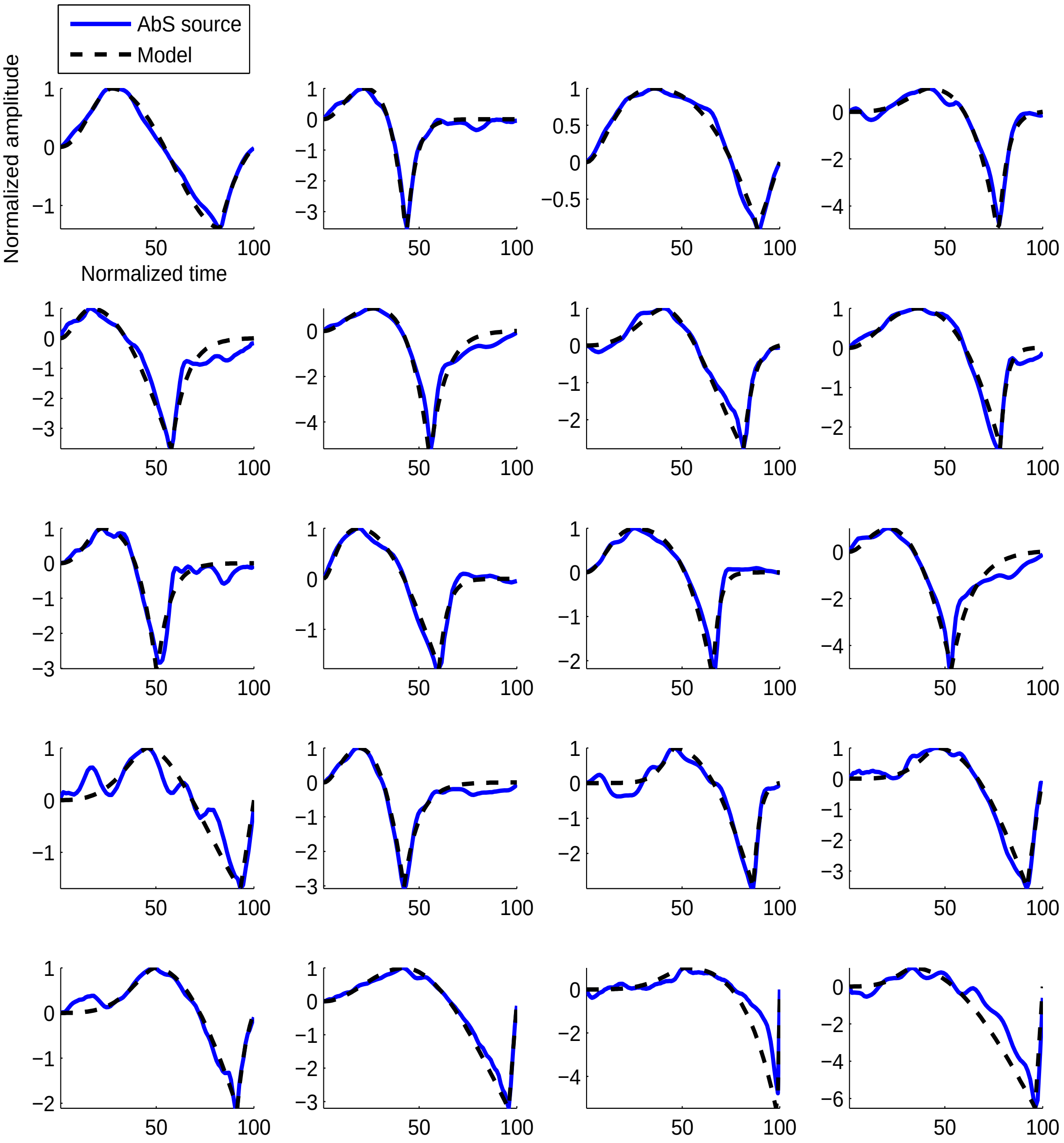


Figure B.1: Model fitting performance of the proposed EE2-LM source model for 20 male subjects.

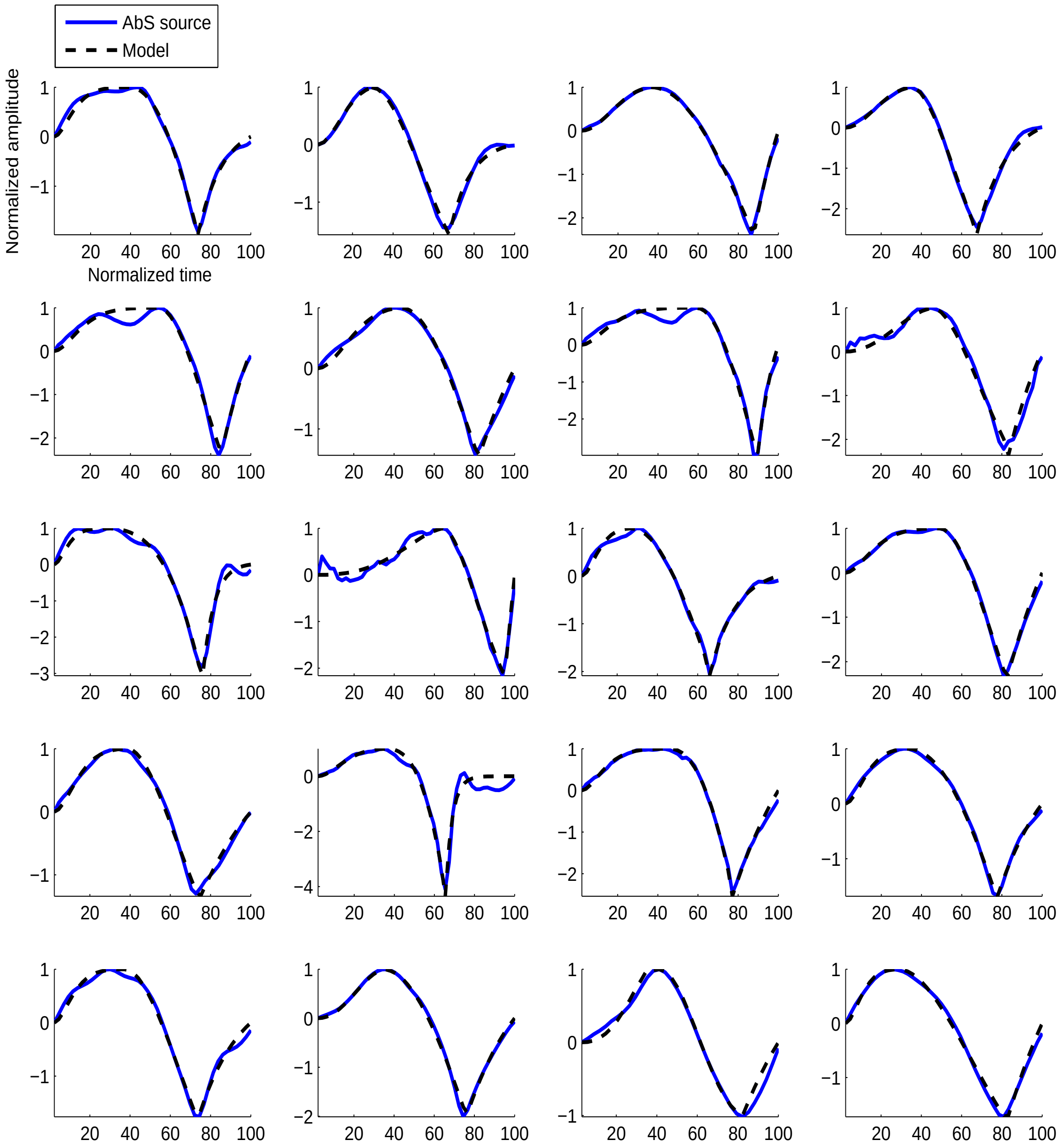


Figure B.2: Model fitting performance of the proposed EE2-LM source model for 20 female subjects.

References

[AA07] Matti Airas and Paavo Alku. "Comparison of multiple voice source parameters in different phonation types." In *Interspeech*, pp. 1410–1413, 2007.

[AAB06] Paavo Alku, Matti Airas, Eva Björkner, and Johan Sundberg. "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity." *J. Acoust. Soc. Am.*, **120**:1052–1062, 2006.

[AB94] R. Adams and L. Bischof. "Seeded region growing." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **16**:641–647, 1994.

[ABV02] P. Alku, T. Bäckström, and E. Vilkman. "Normalized amplitude quotient for parametrization of the glottal flow." *J. Acoust. Soc. Am.*, **112**:701–710, 2002.

[Air08] M. Airas. "TKK Aparat: An environment for voice inverse filtering and parameterization." *Logopedics Phoniatrics Vocology*, **33**:49–64, 2008.

[Alk92] P. Alku. "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering." *Speech Commun.*, **11**:109–118, 1992.

[Alk03] P. Alku. "Parameterisation methods of the glottal flow estimated by inverse filtering." In *VOQUAL*, pp. 81–87, 2003.

[AMY09] Paavo Alku, Carlo Magi, Santeri Yrttiaho, Tom Bäckström, and Brad Story. "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering." *J. Acoust. Soc. Am.*, **125**:3289, 2009.

[Bak92] Ronald J. Baken. "Electroglottography." *J. Voice*, **6**:98–110, 1992.

[BAV02] Tom Bäckström, Paavo Alku, and Erkki Vilkman. "Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range." *IEEE transactions on speech and audio processing*, **10**:186–192, 2002.

[Ber58] J.W. van den Berg. "Myoelastic-aerodynamic theory of voice production." *J. Speech Hear. Res.*, **1**:227–244, 1958.

[BFB04] Pascal Belin, Shirley Fecteau, and Catherine Bédard. "Thinking the voice: neural correlates of voice perception." *Trends in Cognitive Sciences*, **8**:129–135, 2004.

[BHF87] D.M. Bless, M. Hirano, and R.J. Feder. “Videostroboscopic evaluation of the larynx.” *Ear, nose, & throat journal*, **66**:289, 1987.

[BP95] P. Busby and G. Plant. “Formant frequency values of vowels produced by preadolescent boys and girls.” *J. Acoust. Soc. Am.*, **97**:2603–2606, 1995.

[CC90] Raymond H. Colton and Edward G. Conture. “Problems and pitfalls of electroglottography.” *J. Voice*, **4**:10–24, 1990.

[CC95] K.E. Cummings and M.A. Clements. “Glottal models for digital speech processing: A historical survey and new results.” *Digital Signal Processing*, **5**:21–42, 1995.

[CKG13] Gang Chen, Jody Kreiman, Bruce R. Gerratt, Juergen Neubauer, Yen-Liang Shue, and Abeer Alwan. “Development of a glottal area index that integrates glottal gap size and open quotient.” *J. Acoust. Soc. Am.*, **133**:1656–1666, 2013.

[CKS11] G. Chen, J. Kreiman, Y.-L. Shue, and A. Alwan. “Acoustic Correlates of Glottal Gaps.” In *Interspeech*, pp. 2673–2676, 2011.

[CL91] D.G. Childers and C.K. Lee. “Vocal quality factors: Analysis, synthesis, and perception.” *J. Acoust. Soc. Am*, **90**:2394–2410, 1991.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm` (Last viewed Aug. 1, 2009).

[CS95] B. Cranen and J. Schroeter. “Modeling a leaky glottis.” *J. Phonetics*, **23**:165–177, 1995.

[CSK12] G. Chen, Y.-L. Shue, J. Kreiman, and A. Alwan. “Estimating the voice source in noise.” In *Interspeech*, pp. 1600–1603, 2012.

[DAA14] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana. “Glottal source processing: From analysis to applications.” *Computer Speech & Language*, 2014. http://dx.doi.org/10.1016/j.csl.2014.03.003.

[DBL03] M. Döllinger, T. Braunschweig, J. Lohscheller, U. Eysholdt, and U. Hoppe. “Normal voice production: computation of driving parameters from endoscopic digital high speed images.” *Meth. Inf. Med.*, **42**:271–276, 2003.

[DBP07] Mark R. Draper, Barbara Blagnys, and Don J. Premachandra. “To ’EE’ or not to ’EE’.” *J. Otolaryngology*, **36**:191–195, 2007.

[DLS11] Michael Döllinger, Jörg Lohscheller, Jan Svec, Andrew McWhorter, and Melda Kunduk. "Support Vector Machine Classification of Vocal Fold Vibrations Based on Phonovibrogram Features." In Farzad Ebrahim, editor, *Advances in Vibration Analysis Research*, pp. 435–456. InTech, Croatia, 2011.

[DPB07] Dimitar D. Deliyski, Pencho P. Petrushev, Heather Shaw Bonilha, Terri Treman Gerlach, Bonnie Martin-Harris, and Robert E. Hillman. "Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution." *Folia Phoniatrica et Logopaedica*, **60**:33–44, 2007.

[Esp10] C.M. Esposito. "The effects of linguistic experience on the perception of phonation." *J. Phonetics*, **38**:306–316, 2010.

[Fan70] G. Fant. *Acoustic theory of speech production.* Mouton, The Hague, Paris, 2nd edition, 1970. pp. 15-20.

[Fan76] G. Fant. "Vocal tract energy functions and non-uniform scaling." *Journal of the Acoustical Society of Japan*, **11**:1–18, 1976.

[Fan82] G. Fant. "Preliminaries to analysis of the human voice source." *STL-QPSR*, **4**:1–27, 1982.

[Fan93] G. Fant. "Some problems in voice source analysis." *Speech Commun.*, **13**:7–22, 1993.

[Fan95] G. Fant. "The LF-model revisited. Transformations and frequency domain analysis." *STL-QPSR*, **36**:119–156, 1995.

[Far40] D.W. Farnsworth. "High-speed motion pictures of the human vocal cords." *Bell Lab Rec*, **18**:203–8, 1940.

[Fis67] E. Fischer-Jorgensen. "Phonetic analysis of breathy (murmured) vowels in Gujarati." *Indian Linguist*, **28**:71–139, 1967.

[FL86] H. Fujisaki and M. Ljungqvist. "Proposal and evaluation of models for the glottal source waveform." In *ICASSP*, pp. 1605–1608, 1986.

[FLL85] G. Fant, J. Liljencrants, and Q. Lin. "A four-parameter model of glottal flow." *STL-QPSR*, **4**:1–13, 1985.

[FMS01] M. Fröhlich, D. Michaelis, and H.W. Strube. "SIM - simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals." *J. Acoust. Soc. Am.*, **110**:479–488, 2001.

[GK10] B. Gerratt and J. Kreiman. "A spectral-slope compensated scale for measuring perception of vocal aperiodicity." *J. Acoust. Soc. Am.*, **127**:2022–2022, 2010.

[GL01] S. Granqvist and P.-Å. Lindestad. "A method of applying Fourier analysis to high-speed laryngoscopy." *J. Acoust. Soc. Am.*, **110(6)**:3193–3197, 2001.

[Gra03] S. Granqvist. "The visual sort and rate method for perceptual evaluation in listening tests." *Logopedics Phonatrics Vocology*, **28**:109–116, 2003.

[Han97] H. M. Hanson. "Glottal characteristics of female speakers: Acoustic correlates." *J. Acoust. Soc. Am.*, **101**:466–481, 1997.

[HC99] H. M. Hanson and E. S. Chuang. "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data." *J. Acoust. Soc. Am.*, **106**:1064–1077, 1999.

[HCE94] J. Hillenbrand, R.A. Cleveland, and R.L. Erickson. "Acoustic Correlates of Breathy Vocal Quality." *J. Speech Hear. Res.*, **37**:769–778, 1994.

[HdD01] N. Henrich, C. d'Alessandro, and B. Doval. "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data." In *Eurospeech*, pp. 47–50, 2001.

[HdD04] Nathalie Henrich, Christophe d'Alessandro, Boris Doval, and Michele Castellengo. "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation." *J. Acoust. Soc. Am.*, **115**:1321–1332, 2004.

[HGW84] David G. Hanson, Bruce R. Gerratt, and Paul H. Ward. "Cinegraphic observations of laryngeal function in Parkinson's disease." *The Laryngoscope*, **94**:348–353, 1984.

[HHP88] E. Holmberg, R. Hillman, and J. Perkell. "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice." *J. Acoust. Soc. Am.*, **84**:511–529, 1988.

[HHP89] Eva B. Holmberg, Robert E. Hillman, and Joseph S. Perkell. "Glottal airflow and transglottal air pressure measurements for male and female speakers in low, normal, and high pitch." *J. Voice*, **3**:294–305, 1989.

[HHP95] E.B. Holmberg, R.E. Hillman, J.S. Perkell, P. Guiod, and S.L. Goldman. "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice." *J. Speech Hear. Res.*, **38**:1212–1223, 1995.

[HKK88] M. Hirano, K. Kiyokawa, and S. Kurita. *Laryngeal muscles and glottic shaping. In O. Fujimura (Ed.), Vocal fold physiology: Vol. 2. Voice* *production. Mechanisms and functions.* Raven Press, New York, 2nd edition, 1988.

[HKN83] M. Hirano, J. Kurita, and T. Nakahima. "Vocal fold Physiology: Contemporary Research and Clinical Issues." In D. Bless and J. Abbs, editors, *Growth development, and aging of human vocal folds*, pp. 22–43, San Diego, CA, 1983. College Hill Press.

[HLW03] S. Hertegård, H. Larsson, and T. Wittenberg. "High-speed imaging: applications and development." *Logopedics Phonatrics Vocology*, **28**:133–139, 2003.

[HM95] J.W. Hawks and J.D. Miller. "A formant bandwidth estimation procedure for vowel synthesis." *J. Acoust. Soc. Am.*, **97**:1343–1344, 1995.

[HM07] M. Howe and R. McGowan. "Sound generated by aerodynamic sources near a deformable body, with application to voiced speech." *J. of Fluid Mech.*, **592**:36–392, 2007.

[HMW66] Harry Hollien, Paul Moore, Ronald W. Wendahl, and John F. Michel. "On the nature of vocal fry." *J. Speech Lang. Hear. Res.*, **9**:245, 1966.

[Hol74] Harry Hollien. "On vocal registers." *J. Phonetics*, **2**:125–143, 1974.

[HS72] H. Hollien and T. Shipp. "Speaking fundamental frequency and chronologic age in males." *J. Speech Hear. Res.*, **15**:155–159, 1972.

[Huf87] Marie K. Huffman. "Measures of phonation type in Hmong." *J. Acoust. Soc. Am.*, **81**:495, 1987.

[IA04] M. Iseli and A. Alwan. "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation." In *ICASSP*, pp. 669–672, 2004.

[IIH11] Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. "Improved acoustic characterization of breathy and whispery voices." In *Interspeech*, pp. 2965–2968, 2011.

[ISA07] M. Iseli, Y.-L. Shue, and A. Alwan. "Age, sex, and vowel dependencies of acoustic measures related to the voice source." *J. Acoust. Soc. Am.*, **121**:2283–2295, 2007.

[JBM87] H.R. Javkin, N. Antoñanzas Barroso, and I. Maddieson. "Digital inverse filtering for linguistic research." *J. Speech Hear. Res.*, **30**:122–129, 1987.

[KAG10] J. Kreiman, N. Antoñanzas-Barroso, and B.R. Gerratt. "Integrated software for analysis and synthesis of voice quality." *Behavior Research Methods*, **42**:1030–1041, 2010.

[KCP98] H. Kawahara, A. de Cheveigné, and R.D. Patterson. "An instantaneous–frequency–based pitch extraction method for high quality speech transformation: revised TEMPO in the STRAIGHT–suite." In *ICSLP*, 1998. paper 0659.

[KG13] John Kane and Christer Gobl. "Wavelet maxima dispersion for breathy to tense voice discrimination." *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**:1170–1179, 2013.

[KGB07] J. Kreiman, B. Gerratt, and N. Antoñanzas Barroso. "Measures of the glottal source spectrum." *J. Speech Lang. Hear. Res.*, **50**:595–610, 2007.

[KGC12] J. Kreiman, B.R. Gerratt, G. Chen, M. Garellek, and A. Alwan. "Perceptual evaluation of source models." *J. Acoust. Soc. Am*, **132**:2088, 2012.

[KGD10] Jody Kreiman, Bruce R. Gerratt, and Sameer ud Dowla Khan. "Effects of native language on perception of voice quality." *J. Phonetics*, **38**:588–593, 2010.

[KH73] Y. Koike and M. Hirano. "Glottal-area time function and subglottal-pressure variation." *J. Acoust. Soc. Am.*, **54**:1618–1627, 1973.

[KHd12] Sevasti-Zoi Karakozoglou, Nathalie Henrich, Christophe d'Alessandro, and Yannis Stylianou. "Automatic glottal segmentation using local-based active contours and application to glottovibrography." *Speech Commun.*, **54**:641–654, 2012.

[Kit85] P. Kitzing. "Stroboscopy–a pertinent laryngological examination." *J. Otolaryngology*, **14(3)**:151–157, 1985.

[KK90] D.H. Klatt and L.C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." *J. Acoust. Soc. Am.*, **87**:820–857, 1990.

[KKG10] J. Kane, M. Kane, and C. Gobl. "A spectral LF model based approach to voice source parameterisation." In *Interspeech*, pp. 2606–2609, 2010.

[KMC99] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds." *Speech Commun.*, **27**:187–207, 1999.

[Kro93] G. de Krom. "A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals." *J. Speech Hear. Res.*, **36**:254–266, 1993.

[KS11] Jody Kreiman and Diana Van Lancker Sidtis. *Foundations of voice studies: an interdisciplinary approach to voice production and perception.* Wiley-Blackwell, 2011.

[KSC12] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. Gerratt, J. Neubauer, and A. Alwan. "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation." *J. Acoust. Soc. Am.*, **132**:2625–2632, 2012.

[Lav80] J. Laver. *The Phonetic Description of Voice Quality.* Cambridge University Press, Cambridge, 1980.

[LET08] J. Lohscheller, U. Eysholdt, H. Toy, and M. Döllinger. "Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics." *Medical Imaging, IEEE Transactions on*, **27**:300–309, 2008.

[LFM01] Irma M. Verdonck-de Leeuw, Joost M. Festen, and Hans F. Mahieu. "Deviant vocal fold vibration as observed during videokymography: the effect on voice quality." *J. Voice*, **15**:313–322, 2001.

[LHL00] H. Larsson, S. Hertegård, P.-Å. Lindestad, and B. Hammarberg. "Vocal fold vibrations: high-speed imaging, kymography and acoustic analysis: a preliminary report." *Laryngoscope*, **110**:2117–2122, 2000.

[Lin89] C. E. Linke. "A study of pitch characteristics of female voices and their relationship to vocal effectiveness." *Folia Phoniatr*, **25**:173–185, 1989.

[Lin92] S.E. Linville. "Glottal Gap Configurations in Two Age Groups of Women." *J. Speech Hear. Res.*, **35**:1209–1215, 1992.

[LPN99] S. Lee, A. Potamianos, and S. Narayanan. "Acoustics of children's speech: Developmental changes of temporal and spectral parameters." *J. Acoust. Soc. Am.*, **105**:1455–1468, 1999.

[LTR07] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger. "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos." *Medical Image Analysis*, **11**:400–413, 2007.

[Maa11] Laurens van der Maaten. "Matlab Toolbox for Dimensionality Reduction (Version: 0.7.2b).", 2011. `http://homepage.tudelft.nl/19j49/Matlab\_Toolbox\_for\_Dimensionality\_Reduction.html` (Last viewed Oct. 1, 2011).

[MC04] E. Moore and M. Clements. "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information." In *ICASSP*, pp. 101–104, 2004.

[McG88] R.S. McGowan. "An aeroacoustic approach to phonation." *J. Acoust. Soc. Am.*, **83**:696–704, 1988.

[MDQ11] D.D. Mehta, D.D. Deliyski, T.F. Quatieri, and R.E. Hillman. "Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings." *J. Speech Lang. Hear. Res.*, **54**:47–54, 2011.

[MDZ10] Daryush D. Mehta, Dimitar D. Deliyski, Steven M. Zeitels, Thomas F. Quatieri, and Robert E. Hillman. "Voice production mechanisms following phonosurgical treatment of early glottic cancer." *Ann. Otol. Rhinol. Laryngol.*, **119**:1–9, 2010.

[MLU96] J.D. Miller, S. Lee, R.M. Uchanski, A.F. Heidbreder, B.B. Richman, and J. Tadlock. "Creation of two children's speech databases." In *ICASSP*, pp. 849–852, 1996.

[MRB83] M. Morrison, L Rammage, G. Belisle, B Pullan, and H. Nichol. "Muscular tension dysphonia." *J. Otolaryngology*, **12**:302–306, 1983.

[MZQ11] Daryush D. Mehta, Matías Zañartu, Thomas F. Quatieri, Dimitar D. Deliyski, and Robert E. Hillman. "Investigating acoustic correlates of human vocal fold vibratory phase asymmetry through modeling and laryngeal high-speed videoendoscopy." *J. Acoust. Soc. Am.*, **130**:3999–4009, 2011.

[NM00] Seiji Niimi and Mamiko Miyaji. "Vocal fold vibration and voice quality." *Folia Phoniatr. Logop.*, **52**:32–38, 2000.

[OSK98] K. Omori, D.H. Slavit, A. Kacker, and S.M. Blaugrund. "Influence of Size and Etiology of Glottal Gap in Glottic Incompetence Dysphonia." *The Laryngoscope*, **108**:514–518, 1998.

[PB52] G. E. Peterson and H. L. Barney. "Control methods used in a study of the vowels." *J. Acoust. Soc. Am.*, **24**:175–184, 1952.

[POA01] T. L. Perry, R. N. Ohde, and D. H. Ashmead. "The acoustic bases for gender identification from children's voices." *J. Acoust. Soc. Am.*, **109(6)**:2988–2988, 2001.

[PQR99] Michael D. Plumpe, Thomas F. Quatieri, and Douglas A. Reynolds. "Modeling of the glottal flow derivative waveform with application to speaker identification." *Speech and Audio Processing, IEEE Transactions on*, **7**:569–586, 1999.

[Ros71] A. Rosenberg. "Effects of the glottal pulse shape on the quality of natural vowels." *J. Acoust. Soc. Am.*, **49**:583–590, 1971.

[Rot73] M. Rothenberg. “A new inverse-filtering technique for deriving the glottal air flow waveform during voicing.” *J. Acoust. Soc. Am.*, **53**:1632–1645, 1973.

[Rot81] M. Rothenberg. *Acoustic interaction between the glottal source and the vocal tract, Vocal fold physiology.* University of Tokyo Press, Tokyo, 1981. pp. 305–323.

[RS07] Lawrence R. Rabiner and Ronald W. Schafer. “Introduction to digital speech processing.” *Foundations and trends in signal processing*, **1**:1–194, 2007.

[SA10] Y.-L. Shue and A. Alwan. “A new voice source model based on high-speed imaging and its application to voice source estimation.” In *ICASSP*, pp. 5134–5137, 2010.

[SCA10] Y.-L. Shue, G. Chen, and A. Alwan. “On the Interdependencies between Voice Quality, Glottal Gaps, and Voice-Source related Acoustic Measures.” In *Interspeech*, pp. 34–37, 2010.

[SFM05] Johan Sundberg, Ellinor Fahlstedt, and Anja Morell. “Effects on the glottal voice source of vocal loudness variation in untrained female and male voices.” *J. Acoust. Soc. Am.*, **117**:879–885, 2005.

[SH96] Agaath M.C. Sluijter and Vincent J. van Heuven. “Acoustic correlates of linguistic stress and accent in Dutch and American English.” In *ICSLP*, volume 2, pp. 630–633, 1996.

[Shu10a] Y.-L. Shue. *The Voice Source in Speech Production: Data, Analysis and Models.* PhD thesis, University of California Los Angeles, 2010.

[Shu10b] Y.-L. Shue. “VoiceSauce: a program for voice analysis.”, 2010. `http://www.seas.ucla.edu/spapl/voicesauce/` (Last viewed Apr. 1, 2012).

[SI08] Y.-L. Shue and M. Iseli. “The role of voice source measures on automatic gender classification.” In *ICASSP*, pp. 4493–4496, 2008.

[SIK10] K. Sakakibara, H. Imagawa, M. Kimura, H. Yokonishi, and N. Tayama. “Modal Analysis of Vocal Fold Vibrations Using Laryngotopography.” In *Interspeech*, pp. 917–920, 2010.

[Sj04] K. Sjölander. “The Snack Sound Toolkit.” KTH Stockholm, Sweden, 2004. `http://www.speech.kth.se/snack/` (Last viewed Aug. 1, 2009).

[SKA09] Y.-L. Shue, J. Kreiman, and A. Alwan. "A novel codebook search technique for estimating the open quotient." In *Interspeech*, pp. 2895–2898, 2009.

[SL90] M. Södersten and P.-A. Lindestad. "Glottal closure and perceived breathiness during phonation in normally speaking subjects." *J. Speech Hear. Res.*, **33**:601–611, 1990.

[Son59] Bertil Sonesson. "A method for studying the vibratory movements of the vocal cords." *J. Laryngol. Otol*, **73**:732–737, 1959.

[SS96] Jan G. Švec and Harm K. Schutte. "Videokymography: high-speed line scanning of vocal fold vibration." *J. Voice*, **10**:201–205, 1996.

[SS11] Robin A. Samlan and Brad H. Story. "Relation of Structural and Vibratory Kinematics of the Vocal Folds to Two Acoustic Measures of Breathy Voice Based on Computational Modeling." *J. Speech Lang. Hear. Res.*, **54**:1267–1283, 2011.

[ST95] B. Story and I. Titze. "Voice simulation with a body cover model of the vocal folds." *J. Acoust. Soc. Am.*, **97**:1249–1260, 1995.

[Ste98] K. N. Stevens. *Acoustic Phonetics.* MIT Press, Cambridge, MA, 1998. pp. 55–126.

[STH96] B. Story, I. Titze, and E. Hoffman. "Vocal tract area functions from magnetic resonance imaging." *J. Acoust. Soc. Am.*, **100**:537–554, 1996.

[Sto81] M. Stoicheff. "Speaking fundamental frequency characteristics of non-smoking female adults." *J. Speech Hear. Res.*, **24**:437–441, 1981.

[Sto12] B. Story. "LeTalker: Lumped-element model of phonation.", 2012. `http://sal.shs.arizona.edu/~bstory/LeTalkerMain.html` (Last viewed Mar. 1, 2013).

[SV96] Agaath M.C. Sluijter and Vincent J. Van Heuven. "Spectral balance as an acoustic correlate of linguistic stress." *J. Acoust. Soc. Am.*, **100**:2471, 1996.

[SY09] G. Seshadri and B. Yegnanarayana. "Perceived loudness of speech based on the characteristics of glottal excitation source." *J. Acoust. Soc. Am.*, **126**:2061–2071, 2009.

[Tit87] I. R. Titze. "Physiology of the female larynx." *J. Acoust. Soc. Am.*, **82**:90, 1987.

[Tit89] I. R. Titze. "Physiologic and acoustic differences between male and female voices." *J. Acoust. Soc. Am.*, **85**:1699–1707, 1989.

[Tit06] I. Titze. “Theoretical analysis of maximum flow declination rate versus maximum area declination rate in phonation.” *J. Speech Lang. Hear. Res.*, **49**:439–447, 2006.

[Tit08a] I. Titze. “Nonlinear source–filter coupling in phonation: Theory.” *J. Acoust. Soc. Am.*, **123**:2733, 2008.

[Tit08b] I.R. Titze. “Nonlinear source-filter coupling in phonation: Theory.” *J. Acoust. Soc. Am.*, **123**:2733–2749, 2008.

[TLM58] Rolf Timcke, Hans von Leden, and Paul Moore. “Laryngeal Vibrations: Measurements of the Glottic Wave. Part I. The Normal Vibratory Cycle.” *AMA Archives of Otolaryngology*, **68**:1–9, 1958.

[TRP08] I.R. Titze, T. Riede, and P. Popolo. “Nonlinear source-filter coupling in phonation: Vocal exercises.” *J. Acoust. Soc. Am.*, **123**:1902–1915, 2008.

[TS97] I. Titze and B. Story. “Acoustic interactions of the voice source with the lower vocal tract.” *J. Acoust. Soc. Am.*, **101**:2234–2243, 1997.

[TS02] I. Titze and B. Story. “Rules for controlling low-dimensional vocal fold models with muscle activation.” *J. Acoust. Soc. Am.*, **112**:1064–1076, 2002.

[TWM99] M. Tigges, T. Wittenberg, P. Mergell, and U. Eysholdt. “Imaging of vocal fold vibration by digital multiplane kymography.” *Comput. Med. Imaging Graph.*, **23(6)**:323–330, 1999.

[UMH13] Jakob Unger, Tobias Meyer, Christian T. Herbst, W. Tecumseh S. Fitch, Michael Döllinger, and Jörg Lohscheller. “Phonovibrographic wavegrams: Visualizing vocal fold kinematics.” *J. Acoust. Soc. Am.*, **133**:1055–1064, 2013.

[Vel98] Raymond Veldhuis. “A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation.” *J. Acoust. Soc. Am.*, **103**:566–571, 1998.

[WC91] K. Wu and D. G. Childers. “Gender recognition from speech. Part I: coarse analysis.” *J. Acoust. Soc. Am.*, **90**:1828–1840, 1991.

[YAK05] Yuling Yan, Kartini Ahmad, Melda Kunduk, and Diane Bless. “Analysis of vocal-fold vibrations from high-speed laryngeal images using a Hilbert transform-based methodology.” *J. Voice*, **19**:161–175, 2005.

[YCB06] Y. Yan, X. Chen, and D. Bless. “Automatic tracing of vocal-fold motion from high-speed digital images.” *Biomedical Engineering, IEEE Transactions on*, **53**:1394–1400, 2006.

[Zha08] Zhaoyan Zhang. “Influence of flow separation location on phonation onset.” *J. Acoust. Soc. Am.*, **124**:1689–1694, 2008.

[ZKG13] Zhaoyan Zhang, Jody Kreiman, Bruce R. Gerratt, and Marc Garellek. “Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models.” *J. Acoust. Soc. Am.*, **133**:453–462, 2013.