Acoustic Models for the Analysis and Synthesis of the Singing Voice

A Dissertation
Presented to
The Academic Faculty

by

Matthew E. Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Electrical and Computer Engineering
Georgia Institute of Technology
March 2005

Acoustic Models for the Analysis and Synthesis of the Singing Voice

Approved by:

Professor Mark A. Clements, Co-Advisor
School of ECE
Georgia Institute of Technology

Professor Monson H. Hayes, III
School of ECE
Georgia Institute of Technology

Professor Mark J. T. Smith, Co-Advisor
School of ECE
Georgia Institute of Technology

Professor Mary A. Ingram
School of ECE
Georgia Institute of Technology

Professor Thomas P. Barnwell, III
School of ECE
Georgia Institute of Technology

Professor Aaron Bobick
College of Computing
Georgia Institute of Technology

Date Approved: 17 March 2005

To my family,

for their constant love and support.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God whose endless supply of strength, wisdom, patience, and faith has made all of my accomplishments possible.

I owe my deepest gratitude to my advisors, Professors Mark J. T. Smith and Mark Clements, for their consistent guidance, knowledgable recommendations, and multi-faceted support which were invaluable for the completion of my research but also for teaching me to be be a better researcher, scientist, and teacher. My thanks also go to Professor Tom Barnwell for contributing his unique perspective and instilling in me a sense of appreciation for the rich history and tradition of speech and signal processing. I would also like to thank Professor Monson Hayes for serving on my reading committee and for letting me beat him up on the tennis court. I am very grateful to Professors Mary Ingram and Aaron Bobick for graciously serving on my defense committee. Special thanks go out to Dr. Bryan George and the late Dr. Michael Macon for our many helpful discussions on sinusoidal modelling.

On a personal note, I would like to express my appreciation to the many past and present colleagues I've had at CSIP, particularly Robert Morris, Tami Randolph, Adriane Durey, Spence Whitehead, Vince Emanuele, Mike Farrell, and our beloved flag football teams Grad Attack I, II & III. I would especially like to thank Paul Hong and Kevin Chan for their remarkable friendship and generosity. Our many lunch conversations taught me many important lessons but mostly how everything in life can be expressed as either a mathematical formula, a few unix commands, or both.

We have a truly dedicated staff at CSIP and I would like to thank each of them for their tireless effort behind the scenes to make all of our work possible, especially Christy, Kay, and Charlotte.

I have been fortunate enough to have made some great friends who have provided me with an incredible amount of support, even if they didn't quite understand the intricacies of signal processing. I owe a great deal of love and thanks to these special people, especially Tom, Ben, Cozad, BJ, Sergio, Scott, Bryan, Rahul, and Antrach.

I would like to give a special thanks to Jenny, my fiancée, who contributed not only priceless hours of her vocal talents to this project but also her unwavering love and faith. Together, we've been though both triumphs and tragedies, but her support and prayers remained constant.

Finally, I would like to thank my extended family for their eagerness to support me in any way they could. I can not begin to express the gratitude I have for my parents because of their incredible love and support. From the beginning, they inspired me to pursue engineering but also taught me the value of dedication, hard work, patience, and faith. Throughout my life, their words of wisdom reminded me that with God, anything is possible...even a Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Throughout our history, the singing voice has been a fundamental tool for musical expression. While analysis and digital synthesis techniques have been developed for normal speech, few models and techniques have been focused on the singing voice. The central theme of this research is the development of models aimed at the characterization and synthesis of the singing voice. First, a spectral model is presented in which asymmetric generalized Gaussian functions are used to represent the formant structure of a singing voice in a flexible manner. Efficient methods for searching the parameter space are investigated and challenges associated with smooth parameter trajectories are discussed. Next a model for glottal characterization is introduced by first presenting an analysis of the relationship between measurable spectral qualities of the glottal waveform and perceptually relevant time-domain parameters. A mathematical derivation of this relationship is presented and is extended as a method for parameter estimation. These concepts are then used to outline a procedure for modifying glottal textures and qualities in the frequency domain.

By combining these models with the *Analysis-by-Synthesis/Overlap-Add* sinusoidal model, the spectral and glottal models are shown to be capable of characterizing the singing voice according to traits such as level of training and registration. An application is presented in which these parameterizations are used to implement a system for singing voice enhancement. Subjective listening tests were conducted in which listeners showed an overall preference for outputs produced by the proposed enhancement system over both unmodified voices and voices enhanced with competitive methods.

CHAPTER 1

INTRODUCTION

The singing voice lies at the very heart of musical expression. Through the combination of music, lyrics, and emotion, the singing voice is able to convey powerful sentiments and thoughts in a manner that entertains listeners across all cultures. The versatility of the singing voice is reflected in its application to all genres of music, from opera to rock-and-roll. For countless years, society has had an appreciation for good singing voices and trained performing artists. However, our understanding of how to model, enhance, and synthesize singing electronically is presently quite limited. The concept of digitally synthesizing a good singing voice or improving the vocal quality of a poor one has only begun to attract the attention of researchers. This challenge, however, has produced many more questions than answers. While efforts aimed at synthesizing other musical instruments have produced realistic and natural sounding results, the singing voice has yet to be convincingly simulated with synthesis techniques. This is mainly attributed to the complex nature of the singing voice production mechanism. By careful positioning of the many organs of the vocal apparatus (jaw, lips, tongue, etc.), singers are able to produce an incredibly wide variety of sounds. Even small perturbations of any of these components can vastly alter the acoustic properties of the produced waveform as well as a listener's perceptual response.

For several years, recording artists and producers have taken advantage of basic speech synthesis methods for making limited enhancements to recorded voices. The karaoke industry has also incorporated many of these features into their machines. Many of these modification techniques are based on *wavetable synthesis* methods such as *pitch-synchronous overlap-add* (PSOLA) [27, 56]. PSOLA operates by sampling windowed portions of the original signal and then resynthesizing them with a basic overlap-add procedure. Time

scaling is performed by deleting or repeating windowed sections prior to the overlap-add procedure. Pitch-scale modifications are also possible by adjusting the spacing between overlapped windows during resynthesis. Methods of this type have been a popular choice mainly because of their simplicity and capability of high fidelity playback. However, these methods offer only crude modifications that often result in objectionable artifacts [80]. The nature of singing voice synthesis places a high demand on a natural-sounding, artifact-free synthesis procedure.

The interest in efficient and flexible speech models led to the development of a class of *sinusoidal models* in the mid-1980s. Sinusoidal models were initially explored by McAulay and Quatieri [47,49] as well as Marques and Almeida [46] and shown to be an effective representation for speech. By representing a voiced waveform as a sum of sinusoidal components, sinusoidal models have found uses in a wide range of applications. Later work with this model showed the potential for time-scale modification and pitch alteration [62, 64].

An extension to McAulay and Quatieri's work was developed by George and Smith [24–26]. The Analysis-by-Synthesis/Overlap-Add (ABS/OLA) model is based on the combination of a block overlap-add sinusoidal representation and an analysis-by-synthesis parameter estimation technique. ABS/OLA performs synthesis by employing an efficient FFT implementation. Improvements to the prosody modification techniques of this system were implemented by Macon and applied to text-to-speech and singing voice synthesis (LYRICOS) applications [40–43]. The LYRICOS system uses sinusoidal-modeled segments from an inventory of collected singing voice data to drive a concatenation-based synthesis engine.

These are only a few of the models that have been advanced in the past for synthesizing voiced song. While these methods are capable of performing some modifications to a singing voice, such as time and pitch-scale modifications, little has been done in the way of parameterizing the characteristics associated with singing in a way that allows one to digitally transform a poor singer into a good one. For this goal to be realized, a method for

characterizing the voice production mechanism must be considered so that differences in the production of singing voices of varying styles and qualities can be characterized. These are important keys that enable us to take steps toward our ultimate goal of enhancing and synthesizing a singing voice.

The source-filter model for speech is based on a simplified human voice production system where no interaction between the source and vocal tract is assumed. According to this model, the simplified human voice production system is decomposed into three elements: glottal source, vocal tract, and radiation impedance. This is illustrated in Figure 1. The radiation impedance is typically approximated with a simple differentiation filter. Since both the vocal tract filter, $V(z)$, and the radiation filter are linear time-invariant (over short frames), the glottal source and radiation impedance can be combined to form the glottal derivative waveform, $U'(z)$. The result of these manipulations is a source-filter synthesis model in which the human voice is modeled as a vocal tract filter excited by a glottal excitation.

Because of the separable nature of the source-filter model, characterization and enhancement of the singing voice can be performed on the individual components of the voice production mechanism. Both the vocal tract filter and glottal excitation have been shown to be very different in their composition and thus require different techniques for analysis and modification.

In the source-filter representation, the vocal tract is commonly modeled as an acoustic tube of varying diameter [18]. This model is further simplified by dividing it into cylindrical sections of equal width. Depending on the shape of the acoustic tube, a sound wave traveling through it will be reflected in a certain way so that interferences will generate resonances at certain frequencies. These resonances are called *formants*. Their location largely determines the speech sound that is heard as well as its vocal quality [10].

The ability to manipulate the characteristics of the vocal tract is largely dependent on the formant structure of the vocal tract spectrum. Formant characteristics have long been

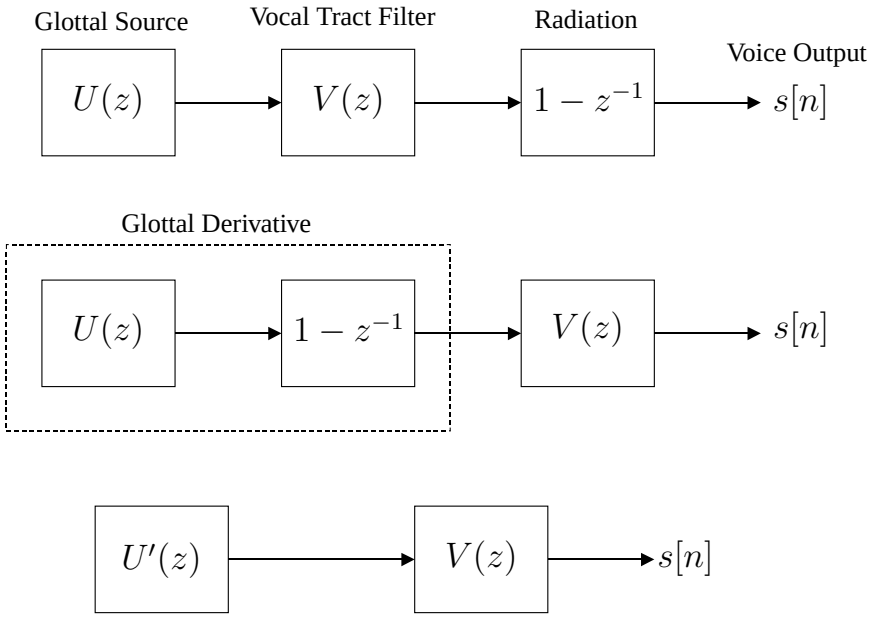


Figure 1: Illustration of the source-filter model for speech production. Because the original model (top) is composed of filters that are linear time-invariant, the glottal source filter ($U(z)$) and the radiation filter can be combined (middle) to form a filter representing the glottal derivative ($U'(z)$) which serves as the excitation to the vocal tract filter.

known to hold importance for recognition and intelligibility, but research has shown that various vocal qualities in singing can also be derived. One example is the *singer's formant*, which is a clustering of the 3rd, 4th, and 5th formants. This merged formant causes a perceptual *ringing* quality in a singer's voice [73].

Modification of formant structure can be performed in a number of ways. All-pole models such as LPC offer formant modification through the shifting and scaling of pole locations. Other methods modify the spectral envelope with functions that warp the envelope along the frequency and/or amplitude axes. These methods, however, are capable of making only limited modifications and offer little control over important formant characteristics. For example, pole modification does not allow a particular formant's bandwidth and amplitude to be easily controlled independently.

Vocal qualities in singing, however, are not solely based on the characteristics of the

vocal tract. The glottal excitation has a significant impact on the vocal textures of a singer's voice. Glottal characteristics have been shown to be correlated with various voicing modes ranging from *pressed* to *breathy*. Further studies [8] have outlined relationships between glottal source characteristics and stress levels in speech.

A model for the glottal excitation that is both accurate and flexible enough for modifications is a crucial component to an effective singing voice enhancement system. Several models for the glottal source waveform have been proposed that can accurately capture glottal characteristics in either the time or frequency domains [13, 20, 66, 85]. However, methods for using such models to effectively enhance the perceptual characteristics of a singing voice have yet to be discovered.

1.1 Research Overview

This thesis presents a multi-fold approach to parameterizing and modifying vocal qualities. As shown in Figure 2, the components of the source-filter model are modified on an individual basis. First, a new spectral model for modifying the formant structure of the vocal tract is investigated. Current methods for spectral modification have been shown to provide only a low level of control over important formant characteristics. Additionally, an algorithm for identifying glottal characteristics in the frequency domain is presented that is used to modify the source excitation in an effort to control the vocal texture of a singing voice waveform. These two modification methods, operating within the context of the source-filter model, are combined with prosodic modifications for correcting pitch and inserting vibrato to perform natural-sounding enhancements to a singing voice. Furthermore, techniques presented in this system are also capable of providing detailed characterization of a particular singing voice so that it can be classified according to skill, style, gender, register, and vocal texture.

Through the presentation of these methods and their application to singing voice enhancement as well as results of subjective human listening tests, it will be shown that the

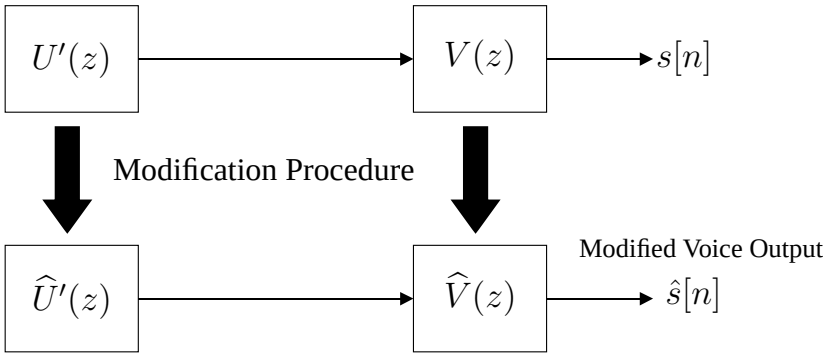


Figure 2: In the presented modification procedure, each component of the source-filter model is modified independently of one another.

models presented in this work can provide a framework for high-quality modifications to the human voice, offering advantages over competing methods.

This thesis is organized as follows:

Chapter 2 presents a brief background of the singing voice and outlines characteristics that differentiate the voice of a trained singer from that of an untrained singer.

Chapter 2 provides background information on spectral modeling and modification techniques. Basic methods for estimating the spectral envelope are presented, followed by an analysis of various modification algorithms.

Chapter 3.2.2 begins with an overview of the proposed spectral modification process. The ABS/OLA sinusoidal model, which provides the framework for the modification algorithm, is then presented. The actual modification procedure and its implementation are then discussed in detail.

Chapter 4.4.1 presents an overview of time-domain glottal flow models as well as a derivation of equations relating these model parameters to frequency-domain characteristics.

Chapter 5.2.2 further explores the relationship between glottal parameters in the time and frequency domains. Based on these findings, a frequency-domain parameter estimation technique is proposed that is able to capture important time-domain characteristics. This chapter also presents a proposed technique for glottal modification.

Chapter 6.3 provides a detailed analysis of a set of recorded waveforms sung by singers with extensive professional training in a classical style and singers with no previous experience or training. The proposed models are used to show characteristics and differences among the groups of singers.

Chapter 7.3 details an implementation of the proposed techniques for enhancing the singing voice of untrained singers base on the findings in Chapter 6.3. The results of subjective listening tests quantifying the performance of the enhancement system are also provided.

Chapter 9 concludes the thesis with a summary of contributions and future work.

CHAPTER 2

BACKGROUND: THE SINGING VOICE

To improve the vocal quality of the singing voice, characteristics must be identified that differentiate "good" singers' voices from "poor" ones. This, however, can be a very subjective endeavor, especially when considering a wide variety of singing styles. For example, the voice of a singer trained in the tradition of the musical theatre (also known as the *belt* voice [17, 51]) may not be considered proficient for an operatic performance of the Western classical tradition. By and large, the majority of singing voice research has focused on singing based on traditional methodologies for vocal training. This produces a style of singing that is most commonly referred to as the *classical* voice. Researchers have identified a number of characteristics in classically trained singers' voices that are commonly absent or less evident in the voices of untrained singers.

Vibrato

Vibrato occurs in most Western opera and concert singing and often in popular singing as well. Vibrato can be described as a nearly sinusoidal modulation of the fundamental frequency during voiced segments. The rate of vibrato is typically 5-8 Hz, and the modulation depth varies between ± 50 and ± 150 cents (where 1200 cents = 1 octave) [74]. Although the rate and depth of vibrato may vary from singer to singer or from genre to genre, there is an acceptable range among trained singers. Studies have shown that the voices of trained singers exhibit vibrato with greater depth and regularity than for those of untrained singers [7]. Additionally, the presence of vibrato has been shown to be directly correlated with the perception of vocal beauty. Robison [65] found that baritones with the most aesthetically pleasing voices maintained vibrato in their tones more than 80% of the time.

The pitch contours for the vowel */o/* sung by both a trained singer and an untrained singer are shown in Figure 3. Both signals clearly show vibrato-like fluctuations, but the depth and consistency are much greater in the contour of the trained singer.

In addition to frequency modulation, vibrato has been shown empirically to have an associated modulation in amplitude as well as spectral shape [44]. The perceptual effects of these amplitude modulations, however, are secondary to the frequency modulation effects.

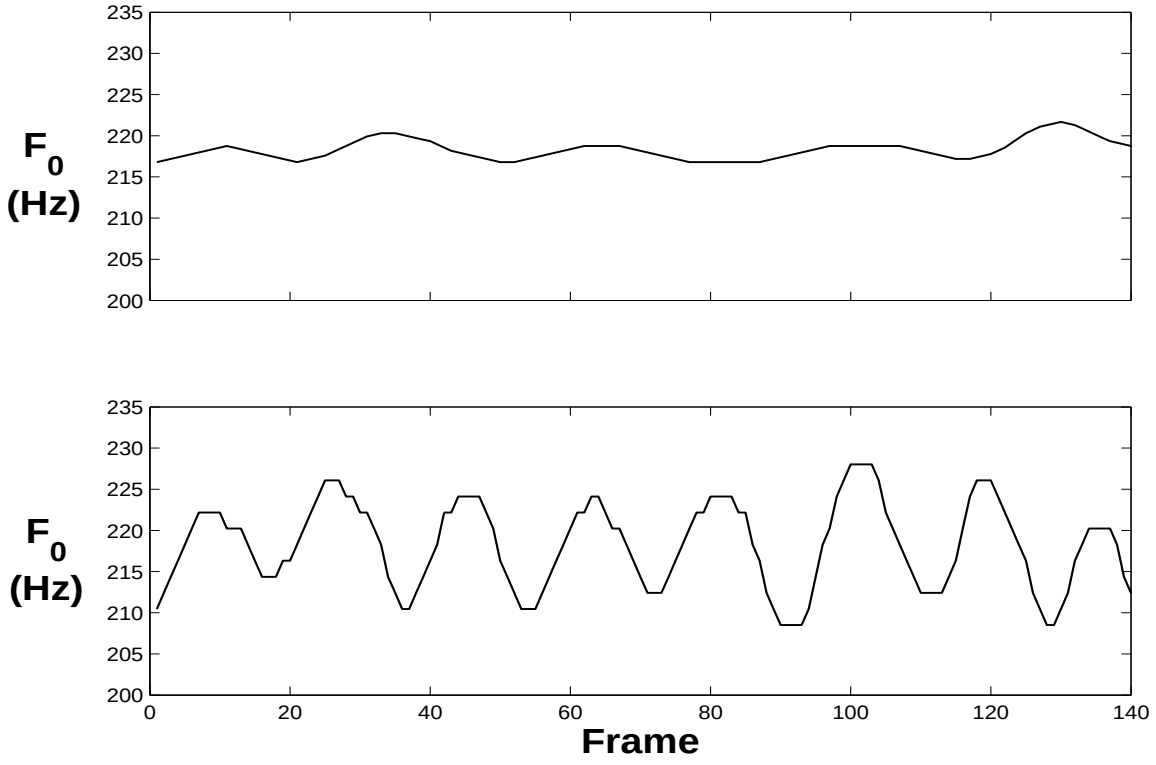


Figure 3: Fundamental frequency tracks of the vowel */o/* for an untrained singer (top) and a trained singer (bottom).

The Singer's Formant

Trained singers (especially males) often create a resonance in the range of 3000 to 5000 Hz by employing a technique in which the larynx is lowered. Acoustically, this

results in a clustering of the third, fourth, and sometimes fifth formants. This resonance, referred to as the *singer's formant*, adds a perceptual loudness that allows a singer's voice to be heard over a background accompaniment [72]. This phenomenon coincides with a perceptual *ring* in a singer's voice. According to Sundberg, the singer's formant is generated as a result of an intentional lowering of the larynx, which leads to a wider pharynx.

Figure 4, which presents the spectral envelopes of the two aforementioned singers averaged over time, clearly illustrates the singer's formant centered at approximately 3000 Hz. Ekholm [16] found that the presence of the singer's formant in the voice, much like vibrato, is strongly correlated with the perception of vocal beauty.

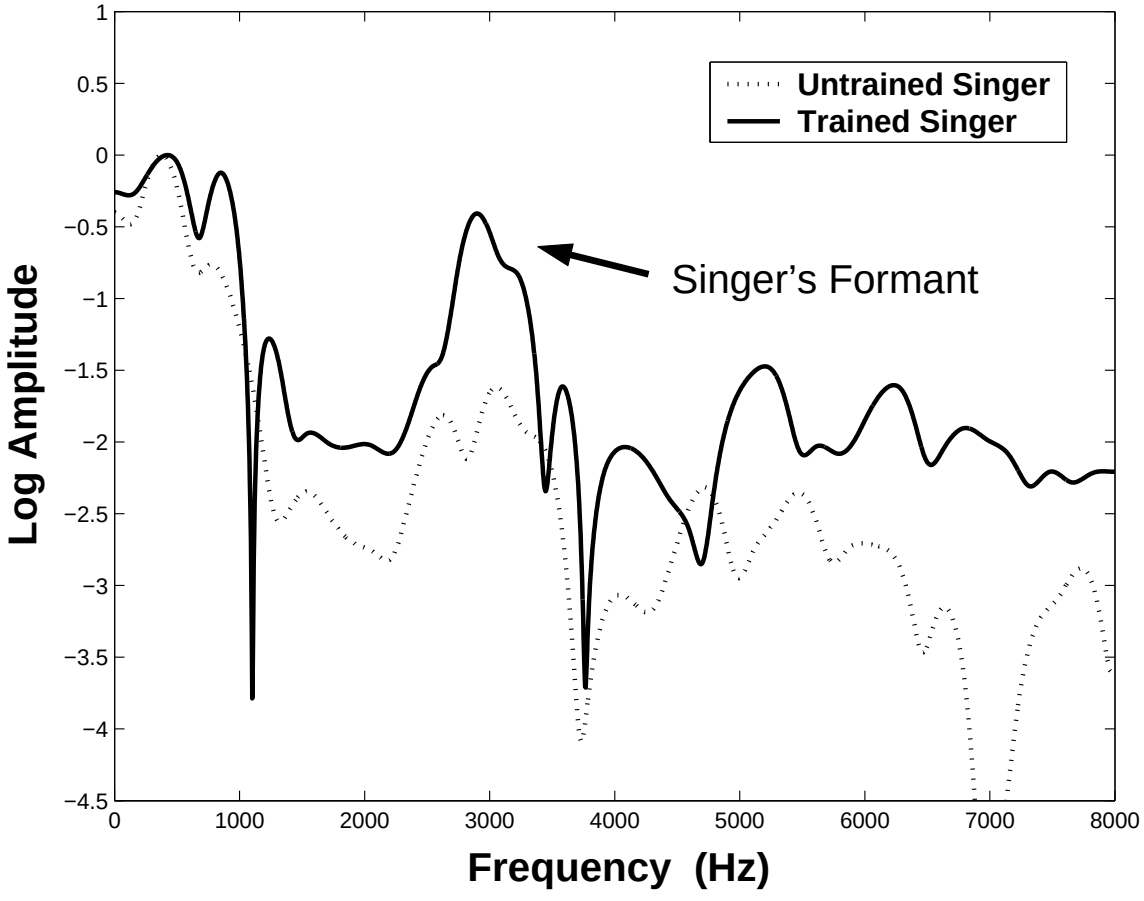


Figure 4: Average spectral envelopes for the vowel */o/* for a trained singer and an untrained singer. The arrow indicates the region of the singer's formant.

Timbre/Tone Color

In general, sounds may be characterized by pitch, loudness, and quality. *Timbre* describes those characteristics of sound that allow the ear to distinguish sounds that have the same pitch and loudness. Timbre is then a broad term for the distinguishable characteristics of a tone. In singing, it is often referred to as *tone color*.

Many perceptual characteristics of timbre are evident in the spectral content or formant structure of a singer's voice. Trained singers often modify the formant structure of their voice in order to add certain desirable characteristics. For example, a lowered second formant results in a "darker" voice–often referred to as "covered"–while a raised second formant produces a "brighter" voice [71]. The covered voice reflects a softer sound, which is desired by singers trained in the Western classical tradition, unlike broadway and pop singers who typically use a brighter voice.

Another interesting phenomenon has been documented in trained female singers' voices. Female singers often shift their first formant to match the fundamental frequency when the fundamental rises above the first formant [72, 73]. This has the effect of increasing the intelligibility of a vowel sound for a performing artist.

Glottal Source

The glottal source waveform has been shown to possess certain qualities that have a great impact on the acoustic characteristics of voiced speech. These characteristics play a large part in determining the individuality as well as vocal qualities of a spoken or sung voice. Applications in the synthesis and enhancement of speech or singing voice require improved naturalness and a higher level of control over these vocal qualities. Identifying and modeling the glottal waveform attributes that produce these qualities can help improve the performance of these applications as well as others such as speech recognition, speaker identification, and voice pathology identification. Such advances can also be used to gain a further understanding of the

glottal characteristics of normal speakers.

The glottal source is typically characterized in either the time or frequency domains. In the frequency domain, studies have aimed at identifying glottal characteristics that can be used to describe perceptual cues. Much of this research has identified spectral characteristics affiliated with vocal qualities (i.e., breathiness, pressed) or sex. Klatt and Klatt [34] and Hanson [28] found that the main spectral parameters necessary for synthesizing natural-sounding voices with different qualities are spectral tilt, bandwidth of the first formant, relative noise levels, and amplitudes of the first few harmonics. While these characteristics have been found to be of great use in identifying perceptual qualities of the voice, they have yet to be successfully used to produce these qualities in synthesis applications.

While several frequency-domain glottal models have been proposed based on these findings, they typically model the glottal waveform with parameters in which the correlation with time domain parameters is unclear. In order to retain the temporal information of the glottal waveform, it is important to identify spectral correlates for time-domain glottal parameters.

In the time domain, the glottal waveform is typically characterized by various measures such as the fundamental period, open quotient, and glottal asymmetry–usually denoted T_0, O_q, and α, respectively. Measures such as these have proven to be integral in characterizing vocal effort, prosodic variations and a wide variety of vocal qualities. The values of these parameters can vary, depending on the configuration of the glottal mechanism. Several glottal flow models have been developed in which many important glottal characteristics such as these are parameterized in the time domain. Cummings and Clements denoted this broad class of models as *non-interactive parametric glottal models* [9]. These models are based on the assumption that the glottal source and vocal tract are linearly separable and that no interaction occurs between the two. Examples of models of this type were proposed

by Rosenberg [66] and Klatt & Klatt [34] (KLGLOTT88). The effectiveness of these models lies in their ability to capture timing relationships that have been shown to have an important perceptual impact on speech signals. While these two particular models assume an abrupt closure of the glottis, other models such as those developed by Fant, Liljencrants, Lin [20] (LF model), and Veldhuis [85] (R++) provide an additional parameter that describes the *return phase* of a glottal cycle. This parameter provides an increased level of flexibility that enables a better fit of the glottal derivative waveform.

While synthesis systems based on either time-domain or frequency-domain parameterizations of the glottal source have had mixed success at producing vocal qualities and textures, a method of utilizing both sets of parameters in a single domain might provide a significant improvement.

Vocal Registration

Vocal registration has been the subject of a vast amount of research but has remained largely controversial in its precise definition and underlying mechanisms. There is general agreement, however, that a register is a series of adjacent tones on the scale that (a) sounds equal in timbre and (b) is felt to be produced in a similar manner.

A register covers a certain frequency range, but adjacent registers do overlap, so it is often possible for the same note to be sung in two different registers. Trained singers have traditionally been taught to impose a smooth transition between registers by "blending" them during transitional regions. This is referred to as *passagio*. The voices of untrained singers often contain register "breaks," which are sharp shifts from one register to another.

The male voice is often distinguished as having three registers, normally referred to as *chest*, *head*, and *falsetto*. A male singer normally sings in the chest and head registers but will commonly oscillate in and out of the falsetto register when breaks occur. The falsetto register for males can also be thought of being used when trying to imitate a female voice.

The literature normally identifies three registers [52, 73] in the female voice: *chest*, *middle*, and *head*. However, some females are capable of singing in a special mode at the upper frequency range in what is referred to as the *whistle* register.

Registers are generally assumed to correspond to voice source properties as determined by the muscular tuning of the vocal folds, particularly the vocal ligament. Hence, they should be basically independent of vocal tract resonance. A number of studies have used various glottal models to identify modalities in glottal behavior that can be used to determine register shifts as well as register-specific characteristics in a singer's voice [31, 77–79].

CHAPTER 3

BACKGROUND: SPECTRAL MODELS AND MODIFICATION TECHNIQUES

Before attempting to develop a model for singing voice analysis and synthesis, it is important to have an understanding of the issues that are specific to this particular task. The singing voice has been shown to be very different from normal speech, and thus speech processing techniques that were designed for general speech are not always suitable for the singing voice. However, many of these techniques can at least provide a basis from which algorithms specific to the analysis, synthesis, and modification of the singing voice can be derived. As mentioned earlier, the source-filter model for speech production can be divided into two components that can be independently modeled and modified. This chapter focuses on the vocal tract filter and its spectral representations. An investigation of current existing methods for spectral modeling and modification for speech is presented.

Methods for spectral modification are targeted at altering the perceived characteristics of a speaker's or singer's voice. In the singing voice, this can be thought of as altering the timbre or tone color by controlling the underlying formant structure that resides in the spectrum. This is generally performed by first identifying a parametric spectral model for the voice and then systematically adjusting parameters in order to modify vocal qualities. Spectral modification can serve a variety of alternate applications such as speaker normalization and voice conversion. The goal of this work, however, is to develop a spectral model and modification algorithm to enhance vocal qualities in the singing voice. Figure 5 shows a block diagram of the analysis, modification, and synthesis procedures.

There are a number of techniques for estimating the vocal tract response in the source-filter model. This filter is often defined in frequency as a *spectral envelope*. While there are several methods for accomplishing this, we will discuss a few basic methods from which

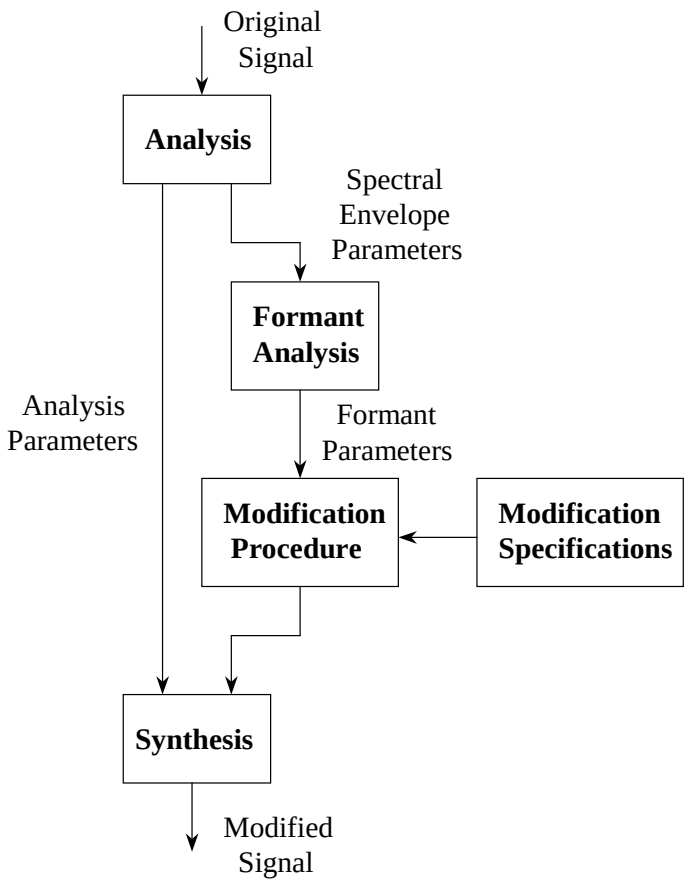


Figure 5: Block diagram of the analysis, modification, and synthesis procedures for modifying vocal qualities.

most alternative methods are derived.

3.1 Spectral Envelope Estimation

In computing the spectral envelope of a signal for the purposes of spectral modification, there are a number of factors that should be considered in choosing a proper method for estimation.

Accuracy It is important for a spectral envelope to provide a proper fit to the magnitude spectrum. A spectral envelope should fit the peaks of the partials without tracing any one of them individually.

Smoothness A certain level of smoothness is desired for a spectral envelope. In frequency, a spectral envelope should give a general idea of the distribution of the signal's energy. Spectral envelopes should also possess consistency from frame to frame. Excessive oscillations or the presence of discontinuities can lead to unnatural qualities in synthesized waveforms.

Flexibility An effective spectral envelope estimation technique must be capable of handling a wide range of signals with varying characteristics. In speech and signing, a variety of vocal tract configurations are possible, as well as sounds that contain a mixture of both harmonic and noisy contents.

3.1.1 Linear Predictive Coding (LPC)

LPC is an early method originally developed for speech coding and compression. Because of the special properties of this method, it can also be used for spectral envelope estimation. LPC represents the spectral envelope as an all-pole filter. This representation is based on the concatenated lossless acoustic tube model.

As discussed earlier, an acoustic tube representation is commonly used to model the vocal tract. The acoustic tube model, however, omits certain complexities of the vocal tract and is thus not a perfect model. The concatenation of acoustic tubes typically does not account for the effects of the nasal tract. This second cavity is shaped very irregularly and introduces additional resonances and anti-resonances (nasal zeros) because of the effect of coupling. While the zeros are not vital for the recognition of the speech sounds, they can lead to problems in formant detection and characterization. Additionally, certain speech sounds like laterals (e.g., */I/*) have a tongue configuration that is not well described by a simple acoustic tube. The acoustic tube model also ignores the viscosity of the walls of the vocal tract as well as any damping that may occur. Despite these drawbacks, the acoustic tube model performs remarkably well in a wide variety of speech analysis and synthesis applications.

The idea behind LPC analysis is to represent each sample of a signal $s[n]$ in the time domain by a linear combination of the p preceding values, $s[n-p]$ through $s[n-1]$, where p is the order of the analysis [45]. The approximated value, $\hat{s}[n]$, is computed from the preceding values and p predictor coefficients, a_i, as follows:

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

For each time frame, the coefficients, a_i, will be computed such that the prediction error, $e[n] = \hat{s}[n] - s[n]$, for this window is minimized. In coding applications, it is sufficient to send the p coefficients, a_i, and the residual signal, $e[n]$, which uses a smaller range of values and can thus be coded with fewer bits. The receiver can then recover the original signal from $e[n]$ and the filter coefficients, a_i.

When the residual signal, $e[n]$, is minimized, the resulting analysis filter serves to flatten the spectrum of the input signal. The transfer function of this filter is given by

$$A(e^{j\omega}) = 1 - \sum_{i=1}^{p} a_i e^{-j\omega i}. \tag{2}$$

Because this filter removes the spectral envelope from the input waveform, it is generally referred to as the *inverse* filter. In a synthesis application, the synthesis filter provides an approximation of the spectral envelope:

$$\frac{1}{A(e^{j\omega})} = \frac{1}{1 - \sum_{i=1}^{p} a_i e^{-j\omega i}}. \tag{3}$$

As can be seen, the synthesis filter (with gain), $G/A(e^{j\omega})$, is an all-pole filter.

The order of the filter is an important parameter that can affect the accuracy of the spectral envelope. As the order decreases, fewer poles are used and the approximation of the spectral envelope becomes coarser. However, the envelope will still reflect the rough distribution of energy in the spectrum. This is illustrated in Figure 6.

In some cases, the LPC spectral envelope will descend down to the level of residual noise in the gap between two harmonic partials. This occurs when the distance in frequency between partials is large, as in high pitched sounds, and the order of estimation is high. This effect is illustrated in Figure 7.

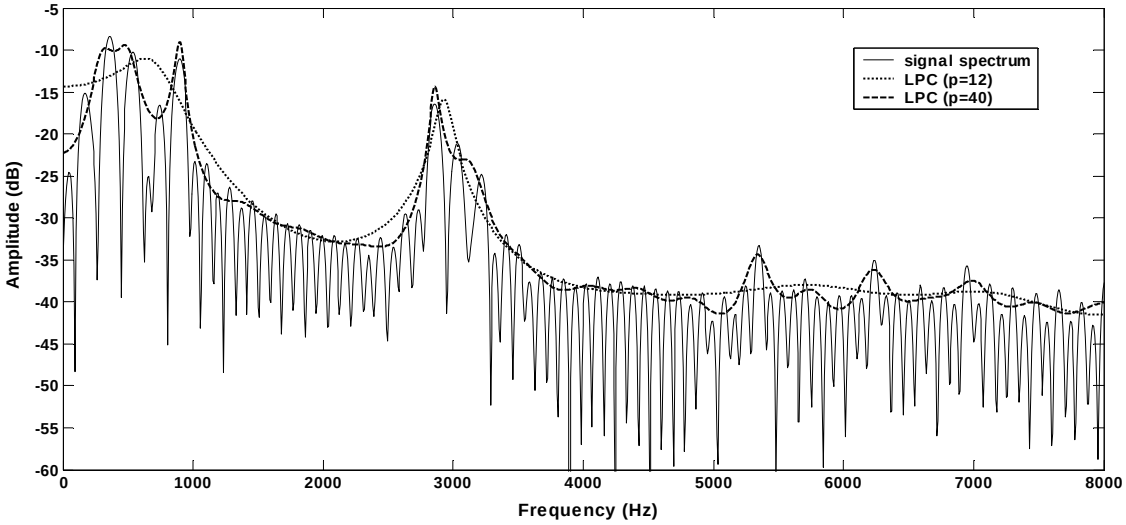


Figure 6: Example of LPC spectral envelopes fitted to the spectrum of a voiced frame. The envelopes are calculated with order p=12 (dotted) and p=40 (dashed).

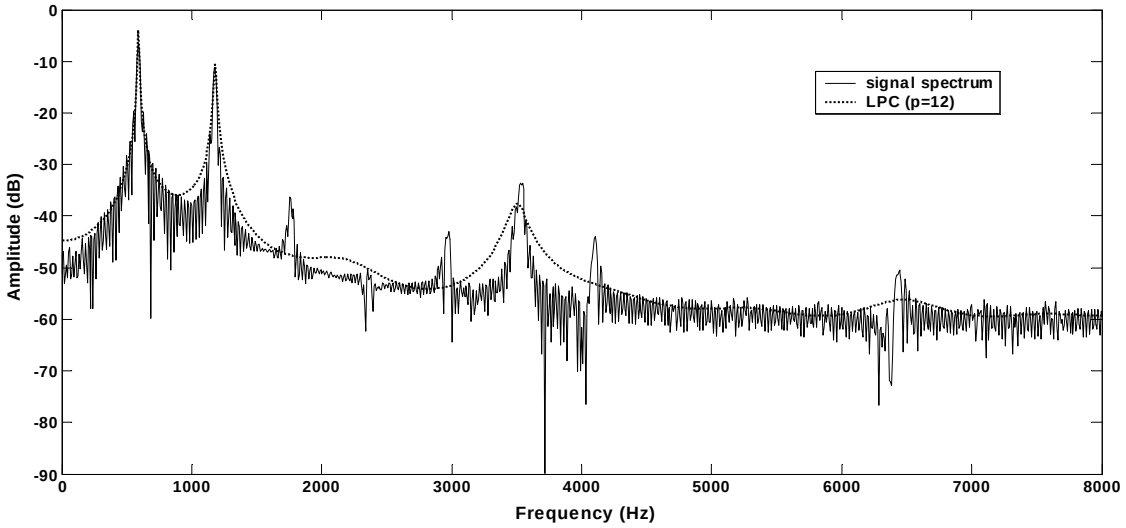


Figure 7: Example of an LPC spectral envelope where the harmonics are spaced too far apart for the LPC order chosen (p=12). The spectral envelope traces unwanted portions of the spectrum in between harmonics.

3.1.2 Cepstrum Spectral Envelope

The cepstrum is a method of speech analysis based on a spectral representation of the signal. According to the source-filter model of speech production, a speech signal $s[n]$ can be expressed as a convolution between a source or excitation signal $e[n]$ and the impulse response of the vocal tract filter $v[n]$:

$$s[n] = e[n] * v[n]. \tag{4}$$

In the frequency domain, this convolution becomes the multiplication of the respective frequency responses:

$$S(e^{j\omega}) = E(e^{j\omega}) \cdot V(e^{j\omega}). \tag{5}$$

Taking the logarithm of the absolute value of the Fourier transforms, the multiplication in (5) is converted to an addition:

$$\log|S(e^{j\omega})| = \log|E(e^{j\omega})| + \log|V(e^{j\omega})|. \tag{6}$$

If we now apply an inverse Fourier transform to the logarithm of the magnitude spectrum, we get the frequency distribution of the fluctuations in the curve of the spectrum, denoted $c[n]$, which is called the *cepstrum* [3, 59]:

$$c[n] = F^{-1}\left[\log|S(e^{j\omega})|\right] = F^{-1}\left[\log|E(e^{j\omega})|\right] + F^{-1}\left[\log|V(e^{j\omega})|\right]. \tag{7}$$

The cepstrum no longer exists in the frequency domain but instead operates in an alternate domain referred to as the *quefrency* domain.

Under the assumption that the source spectrum has only rapid fluctuations (the excitation signal $e[n]$ is a stable, regular oscillation on the order of 10^2 Hz), its contribution to $c[n]$ will be concentrated in its higher regions, while the contribution of $V(e^{j\omega})$ will be the slow fluctuations in the spectrum $S(e^{j\omega})$ and will therefore be concentrated only in the lower part of $c[n]$. This can be seen in Figure 8. Thus, separating the two components is accomplished simply by keeping the first p cepstral coefficients of $c[n]$ and throwing

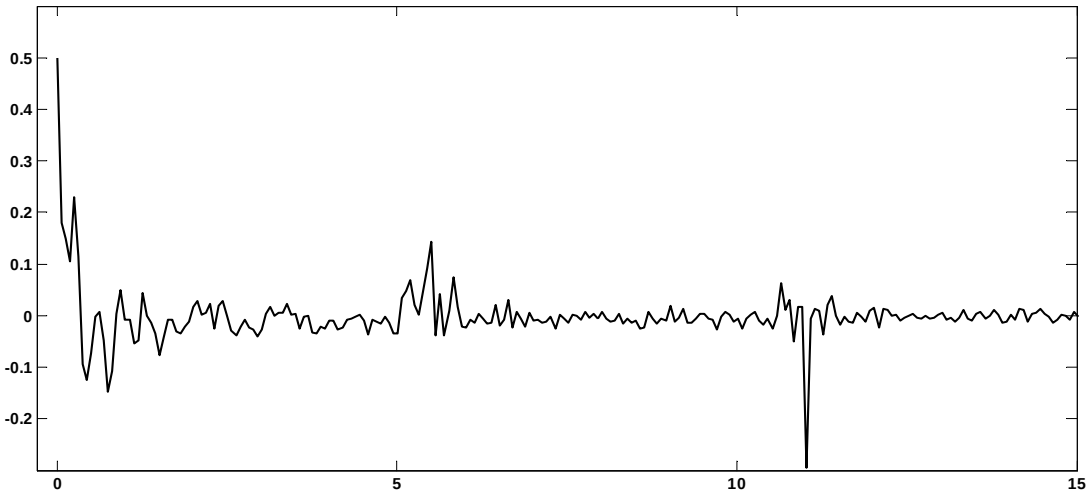


Figure 8: Cepstrum of a voiced frame of speech. The voiced excitation manifests itself as a weighted impulse train, while the vocal tract response is represented by the quickly decaying portion in the lower regions of the "quefrency."

away the remainder. The resulting representation models the low-frequency components or smoothed portion of the spectrum. This interpretation serves as an estimate of the spectral envelope.

There are two disadvantages of the cepstrum method for spectral envelope estimation. First, since the cepstrum is essentially a low pass filtering of the curve of the spectrum, the partial peaks are not always properly linked. Instead, the fluctuations of the spectrum are merely averaged out. This effect is illustrated in Figure 9.

Another disadvantage of the cepstrum method is similar to that of LPC. In cases where both the frequency gap between partials and the estimation order are large, the resulting spectral envelope will trace the residual noise present in the gaps. Figure 10 illustrates this case for the cepstrum estimator.

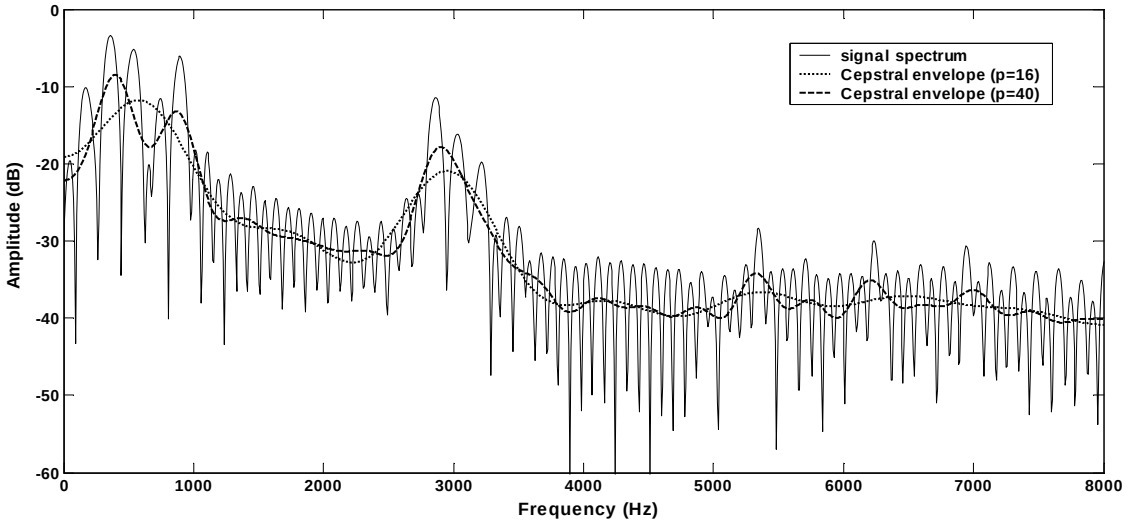


Figure 9: Example of cepstrum spectral envelopes fitted to the spectrum of a voiced frame. The envelopes are calculated with order p=16 (dotted) and p=40 (dashed).

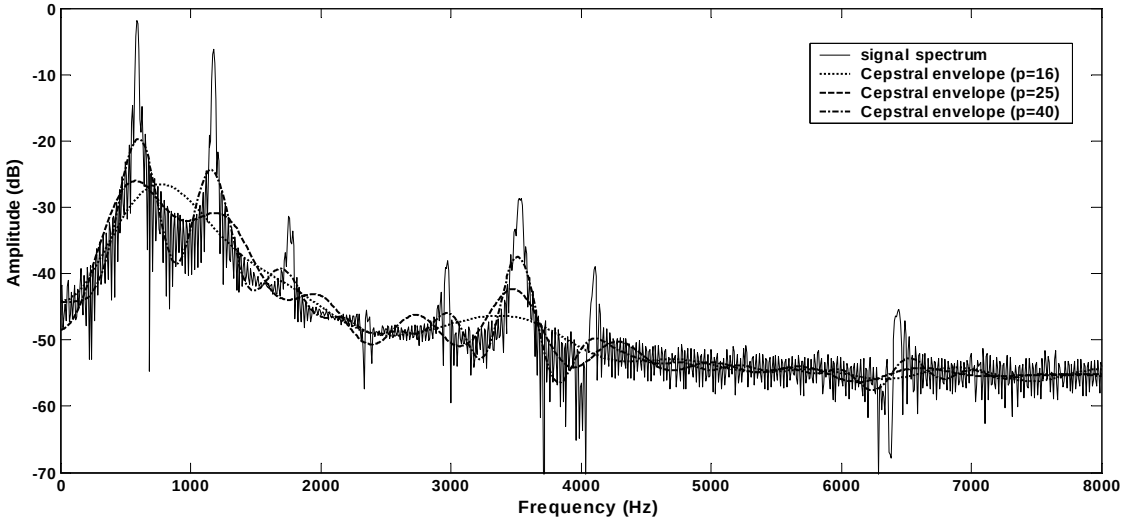


Figure 10: Example of cepstrum spectral envelopes where the harmonics are spaced too far apart. Even with a wide variety of orders chosen, the envelopes do not smoothly connect the harmonic peaks.

3.2 Spectral Envelope Modification

The task of spectral envelope modification is one that has been approached with respect to a variety of applications. While there has been much research aimed at transforming the identity of a speaker, our focus is on improving vocal qualities of the singing voice while maintaining the singer's identity. There are two popular approaches to spectral modification: (1) all-pole (LPC)-based methods of scaling poles by a complex factor in order to alter formant characteristics and (2) frequency warping procedures for modifying the spectral envelope directly.

3.2.1 Pole Modification

When LPC analysis is used to estimate the spectral envelope, formants are assigned to poles and can then be modified to correspond with desired formant locations. The formant structure of the vocal tract can be viewed as a set of cascaded second-order IIR filters of the form

$$F(z) = \frac{1}{(1 - az^{-1})(1 - a^* z^{-1})}. \tag{8}$$

The conjugate pole pair $z = |a|e^{\pm j\angle a}$ and sampling frequency f_s determine the formant frequency F and 3-dB bandwidth B according to

$$F = \frac{f_s}{2\pi}\angle a \text{ Hz}, \tag{9}$$

$$B = -\frac{f_s}{\pi}\ln|a| \text{ Hz}. \tag{10}$$

Formant modifications can be performed by scaling the angle, $\angle a$, and magnitude, $|a|$, of each pole.

Occasionally, when two poles are shifted in frequency too close to one another, only one peak will appear in the spectrum. This is a symptom of *pole interaction*. An example of pole interaction is shown in Figure 11. In part (a) of the figure, the spectral envelope is characterized by pole 1 and pole 2. When pole 2 is shifted to the desired frequency for

the second formant, F_2, as shown in part (b), it is no longer distinguishable from the first formant, F_1.

A number of algorithms have been developed to compensate for pole interaction. Hsiao and Childers [33] define a pole interaction factor that identifies the effect of surrounding poles on a given pole at its center frequency. In a simplified two-pole case, where $z_i = r_i e^{j\phi_i}$ and $z_j = r_j e^{j\phi_j}$, the frequency response of the overall filter at the angle ϕ_i is

$$|H(e^{j\phi_i})|^2 = \frac{1}{(1-r_i)^2} \cdot \Delta|H|_j^2 \tag{11}$$

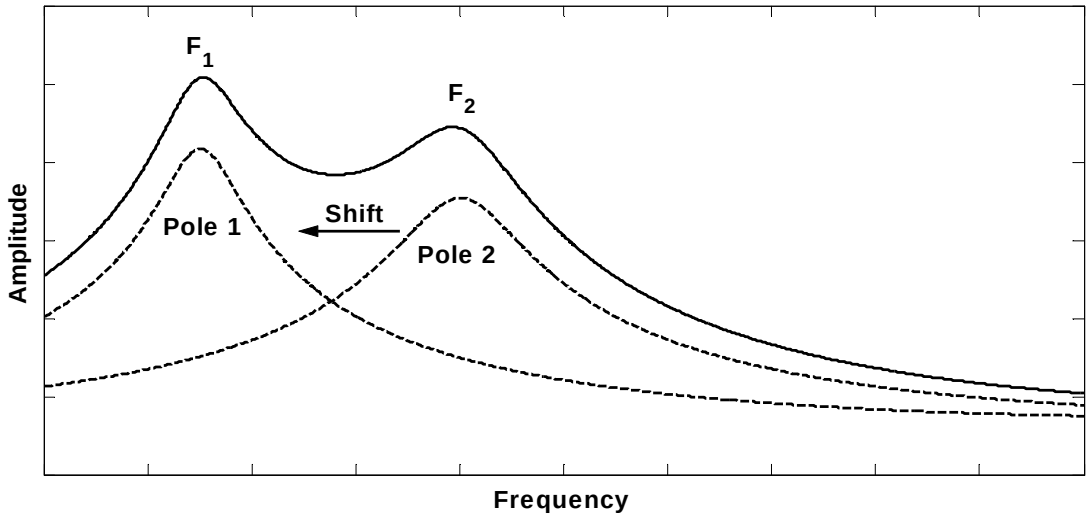
where $\Delta|H|_j^2$ is the pole interaction factor of pole z_j with pole z_i. This factor is defined as

$$\Delta|H|_j^2 = \frac{1}{1 - 2r_j \cos(\phi_i - \phi_j) + r_j^2}. \tag{12}$$

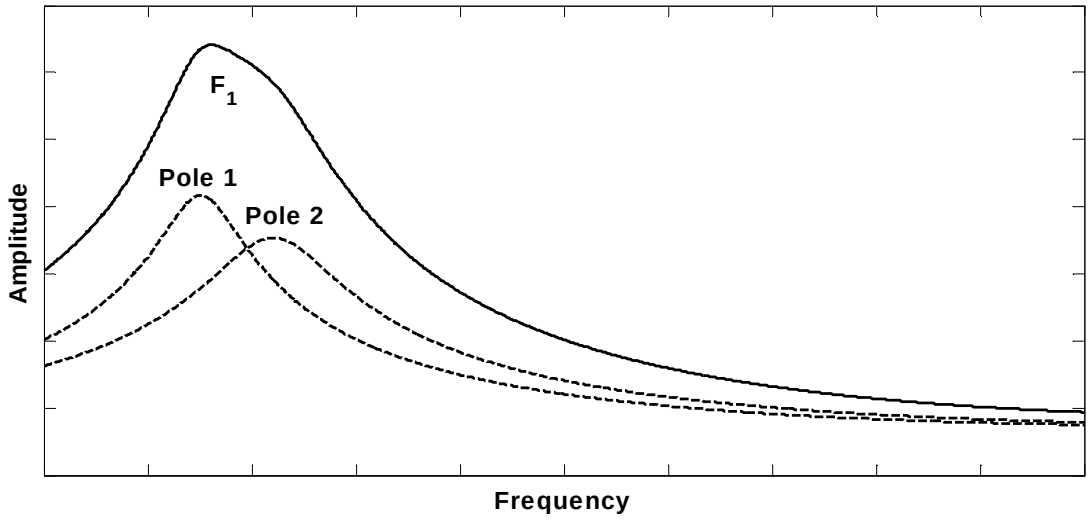
For formant modification, the radii of the poles are scaled in order to achieve desired formant amplitudes in spite of pole interaction. This process is applied iteratively until the spectral deviation, which is defined as the sum of the differences between the resulting formant spectrum and the desired spectrum at the formant frequencies, falls below a threshold. A similar iterative algorithm for overcoming pole interaction during formant modification was also developed by Mizuno, Abe, and Hirokawa [53].

While these methods produce spectral envelopes with desired formant amplitudes at the formant frequencies, one drawback to this technique is that the bandwidth of each formant cannot be controlled. As shown in (10), each formant's bandwidth is dependent on the magnitude of the corresponding pole. Therefore, the amplitude and bandwidth of each formant cannot be independently modified with these procedures.

Recently, modification techniques for transforming the line spectrum pair (LSP) frequencies have been developed [55] that enable a higher level of control over formant characteristics. By taking advantage of the nearly linear relationship between the LSPs and formants, modifications are performed based on desired shifts in formant frequencies and bandwidths.



(a)



(b)

Figure 11: When poles are shifted too close to one another, their respective formants merge and become indistinguishable. This is an example of pole interaction.

3.2.2 Frequency Warping

Frequency warping is a simple method for shifting formants by applying a frequency warping function directly to the spectral envelope. Four parameters typically specify a formant shift. The lower and upper frequencies, f_L and f_U, determine the range of the spectral envelope to be affected. The original formant center frequency and the target center frequency are specified by f_1 and f_2, respectively. The warping function gradually decreases the shift distance as it gets further from F_1. The resulting warping function can either be a piecewise linear function or a smoother realization connecting these parameters. An example of this process is illustrated in Figure 12. Formant bandwidths can also be modified with the use of a warping function. Figure 13 shows a warping function for increasing the bandwidth of a single formant as well as the resulting spectral modification.

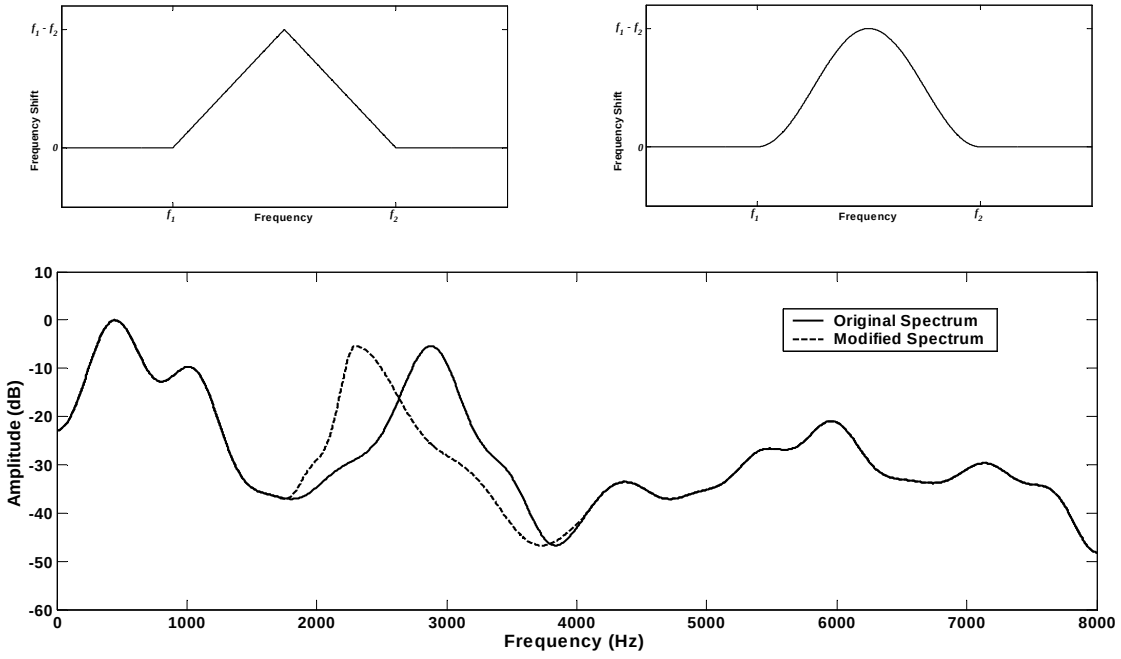


Figure 12: Two possible frequency warping functions are linear (top-left) and Gaussian (top-right). These can be applied to a spectral envelope in order to shift formant locations (bottom).

Turajlic [83] suggests an alternate process in which a frequency shift warping function,

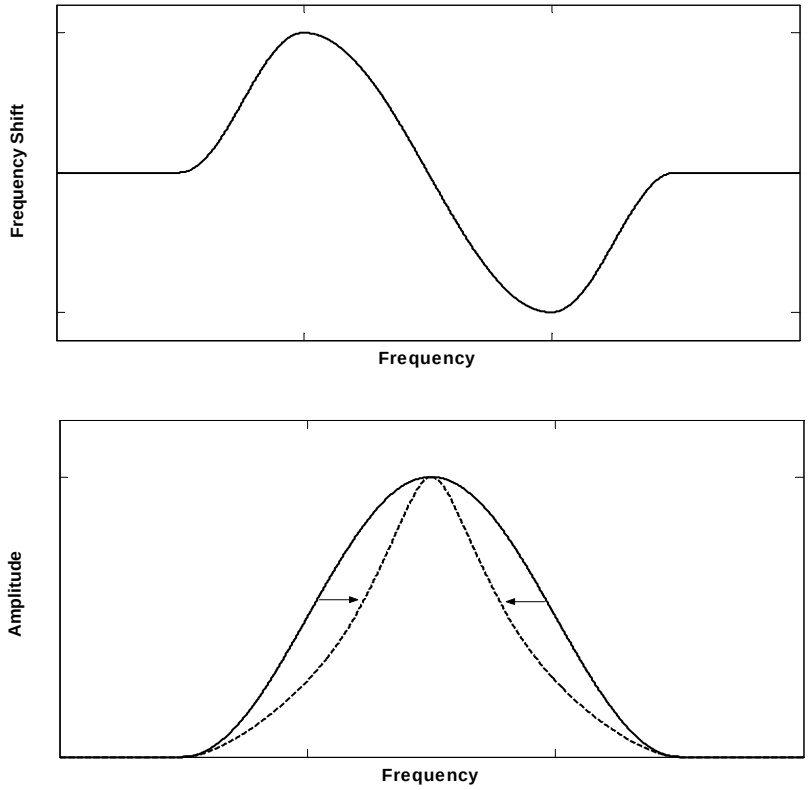


Figure 13: Example of a frequency warping function (top) being applied to alter a formant's bandwidth (bottom).

$\alpha(\omega)$, and a bandwidth warping function, $\beta(\omega)$, are combined with an *intensity shaping function*, $\gamma(\omega)$, which allows modifications to the magnitude of the spectrum. The equation for this process is expressed as

$$\hat{H}(\omega) = \gamma(\omega) \cdot H((\alpha(\omega) * \beta(\omega)) \cdot \omega) \tag{13}$$

where $\hat{H}(\omega)$ is the modified spectral envelope.

Frequency warping methods allow a high level of control over formant characteristics, but only when the original and modified formants are spaced far enough apart so as to be nearly independent of one another. When formants are too close to one another, it is difficult to modify their bandwidths to desirable specifications. This is similar to the pole interaction problem suffered by pole modification techniques. Additionally, frequency warping methods do not allow formants to merge or split as is often desired in formant modification processes. Figure 14 illustrates this phenomenon. In this example, the center frequency of F_3 has been warped from 2300 Hz to 2700 Hz. In doing so, F_3 has failed to split from F_2 (centered at 1800 Hz) and merge with F_4 (centered at 3200 Hz).

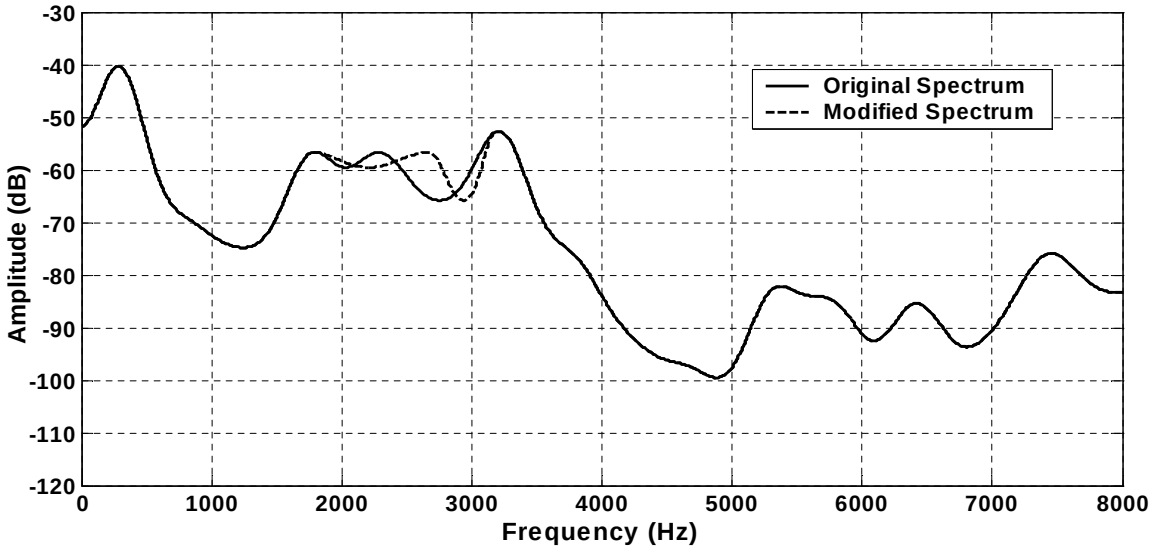


Figure 14: Frequency warping example: the center frequency of F_3 has been warped from 2300 Hz to 2700 Hz, but in doing so, F_3 has failed to split from F_2 (centered at 1800 Hz) and merge with F_4 (centered at 3200 Hz).

CHAPTER 4

A NEW METHOD FOR SPECTRAL MODELING AND MODIFICATION

As described in the previous chapter, spectral modification techniques are capable of performing a variety of modifications to the formant structure that resides within the spectral envelope. However, these techniques are limited by their inability to independently control important formant characteristics such as amplitude and bandwidth. Furthermore, these modifications are only effective when the underlying model for the spectral envelope provides an accurate representation of the formant structure. Any errors such as the ones described in Section 3.2 will render the modification process ineffective and may result in resynthesized outputs with unnatural qualities or artifacts. In this chapter, a new method for spectral modeling and modification is presented that is aimed at overcoming the aforementioned shortcomings of current methods. The block diagram in Figure 15 outlines the proposed spectral modification procedure compared to the general modification procedure.

4.1 Analysis/Synthesis Procedure

The analysis and synthesis procedures provide the spectral modification algorithm with an interface with the actual speech waveforms. The requirements of the analysis and synthesis methods are similar to those of the spectral estimation procedures discussed in Section 3.1. The analysis/synthesis techniques must provide accurate modeling of the dynamic characteristics of the speech production process, as well as flexibility to model and synthesize a wide variety of signals with minimal computational cost.

In order to achieve the requirements of accuracy, flexibility, and computational efficiency, the Analysis-by-Synthesis/Overlap-Add (ABS/OLA) sinusoidal modeling system was chosen to provide the framework for the spectral modification procedure. The ABS/OLA sinusoidal model represents an input signal, $s[n]$, by a sum of equal-length,

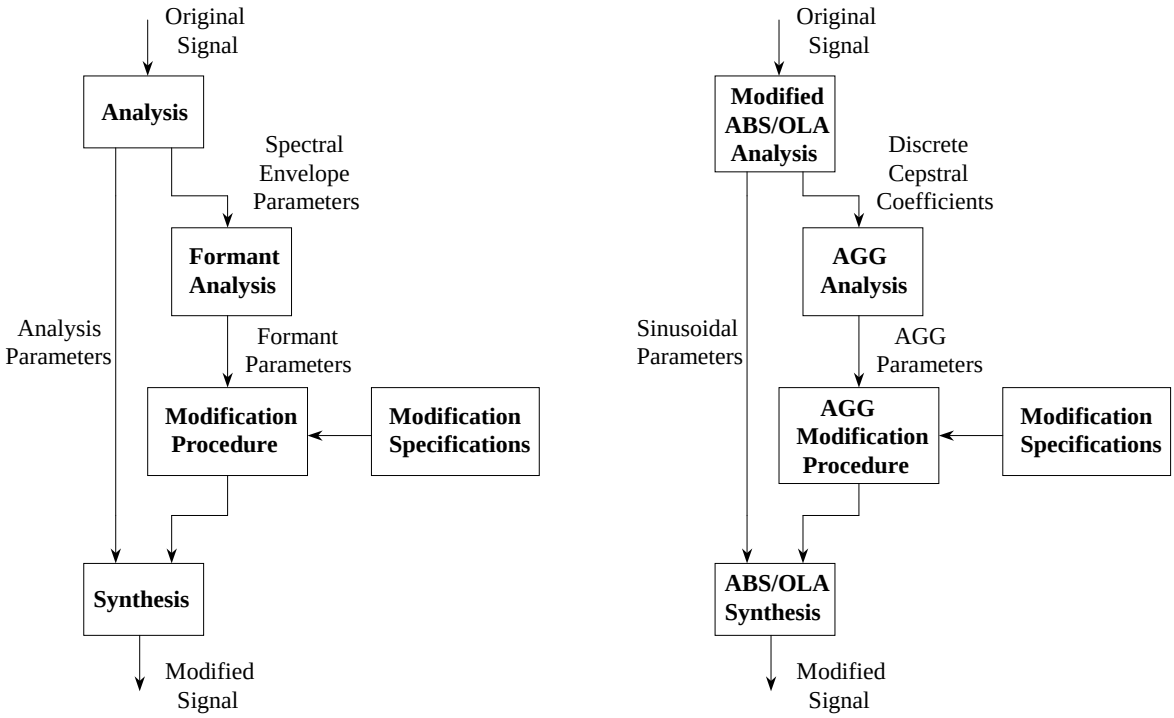


Figure 15: Block diagrams for the original spectral modification procedure (left) and the proposed system (right).

overlapping short-time signal frames, $s_k[n]$. Thus,

$$s[n] = g[n] \sum_k w[n - nN_s] s_k[n], \tag{14}$$

where Ns is the frame length, $w[n]$ is a complementary tapered window function, $g[n]$ is a slowly time-varying gain envelope, and $s_k[n]$ represents the kth frame of the synthesized signal. Each signal contribution, $s_k[n]$, is represented as the sum of a small number of constant-frequency sinusoidal components, given by

$$s_k[n] = \sum_{j=0}^{J-1} A_j^k \cos(\omega_j^k n + \phi_j^k), \tag{15}$$

where J is the number of sinusoidal components in the frame, and A_j^k, ω_j^k, and ϕ_j^k are the sinusoidal amplitudes, frequencies, and phases for the kth frame.

The parameters for each frame are determined by an iterative analysis-by-synthesis procedure designed to minimize a mean-squared error criterion. While the frequencies of the

sinusoids are not restricted to be harmonically related to one another (i.e., exact multiples of the fundamental frequency), only one sinusoid near each harmonic is retained during the analysis. This form of the sinusoidal model is called a "quasiharmonic" representation [24].

The quasi-harmonic nature of the synthesis model implies that the frequency of each sinusoid, ω_j, is at or near a multiple of the fundamental frequency, ω_0, as follows:

$$\omega_j^k = j\omega_0^k + \Delta_j^k, \tag{16}$$

where Δ_j^k is the differential frequency of each component.

The synthesis procedure uses the inverse fast Fourier transform to compute each synthetic contribution, $s_k[n]$, instead of the oscillator functions commonly used in other sinusoidal models [64, 70]. The final synthesized output is computed by applying an overlap-add procedure to the synthesized frames.

High-quality modifications are possible within the framework of the ABS/OLA model [24], which make it particularly attractive for singing voice synthesis. Time scaling is performed by altering the update rate for the model parameters in conjunction with changing the frame duration. Phase constraints are imposed on the sinusoids in order to maintain the phase relationship between the sinusoids within each frame. Pitch modifications are implemented by modifying the frequencies of the sinusoids to be multiples of a scaled fundamental frequency. A phasor interpolation scheme was developed so that these changes could be made while maintaining the spectral shape of the original signal [25].

4.2 Spectral Envelope Estimation

As discussed in Section 3.1, LPC and cepstral analysis both tend to trace the residual noise in the spectrum of voiced speech when partials are spaced far enough apart and the order of estimation is sufficiently high. Because of this shortcoming, an alternative method for the envelope estimation was implemented within the ABS/OLA analysis procedure. This method, developed by Galas and Rodet [21, 22], is known as *generalized discrete cepstral analysis*.

Unlike LPC, which is computed directly from the waveform, and cepstral analysis, which is computed from a spectral representation of the signal with points spaced regularly on the frequency axis, the discrete cepstrum spectral envelope is computed from discrete points in the frequency/amplitude plane. This method is an ideal companion to the sinusoidal model, where the peaks of the sinusoids serve as the necessary discrete points. Assuming that the sinusoids accurately model the partials of voiced speech without regard to the residual noise, the discrete cepstrum will not suffer from the problem of tracing the spectrum in the frequency region between partials. Instead, it will produce a smooth spectral envelope that links the peaks of the partials. The methodology for determining the discrete cepstrum is as follows.

Given a set of spectral peaks with amplitudes x_i at frequencies ω_i, for $i = [1, ..., n]$, a magnitude spectrum, $X(e^{j\omega})$, is defined as

$$X(e^{j\omega}) = \sum_{i=1}^{n} x_i \delta(\omega - \omega_i). \tag{17}$$

$X(e^{j\omega})$ is considered to be the combined frequency response of the source spectrum, $S(e^{j\omega})$, and a filter transfer function, $P(e^{j\omega})$, as follows:

$$X(e^{j\omega}) = S(e^{j\omega}) \cdot P(e^{j\omega}). \tag{18}$$

The source spectrum is given by

$$S(e^{j\omega}) = \sum_{i=1}^{n} s_i \delta(\omega - \omega_i), \tag{19}$$

where s_i are the source amplitudes at the same frequencies as ω_i in $X(e^{j\omega})$. The filter transfer function is modeled by

$$P(e^{j\omega}) = \prod_{i=0}^{p} e^{c_i \cos(\omega_i)}, \tag{20}$$

where c_i are the filter parameters.

Assuming a flat source spectrum, $S(e^{j\omega}) = 1$, for all ω, the filter parameters, c_i, must be determined so that the quadratic error, E, between the log spectra is minimized. This

error criterion is given as

$$E = \sum_{i=1}^{n} \left[\log s_i |P(e^{j\omega_i})| - \log x_i)\right]^2 . \tag{21}$$

Determining the cepstral coefficients is accomplished simply by solving the following matrix equation for c.

$$Ac = b, \tag{22}$$

where A is a matrix of order $p + 1$ whose components are given by

$$a_{ij} = \sum_{k=1}^{n} \cos(\omega_k i) \cos(\omega_k j), \tag{23}$$

and b is the column vector given by

$$b_i = \sum_{k=1}^{n} \log \frac{x_k}{s_k} \cos(\omega_k i). \tag{24}$$

Because the resulting matrix is symmetric, it can be solved efficiently.

4.3 Spectral Modeling and Modification Using Asymmetric Generalized Gaussian Functions

The proposed approach for spectral modeling and modification represents the formant structure of a speech waveform as a weighted sum of asymmetric generalized Gaussian functions [38, 39]. The discrete vocal tract response $V[k]$ is approximated as

$$V[k] = \sum_{m=0}^{M} A_m G_m[k], \tag{25}$$

where the generalized Gaussian function G[k] is specified as

$$G[k] = \begin{cases} \exp\left[-\left(\frac{|k-\mu|}{\beta^l}\right)^{\alpha^l}\right], & \text{for } k \leq \mu, \\ \\ \exp\left[-\left(\frac{|k-\mu|}{\beta^r}\right)^{\alpha^r}\right], & \text{for } k > \mu. \end{cases} \tag{26}$$

The discrete frequency index and the center frequency are given by k and μ, respectively. These functions independently parameterize the width and shape of the left and right sides

of the generalized Gaussian function. The spectral width parameter, β, dictates the bandwidth of each formant, while α is a shaping parameter that controls the rate of decay. Figure 16 illustrates the effect of these parameters.

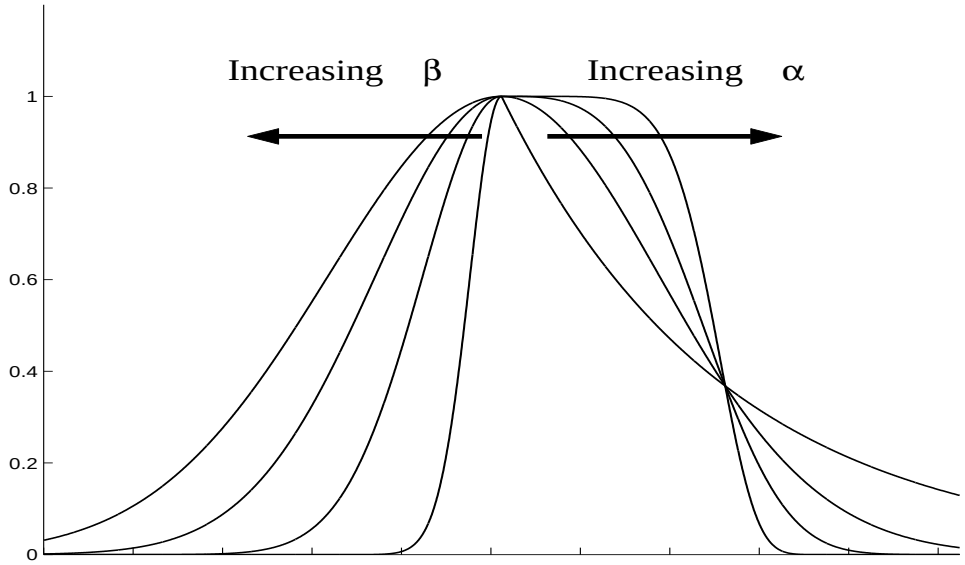


Figure 16: Asymmetric generalized Gaussian functions with increasing width parameters (β) to the left and increasing shape parameters (α) to the right.

The final vocal tract response is obtained by estimating the generalized Gaussian formant structure with a high-order cepstral approximation. The purpose of this is to couple minimum phase characteristics with the magnitude spectrum. When no spectral modifications are applied, the final vocal tract response should closely fit the sinusoidal parameters. Figure 17 shows an example of asymmetric generalized Gaussian functions fit to a spectral envelope.

The flexibility of the asymmetric generalized Gaussian functions ensures an accurate fit to the spectral envelope and enables intuitive and independent modification of each formant's frequency, amplitude, bandwidth, and shape. This provides a high level of control over the formant structure of a singer's voice.

Before formant modification can be performed, each formant must be mapped to a

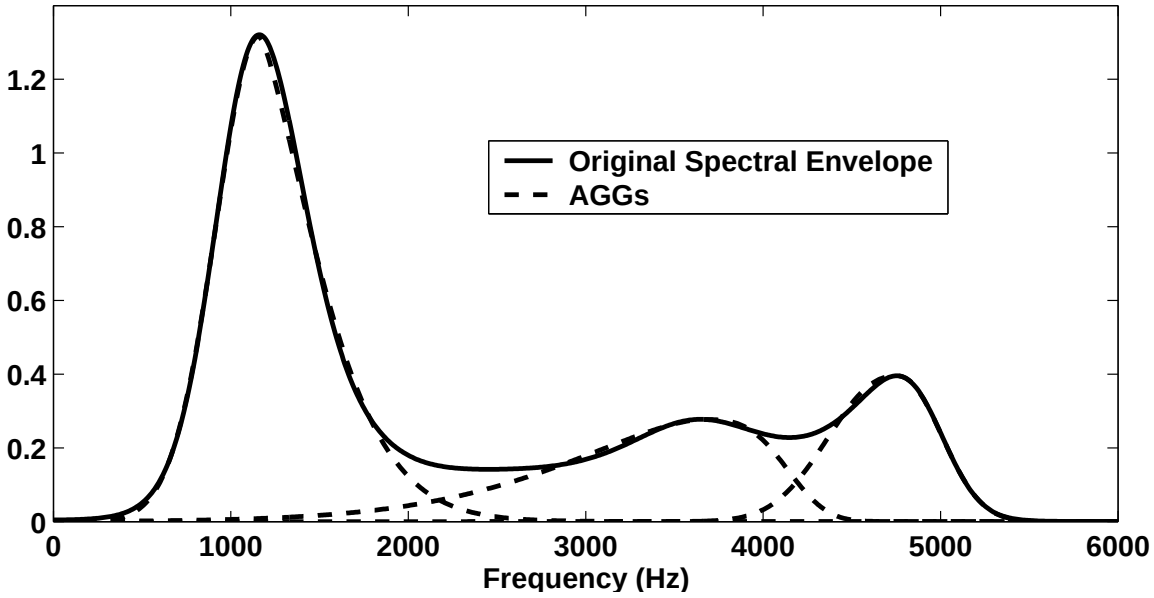


Figure 17: Asymmetric generalized Gaussian functions fitted to a spectral envelope.

particular Gaussian function. Errors can often occur when trying to assign smooth formant trajectories to continuously varying spectral shapes. Formants can merge, split, and sometimes disappear. Since formant changes occur relatively slowly over time, a formant tracking system was implemented to perform the mapping within each frame, as well as to form tracks across frames. The process is based on McAulay and Quatieri's peak matching algorithm for tracking harmonics [49]. A cost function is employed that is based on proximity in frequency and difference in amplitude. Formant tracks are derived such that the cost function is minimized. "Births" and "deaths" of formant tracks are allowed to account for the possibility of the number of distinguishable formants changing from frame to frame.

4.4 Parameter Estimation Using the Expectation Maximization Algorithm

The task of estimating parameters for a set of coexisting asymmetric generalized Gaussian functions must be carefully considered because of the large search space involved. Therefore, an implementation of the Expectation Maximization (EM) algorithm is used to

determine the function parameters optimally.

The Expectation Maximization algorithm is a broadly applicable algorithm used to maximize the log-likelihood from incomplete data by iteratively maximizing the expectation of the log-likelihood of the complete data [11]. This particular type of statistical estimation algorithm in an unsupervised learning algorithm, which means only the data sample is observed while the class or state of the data sample remains unobservable.

The EM algorithm is composed of an expectation step (E-step) and a maximization step (M-step). The E-step estimates the unobserved data based on the current estimated distribution parameters. The M-step then updates the parameters with a maximum-likelihood estimate based on the data derived from the previous E-step. These two steps are repeated until convergence is reached.

In the proposed application of the EM algorithm, the spectral envelope, $X(e^{j\omega_n})$, of a speech waveform is viewed as a probability distribution, $P(x_k)$, where x_k are the bin numbers $(k = 1, ..., N)$. The purpose of the algorithm is to approximate $P(x_k)$ with a mixture of asymmetric generalized Gaussian functions.

For our model, we let x_k be the observed incomplete data and (x_k, y_k) be the complete data, where y_k is an unobservable integer between 1 and s, indicating the corresponding component density, $f(x_k|y_k, \phi_{y_k})$, and mixing parameter, ω_{y_k}, of the mixture pdf. The next step is to compute the expectation of the log-likelihood given the complete data. The *Q* function serves to represent this expectation for multiple observed data, $\mathbf{X} = \{x_1, \ldots, x_n\}$, and multiple unobserved data, $\mathbf{Y} = \{y_1, \ldots, y_n\}$. It is assumed that a parametric family of mixture probability density functions is given and that a particular $\overline{\mathbf{\Phi}}$ is the parameter value to be estimated.

Given the probability of multiple observed data, the log-likelihood of the observable

data is as follows:

$$\begin{aligned} L(\mathbf{X}, \mathbf{\Phi}) &= \log \left[\prod_{k=1}^{N} f(\mathbf{x}_k|\mathbf{\Phi})^{P(\mathbf{x}_k)} \right] \\ &= \sum_{k=1}^{N} P(x_k) \log f(\mathbf{x}_k|\mathbf{\Phi}). \end{aligned} \tag{27}$$

The log-likelihood of a single complete data point, (x_k, y_k), is

$$f(x_k, y_k|\mathbf{\Phi}) = \omega_{y_k} f(x_k|y_k, \phi_{y_k}), \tag{28}$$

where ω_{y_k} is the *a priori* probability (the mixture weight). The log-likelihood of an incomplete data point, x_k, is given by

$$f(x_k|\phi) = \sum_{y_k} f(x_k, y_k|\mathbf{\Phi}), \tag{29}$$

and the posterior probability is

$$P(y_k|x_k, \mathbf{\Phi}) = \frac{\omega_{y_k} f(x_k|y_k, \phi_{y_k})}{\sum_{y_k} \omega_{y_k} f(x_k|y_k, \phi_{y_k})}. \tag{30}$$

The *Q* function can then be formulated as

$$Q(\mathbf{\Phi}, \overline{\mathbf{\Phi}}) = \sum_{k=1}^{N} P(x_k) \left\{ \sum_{y_k} P(y_k|x_k, \phi_{y_k}) \log \left[\overline{\omega}_{y_k} f(x_k|y_k, \overline{\phi}_{y_k}) \right] \right\}. \tag{31}$$

Since the inner summation is over all y_k and $y_k \in \{1 \ldots s\}$ for each k, y_k can be denoted by i. For example, if the k^{th} sample was generated by the i^{th} mixture, then $y_k = i$. By expanding the log terms, the *Q* function can be reformulated as

$$\begin{aligned} Q(\mathbf{\Phi}, \overline{\mathbf{\Phi}}) = &\sum_{i} \left\{ \sum_{k=1}^{N} P(x_k) P(i|x_k, \phi_i) \right\} \log \overline{\omega}_i \\ &+ \sum_{i} \left\{ \sum_{k=1}^{N} P(x_k) P(i|x_k, \phi_i) \log f(x_k|i, \overline{\phi}_i) \right\}. \end{aligned} \tag{32}$$

It is important to distinguish between the first and second arguments of the *Q* function, $\mathbf{\Phi}$ and $\overline{\mathbf{\Phi}}$. $\mathbf{\Phi}$ is a conditioning argument to the expectation and is regarded as fixed and known at every E-step. The second argument, $\overline{\mathbf{\Phi}}$, conditions the likelihood of the complete data. During the M-step, a value for $\overline{\mathbf{\Phi}}$ is determined such that the *Q* function is maximized.

The asymmetric Generalized Gaussian function is the probability density that is used for this particular model and is given by

$$f(x_k|i,\phi_i) = \begin{cases} \gamma \exp\left[-\left(\frac{|x-\mu_i|}{\alpha_i^L}\right)^{\beta_i^L}\right], & x \leq \mu_i, \\ \gamma \exp\left[-\left(\frac{|x-\mu_i|}{\alpha_i^R}\right)^{\beta_i^R}\right], & x > \mu_i. \end{cases} \tag{33}$$

Once the *Q* function has been determined, the M-step is completed by maximizing each term in (32) for each i with respect to $\overline{\omega}_i$ and $\overline{\phi}_i$. This is accomplished by solving for

$$\frac{\partial Q(\mathbf{\Phi}, \overline{\mathbf{\Phi}})}{\partial \overline{\phi}_i} = \sum_{k=1}^{N} P(x_k) P(i|x_k, \phi_i) \frac{\partial}{\partial \overline{\mathbf{\Phi}}} \left[\log f(x_k|i, \overline{\mathbf{\Phi}})\right] = 0. \tag{34}$$

However, because of the asymmetric nature of this particular probability density function, μ is not the true mean and the left and right α_i terms are unrelated to the standard deviation. Therefore it is impossible to determine a closed-form solution for (34). Subsequently, it must be solved numerically with respect to $\alpha_i^L,\ \alpha_i^R,\ \beta_i^L,\ \beta_i^R,$ and μ_i. This, however, is a much simpler optimization than estimating parameters for all of the generalized Gaussian functions simultaneously.

In summary, the EM algorithm can be utilized to fit a set of probability density functions to a frequency spectrum by maximizing the log-likelihood of the observed data, $L(\mathbf{X}, \mathbf{\Phi})$, in the following manner:

1. Choose an initial estimate $\mathbf{\Phi}$.
2. **E-step** : Compute $Q(\mathbf{\Phi}, \overline{\mathbf{\Phi}})$ based on the given $\mathbf{\Phi}$.
3. **M-step** : Determine $\overline{\mathbf{\Phi}} = \underset{\overline{\mathbf{\Phi}}}{argmax} Q(\mathbf{\Phi}, \overline{\mathbf{\Phi}})$.
4. Set $\mathbf{\Phi} = \overline{\mathbf{\Phi}}$ and repeat steps 2-4 until convergence is reached.

4.4.1 Initialization of the EM Algorithm

The EM algorithm has been shown to provide an increase in the likelihood function after every iteration. Furthermore, it is guaranteed to converge on a local maximum. Despite this

property, the EM algorithm is not guaranteed to converge to a global maximum. Therefore, it is important to provide the algorithm with a proper initial estimate.

A number of methods were investigated for initializing the EM algorithm. These include

- Peak picking the spectral envelope to initialize the center frequencies and weights. The width and shape parameters were set to equal values.
- Using the estimated parameters from the previous frame to initialize the current frame.
- Employing a formant tracking scheme to determine the formant frequencies and using the corresponding spectral envelope magnitudes to initialize the weights.

The first method often missed formants that were too close to another formant so as to not exhibit a peak in the spectral envelope. The second method tended to produce errors because two Gaussian functions would occasionally converge to a single formant. Figure 18 shows examples of these errors produced by the first two initialization methods. In both cases, the formant at 1000 Hz is missed by the estimation process.

While not the most efficient of the three methods, a formant tracking algorithm employed to initialize the parameter estimation process provided the most accurate and consistent results. The formant estimation method employed was originally formulated by Schafer and Rabiner [69]. This particular method was chosen because of its ability to determine formant parameters directly from cepstral coefficients.

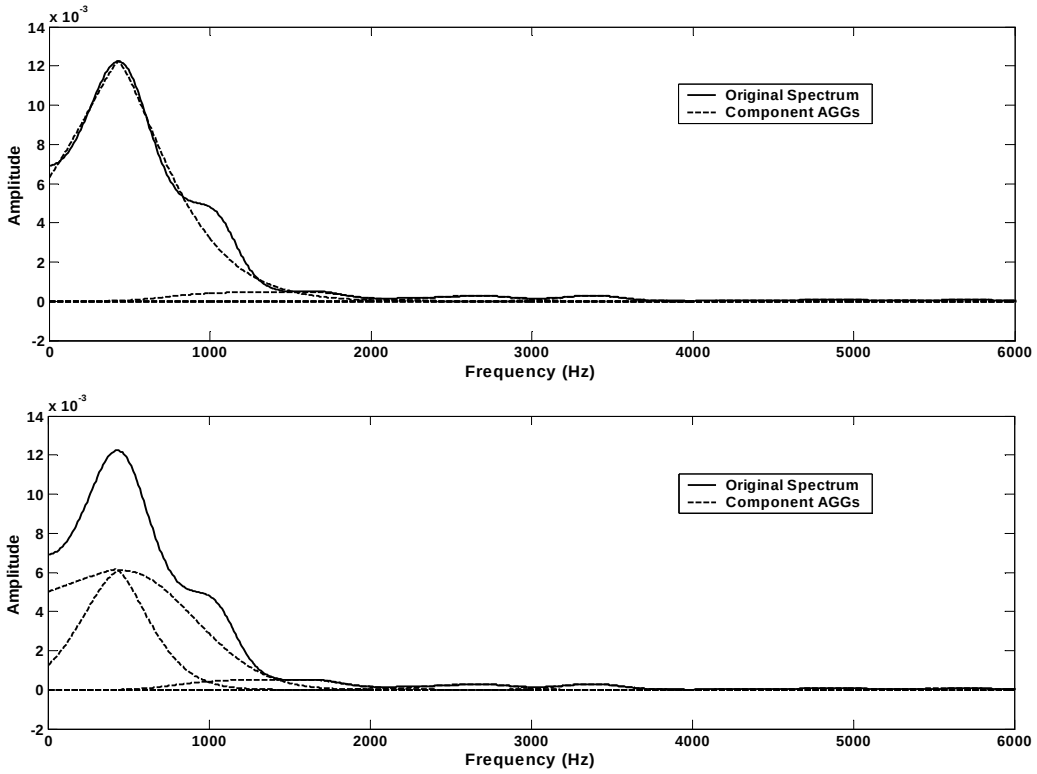


Figure 18: Examples of errors in the Expectation Maximization procedure for fitting asymmetric generalized Gaussians to a spectral envelope. Top: missed formant at 1000 Hz. Bottom: double formant at 400 Hz.

CHAPTER 5

BACKGROUND: THE GLOTTAL VOICE SOURCE

The signal produced when the vocal folds vibrate the airstream from the lungs is commonly known as the glottal voice source. This intermediate waveform serves as the canvas onto which the vocal tract imprints its own characteristics in the form of resonances and timbre. While many basic speech synthesis techniques assume that the excitation to the vocal tract is an impulse train and that the glottal spectral shaping is included in the overall vocal tract filter model, it has been shown that an appropriate model for the glottal source can greatly improve naturalness and vocal quality. An important study by Rosenberg [66] showed that the use of a more natural glottal pulse shape resulted in synthetic speech with significantly improved quality over those that were produced with simpler glottal pulse models such as impulses, triangular waves, or square waves. This work, along with many others, provided the motivation for the development of glottal models of appropriate complexity that could effectively capture the characteristics of the glottal source for more natural speech synthesis.

5.1 Glottal Flow Models

Several glottal source models have been proposed for modeling the glottal derivative waveform. While some glottal flow models such as the KLGLOTT88 offer a simple and efficient method for representing a glottal wave period, we have chosen to use more complex models that offer greater flexibility as well as accuracy. The LF and R++ glottal flow models are both five-parameter models capable of modeling smooth closure of the glottis as well as asymmetric glottal flow pulse shapes. Other models assume an abrupt closure that can cause a skewing of the spectral tilt when this assumption does not hold. Fant and his colleagues note that even a slight departure from abrupt closure results in a significant increase in the roll-off of the glottal spectrum [20].

5.1.1 LF model

Originally proposed by Fant and his colleagues [20], the Liljencrants-Fant (LF) model is a representation of the glottal flow derivative. In the source-filter model of the speech production mechanism, the glottal flow derivative serves as the excitation for the vocal tract filter. The actual glottal flow waveform, representing the volume velocity of air traveling through the glottis, can be calculated by integrating the glottal derivative over a single period. The five independent parameters of the LF model are

- T_0 : fundamental period
- T_e : instant of maximum excitation
- T_p : instant of maximum glottal flow
- T_a : return phase constant
- E_e : amplitude at instant of maximum excitation

The model divides the glottal cycle into two distinct phases, with the boundary being marked by the instant of glottal closure, which is where maximum excitation occurs. The amplitude of the glottal derivative at this point, E_e, also marks the point of steepest decent in the glottal waveform.

The LF model is parameterized by the following equation:

$$g(t) = \begin{cases} -E_e e^{a(t-T_e)} \frac{\sin(\pi t/T_p)}{\sin(\pi T_e/T_p)}, & 0 \leq t \leq T_e, \\ \\ \frac{-E_e}{\epsilon T_a} \left[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_0-T_e)} \right], & T_e < t \leq T_0. \end{cases} \tag{35}$$

The first segment of the LF model characterizes the derivative glottal flow from the glottal opening to the glottal closure instant. During this period, the glottis is considered to be open and is thus denoted the *open phase*. This portion of the glottal derivative is modeled as an increasing exponential modulated by a sinusoid. The second segment, the

closed phase, characterizes the closure of the glottis as a decreasing exponential. Figure 19 illustrates the LF model for a single glottal cycle. The effects of the parameters on the waveform are also indicated.

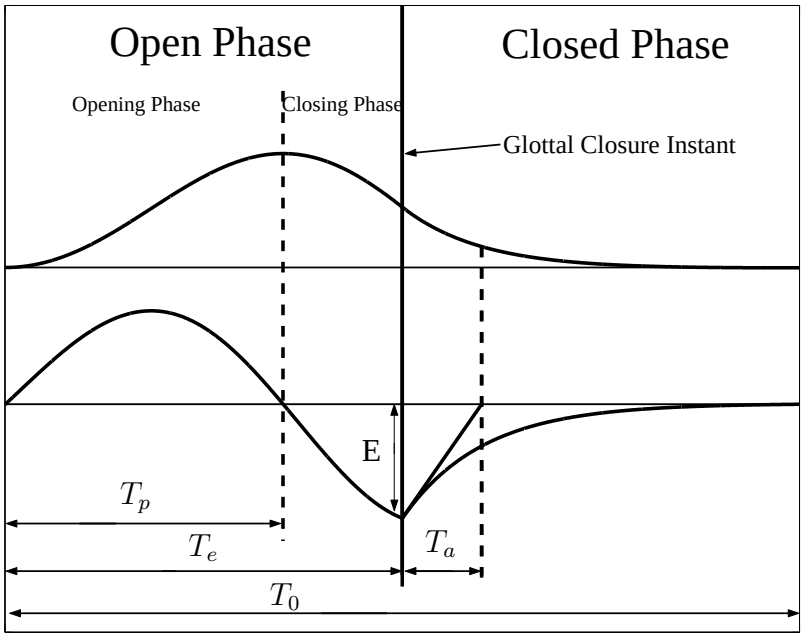


Figure 19: Parameters for glottal flow models and how they relate to glottal waveforms (top) and their derivatives (bottom).

The internal parameters of the LF model equation, a and ϵ, are determined by solving the constraint equations,

$$\epsilon T_a = 1 - e^{-\epsilon(T_0 - T_e)} \tag{36}$$

and

$$\frac{1}{a^2 + \left(\frac{\pi}{T_p}\right)^2}\left[e^{-aT_e}\frac{\pi/T_p}{\sin(\frac{\pi T_e}{T_p})} + a - \frac{\pi}{T_p}\cot(\frac{\pi T_e}{T_p})\right] = \frac{T_0 - T_e}{e^{\epsilon(T_0 - T_e)} - 1} - \frac{1}{\epsilon} \tag{37}$$

These constraints ensure the continuity of the glottal derivative at the boundary points as well as the glottal closure instant.

The open quotient, O_q, is a significant measure that defines the ratio of the open phase duration to the fundamental period. Because this portion of the glottal cycle is represented

by the first segment in the LF model, the open quotient is calculated as

$$O_q = \frac{T_e}{T_0}. \tag{38}$$

The open phase can be further divided into two segments, the *opening phase* and *closing phase*. The division of these two segments is marked by the instant of maximum glottal flow (T_p). The asymmetry coefficient, α, defines the ratio of the opening phase duration to the length of the entire open phase. Thus,

$$\alpha = \frac{T_p}{T_e}. \tag{39}$$

The duration required for the glottis to reach full closure is characterized by the parameter T_a. This period is also known as the *return phase*. An alternative form of this parameter, Q_a, is often used that describes the ratio of the return phase to the closed phase. Q_a is calculated as

$$Q_a = \frac{T_a}{T_0 - T_e}. \tag{40}$$

O_q, α, and Q_a have been shown to be indicators of perceptual qualities. As we will see later in this document, these measures also have a great impact on the spectrum of the glottal source. Although these measures are not explicit parameters of the LF model, they are closely related to the LF parameters and can be calculated using the above equations.

5.1.2 R++ model

Veldhuis proposed the R++ model for the glottal derivative [85] as an extension to a polynomial model originally proposed by Rosenberg [66]. These extensions were designed to increase the flexibility of the original model by incorporating control over both the return phase and asymmetry of a glottal pulse. The independent parameters for this model (T_0, T_e, T_p, T_a, A) are equivalent to those of the LF model with the exception of the parameter A, which is an amplitude coefficient that scales the maximum glottal flow.

Like the LF model, this model offers a flexible and accurate representation of the glottal flow derivative waveform. The glottal derivative of a single cycle is segmented into open

and closed phases as in the LF model, but the R++ model uses a third-order polynomial to represent the waveform during the open phase. The closed phase of the glottal derivative cycle is modeled with a decreasing exponential. The formulation of the R++ model is given by

$$g(t) = \begin{cases} 4At(T_p - t)(T_x - t), & 0 \leq t \leq T_e, \\ \\ g(T_e)\frac{e^{-(t-T_e)/T_a} - e^{-(T_0-T_e)/T_a}}{1-e^{-(T_0-T_e)/T_a}}, & T_e < t \leq T_0. \end{cases} \tag{41}$$

The parameter T_x in the equation is calculated as:

$$T_x = T_e \left(1 - \frac{\frac{1}{2}{T_e}^2 - T_e T_p}{2{T_e}^2 - 3T_e T_p + 6T_a(T_e - T_p)D(T_0, T_e, T_a)} \right), \tag{42}$$

where

$$D(T_0, T_e, T_a) = 1 - \frac{(T_o - T_e)/T_a}{e^{(T_0-T_e)/T_a} - 1}. \tag{43}$$

Because the timing parameters are the same as those of the LF model, the formulas for the open quotient, asymmetry coefficient, and return phase parameter for the R++ model are identical.

5.2 Spectral analysis of time-domain glottal flow models

5.2.1 Frequency-domain representations

In this section, spectra of the glottal flow models presented previously are examined in an effort to determine the relationship between the time-domain parameters and frequency-domain characteristics of these models. In order to accomplish this, it is useful to derive analytic formulas for the frequency-domain representations of these glottal derivative models (LF and R++). This is performed by employing properties of the continuous-time Fourier Transform. Specifically, a frequency representation of the glottal flow waveform can be calculated by using the integral property of the Fourier Transform:

$$\int_{-\infty}^{t} x(t)dt \Longleftrightarrow \frac{1}{j\Omega}X(\Omega) + \pi X(0)\delta(\Omega). \tag{44}$$

The spectra of each segment of the LF model are derived independently of one another and summed in the frequency domain:

$$G^{LF}(\Omega) = G_1^{LF}(\Omega) + G_2^{LF}(\Omega) \tag{45}$$

where

$$G_1^{LF}(\Omega) = \frac{-E_e}{\sin(\frac{\pi T_e}{T_p})}\left[\frac{e^{-aT_e}(\frac{\pi}{T_p})}{(j\Omega-a)^2+(\frac{\pi}{T_p})} - \frac{e^{-j\Omega T_e}\cos(\frac{\pi T_e}{T_p})\frac{\pi}{T_p}}{(j\Omega-a)^2+(\frac{\pi}{T_p})} - \frac{e^{-j\Omega T_e}\sin(\frac{\pi T_e}{T_p})(j\Omega-a)}{(j\Omega-a)^2+(\frac{\pi}{T_p})}\right] \tag{46}$$

$$G_2^{LF}(\Omega) = \frac{-E_e}{\epsilon T_a}\left[\frac{e^{-j\Omega T_e}}{j\Omega+\epsilon} - \frac{e^{-\epsilon(T_0-T_e)}e^{-j\Omega T_0}}{j\Omega+\epsilon} - \frac{e^{-\epsilon(T_0-T_e)}(e^{-j\Omega T_e}-e^{-j\Omega T_0})}{j\Omega}\right] \tag{47}$$

A similar derivation is used to determine the spectrum of the R++ model:

$$G^{R++}(\Omega) = G_1^{R++}(\Omega) + G_2^{R++}(\Omega) \tag{48}$$

where

$$G_1^{R++}(\Omega) = 4K\left[-W'''(\Omega) - (T_p+T_x)W''(\Omega) - T_pT_xW'(\Omega)\right] \tag{49}$$

$$G_2^{R++}(\Omega) = \frac{g^{R++}(T_e)}{1-e^{\frac{-(T_0-T_e)}{T_a}}}\left\{\frac{e^{-j\Omega T_e} - e^{\frac{-(T_0-T_e)}{T_a}}e^{-j\Omega T_0}}{j\Omega+1/T_a} - \frac{e^{\frac{-(T_0-T_e)}{T_a}}(e^{-j\Omega T_e}-e^{-j\Omega T_0})}{j\Omega}\right\} \tag{50}$$

and

$$W(\Omega) = \frac{1}{j\Omega}(1-e^{-j\Omega Te}). \tag{51}$$

The spectral formulations for each of these glottal models can be expressed in terms of O_q, α, and Q_a by substituting (38)-(40).

Figure 20 illustrates an example of spectra generated using each of the models with a common parameter set. Glottal derivative waveforms generated with identical parameters by each model are shown with their corresponding frequency-domain representations. Although not identical, the spectra of the two models are in relative agreement in capturing the spectral shape of the glottal excitation. The spectrum of the glottal source generally takes

the shape of a low-pass filter. Because of the apparent resonant frequency and asymptotic roll-off in high frequencies, the spectrum of the glottal derivative is generally described as a *glottal formant*.

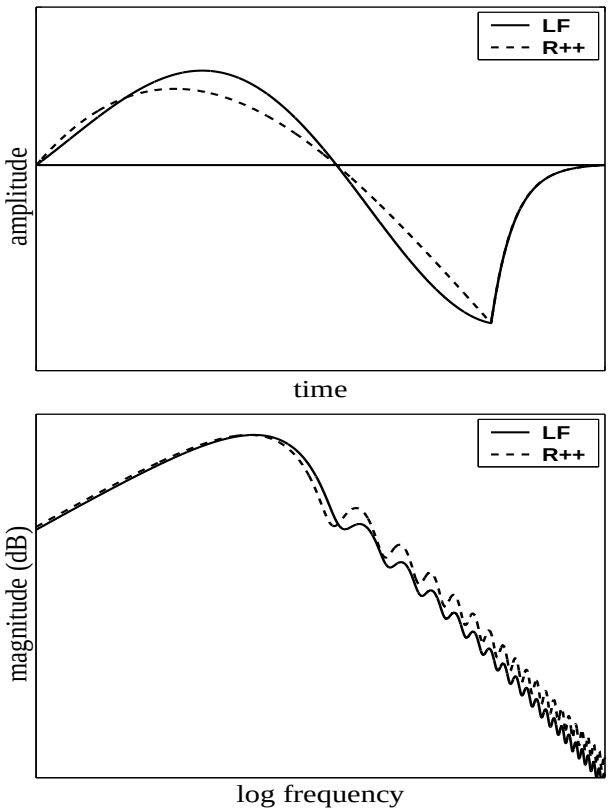


Figure 20: Glottal waveform derivatives produced using the LF and R++ models (top) and their corresponding spectra (bottom). ($O_q = 0.6, \alpha = 0.66, Q_a = 0.3$)

5.2.2 Spectral correlates with time-domain parameters

The spectral derivations presented above can be used to illustrate the effects of the time-domain parameters of the glottal flow models in the frequency domain. For this section, the LF model will be used instead of the R++ model. Although the LF model requires a larger number of calculations and requires two constraint equations to be resolved, it allows for a greater range of values for the asymmetry coefficient than the R++ model, where α is restricted to a relatively narrow range.

The analysis presented in this section only considers the effects of varying the time-domain parameters of the glottal model on the magnitude spectrum. The phase spectrum of the glottal waveform will be analyzed later in this paper.

The parameter E_e determines the amplitude of the glottal derivative at the glottal closure instant. E_e can also be described as the closing rate of the glottal flow waveform. Sundberg found that E_e showed a strong correlation with loudness of phonation [73]. By inspection of (46) and (47), the spectral effects of varying E_e are evident. When the other model parameters are kept constant, E_e serves as a scalar for the spectrum of the glottal source. Varying the value of E_e scales the spectrum equally across all frequencies. This parameter gives the LF model the flexibility to match the vocal intensity of a glottal waveform.

Figure 21 illustrates the spectral effect of varying the open quotient, O_q, while keeping the remaining parameters constant. These plots show synthesized periods of the glottal derivative with varying O_q values and their corresponding spectra. As can be seen, lowering the open quotient results in an upward frequency shift of the glottal formant. This energy increase in higher frequencies has been linked to perceived increases in the loudness and brightness of a voiced waveform. Several studies have also noted that the type of phonation (e.g., pressed, modal, breathy) can have a large effect on the open quotient.

The asymmetric coefficient, α, and the return phase parameter, Q_a also affect the higher frequencies in the spectrum of the glottal waveform. Instead of shifting the glottal formant, however, increasing α or lowering O_q results in an increase in the bandwidth of the glottal formant. This is shown in Figures 22 and 23. In these plots, α and Q_a are varied while all other parameters are held constant. The increase in bandwidth causes a decrease in the spectral roll-off in higher frequencies. Rothenberg noticed this relationship in voices of trained singers. He observed higher energy in the third and fourth formants of voices with greater levels of glottal asymmetry [68]. He contended that this characteristic is desirable for good singers and would lead to a clearer and more intelligible voice.

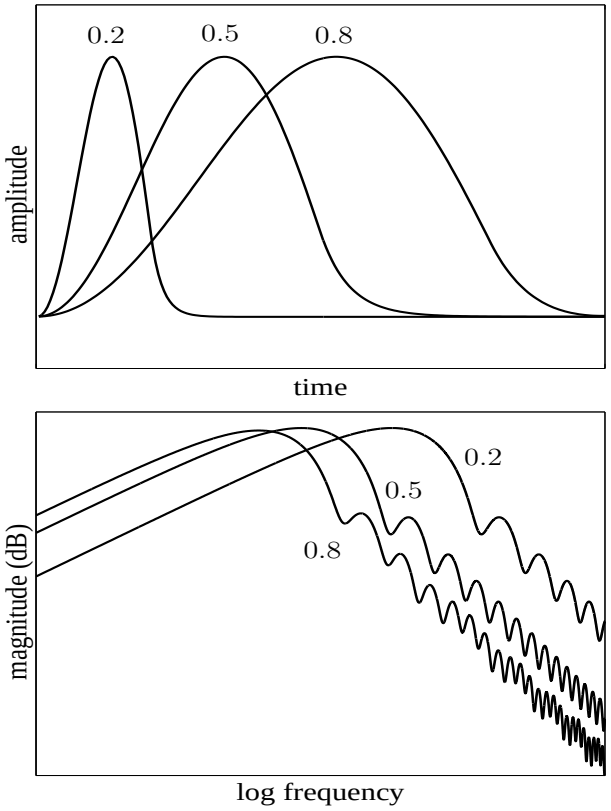


Figure 21: Glottal flow waveforms with varying open quotient values ($O_q = 0.2, 0.5, 0.8$) and the corresponding spectra of the waveform derivatives. All other parameters are held constant.

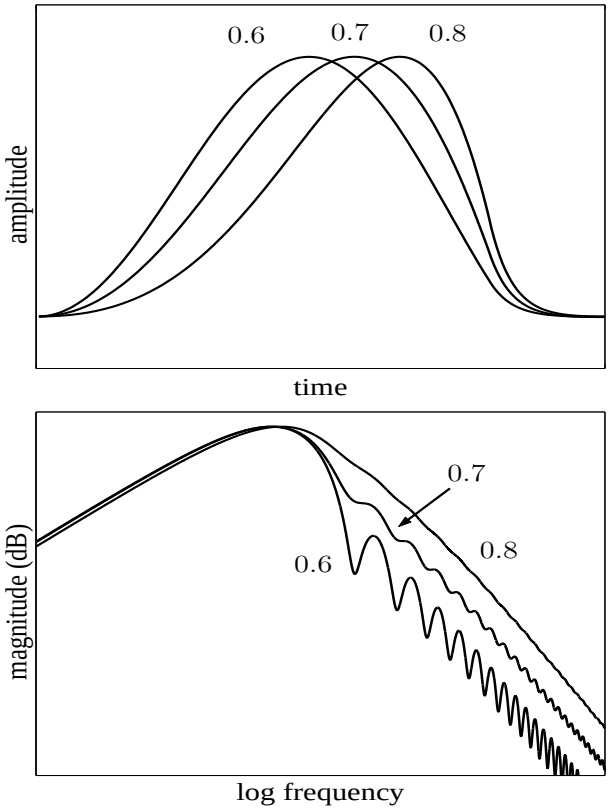


Figure 22: Glottal flow waveforms with varying asymmetric coefficient values ($\alpha = 0.6, 0.7, 0.8$) and the corresponding spectra of the waveform derivatives. All other parameters are held constant.

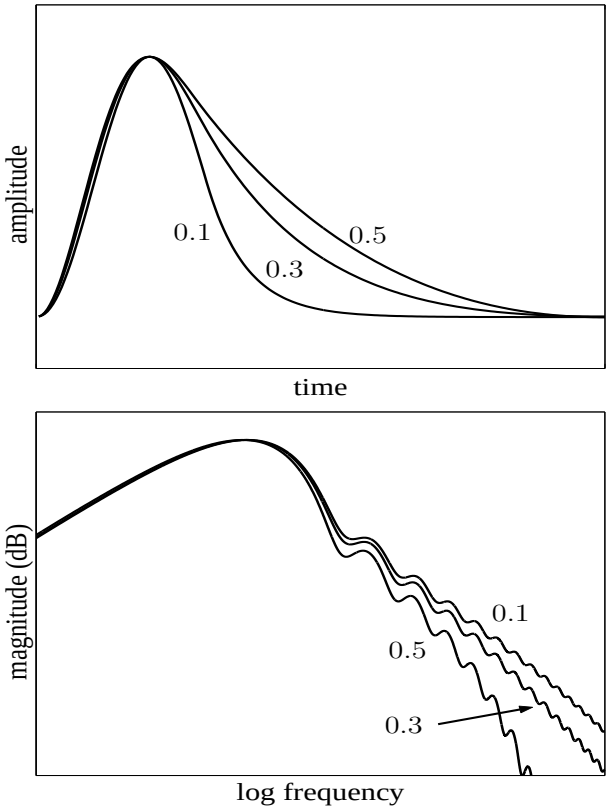


Figure 23: Glottal flow waveforms with varying return phase coefficient values ($Q_a = 0.1, 0.3, 0.5$) and the corresponding spectra of the waveform derivatives. All other parameters are held constant.

CHAPTER 6

NEW FREQUENCY-DOMAIN METHODS FOR GLOTTAL CHARACTERIZATION AND MODIFICATION

The time-domain glottal flow models discussed thus far have been shown to be capable of effectively parameterizing perceptually relevant characteristics of the glottal source. A frequency-domain solution, however, would offer a seamless integration with the ABS/OLA sinusoidal model chosen for the proposed system. Current frequency-domain models fail to provide reliable parameterizations for reliable glottal modification. Thus, it would be desirable to implement a frequency-domain model capable of capturing time-domain features. This chapter continues the investigation of the relationship between time-domain parameters and frequency-domain characteristics introduced in the previous chapter. Based on these findings, new methods for estimating and modifying time-domain glottal parameters in the frequency domain are proposed.

6.1 Analysis of H1-H2* and time-domain parameters*

As has been shown earlier, the parameters O_q, α, and T_a all have an effect on the magnitude of the spectrum of a glottal waveform. In particular, the open quotient, O_q, has a direct effect on the center frequency of the glottal formant. Several experimental studies have supported this by showing a correlation between the O_q and the relative spectral amplitude of the first two harmonics of the glottal derivative waveform (H1*-H2*). The amplitudes of H1 and H2 are typically measured from the spectrum of a windowed frame of the glottal source. The glottal spectrum is usually obtained by one of two methods: inverse filtering the original speech waveform in order to remove the contributions of the vocal tract, or by applying a formula in which the amplification of the glottal source due to the first formant is "corrected" [29].

Based on these results, additional studies have attempted to develop methods for estimating O_q from measured values of H1*-H2*. Sundberg and his colleagues observed a strong pattern between the open quotient and the ratio of the first two harmonics of the glottal flow waveforms for five professional baritone singers [75]. Their analysis resulted in the following relationship based on these measurements:

$$H1* - H2* = 21.5 - 31.1(1 - O_q). \tag{52}$$

It should be noted that in this case, H1*-H2* indicates the relative amplitude of the first two harmonics of the glottal flow waveform and not its derivative.

Fant used the LF model to analyze this relationship between the open quotient and the amplitudes of the first two harmonics [19]. A regression analysis of synthesized data varying O_q was used to derive the expression:

$$H1* - H2* = -6 + 0.27e^{5.5O_q}. \tag{53}$$

While these studies have attempted to define a direct relationship between the open quotient and relative amplitudes of the first two harmonics of the glottal source, our analysis from section 5.2 revealed that the magnitude spectrum of the LF model can be affected by the asymmetry coefficient (α) as well as the return phase parameter (Q_a). This could decrease the accuracy of any estimation of Oq since its relationship to H1*-H2* is not one-to-one.

By using (45), it is possible to measure the effects of varying the time-domain glottal parameters of the LF model on the relative amplitude of the first two harmonics. The frequency-domain representation of the LF model can be viewed as a function of both frequency and time-domain parameters $(T_0, O_q, \alpha, Q_a, E_e)$. If T_0 and E_e are normalized to values of 1, then the fundamental frequency is guaranteed to be: $\Omega_0 = 2\pi$ rad/sec. Therefore, H1 and H2 can be measured by evaluating $G(\Omega, I)$ at frequencies 2π and 4π, respectively. In the equation below, $I = \{O_q, \alpha, Q_a\}$. H1*-H2* can then be calculated as:

$$H1* - H2* = 20\log_{10}\left(\frac{|G(2\pi, I)|}{|G(4\pi, I)|}\right), \tag{54}$$

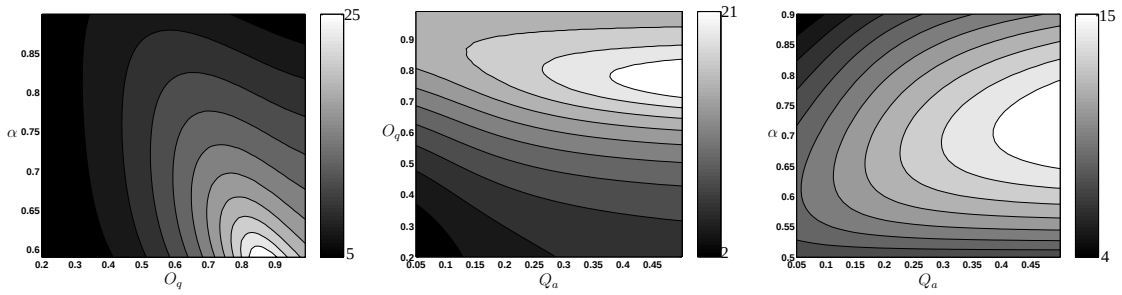


Figure 24: Contour plots of the relative amplitude of the first two harmonics of the glottal flow waveform (H1*-H2*) using the LF model. In each plot, two parameters are varied: (a) α vs. O_q, ($Q_a = 0.3$); (b) O_q vs. Q_a, ($\alpha = 0.66$); (c) α vs. Q_a, ($O_q = 0.7$);

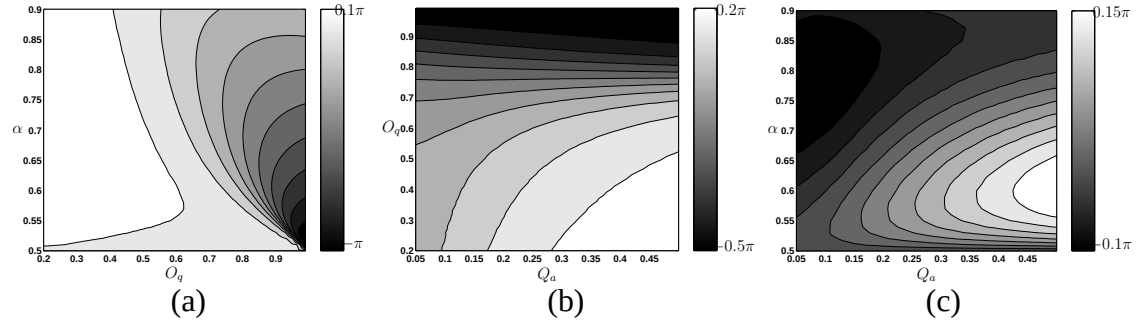


Figure 25: Contour plots of the relative phase of the first two harmonics of the glottal flow waveform ($\Delta\phi_2$) using the LF model. In each plot, two parameters are varied: (a) α vs. O_q, ($Q_a = 0.3$); (b) O_q vs. Q_a, ($\alpha = 0.66$); (c) α vs. Q_a, ($O_q = 0.7$);

The parameters O_q, α, and Q_a can then be systematically varied while resulting values of H1*-H2* are calculated. Figure 24 illustrates the dependence of H1*-H2* on O_q, α, and Q_a using the LF model. In these plots, the contours represent constant values of H1*-H2*. In each of the plots two of the three parameters are varied while all other parameters are held constant.

In part (a), α and O_q are varied while Q_a is held constant. It can be seen that H1*-H2* is largely dependent on O_q over α for low values of O_q. For higher values of O_q, however, α has a much greater influence on H1*-H2* than O_q. It is interesting to note that O_q and α have a wide range of possible values that can produce the same H1*-H2*.

This phenomenon is further illustrated in parts (b) and (c) of Figure 24. In these plots, H1*-H2* is shown as a function of the variable pairs (O_q, Q_a) and (α, Q_a), respectively. The contours clearly indicate that for a given value of H1*-H2*, several pairs of (O_q, Q_a) and (α, Q_a) are possible. It can be seen that the effects of the return phase parameter Q_a on H1*-H2* is very moderate compared to O_q and α.

These plots show that H1*-H2* is not a consistent indicator of the open quotient as was hypothesized in other studies. It can be further generalized that H1*-H2* does not have a one-to-one relationship with any of the three parameters (O_q, α, Q_a).

It is apparent that for many cases it is not possible to estimate O_q, α, or Q_a accurately based on amplitude measurements of the first two harmonics of the glottal source. However, additional information derived from these harmonics exists that can be used to develop an improved parameter estimation technique. This information is provided by the phases of the harmonics. If each harmonic is viewed in time as a sinusoid, then the phase of that harmonic indicates the position of the sinusoid relative to the analysis window. Since the phases of sinusoids vary based on their position relative to the window, it is necessary to use a phase measure that is shift-invariant. Therefore, a relative phase measure of the first two harmonics is calculated by determining the phase of H2 at the point at which the phase of H1 is zero. Given the phases and frequencies of H1 and H2 as (ϕ_1, ϕ_2) and (ω_1, ω_2), respectively, the relative phase of H2 to H1 is calculated as:

$$\Delta\phi_2 = \phi_2 - \phi_1 \frac{\omega_2}{\omega_1}. \tag{55}$$

This measure has the advantage of being invariant to the relative position of the analysis window while yielding a parameter that characterizes the shape of the glottal waveform.

Figure 26 shows the sum of two harmonic sinusoids with a relative amplitude (H1*-H2*) of 20 dB and with relative phase ($\Delta\phi_2$) values of 0 and $\frac{3\pi}{4}$. It can be seen from this figure that the difference in relative phase affects the general shape of a glottal cycle which in turn influences the glottal parameters $O_q, \alpha,$ and Q_a. Using (45) for the spectra of the LF model, the relative phase of H2 to H1 is calculated for a wide range of possible values of

$O_q, \alpha,$ and Q_a. This relationship is illustrated in Figure 25. These plots uncover a complex relationship between the glottal parameters and relative phase of H2 to H1. While $\Delta\phi_2$ tends to be slightly more dependent on O_q and α than Q_a, the effect of the return phase parameter is nonetheless significant. Additionally, it is evident that $\Delta\phi_2$ does not vary with the glottal parameters in the same fashion that (H1*-H2*) does. It is therefore conceivable that the combination of the measured values for relative amplitude (H1*-H2*) and relative phase ($\Delta\phi_2$) can be used to estimate the glottal waveform parameters $O_q, \alpha,$ and Q_a.

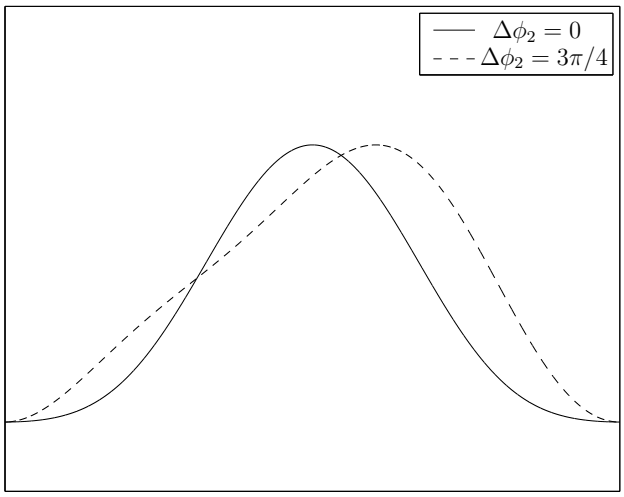


Figure 26: Summation of two sinusoids ($\frac{|H1|}{|H2|} = 10$) with relative phase values of $\Delta\phi_2 = 0$ (solid) and $\Delta\phi_2 = 3\pi/4$ (dashed).

6.2 Parameter Estimation

In this section, a technique is proposed in which frequency-domain parameterizations are used to estimate the time-domain parameters of a glottal waveform. The performance of this method is then measured by examining experimental results obtained from recorded speech waveforms. In an ideal test, the estimates of the time-domain characteristics would be compared to known ground-truth values. However, because of the difficulty in measuring the true volume velocity of air travelling through the glottis, the experimenters were

unable to obtain these true measures and thus determine the absolute accuracy of the procedure. As an alternative, the experimental estimates of the proposed procedure are compared with those obtained using a biologically-inspired reference procedure. This reference procedure combines acoustic and biological measurements with a closed-phase inverse filtering technique for determining glottal flow. It has been shown that closed-phase inverse filtering is highly proficient in extracting glottal flow waveforms from acoustic speech signals when glottal closure instants are located using auxiliary biological sensors [35,60]. By exhibiting a level of consistency between the proposed method and the reference method, it is shown that key time-domain features of the glottal waveform can be estimated from frequency-domain measurements of the first two harmonics.

6.2.1 Estimation Method

The parameter estimation method requires a recorded segment of speech or singing to determine the corresponding glottal characteristics. By measuring the frequencies, amplitudes, and phases of the first two harmonics of the glottal source signal, the parameters can be estimated by incorporating the concepts discussed in the previous section.

Before the estimation can be performed, the measurement of the source parameters requires a separation of the glottal waveform and vocal tract response from the acoustic signal. While several inverse filtering techniques exist, a method based on an algorithm by [54] was chosen for this work. In this procedure, basic approximations of glottal closure regions are identified and a local iterative search is performed in order to estimate the closed-phase portions of the waveform. These segments are then used to produce an estimate of the vocal tract that minimizes the linear predictive error for a frame of speech. An estimate of the glottal derivative can then be obtained by inverse filtering the original speech waveform in either the time or frequency domains.

The next step of the estimation procedure is to determine the amplitudes and phases of the first two harmonics. This is performed using a sinusoidal analysis method such

as that found in [24]. This class of techniques models a speech waveform as a sum of sinusoids. Sinusoidal parameters are typically obtained iteratively on a frame-by-frame basis by minimizing an error criterion. These methods are able to accurately capture the amplitudes, frequencies, and phases of the sinusoids and have the advantage of being able to compensate for the spectral effects of the window used to truncate the waveform.

At this point, the relative phase of the first two measured harmonics is calculated with (55) and used to identify a candidate set of LF parameters. An iterative search is performed to find all the parameter combinations of the LF model that yield a relative phase within $\pi/10$ of the measured value:

$$|[\angle G(2 \cdot \Omega_0, I) - 2 \cdot \angle G(\Omega_0, I)] - \phi_2{}'| < \pi/10. \tag{56}$$

The final estimation of the parameters $O_q, \alpha,$ and Q_a is then made by performing a search for the candidate parameter set whose relative relative amplitude is closest to that of the measured signal as follows,

$$\arg\min_I \left| \frac{|G(\Omega_0, I)|}{|G(2 \cdot \Omega_0, I)|} - \frac{|H1|}{|H2|} \right|. \tag{57}$$

In order to minimize the computational load required for the search processes, a database of relative amplitudes and phases for all parameter combinations can be constructed prior to the estimation procedure. The size of the database can be controlled by imposing limits on the ranges of each of the parameters.

6.2.2 Reference Method

In order to evaluate the performance of the frequency-domain parameter estimation technique, experimental data was collected and parameter estimates were compared to those produced by a baseline time-domain method of measuring the glottal waveform. Accurately determining glottal flow has been a problem of considerable interest for several years. Several sensors have been developed to make physical measurements which estimate glottal flow. Devices such as flow masks [67] and electroglottographs (EGG) have

been shown to provide general approximations for glottal flow. While these techniques provide useful information, they often do not determine the exact information required for certain applications and may also be inherently inaccurate. Glottal inverse filtering is the only waveform-based method of determining glottal flow. Inverse-filtering techniques, however, require an accurate estimate of the closed phase portion of a periodic signal. This estimate is often prone to error. As mentioned earlier, it has been shown that auxiliary biological sensors can be used to locate the closed-phase portions of the acoustic signal and thus produce an accurate signal representing glottal flow. This particular approach was chosen as the front end for the reference method in the experiments conducted.

The auxiliary device used is a General Electromagnetic Sensor (GEMS). The GEMS device is a low-power miniature device consisting of a penetrating radar that can be used to detect the motion in the region of the glottis. When positioned correctly on the exterior of the throat adjacent to the glottis, the output of the radar during voiced speech is a signal from which many important characteristics can be extracted that are useful for speech processing. Studies have shown that reliable estimates of glottal activity can be derived from the output of the GEMS device [4, 5]. While it has been suggested that this signal can also be used to calculate subglottal pressure, we chose to use the GEMS device only to segment closed-phase portions of the acoustic signal. The closed-phase portions of the simultaneously recorded acoustic signal were hand-marked based on the glottal activity recorded by the GEMS. Methods for synchronizing the acoustic and GEMS signals as well as for removing the filter response of the GEMS device are outlined in [4] and were implemented in this experiment. Vocal tract filter parameters were then calculated from these segments and an estimate of glottal excitation was obtained through an inverse-filtering operation. The LF parameters were determined by performing an iterative time-domain optimization based on a minimum squared error criterion. The parameters $O_q, \alpha,$ and Q_a were then calculated from the LF parameters using (38)-(40).

6.2.3 Experimental Setup

Simultaneous recordings of 6 speakers were made using a high-quality condenser microphone and the GEMS device. The corpus consisted of 12 TIMIT sentences recorded by three male speakers in an isolated sound studio. Both signals–acoustic and electromagnetic–were recorded at a sampling rate of 10 kHz.

The acoustic data from the microphone was windowed with overlapping frames and the proposed parameter estimation technique was used to determine estimates for O_q, α, and Q_a for randomly selected voiced frames of data. These estimates were then compared to measurements for the same frames obtained through the multi-sensor inverse-filtering method.

6.2.4 Experimental Results

Figure 27 shows a comparison of the results produced by the two estimation methods for each parameter as well as a reference line indicating perfect correlation of the estimation techniques. For the parameter α, a strong correlation exists ($r = 0.91$) between the two methods, although the frequency-based method consistently yielded slightly lower estimates than the reference method estimates. The comparison of estimates for the parameter O_q revealed a relatively low correlation ($r = 0.63$) with the proposed method while generally producing somewhat higher estimates than the reference method. However, a small number of frames resulted in grossly erroneous estimates of either the minimum ($O_q = 0.2$) or maximum ($O_q = 0.99$) values of the search interval. The removal of these outliers improved the correlation coefficient to $r = 0.87$.

The frequency-based estimation technique produced results for Q_a that were only moderately correlated to the reference estimates ($r = 0.75$). As shown in part (c) of Fig. 27, the correlation between the two estimates for the return phase coefficient, Q_a, is minimal. It is likely that this is partly due to the moderate correlation of Q_a to the relative amplitude and relative phase measures as compared to O_q and α.

The experimental results show encouraging evidence that both the amplitudes and phases of the first two harmonics can be used to provide an accurate estimation of time-domain glottal parameters. Although the return phase parameter showed very little correlation with the reference measurements, the estimates for both the open quotient and asymmetry coefficient showed a significant correlation with the GEMS-based measurements. The tendency of the frequency-based estimates of O_q to be higher than the GEMS estimate and the slightly lower estimates of α were largely related to the observation that the spectra of the estimated glottal waveform showed a higher level of spectral tilt than expected.

One possible factor which may have contributed to the inconsistent results for the return phase parameter, Q_a, is the high level of sensitivity to quantization error which affects this measurement. The analysis presented earlier shows that large variations in the parameter Q_a produce only moderate effects in the relative amplitude and phase measurements.

It should also be noted that for a number of voiced frames, the initial search for parameter sets within the relative phase boundaries produced no candidates. In this situation, no estimation was made for that particular frame. There are a number of factors which could lead to an unsuccessful candidate search. Occasionally, the iterative closed-phase inverse-filtering technique produced an irregular estimate of the glottal source. This was typically caused by an erroneous estimate of the glottal closure instant or an acoustic waveform with a closed-phase segment which was too short to produce an accurate approximation of the vocal tract filter. Incomplete closure of the glottis or a high level of glottal leakage–characteristics of certain modes of phonation, such as breathy or whispered speech–would also cause an inaccurate estimate.

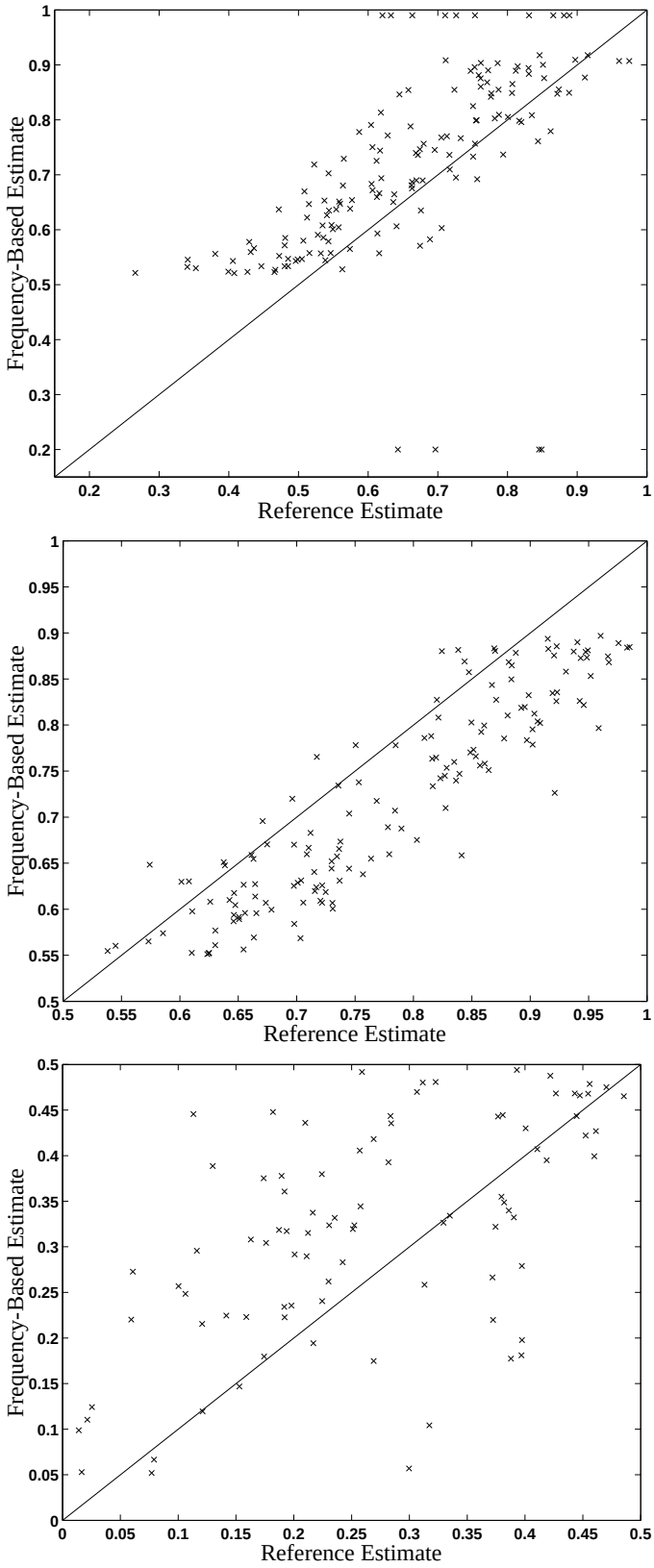


Figure 27: Comparison of frequency-based estimates of time-domain glottal parameters and reference estimates (O_q (top), α (middle), Q_a (bottom)).

6.3 Glottal Modification

The methods outlined in the previous section serve to illustrate the intimate relationship between time-domain glottal characteristics and their frequency-domain counterparts. It was shown that two perceptually relevant time-domain parameters, α and O_q, can be predicted based on frequency domain measurements of the relative amplitudes and phases of the first two harmonics. The purpose of this exercise was not only to establish a link between glottal source parameters in both the time and frequency domains, but also to provide an accurate, yet simple parameterization of glottal excitation. While researchers have typically used either a large set of time-domain parameters or a small set of frequency-domain parameters (spectral tilt, H1*-H2*), we propose a set of parameters that retains characteristics captured by each of these classes of glottal parameters.

The parameters H1*-H2* and $\Delta\phi_2$ have been shown to be useful in analyzing glottal waveforms. However, experiments have shown that altering these parameters do not produce reliable modifications to the glottal source. This is mainly because these parameters are only indicators of broader phenomena. The measurement of H1*-H2* specifies the relative amplitude of the first two harmonics, but in many cases this is merely an indication of the general roll-off of all of the harmonics. This characteristic is often referred to as the *spectral tilt*. While the spectral tilt is a frequency-based characteristic that has been shown to be linked to vocal textures and singing quality, it too is not a parameter that can be solely modified to change the shape of the glottal pulse in a perceptually controlled manner. In a similar fashion as H1*-H2*, the relative phase measure of the first two harmonics, $\Delta\phi_2$, is an indicator of the behavior of the remaining harmonics.

Figure 28 shows the amplitudes and relative phases for the first ten harmonics of the glottal sources of the vowel */a/* sung by an untrained and trained singer. As can be seen, the shape of the glottal pulses are largely correlated with the roll-off of the harmonic amplitudes. This is consistent with earlier studies measuring the spectral tilt of singing voices. Part (c) indicates the relative phases of the harmonics to that of the fundamental. These

phases appear to follow a linear pattern with a consistent slope. The previous analysis of the relative phase of the first two harmonics as well as observations of this general pattern indicate that this slope is also associated with glottal pulse shape. This slope combined with the spectral tilt serves as reliable parameters from which open quotient and asymmetry values can be modified.

These two parameters can be calculated directly from the sinusoidal components as:

$$S.T. = \frac{6.02 \sum_{j=2}^{N_s} A_j \frac{\log\left(\frac{A_j}{A_1}\right)}{\log\left(\frac{\omega_j}{\omega_1}\right)}}{\sum_{j=2}^{N_s} A_j} \tag{58}$$

$$\overline{\Delta\phi} = \frac{\sum_{j=2}^{N_s} A_j \frac{\left(\phi_j - \phi_1 \frac{\omega_j}{\omega_1}\right)}{j-1}}{\sum_{j=2}^{N_s} A_j}, \tag{59}$$

where (A_j, ω_j, ϕ_j) are the amplitudes, frequencies and phases of the component sinusoids, and N_s is the number of sinusoids. While there are several methods methods for measuring and representing spectral tilt, this formulation was developed to utilize the sinusoidal parameters and produce an output with the units of dB/octave. The average relative phase, $\overline{\Delta\phi}$, represents the average slope of the unwrapped phases of the sinusoids relative to the fundamental in units of rad/rad.

Our experiments have shown that modifying these parameters for a singing voice can have a perceptual effect on the vocal texture as well as the perceived quality of the voiced waveform. We contend that it is therefore possible to use this set of parameters to impose vocal quality enhancements and stylistic modifications on a singing voice. The following chapter discusses the implementation of this technique in combination with the proposed spectral modification system to perform stylistic enhancements to specific voices.

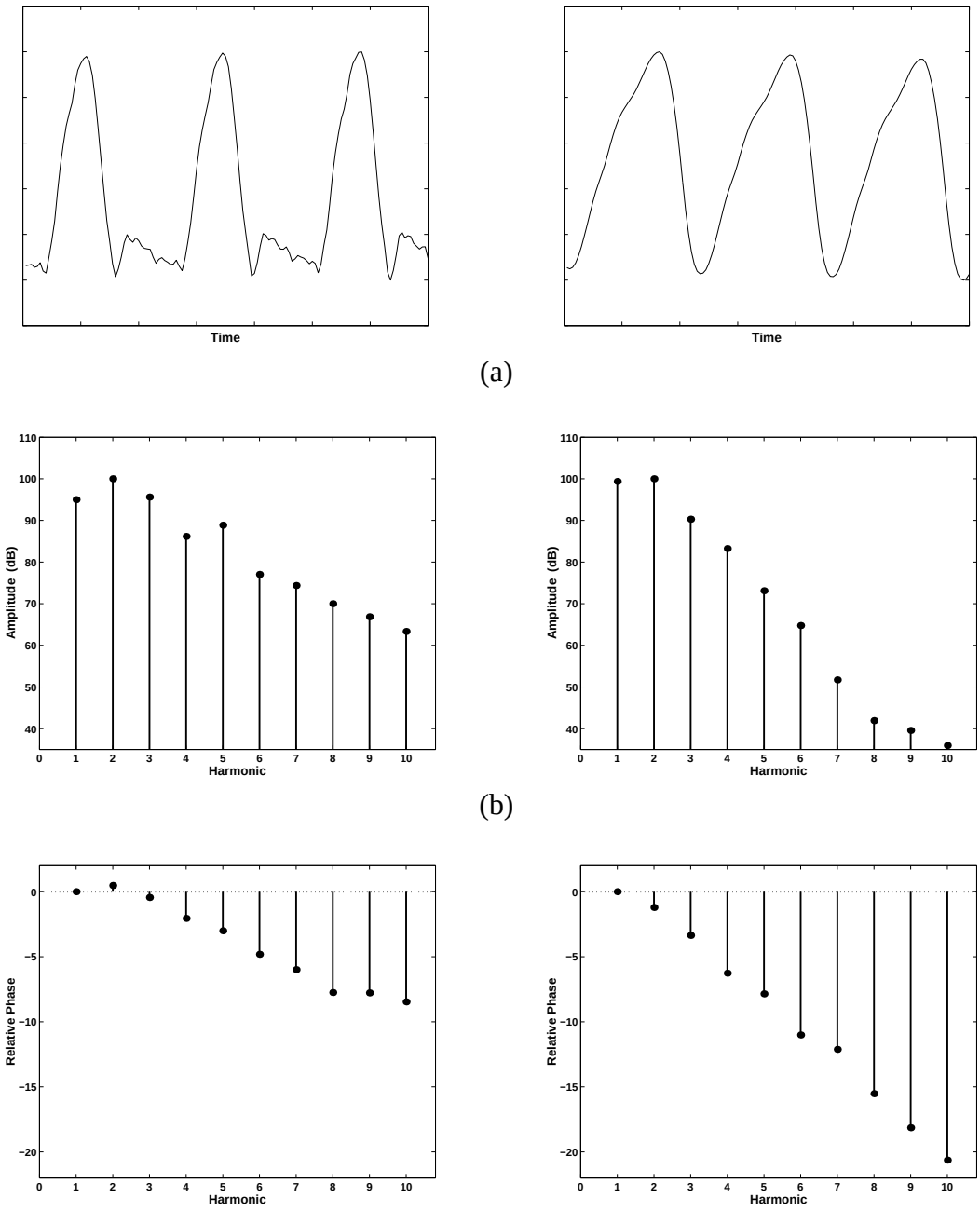


Figure 28: Examples of (a) glottal waveform, (b) harmonic amplitudes, and (c) relative phases for a voiced frame of a trained singer (left) and an untrained singer (right).

CHAPTER 7

CHARACTERIZATION OF THE SINGING VOICE USING THE PROPOSED SPECTRAL AND GLOTTAL MODELS

The spectral and glottal models that have been outlined in the previous chapters have been shown to be capable of parameterizing many important characteristics of the singing voice. The flexibility of these parameters enable various techniques for the modification and transformation of the singing voice. However, as we have demonstrated, the singing voice is a highly complex instrument that can be configured in a virtually limitless number of combinations to produce a desired sound. While a long history of singing voice pedagogy exists for training singers to control these mechanisms in their voices, only recent research has uncovered acoustic qualities that result from training. In order to determine the usefulness of our models for representing these qualities as well as to enable high-quality vocal enhancement, it is necessary to quantify these acoustic qualities using the parameters of the proposed models.

In this chapter, experimental results are presented in which the spectral and glottal models are used in an attempt to characterize the singing voices of trained versus untrained singers. While there are many different genres and styles of singing as well as differing opinions concerning the desired perceptual qualities of singing, the scope of the experiments presented here has been narrowed to focus on the Western classical singing tradition. This style of singing, which includes opera and most classical styles, possesses qualities which are generally agreed upon and have remained consistent throughout it's history. The purpose of this investigation is to illustrate the effectiveness of the proposed models for characterizing vocal qualities which can differentiate overall vocal beauty as well as level of training.

7.1 Experiment 1

In the first experiment, 8 male subjects, ranging in ages from 21 to 45, were recorded singing various segments of the song, *The Star Spangled Banner*. Four of the subjects had no previous training and no previous professional experience. The remaining four subjects were classically trained singers with extensive (10 years or more) singing experience in the performing arts. The segments were chosen such that the notes could be comfortably sung by each of the singers in the chest register. For the subjects involved, this required the segments to be sung in a key such that all notes were sung at F above middle C or lower. All of the subjects were recorded in an isolated studio. The samples were recorded at a sampling rate of 48 kHz and downsampled to 16 kHz for computational efficiency. The samples were then analyzed with both the spectral and glottal models outlined in Chapter 3.2.2 and Chapter 5.2.2.

7.1.1 Spectral Analysis

The spectral modelling technique described in Chapter 3.2.2 represents the spectral envelope of a windowed frame of data as a sum of asymmetric generalized Gaussian functions, each of which is comprised of six parameters, as shown in (26). These six parameters, $[A, \mu, \beta^l, \beta^r, \alpha^l, \alpha^r]$, quantify the amplitude and the center frequency as well as the width and shape of the left and right sides of the function. Before, analyzing the recorded data, voiced portions were segmented and phonetically labelled. A representative subset of these segments were selected and parameterized with the spectral model. Table 1 shows the average parameter values for the phones */a/*, */i/*, and */o/* for each of the subjects. In these cases, the model was limited to parameterizing only the first four formants. As can be seen in the table, several segments are represented with only three asymmetric generalized Gaussian functions. This is due to a single function representing two merged formants, such as the Singer's Formant.

The average values of the frequency and amplitude parameters for the trained singers

versus the untrained singers are given in Table 2. An analysis of the measured model parameters presented in this table reveals a number of patterns differentiating the singers with no previous training or experience from the trained singers. The most prominent difference is the strong, sometimes dominating, presence of the Singer's Formant in the voices of the trained singers. Figures 29,30, and 31 illustrate the asymmetric Gaussian function averages for the phones */a/*, */i/*, and */o/* for the trained and untrained singers. As shown in these figures, there is a consistent Gaussian function with high amplitude in the frequency region of 2500 Hz to 3000 Hz for the trained singers. On average, the formant in this region has an amplitude approximately equal to that of the first formant and more than 3 dB greater than the second formant. The Singer's Formant shows a high level of consistency across all of the trained singers as well as across all of the vowel sounds. Additionally, because no discernable formants are detected in the region of 3000 Hz to 3500 Hz, these results are consistent with the hypothesis that the Singer's Formant is a result of the merging of the upper formants.

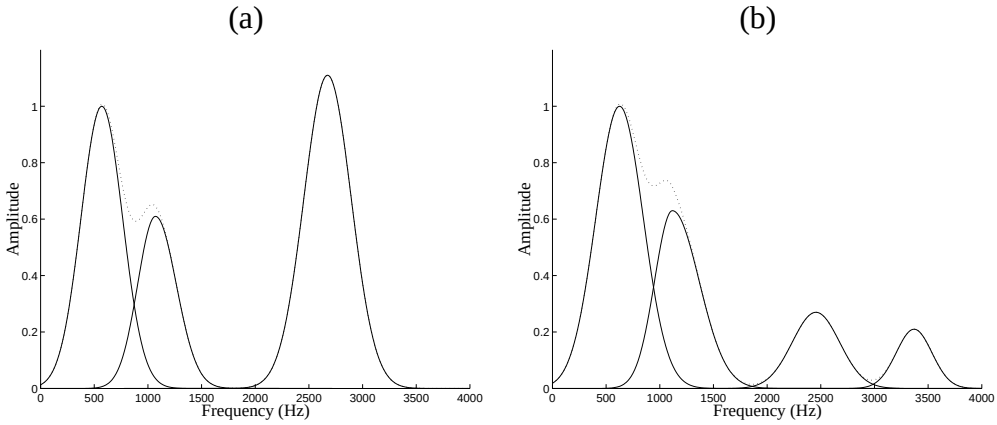


Figure 29: Spectra of the average asymmetric generalized Gaussian functions for trained (left) and untrained (right) singers singing the vowel */a/*.

Conversely, the untrained singers show little energy in the upper formants. The third and fourth formants for these singers have average amplitudes that are 17 dB and 20 dB lower than that of the first formant. They also show no signs of merging or blending. The

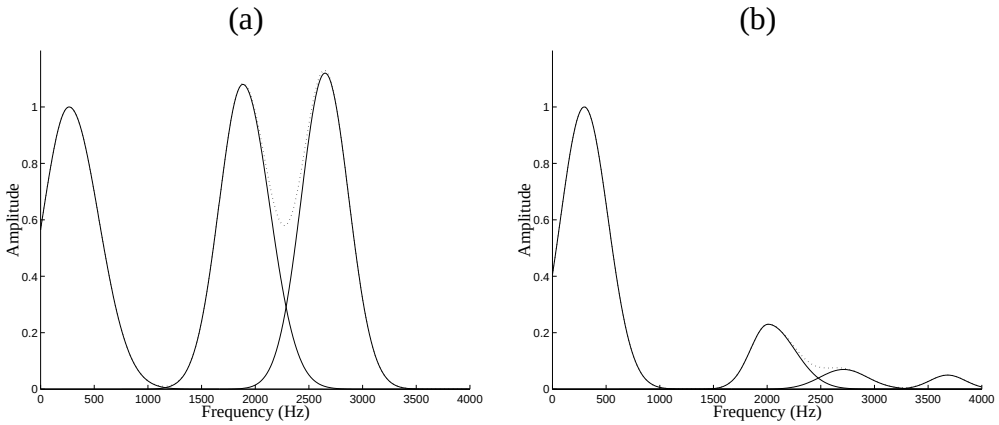


Figure 30: Spectra of the average asymmetric generalized Gaussian functions for trained (left) and untrained (right) singers singing the vowel */i/*.

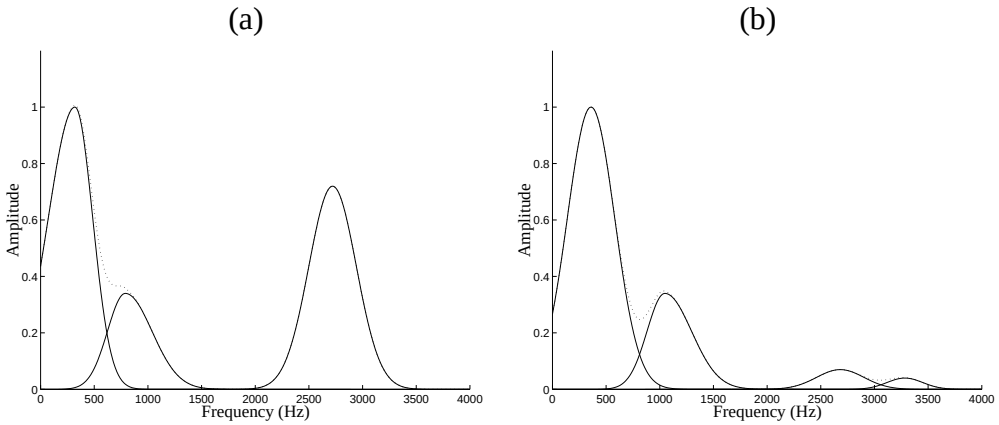


Figure 31: Spectra of the average asymmetric generalized Gaussian functions for trained (left) and untrained (right) singers singing the vowel */o/*.

average center frequencies of these two resonances in the data are separated by more than 80 Hz.

Another pattern that emerged form the analysis parameters concerns the center frequencies of the first two formants. It has been well established in speech processing that these two formants are closely related to the identity of a spoken vowel [10, 57]. Frequency modification of these formants can therefore have an undesirable effect on the intelligibility of the output. However, classically trained singers are often taught to "color" their vowels which results in the modification of the first two formants [73]. This typically serves

two purposes. First, a quality most often described as “dark” or “covered” is obtained by lowering the center frequencies of these formants. Additionally, formant tuning is often performed in order to maximize the energy in the available harmonics.

The mechanism behind this tuning has been largely attributed to a lowering of the larynx during phonation. Sundberg notes that this phenomenon results in a lowering of the formant frequencies because a depressed larynx effectively elongates the vocal tract [73]. Sundberg also performed X-ray examinations in which he showed that a lowering of the larynx has a secondary effect of widening the bottom part of the pharynx. This effect is believed to be responsible for the Singer’s Formant. By widening the pharynx, an impedance mismatch is formed between the laryngeal tube and the lower pharynx, creating a strong resonance that is independent of the remainder of the vocal tract [1, 12, 71].

In order to understand the results of formant modifications induced by trained singers, it is useful to analyze a chart plotting the first two formants for various vowels sung by the subjects. This is given in Figure 32. It can be clearly seen from this figure that the trained singers maintain lower frequencies for both the first and second formants. While it is possible that this could be due to natural factors such as vocal tract length, the data supports claims that the lowering of formant frequencies is a result of an intentional manipulation of the vocal tract mechanism by trained singers.

The amplitude parameters for the first two formants also show an interesting pattern. While the relative amplitude of the second formant to the first is approximately equal for untrained and trained singers for the vowels */a/* and */o/*, the amplitude of the second formant is considerably larger for the vowel */i/* in trained singers. In fact, the second formant in trained singers exhibits greater amplitude (with a relative amplitude of 1.08) than the first formant. This compares to a relative amplitude of 0.23 for untrained singers. It should be noted that the phoneme */i/* is a *front vowel*, whereas */a/* and */o/* are *back vowels*. This terminology refers to the placement of tongue constriction in the oral tract. Front vowels are typically characterized by higher values in center frequencies of the second formant.

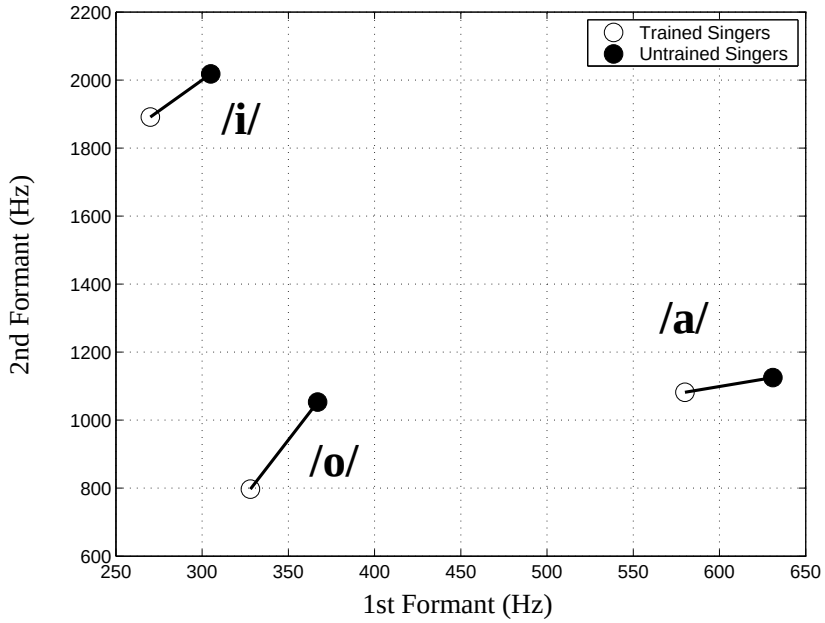


Figure 32: Comparison of center frequencies for the first two formants for trained and untrained singers for the phones */i/*, */a/*, and */o/*.

The data therefore suggests that trained singers are able to provide additional power to this resonance during phonation of front vowels.

One advantage of using the parameters of the asymmetric generalized Gaussian functions is that they enable an examination of the bandwidth characteristics of formants. It has been noted that while the bandwidths of formants likely have an effect on vowel identity, their primary contribution is to a vowel's character or quality [10]. Table 3 reveals a number of patterns concerning the bandwidths of the Gaussian functions. Bandwidths are calculated as the sum of the average values for the β parameter on the left and right sides,

$$BW = \beta^l + \beta^r. \tag{60}$$

This definition of bandwidth is equivalent to finding the frequency range in which the amplitude is within 8.69 dB of the peak.

The bandwidths of the first two formants appear to be slightly greater on a consistent basis for the trained singers than for the untrained singers. Studies have shown that increased formant bandwidths can be linked to a higher level of nasalization during vowel

phonation [10].

The third formant shows similar bandwidths for both groups of singers. This is significant, however, in light of the observation that the amplitude of the third formant is significantly larger for the trained singers. It appears that trained singers can raise the energy in this formant without altering the bandwidth. This is an example of the flexibility of the asymmetric generalized Gaussian model. An all-pole model is not able to independently control the amplitude and bandwidth of a formant in this fashion unless multiple poles are used for a single formant.

7.1.2 Glottal Analysis

The glottal parameterization presented in Chapter 5.2.2 first applies a closed-phase inverse filtering technique and then characterizes the resulting glottal waveform with two frequency-domain parameters. These parameters reflect the spectral tilt of the glottal harmonics as well as the average slope of the relative phases. The selected voiced portions of the recorded subjects samples were analyzed using this technique and average values for the trained and untrained subjects were calculated. These results are given in Table 4. In addition, the time-domain parameters for the open quotient and asymmetry coefficient are calculated using the frequency-domain estimation technique outlined in Section 6.2 based on the relative amplitudes and phases of only the first two harmonics.

As noted in earlier observations (Figure 28), both the spectral tilt and the slope of the relative phase show a significant discrepancy between trained and untrained singers. The averages shown in Table 4 quantify these differences. For all three measured phones, the glottal waveforms of trained singers show lower roll-off by an average of 8.4 dB/octave. The largest average difference occur for the phoneme */i/* (14.7 dB/octave) and the smallest for */o/* (2.1 dB/octave). The average slope of the glottal harmonic phases relative to the fundamental also show a similar pattern. The phases of the untrained singers exhibit a much sharper negative slope than those of the trained singers. These differences range

from 12.6 rad/rad for the phoneme */a/* to 21.9 rad/rad for */o/*.

It was shown earlier that these frequency-domain characteristics exhibit a strong relationship with the time-domain parameters of the glottal waveform. The time-domain estimates of the open quotient (O_q) and asymmetry coefficient (α) are consistent with studies that show a lower O_q and higher α for classically trained singers.

7.2 Experiment 2

A second experiment was conducted in which subjects were asked to perform a vocal exercise known as *arpeggio*. The four classically trained male subjects (T1, T2, T3, T4) from the previous experiment were recorded along with four new untrained subjects (U5, U6, U7, U8) which were not part of the previous experiment. An arpeggio consists of a series of notes which make up a chord. In this particular experiment, the subjects were asked to sing the notes composing the chord A-major using the vowel */a/* in an ascending and descending pattern as shown in Figure 33. These notes consist of A_3, $C_4\sharp$, E_4, and A_4. The A-major chord was chosen for this experiment because the range of notes, from A below middle C (A_3) to A above middle C (A_4), span the boundary between the chest register and the head register for most male singers. The recorded samples in this experiment were recorded and processed with the presented analysis techniques in a manner identical to that in Experiment 1.

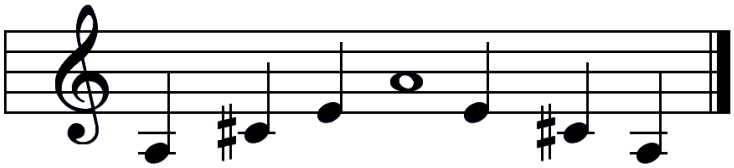


Figure 33: A-major arpeggio.

7.2.1 Spectral Analysis

A number of researchers have claimed that changes in registration in singing are mainly associated with different modalities of glottal excitation. Various studies [2, 31, 76] have used signals obtained from electroglottographs or glottal flow masks to support this claim. However, it is still highly debated whether these modes of excitation are accompanied by changes in the shape of the vocal tract. This spectral analysis is designed to shed some light on this issue as well as identify some differences in registration techniques (or lack thereof) between trained and untrained singers.

The parameters of the asymmetric generalized Gaussian spectral model were calculated for the notes A_3 and A_4 for each of the singers in the experiment. The average values of the frequencies and amplitudes for each of the subjects are given in Table 5. Average values for the complete set of parameters for the trained and untrained subgroups are given in Tables 6 and 7. As expected, the parameters for the notes sung at A_3 reflect very similar patterns to those observed in Experiment 1. The spectral parameters of the trained singers show a prominent singer's formant consisting of a merging of the third and fourth formants. Additionally, the first two formants of the trained singers are lower in frequency and have a slightly greater bandwidths.

When comparing the parameters of the untrained singers for both notes, a few differences can be observed in their respective formant structures. Tables 6 and 7 show a small increase in formant frequencies as well as slightly wider bandwidths in the first two formants. Additionally, the ratio of the second formant amplitude to that of the first increases from 0.59 to 0.88 for A_4. This increase in ratio, however, is significantly more substantial for the trained singers. The average A2/A1 ratio jumps from 0.56 to 4.00, an increase of 614%. This phenomenon may not be solely caused by an increase in the resonance associated with the second formant, but rather by a decrease in the low-frequency resonance associated with the first formant. This supposition is substantiated by an increase in the ratio of the third formant to the first (A3/A1) from 1.25 to 5.96. The discrepancy between

trained and untrained singers shows a high level of consistency across all of the subjects. It has been hypothesized that trained singers are able to reduce the resonant power of the first formant by tuning it so that it lies directly between two harmonics [52].

Figures 34 and 35 further illustrate differences between trained and untrained singers by showing the average spectral models for subjects T3 and U3 for each of the notes. Comparing the spectra of singer T3 for notes A_3 and A_4 shows a significant decrease in the relative amplitude of the Gaussian function modeling the first formant. The spectra of singer U3, however, shows little change in the relative amplitudes of the formants. This is most likely due to an improper registration technique among the untrained singers. The most common result of improper registration is a singer attempting to maintain the chest register at notes that are above the proper range of this particular register [6]. The spectra of singer U3 appears to exemplify this error. The similarity of the spectra for notes that should clearly be sung in different registers indicates a failure to transition from chest to head.

Table 1: Experiment 1: Average AGG frequency and amplitude parameter values for each subject for the phones */a/*, */i/*, and */o/*.

	Subject	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	A1	A2	A3	A4
/a/	T1	562	1063	2664		1	0.47	0.94	
	T2	602	1117	2516		1	0.34	0.71	
	T3	508	1039	2617		1	0.94	2.09	
	T4	648	1109	2930		1	0.70	0.71	
	U1	648	1125	2734	3367	1	0.77	0.34	0.27
	U2	617	1117	2359	3203	1	0.53	0.31	0.22
	U3	656	1172	2453	3633	1	0.52	0.29	0.27
	U4	602	1086	2305	3289	1	0.71	0.12	0.07
/i/	T1	273	1977	2781		1	1.41	0.91	
	T2	258	1914	2586		1	1.05	0.78	
	T3	242	1773	2461		1	1.07	1.90	
	T4	305	1898	2813		1	0.80	0.87	
	U1	289	2133	2938	3852	1	0.14	0.02	0.03
	U2	313	2063	2531	3695	1	0.22	0.06	0.03
	U3	297	1867	2711	3484	1	0.25	0.09	0.08
	U4	320	2008	2695	3734	1	0.29	0.11	0.06
/o/	T1	336	781	2930		1	0.60	0.31	
	T2	297	719	2617		1	0.76	0.80	
	T3	359	852	2508		1	1.06	1.00	
	T4	320	836	2852		1	0.51	0.77	
	U1	336	1102	2734	3313	1	0.25	0.07	0.05
	U2	375	1055	2594	3047	1	0.29	0.06	0.03
	U3	359	1047	2727	3367	1	0.25	0.11	0.07
	U4	398	1008	2688	3438	1	0.54	0.03	0.02

Table 2: Experiment 1: Average AGG frequency and amplitude parameter values for trained and untrained singers for the phones */a/*, */i/*, and */o/*.

	Subjects	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	A1	A2	A3	A4
/a/	Trained	580	1082	2682		1	0.61	1.11	
	Untrained	631	1125	2463	3373	1	0.63	0.27	0.21
/i/	Trained	270	1891	2660		1	1.08	1.12	
	Untrained	305	2018	2719	3691	1	0.23	0.07	0.05
/o/	Trained	328	797	2727		1	0.34	0.72	
	Untrained	367	1053	2686	3291	1	0.34	0.07	0.04

Table 3: Experiment 1: Average AGG width and shape parameter values for trained and untrained singers for the phones */a/*, */i/*, and */o/*.

		(left/right)			
	Subjects	$\beta 1$	$\beta 2$	$\beta 3$	$\beta 4$
/a/	Trained	297/273	305/266	281/344	
	Untrained	258/219	211/281	305/242	289/273
/i/	Trained	688/391	313/391	430/352	
	Untrained	750/375	344/313	438/313	242/266
/o/	Trained	344/219	234/250	250/313	
	Untrained	320/211	211/188	313/227	219/180

		(left/right)			
	Subjects	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$
/a/	Trained	1.75/2.11	2.45/1.54	1.96/1.87	
	Untrained	1.77/2.08	2.34/1.79	1.84/1.85	2.04/1.87
/i/	Trained	1.56/1.83	1.94/2.43	2.83/1.98	
	Untrained	1.72/1.97	1.81/1.71	2.74/1.94	1.77/1.81
/o/	Trained	1.34/2.41	2.82/1.91	1.74/1.58	
	Untrained	1.45/1.94	1.43/1.64	1.29/1.85	1.76/1.78

Table 4: Experiment 1: Average glottal parameters of trained and untrained singers for the phones */a/*, */i/*, and */o/*.

	Subjects	Spectral Tilt (dB/octave)	$\overline{\Delta\phi}$ (rad/rad)	O_q	Asymmetry Coeff. (α)
/a/	Trained	-12.4	-7.1	0.68	0.69
	Untrained	-20.7	-19.7	0.71	0.56
/i/	Trained	-9.5	-9.4	0.64	0.81
	Untrained	-24.2	-23.1	0.75	0.59
/o/	Trained	-16.1	-8.3	0.67	0.77
	Untrained	-18.2	-30.2	0.72	0.67

Table 5: Experiment 2: Average AGG frequency and amplitude parameter values for each singer for the notes A_3 and A_4.

	Subject	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	A1	A2	A3	A4
A3	T1	617	1070	2710		1	0.37	1.00	
	T2	625	1171	2460		1	0.26	0.74	
	T3	492	1078	2593		1	0.97	2.59	
	T4	711	1164	2984		1	0.64	0.66	
	U5	648	1109	2789	3320	1	0.76	0.39	0.32
	U6	625	1093	2289	3328	1	0.49	0.23	0.19
	U7	617	1164	2468	3710	1	0.55	0.15	0.04
	U8	641	1132	2507	3398	1	0.56	0.26	0.12
A4	T1	656	1102	2563		1	1.21	2.91	
	T2	672	1086	2336		1	5.13	2.54	
	T3	609	1328	2656		1	3.63	14.50	
	T4	641	1133	2578		1	6.05	3.89	
	U5	773	1273	2953	3469	1	1.56	0.52	0.30
	U6	688	1156	2328	3305	1	0.52	0.38	0.09
	U7	680	1289	2594	3758	1	0.64	0.15	0.06
	U8	719	1203	2664	3523	1	0.82	0.26	0.12

Table 6: Experiment 2: Average AGG frequency and amplitude parameter values for the trained and untrained singers for the notes A_3 and A_4.

	Subjects	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	A1	A2	A3	A4
A3	Trained	611	1121	2688		1	0.56	1.25	
	Untrained	633	1125	2514	3439	1	0.59	0.26	0.17
A4	Trained	645	1162	2533		1	4.00	5.96	
	Untrained	715	1230	2635	3514	1	0.88	0.33	0.14

Table 7: Experiment 2: Average AGG bandwidth and shape parameter values for the trained and untrained singers for the notes A_3 and A_4.

		(left/right)			
	Subjects	$\beta 1$	$\beta 2$	$\beta 3$	$\beta 4$
A3	Trained	258/281	242/258	273/289	
	Untrained	258/242	227/234	273/242	250/258
A4	Trained	305/250	289/250	305/367	
	Untrained	344/242	250/266	313/258	328/305

		(left/right)			
	Subjects	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$
A3	Trained	1.25/2.09	2.68/1.76	1.86/1.85	
	Untrained	1.18/2.28	2.90/1.59	1.19/1.94	2.11/1.59
A4	Trained	1.42/2.02	1.48/1.86	1.65/1.46	
	Untrained	1.16/2.50	2.77/1.83	1.73/2.02	2.53/1.66

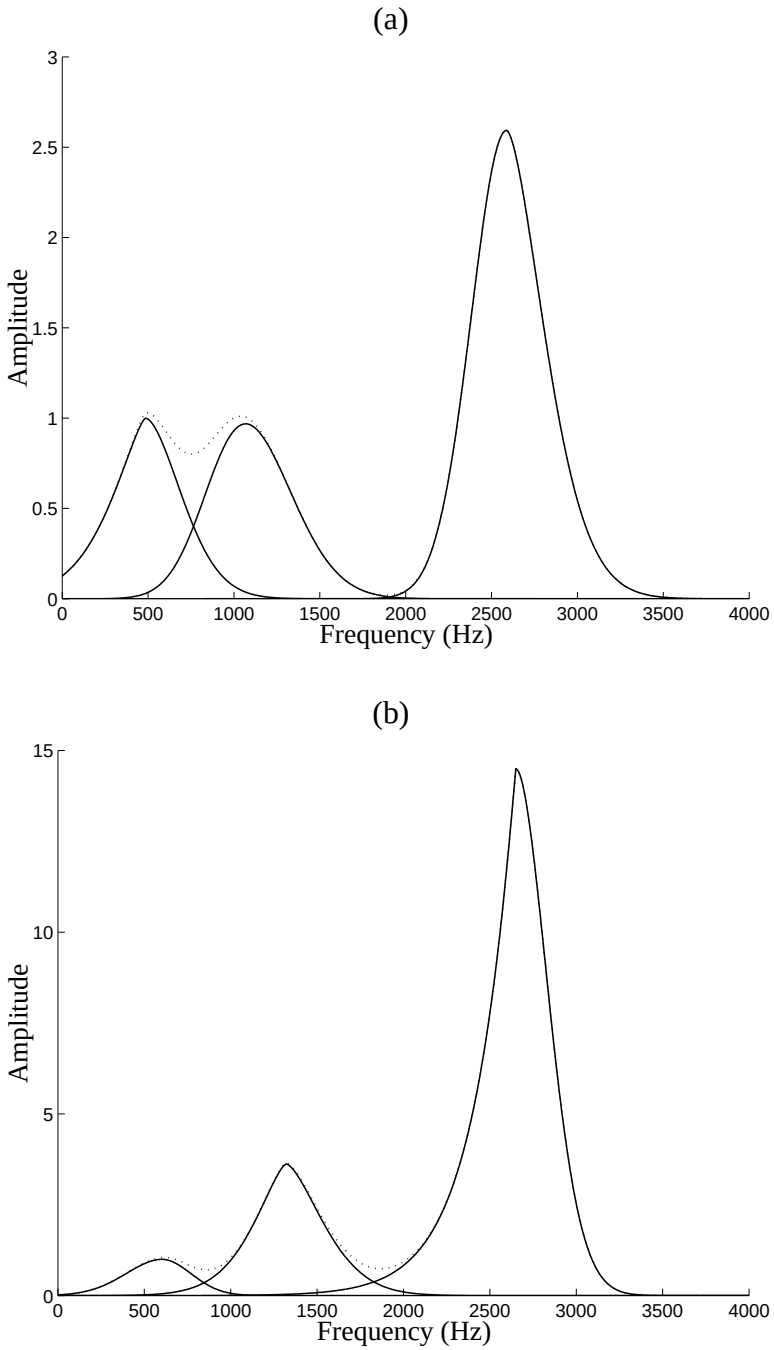
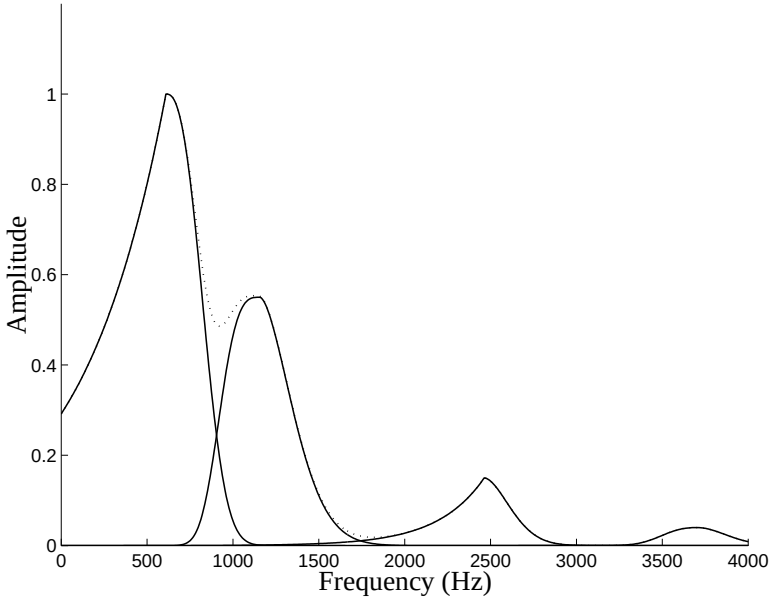


Figure 34: Spectra of the average asymmetric generalized Gaussian functions for the trained singer T3 for the notes A_3 (left) and A_4 (right).

(a)



(b)

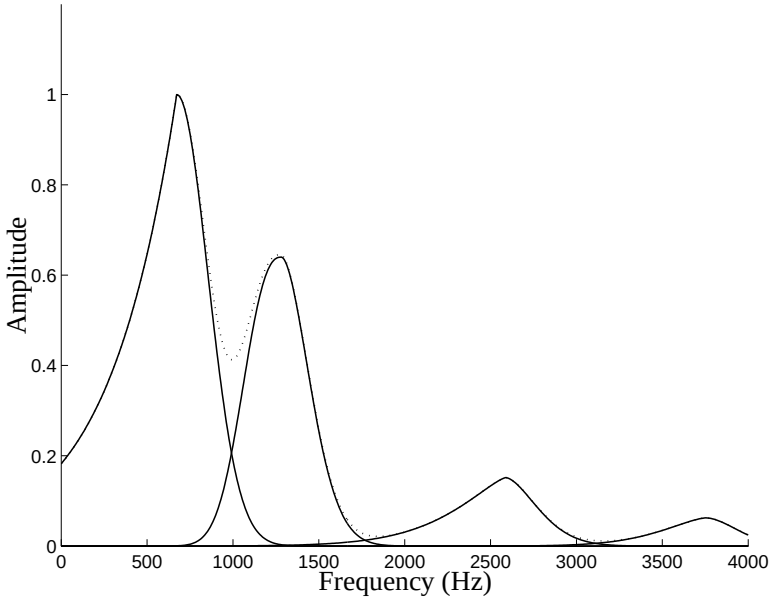


Figure 35: Spectra of the average asymmetric generalized Gaussian functions for the untrained singer U3 for the notes A_3 (left) and A_4 (right).

7.2.2 Glottal Analysis

As discussed earlier, it has been widely thought that a strong link exists between singing registers and distinct modes of glottal production. These modes are typically implemented by controlling the length and thickness of the vocal folds. In the chest register, the vocal folds are relatively short and thick. Vibration occurs over the entire length of the vocal fold with a vertical phase difference, whereas in the head register, the mass and vibratory length are reduced [86] and there is little to no vertical phase difference. In this mode, the glottal flow signal exhibits a more symmetric quality.

An analysis of the glottal parameters for the trained and untrained singers is performed by applying the presented glottal model and estimation techniques to the recorded waveforms. Table 8 reflects the average parameters for each group of singers. The measured parameters (spectral tilt, average relative phase slope, open quotient, and asymmetric coefficient) for the note A_3 coincide closely with those calculated in Experiment 1. The parameters for A_4, however, show a much smaller discrepancy between trained and untrained singers. While the untrained singers exhibit glottal characteristics very similar to those sung in A_3, the trained singers show an increase in spectral tilt, $\overline{\Delta\phi}$, and O_q along with a decrease in asymmetry.

These results support the idea that trained singers are able to maintain separate modes of glottal production. Glottal flow during chest registration versus head registration shows less spectral roll-off in terms of amplitude as well as relative phase of the harmonics. This corresponds in the time domain to a larger open quotient and a more asymmetric glottal pulse. Although additional registers (i.e. falsetto) for males do exist, chest and head are the two that are most commonly used in classical singing are are also the most easily identified [52].

The untrained singers show little change in glottal characteristics when comparing the two notes. This supports the hypothesis that untrained singers are typically incapable of transitioning from one register to the next, thus attempting to sing all notes in the chest

register until the voice "breaks" into falsetto. However, it it interesting to note that the glottal characteristics of the untrained singers in this experiment are close to those of the trained singers singing in the head register and not the chest register.

Table 8: Experiment 2: Average glottal parameters of trained and untrained singers for the notes A_3 and A_4.

	Subjects	Spectral Tilt (dB/octave)	$\overline{\Delta\phi}$ (rad/rad)	O_q	Asymmetry Coeff. (α)
A3	Trained	-14.3	-8.7	0.63	0.71
	Untrained	-22.1	-18.8	0.68	0.53
A4	Trained	-18.5	-16.3	0.74	0.55
	Untrained	-23.5	-19.7	0.67	0.56

7.3 Conclusion

The two experiments presented in this chapter have exemplified the ability of the proposed spectral and glottal models to parameterize a singing voice signal so that various characterizations including level of training and mode of registration can be identified and measured. Previous studies have also been conducted using portions of the model to characterize various singing styles, such as Broadway belt [38] and country/western [37].

The ability to capture parameters associated with vocal qualities can be applied to a number of applications such as classification or vocal training feedback tools. The next chapter examines our attempt to combine these models with the results from experiments 1 & 2 to perform modifications to the voices of untrained singers in a manner that would enhance their vocal qualities.

CHAPTER 8

CLASSICAL ENHANCEMENTS TO THE SINGING VOICE

The previous chapter detailed experiments for using the proposed spectral and glottal models for identifying acoustic characteristics and model parameters that differentiate trained singers from untrained singers according to the western classical style of singing. This chapter explores an analysis/modification/synthesis application which is designed to use this knowledge to perform classical enhancements on the singing voices of untrained singers. Details of experiments used to evaluate the performance of this system are also provided. The advantage of an analysis/synthesis system is that a synthesized output is produced which can be used in human listening experiments to validate the system perceptually, which is arguably the most important metric. However, as mentioned in the introduction, music is highly subjective and the musical quality of a singers voice is not as easily agreed upon as the intelligibility of a spoken word.

Seventeen male subjects with no previous training or professional experience were asked to sing an arpeggio exercise identical to the one presented in the previous chapter. This exercise was sung with the vowel */a/* in the notes [A_3, $C_4\sharp$, E_4, A_4]. The subjects were allowed to listen to the correct notes immediately before singing in order to reduce any errors in pitch. For subjects whose comfortable range did not coincide with these notes, the arpeggio was raised or lowered one half note at a time until the subject was able to complete the task comfortably. This exercise was repeated and recorded five times. Recordings were made in a sound-proof studio with the amplified microphone output recorded directly to disk at 48 kHz. The data was then downsampled to a sampling rate of 16 kHz to reduce the computational requirements. Three of the five segments were then randomly chosen for each singer and analyzed offline with the proposed spectral and glottal modeling techniques

using 25 ms frames tapered with a Hamming window updated every 10 ms.

8.1 Spectral and Glottal Modifications

The vocal enhancements were performed so that the characteristics of the untrained singers mimicked those of the trained singers that were identified in the previous chapter. These modifications are summarized as follows:

- decrease frequencies of first two formants
- increase bandwidths of first two formants
- merge third and fourth formants and increase their amplitudes while maintaining their bandwidths
- for notes at or above F_4:
  - decrease amplitude of first formant
  - decrease the open quotient of the glottal flow waveform
  - increase asymmetry in the glottal flow waveform

The degree to which each of these modifications was performed was determined so that the modified output possesses parameters that are equivalent to the average of the source and target parameters. According to this procedure, the source parameters are updated each frame. The target parameters remain constant throughout the proper register (chest, head) and are determined by using the median values for the trained singers in Experiment 2 discussed in the previous chapter. Register transition regions were identified prior to modification based on target notes crossing the boundary between E_4 and F_4. The target parameters for these regions were then linearly interpolated between the target parameters for each register. These regions are illustrated in the pitch contour shown in Figure 36.

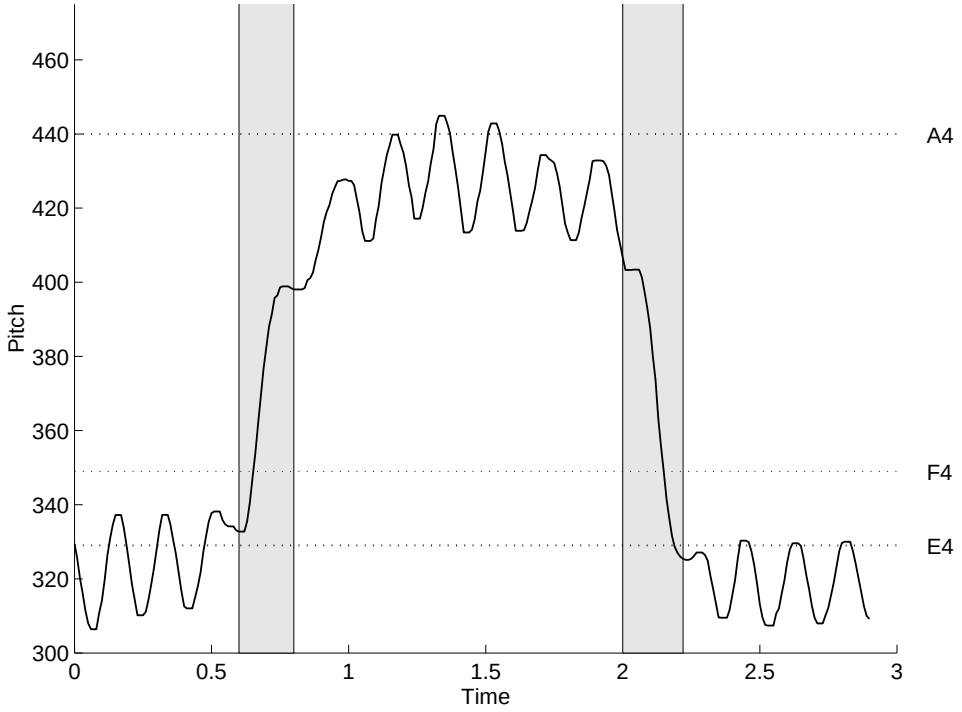


Figure 36: Example of identified register transition regions for a given pitch contour.

8.1.1 Competing Methods

In order to test the performance of the proposed enhancement, competing methods were also included in the experiments. The algorithms for spectral and glottal modification were implemented in a modular fashion so that they could be seamlessly interchanged with the proposed methods. The framework of the enhancement system is provided by the AbS/OLA sinusoidal model described in Chapter 3.2.2.

In determining a competing spectral method, a number of commercial products for singing voice processing were investigated. While a number of products exist that perform varying levels of spectral modification, they are mostly data-driven models in which a source singer's spectral envelope is replaced with that of a target singer. This has the effect of changing the identity of the singer, which is not desired in this particular application. Other products such as the Yamaha *Vocaloid* [87] implement rule-based synthesis in which lyrics and a musical score serve as the only inputs. The identity of the singer is solely based

on the loaded database of units which are selected and concatenated by the system.

The competing spectral modification method that was chosen for this experiment has been used in various forms in a number of leading singing voice enhancement systems, such as the TC Helicon *Voice Modeler* series of products. This technique is a frequency warping method based on the algorithm presented in Section 3.2.2. This method combines a frequency shift warping function, $\alpha(\omega)$, with an intensity shaping function, $\gamma(\omega)$, in order to control the formant frequencies and amplitudes. Bandwidth warping functions, $\beta(\omega)$, were also included for altering formant bandwidths. As shown in 61, these functions can be combined to modify a spectral envelope, $H(\omega)$, according to

$$\hat{H}(\omega) = \gamma(\omega) \cdot H((\alpha(\omega) * \beta(\omega)) \cdot \omega) \tag{61}$$

where $\hat{H}(\omega)$ is the modified spectral envelope.

While this particular method offers independent control over the formant frequencies, amplitudes, and bandwidths, it's effectiveness is only maximized when there is no formant interaction. This is only the case when formants are spaced far enough apart in frequency that they can be modified separately.

Although many techniques have been put forth for synthesizing glottal excitation waveforms based on a synthesis-by-rule approach, very few analysis/synthesis methods have been developed for the modification of an existing glottal signal in a meaningful way. Therefore, it is difficult to perform a comparative evaluation. However, the techniques proposed in this thesis are based on earlier studies which linked open quotient values to the relative amplitude of the first two harmonics of the glottal source [19, 23]. The open quotient has been shown by many studies–including the one presented in the previous chapter–to be highly correlated with various vocal qualities and textures. This implies a simple method for glottal modification. By modifying the relative amplitudes of the first two harmonics, and not accounting for phase relationships as proposed in Chapter 5.2.2, modifications which may be equivalent to modifying the open quotient may be possible. This technique was implemented and included in the experiment for a comparative evaluation with the

proposed glottal modification procedure.

8.2 Additional Modifications: Pitch and Vibrato

As mentioned in Chapter 3.2.2, the ABS/OLA sinusoidal model that is used in the proposed system is capable of natural sounding pitch-scaling using a phasor interpolation scheme [25]. This technique can be applied on a frame-by-frame basis, so that specific time-varying changes to the pitch contour of a waveform can be made. When enhancing the pitch of an untrained singer's voice, the two main aspects to be considered are note errors and vibrato.

While it is possible to modify the pitch contour of the original singer's waveform to the exact notes as prescribed by the musical score, the result is usually not a natural sounding waveform. Care must be taken to maintain the prosodic features of a voice to maintain the unique qualities of a singer's voice as well as it's individuality. Therefore, pitch corrections are performed gradually by allowing transition periods in between notes and applying slowly varying modifications. These tolerances may vary from singer to singer and often need to be adjusted to fit a particular singer's attributes. However, the result of such corrections is a pitch-corrected voice waveform that retains the singer's vocal qualities.

Vibrato is a highly important factor in the enhancement of the singing voice. As mentioned in Chapter 2, the presence of vibrato is present in nearly all trained singers' voices and is strongly correlated with the perception of vocal beauty [16]. The insertion or modification of vibrato requires a specification of the *rate* and *extent* of the sinusoidal oscillations of the fundamental frequency. Studies have shown that these parameters can vary based on the individual singer as well as the note being sung. However, one common characteristic of trained singers is the regularity of the vibrato cycles [50]. Untrained singers typically sing with little to no vibrato or with vibrato of varying rate. Another observation of trained singers regards the extent of vibrato. Prame [61] noted that vibrato extent tended to increase throughout the duration of a sustained note in trained singers' voices.

Based on these studies as well as our own observations, vibrato insertion is implemented as a frame-based frequency modulation. A desired pitch contour is outlined based on the score as well as target vibrato characteristics. The modulation function is modeled as a sinusoid with an amplitude envelope that is an increasing piecewise-linear function as shown in Figure 37. An example of the resulting modifications to the pitch contour of an untrained singer's voice is illustrated in Figure 38.

It has been observed that modulating the fundamental frequency with a sinusoid with constant frequency results in an unnatural sounding waveform. Although trained singers exhibit higher degrees of regularity, slight fluctuations in frequency nonetheless exist. Therefore, the modulating sinusoid of the vibrato model is phase modulated with a random noise signal that has been lowpass filtered. Vibrato is thus formulated by modifying the frame-based fundamental frequency ($F_0[n]$) as

$$F_0[n] = F_0[n] + a[n] \cdot \sin\left(2\pi\omega_v n + r[n]\right), \tag{62}$$

where $a[n]$ is the piecewise-linear amplitude envelope, ω_v is the vibrato rate (5-7 Hz), and $r[n]$ is the lowpass noise signal.

Vibrato onset time can also be modeled as a simple delay from the onset of the note to the onset of the pitch oscillation. While vibrato onset time has been demonstrated in a number of classical singers, the majority of classical singers have exhibited very short to no onset times. Others have hypothesized that longer onset times are common in other styles of singing such as Broadway belt [36].

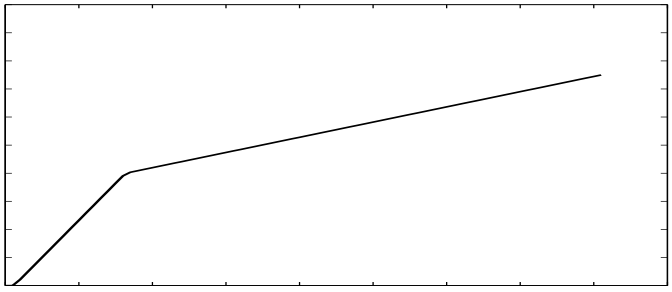


Figure 37: Piecewise-linear function used to shape the amplitude of the vibrato inserted into untrained singers' voices.

8.3 Listening Experiments

In order to measure the usefulness of the proposed methods for singing voice enhancement, it is necessary to determine a procedure for testing and evaluation. In addition to determining the level and nature of enhancement that can be achieved with the techniques proposed here, a comparison with competitive algorithms must be documented. However, it is clear that the testing and evaluation of the proposed system requires a unique methodology. There are many challenges to obtaining a consistent and reliable evaluation of the singing voice. Additionally, the testing of synthesized waveforms provides issues that must be addressed during testing. Issues such as the presence of artifacts and naturalness are important in determining the success of synthesis applications.

One of the challenges of subjective vocal quality evaluations is obtaining a sufficient number of experts to serve as evaluators. Typically, this type of resource is only available at large music institutions or conservatories. However, studies have shown that there is some value in using non-expert evaluators. Ekholm [15] conducted a study in which a group of vocal experts and a group of students evaluated the same set of vocal performances according to the twelve factors identified in [88]. Their results show that while the inter- and intrajudge reliability were higher for the group of vocal experts than for the students, there

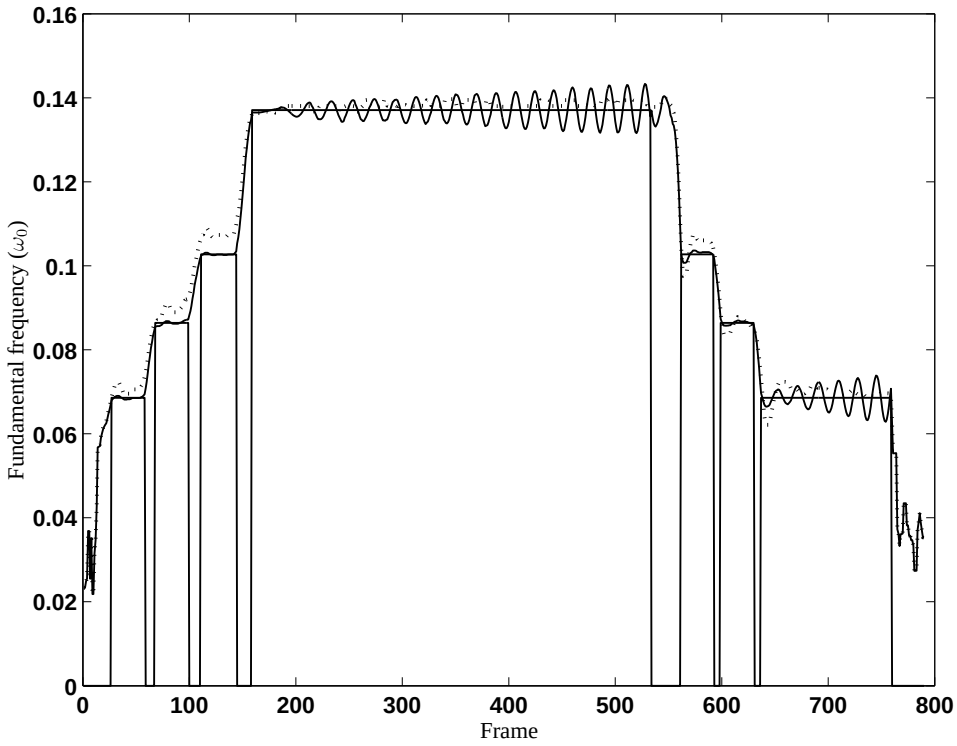


Figure 38: Prosodic modification: the original pitch contour (dotted) pitch-scaled to the correct pitch (solid bars) and vibrato inserted.

agreement within the group of students and between groups was statistically significant. Therefore, due to the inability to obtain the services of a sufficient number of experts, the following methodology was used for the listening experiments.

8.3.1 Methodology

Ten non-expert listeners were asked to take a series of AB preference tests comparing several pairs of waveforms. Prior to any processing, all samples were pitch-corrected with the ABS/OLA phasor-interpolation scheme. Each subject was asked to compare ten samples for each comparison condition. The order of the pairs as well as the elements of each pair were selected randomly for each subject. Each subject was asked to select sample "A" or "B" based on two different criteria. These criteria are: "overall musical quality" and "naturalness or freedom from artifacts."

8.3.2 Comparison Results

The results of the comparison test are shown in Table 9. In this table "Glot1" refers to the proposed glottal modification method, and "Glot2" refers to the competing method described earlier in this chapter. A detailed analysis of each of the testing conditions provides an assessment of the performance of the proposed methods.

Table 9: Results of AB comparison tests for each testing condition.

		(% preferring B)	
Condition A	Condition B	Overall Musical Quality	Naturalness
Original	AGG	73 %	38 %
Original	Glot1	56 %	32 %
Original	AGG/Glot1	62 %	42 %
Original	AGG/Glot1/Vibrato	86 %	46 %
Original	FWarp	53 %	41 %
Original	Glot2	51 %	47 %
Original	FWarp/Glot2	56 %	33 %
FWarp	AGG	69 %	63 %
Glot2	Glot1	56 %	42 %
FWarp/Glot2	AGG/Glot1	65 %	54 %
AGG/Glot1	AGG/Glot1/Vibrato	68 %	56 %

Original vs. Proposed Methods The results comparing the proposed spectral and glottal modification methods versus the original waveforms (with pitch-corrections applied) show a significant preference in overall musical quality for the modification methods. Glottal modifications exhibited less improvement (56% preferred) than the AGG spectral modifications (73% preferred). Combining the two methods did not produce an improvement greater than either of the two (62% preferred), but nonetheless improved overall quality. The most significant improvement (85% preferred), however, occurred when vibrato modifications were incorporated into the combined

spectral/glottal modifications. All of these results were determined to be statistically significant ($p < 0.05$).

The breakdown by-singer in Figure 39 shows a number of interesting patterns. It can be seen in Part (a) of the figure that singers S3 (40%) and S8 (36%) showed significantly lower scores for overall quality when AGG spectral modifications were performed. Upon review of these samples, a number of artifacts were noted during some of the sustained portions. These distortions manifested themselves as short discontinuities in the perceptual quality of the waveforms. Because they were also present in the synthesized samples using the frequency warping procedure, it was concluded that the source of the distortions was in the sinusoidal model. Further analysis showed that these errors were due to a misalignment between frames during sinusoidal synthesis. This was caused by errors in the estimate of the pitch pulse onset time for certain frames.

The by-singer results for the proposed glottal modification method are shown in Part (b) of Figure 39. These results show a consistent level of preference with the exception of singers S2 (35%), S3 (30%), and S14 (27%) whose scores were significantly lower than the remaining singers. Informal questioning after the test showed that listeners found these samples to have a “rough” quality that markedly decreased the musical quality.

Part (c) of Figure 39 shows the overall musical quality results for each singer when both the spectral and glottal modifications are performed. The scores for singers S2 (41%), S3 (25%), S8 (37%), and S14 (25%) are all more than one standard deviation ($\sigma = 20.8$) below the mean ($\mu = 62$). This clearly shows that any degradation due to the proposed spectral or glottal modifications will likewise corrupt the output when both modifications are performed.

Original vs. Competing Methods Table 9 shows that the competing methods for spectral and glottal modifications resulted in an increase in preference over the absence of modifications for overall musical quality. The improvements for these methods (frequency warping - 53%, H1H2 glottal modification - 51%, both - 56%), however, were slight.

The by-singer results given in Figure 40 show that the spectral and glottal modifications resulted in significantly lower scores for the same singers as with the proposed methods. As noted earlier, the distortions for singers S3 and S8 after spectral modifications were due to inaccuracies associated with the sinusoidal model and not necessarily the spectral modification procedures. It is also logical that the same singers (S2, S3, S14) fared worse with both glottal modification methods since the methods are similar in nature. It can be said, however, that the competing methods show a much more consistent pattern across singers than the proposed methods. This is verified by a comparison of the standard deviation values for the by-singer tests as shown in the table below.

	Standard Deviation (σ)	
	Proposed Methods	Competing Methods
Spectral Modification	16.6	7.0
Glottal Modification	14.1	9.7
Both	20.8	13.8

Proposed Methods vs. Competing Methods A third set of comparisons was implemented using human listeners in which the proposed models were tested against the competing methods. Table 9 shows a significant level of preference for the proposed models in all three comparisons. Figure 41 provides the by-singer results for the comparison.

When only spectral modeling was applied, the AGG spectral model was preferred by a statistically significant 69% of listeners ($p < 0.05$) over the frequency warping method. The by-singer results show that the AGG method was preferred by the

listeners for 15 of the 17 singers. It is curious to note that the two singers (S3, S8) who scored lower when compared to the frequency warping method were the two singers who suffered from the errors in pitch pulse onset time estimation. Although, these errors were common to both methods, it is apparent that the frequency warping modifications mitigated the distortions somewhat compared to the AGG method.

The glottal modifications showed a somewhat less significant level of preference for the proposed method (56%). Listeners showed a preference for 11 of the 17 speakers. Many listeners reported having difficulty in distinguishing between the samples in these comparisons.

The combined experiment where both spectral and glottal modifications were simultaneously performed shows a preference pattern that closely matches the results for the spectral modifications. A 65% preference for the proposed methods was determined with the by-singer breakdown showing a much higher correlation with the spectral comparison than the glottal comparison. This pattern implies a greater effect of the spectral modifications on the perceptual quality of the synthesized waveform than the glottal modifications.

Effects of Vibrato A final set of experiments was conducted to determine the performance of the vibrato model. Table 9 shows the results of two additional comparisons that were tested, (1) the unmodified waveform versus the proposed spectral, glottal, and vibrato modifications, and (2) a comparison of the proposed spectral and glottal modifications with and without vibrato modifications.

The preference results show that the vibrato modifications have a substantial effect on the perceived overall musical quality. Listener preference increased from 62% to 86% when vibrato modifications were added to spectral and glottal modifications. A direct comparison between these two conditions showed a 68% preference for the vibrato modified output. This is not a surprising result since several past studies have

demonstrated the importance of vibrato to the perception of vocal beauty [44, 65].

Naturalness Listener preference results based on naturalness and freedom from artifacts or distortions show a decrease in this measure when any of the modifications were performed and compared to the unmodified waveform. This can be seen in Table 9. In some cases, however, this preference was slight (Original vs. AGG/Glot1/Vibrato - 46%, Original vs. Glot2 - 47%). When comparing the proposed methods to the competing methods, the proposed methods fared better in naturalness for the spectral-only modifications (63%) and the spectral/glottal combination (54%). Listeners found less naturalness in the proposed glottal model when compared to the competing model by a small margin (54%).

Vibrato modifications did appear to actually increase the naturalness of the synthesized output. When vibrato was added and compared to the spectral/glottal modification combination, it was preferred 56% of the time. However, informal comments by listeners revealed that the addition of vibrato to certain voices resulted in an extremely unnatural sound.

8.3.3 Discussion

The results of the subjective comparison indicate that, taken as a whole, the pool of 10 listeners preferred the proposed methods for spectral and glottal modifications over both the unmodified singing voices and the modifications using the competing methods. This suggests that for the application of enhancing the voices of untrained singers, the spectral and glottal methods outlined in this thesis offer a viable solution.

It should be emphasized that the overall performance of the enhancement system is reliant on a number of interdependent modules. As was the case with pitch pulse onset time estimation, any errors in a single module results in distortions or artifacts. During the course of this experiment, a number of common errors have been encountered that degraded the performance of the modification system. Some errors were alleviated by

adjusting various parameters, but others persisted. A few of these issues are detailed here:

- In order to enable the AGG spectral modification method or frequency warping algorithm to be effective in performing formant modifications, it is necessary to employ an accurate formant tracking system. It was noted in Chapter 3.2.2 that the AGG model can be used to refine initial formant estimates but its performance is largely affected by the performance of the initial estimate provided by a formant tracker. Performance of the AGG spectral model improved somewhat by median filtering the formant estimates, thus smoothing noisy tracks. Excessive formant smoothing in the AGG spectral model, however, resulted in reverberant effects in the synthesized output. The frequency warping modification procedure also improved when smoothing was applied but was also found to be more robust to inaccuracies in formant tracking.

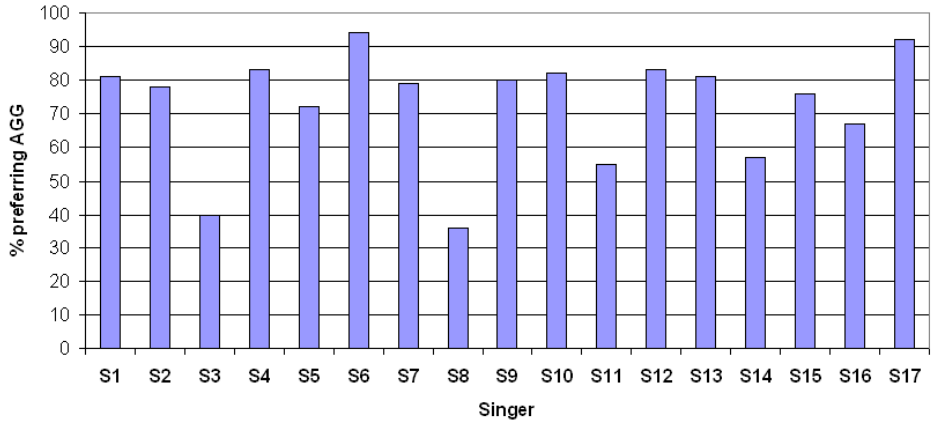
- Pitch doubling and halving errors occasionally occurred during the analysis of the samples. This impacted the spectral estimate which caused slight perceptual variations in the modified waveforms. This issue was also alleviated somewhat by median filtering but only when errors occurred in short spurts. Errors spanning several consecutive frames persisted after filtering. It could be argued, however, that in a singing voice enhancement application, the correct melody would be known beforehand and thus could be utilized to prevent pitch doubling and halving. However, it was found that some untrained singers would occasionally, albeit rarely, err in pitch by an entire octave or more.

- During the processing stage of the modification experiments, it was found that improved performance could be attained by manually adjusting the target parameters for each of the spectral methods. These adjustments often varied from singer to singer. However, these target parameters that were empirically determined for the AGG and frequency warping methods often did not coincide. Because the purpose of the experiment was to objectively compare the ability of each method to perform

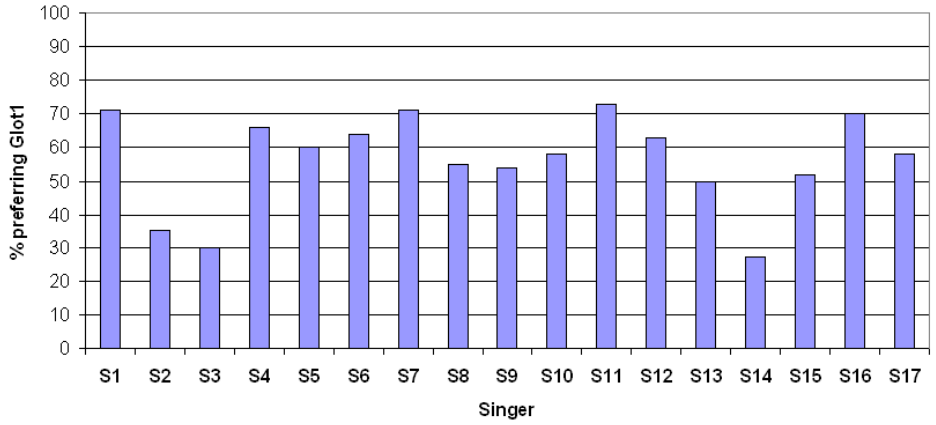
modifications based on the analysis performed in Chapter 6.3, the target parameters were not altered from the original specifications.

The results of the listening tests show that the proposed spectral and glottal modification algorithms enable a number of important characteristics of trained singers to be parameterized and implemented in an enhancement system. However, it is clear that there are still some aspects of trained singers' voices that are not encompassed by these models. Informal comparisons show that even the most highly rated enhanced samples do not match the vocal beauty of the voices of professional classical singers. These observations imply that all of the qualities of vocal beauty can not be solely described by a static model. While it is agreed upon that prosodic features contribute to the perception of beauty or level of training, models that are able to effectively capture these characteristics in singing have yet to be formulated. Although our attempts at separately enhancing voiced segments based on registration were able to capture some of the time-varying characteristics of singing, the complex nature of the modalities of the singing voice must be further investigated to truly enhance the singing voice.

(a)



(b)



(c)

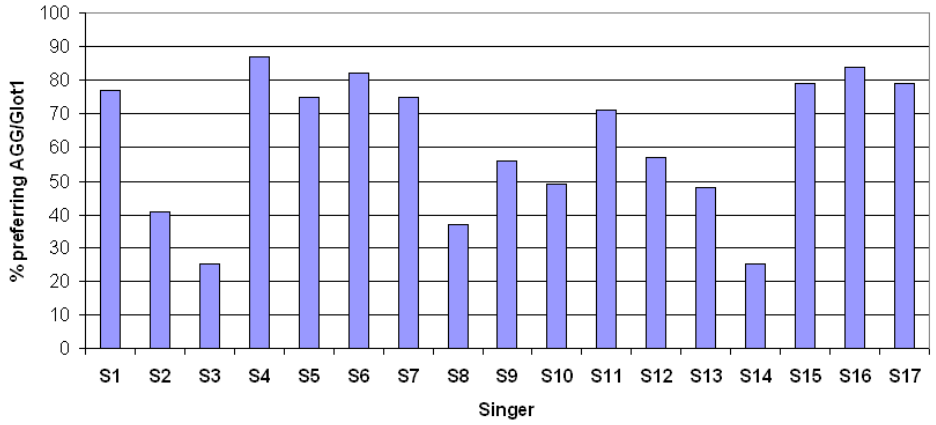
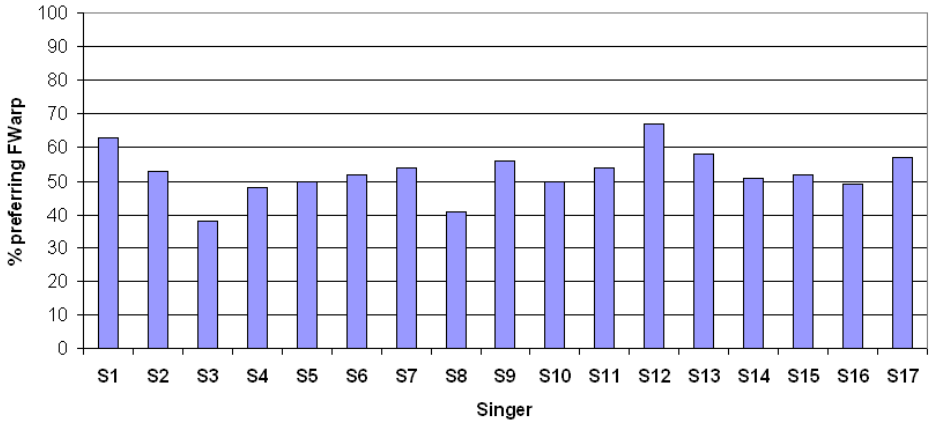
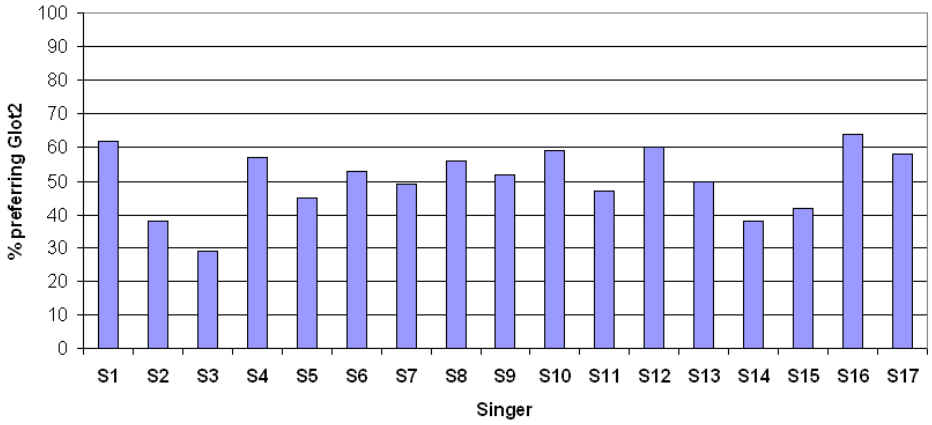


Figure 39: By-singer results for (a) unmodified vs. AGG modifications, (b) unmodified vs. glottal modifications, and (c) unmodified vs. AGG/glottal modifications.

(a)



(b)



(c)

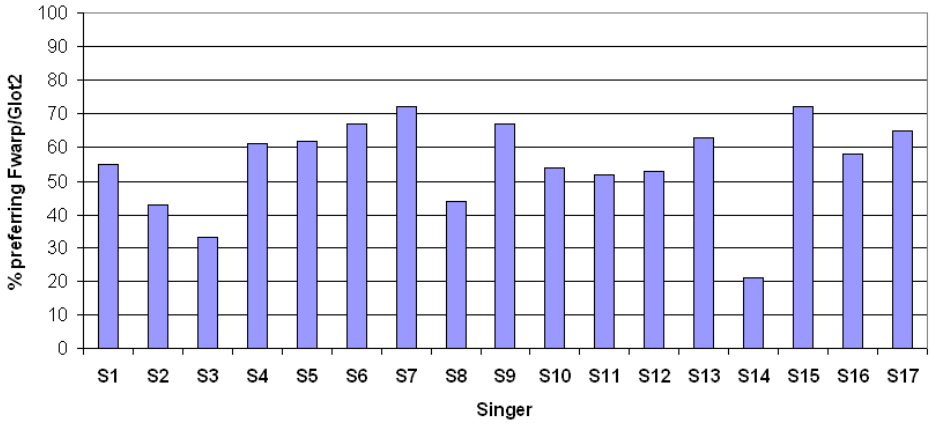
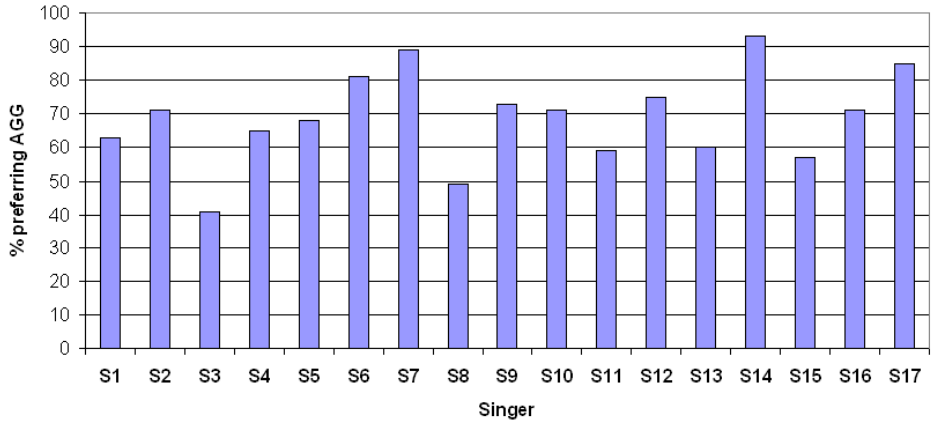
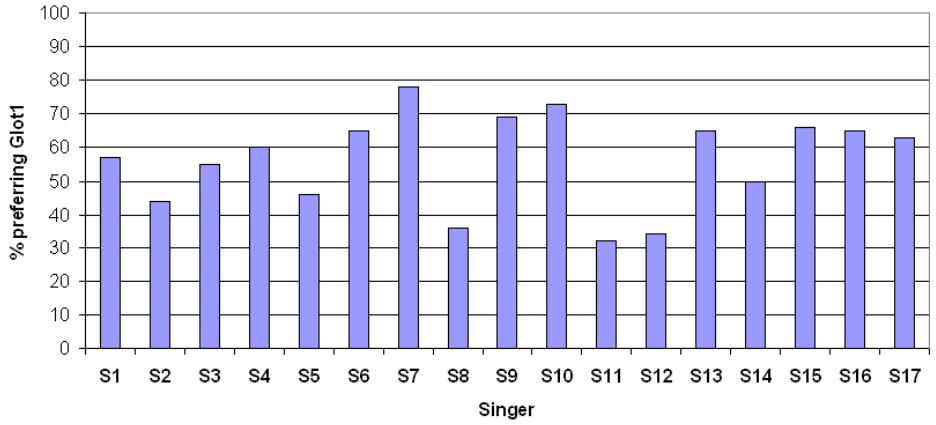


Figure 40: By-singer results for (a) unmodified vs. frequency warping modifications, (b) unmodified vs. H1H2 glottal modifications, and (c) unmodified vs. frequency warping/glottal modifications.

(a)



(b)



(c)

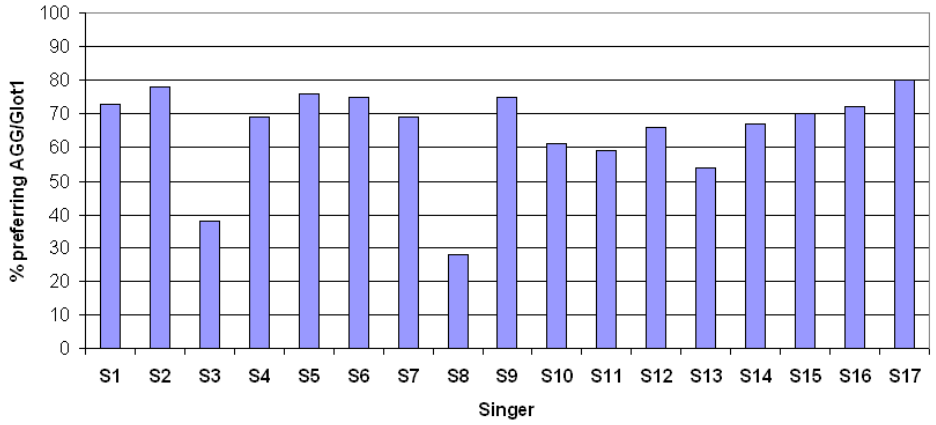


Figure 41: By-singer results for (a) frequency warping vs. AGG modifications, (b) H1H2 vs. the proposed glottal modifications, and (c) frequency warping/H1H2 vs. AGG/glottal modifications.

CHAPTER 9

CONCLUSIONS

In this research, a set of models for modeling the spectral and glottal characteristics of the singing voice has been developed and used to characterize and modify certain vocal qualities of the singing voice.

First, a spectral model was described in which the formant structure of a voice is modeled as a sum of asymmetric generalized Gaussian functions. The use of these functions is advantageous over traditional methods because of their inherent flexibility. Most notably, the ability to independently control the amplitude and bandwidth of each Gaussian function in an asymmetric fashion enables an accurate estimate of the spectral envelope as well as a wide variety of modifications.

A glottal model based on a frequency-domain derivation of time-domain glottal flow models was also developed. It was shown that important time-domain characteristics of the glottal source can be captured in the time domain by parameterizing the relative amplitudes and phases of the harmonics in a voiced signal. This discovery led to a method for accurately estimating perceptually important time-domain parameters in the frequency domain. Experimental evidence was presented in which these estimates showed a high level of correlation with estimates produced by a closed-phase inverse filtering technique using a physical microradar device to accurately measure the glottal closure instants.

An additional technique for modeling vibrato was also formulated in which vibrato could be naturally infused into a singer's voice.

An experimental study was presented in which these models were used to capture and parameterize those characteristics that differentiate singing produced by singers with no training and those with extensive classical training. It was shown that a number of spectral

and glottal patterns could be discerned using the proposed models. Additionally, differences in registration, or the modes of singing often used by trained singers, were illustrated with these models. Evidence that changes in registration involve changes in spectral as well as glottal characteristics were presented, thus supporting a largely minority opinion that registration is independent of the vocal tract.

Finally, an application for enhancing the characteristics of untrained singers was developed and subjectively tested against unmodified sources and competing algorithms. It was shown that the proposed models are capable of providing an improved framework for high-quality vocal modifications.

9.1 Contributions

Contributions of the proposed work include the following:

- Development of a spectral model based on asymmetric generalized Gaussian functions for parameterizing and modifying the formant structure of the spectral envelope.
- Development of a method for estimating the parameters of the spectral model using the Expectation-Maximization algorithm.
- A theoretical analysis of the frequency-domain characteristics of two time-domain glottal flow models (LF, R++).
- Development of a parameter estimation technique for determining time-domain glottal characteristics in the frequency domain.
- Development of a model for characterizing perceptually relevant glottal characteristics based on amplitude and phase characteristics of the harmonics of the glottal source.
- Development of a model for pitch and vibrato correction based on the ABS/OLA sinusoidal model as a pitch-scaling engine.

- Implementation of the proposed spectral, glottal, and pitch modification methods for the classical enhancement of an untrained singer's voice.
- Various extensions to the existing glottal estimation method in [54]:
  - Improvement of initial estimation of the glottal closure instant using group delay methods.
  - Improvement of averaging technique for vocal tract estimation.

9.2 Future Work

Algorithm robustness Perhaps the best way in which the spectral and glottal models could be improved for all applications is by the development of more robust methods for the estimation of various parameters such as pitch pulse onset time, voicing, formant frequencies, and pitch. Each one of these is a challenge in itself and has much room for improvement. Although several ad hoc methods for dealing with isolated errors were implemented, these techniques are far from capable of dealing with more serious errors that can be common occurrences in singing.

Genre classification In this thesis an investigation was performed which identified a set of characteristics describing a classical style of singing. While we have conducted studies regarding other styles [37, 38], a wide-ranging examination of genres as well as genders could utilize the spectral and glottal models to their fullest capabilities and provide significant insight of the characteristics of singing for both the music and research communities.

Database collection One of the most challenging aspects of this research has been the inability to obtain a sufficient variety of studio-quality isolated singing voice samples. Any future comprehensive study in singing will require such a database spanning gender, genre, vocal range, and level of training. Additionally, such a database would

be invaluable for using the proposed models for some type of statistical classifier that would require a relatively large amount of training data.

Prosody models in singing synthesis Spectral and glottal characteristics encompass a large portion of a singer's characteristics, but time-varying changes or prosodic features are also significant contributors. The vibrato model proposed in this thesis is capable of capturing some of these features, but it nonetheless fairly rudimentary. In order for a complete model of the singing voice to be attained, it is necessary for a more complex model capable of capturing the dynamic features of singing to be developed.

Alternate applications Singing voice enhancement and genre classification are only a few of the many applications that can utilize the models developed in this work. The models developed in this work open up a myriad of possible applications that can be used to transform, segment, identify, compress, or even train the singing voice.

9.3 Concluding Remarks

In this work, we have attempted to answer the difficult question: What gives a singing voice its unique qualities? Although this question may never be completely answered, it has been shown that it is not feasible to simply use methods originally designed for normal speech processing. There may never be a sufficient level of agreement among the music community as to what makes a singing voice "good" or "bad," but through the models and methods presented here, a set of tools has been developed that can hopefully be used to take steps toward clarifying the picture.

REFERENCES

[1] BENADE, A. H., *Fundamentals of Musical Acoustics*, ch. 19. New York: Dover Publications, Inc., 1990.

[2] BJORKNER, E., SUNDBERG, J., CLEVELAND, T., and STONE, E., "Voice source characteristics in different registers in classically trained female musical theatre singers," *TMH-QPSR, KTH*, vol. 46, pp. 1–11, 2004.

[3] BOGERT, B., HEALY, M., and TUKEY, J., "The quefrency analysis of time series for echoes," in *Proc. Symposium on Time Series Analysis* (ROSENBLATT, M., ed.), pp. 209–243, New York: John Wiley and Sons, 1963.

[4] BURNETT, G. C., *The physiological basis of glottal electromagnetic micro power sensors (GEMS) and their use in defining an excitation function for the human vocal tract*. PhD thesis, U. C. Davis, 1999.

[5] BURNETT, G. C., HOLZRICHTER, J. F., GABLE, T. J., and NG, L. C., "The use of glottal electromagnetic micropower sensors (GEMS) in determining a voiced excitation function," in *138th Meeting of the Acoustical Society of America*, (Columbus, Ohio), ASA, 1999.

[6] CALLAGHAN, J., *Singing and Voice Science*. San Diego, CA: Singular Publishing Group, 2000.

[7] COOK, P., "Pitch, periodicity, and noise in the voice," in *Music, Cognition, and Computerized Sound*, pp. 195–208, M.I.T. Press, 1999.

[8] CUMMINGS, K. E., *Analysis, Synthesis, and Recognition of Stressed Speech*. PhD thesis, Georgia Institute of Technology, 1992.

[9] CUMMINGS, K. E. and CLEMENTS, M. A., "Glottal models for digital speech processing: a historical review and new results," *Digital Signal Processing: A Review Journal*, vol. 5, no. 1, pp. 21–42, 1995.

[10] DELLER, J. R., PROAKIS, J. G., and HANSEN, J. H. L., *Discrete-Time Processing of Speech Signals*. Upper Saddle River, New Jersey: Prentice Hall, 1987.

[11] DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, vol. 39, pp. 1–38, 1977.

[12] DOSCHER, B. M., *The Functional Unity of the Singing Voice*. Metuchen, NJ: Scarecrow Press, 2nd ed., 1994.

[13] DOVAL, B. and D'ALESSANDRO, C., "The spectrum of glottal flow models," *LIMSI report*, 1999. 99-07.

[14] DOVAL, B., D'ALESSANDRO, C., and DIARD, B., "Spectral methods for voice source parameter estimation," in *EUROSPEECH 97*, vol. 1, (Rhodes, Greece), pp. 533–536, ISCA, 1997.

[15] EKHOLM, E., *The Effect of Guided Listening on Evaluation of Solo Vocal Performance*. PhD thesis, McGill University, 1994.

[16] EKHOLM, E., PAPAGIANNIS, G., and CHAGNON, F., "Relating objective measurements to expert evaluation of voice quality in western classical singing: critical perceptual parameters," *Journal of Voice*, vol. 12, no. 2, pp. 182–196, 1998.

[17] ESTILL, J., "Belting and classic voice quality: Some physiological differences," in *Medical Problems of Performing Artists*, pp. 37–43, Mar. 1988.

[18] FANT, G., *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.

[19] FANT, G., "The voice source in connected speech," *Speech Communication*, vol. 22, pp. 125–139, 1997.

[20] FANT, G., LILJENCRANTS, J., and LIN, Q., "A four-parameter model of glottal flow," *STL-QPSR*, vol. 85, no. 2, pp. 1–13, 1985.

[21] GALAS, T. and RODET, X., "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: application to musical sound signals," in *Proceedings of the International Computer Music Conference*, Sept. 1990.

[22] GALAS, T. and RODET, X., "Generalized discrete cepstral analysis for deconvolution of source-filter systems with discrete spectra," in *IEEE Workshop on Applications of Signal Processing of Audio and Acoustics*, (New Paltz, New York), Oct. 1991.

[23] GAUFFIN, J. and SUNDBERG, J., "Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech and Hearing Research*, vol. 32, pp. 556–565, 1989.

[24] GEORGE, E. B. and SMITH, M. J. T., "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions of Speech and Audio Processing*, vol. 5, pp. 389–406, Sept. 1997.

[25] GEORGE, E. B., *An Analysis-by-Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Processing*. PhD thesis, Georgia Institute of Technology, 1991.

[26] GEORGE, E. B. and SMITH, M. J. T., "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, June 1992.

[27] HAMON, C., MOULINES, E., and CHARPENTIER, F., "A diphone synthesis system based on time-domain prosodic modifications of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 238–241, 1989.

[28] HANSON, H. M., *Glottal characteristics of female speakers*. PhD thesis, Harvard University, 1995.

[29] HANSON, H. M., "Glottal characteristics of female speakers: Acoustic correlates," *JASA*, vol. 101, no. 1, pp. 466–481, 1997.

[30] HENRICH, N., D'ALESSANDRO, C., and DOVAL, B., "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data," in *EUROSPEECH 2001*, (Aalborg, Denmark), ISCA, 2001.

[31] HENRICH, N., ROUBEAU, B., and CASTELLENGO, M., "On the use of electroglottography for characterisation of the laryngeal mechanisms," in *Proceedings of the Stockholm Music Acoustics Conference*, (Stockholm, Sweden), pp. 455–458, 2003.

[32] HOLMBERG, E. B., HILLMAN, R. E., PERKELL, J. S., GUIOD, P. C., and GOLDMAN, S. L., "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech and Hearing Research*, vol. 38, pp. 1212–1223, 1995.

[33] HSIAO, Y. S. and CHILDERS, D. G., "A new approach to formant estimation and modification based on pole interaction," in *Proceedings of the Thirtieth Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 783–787, Nov. 1996.

[34] KLATT, D. H. and KLATT, L. C., "Analysis, synthesis and perception of voice quality variations among female and male talkers," *JASA*, vol. 87, no. 2, pp. 820–857, 1990.

[35] KRISHNAMURTHY, A. and CHILDERS, D., "Two-channel speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 730–743, Aug. 1986.

[36] LEBORGNE, W., *Defining the Belt Voice: Perceptual Judgments and Objective Measures*. PhD thesis, University of Cincinnati, 2001.

[37] LEE, M. and SMITH, M. J. T., "Analysis and enhancement of country singing," in *145th Meeting of the Acoustical Society of America*, (Nashville, TN), ASA, 2003.

[38] LEE, M. and SMITH, M. J. T., "Spectral modelling of the singing voice using asymmetric generalized gaussian functions," in *Proceedings of the Stockholm Music Acoustics Conference*, (Stockholm, Sweden), pp. 483–486, 2003.

[39] LEE, M. and SMITH, M. J. T., "Spectral modification for digital singing voice synthesis using asymmetric generalized gaussians," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 260–263, Apr. 2003.

[40] MACON, M., *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, 1996.

[41] MACON, M. and CLEMENTS, M. A., "Speech concatenation and synthesis using and overlap-add sinusoidal model," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 361–364, May 1996.

[42] MACON, M. and CLEMENTS, M. A., "Sinusoidal modeling and modification of unvoiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 557–560, Nov. 1997.

[43] MACON, M., JENSEN-LINK, L., OLIVERIO, J., CLEMENTS, M. A., and GEORGE, E. B., "A singing voice synthesis system based on sinusoidal modeling," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 435–438, May 1997.

[44] MAHER, R. and BEAUCHAMP, J., "An investigation of vocal vibrato for synthesis," *Applied Acoustics*, vol. 30, pp. 219–245, 1990.

[45] MAKHOUL, J., "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, 1975.

[46] MARQUES, J. S. and ALMEIDA, L. B., "Sinusoidal modeling of speech: Representation of unvoiced sounds with narrow-band basis functions," in *Proceedings of EUSIPCO-88: Fourth European Signal Processing Conference*, pp. 891–894, 1988.

[47] MCAULAY, R. J. and QUATIERI, T. F., "Magnitude-only reconstruction using a sinusoidal speech model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 27.6.1–27.6.4, Apr. 1984.

[48] MCAULAY, R. J. and QUATIERI, T. F., "Phase modeling and its application to sinusoidal transform coding," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 1713–1715, Apr. 1986.

[49] MCAULAY, R. J. and QUATIERI, T. F., "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.

[50] MCLANE, M., "Artistic vibrato and tremolo: a survey of literature," *Journal of Research in Singing*, vol. 8, no. 2, pp. 21–43, 1985.

[51] MILES, B. and HOLLIEN, H., "Whither belting?," *Journal of Voice*, vol. 4, no. 1, pp. 54–70, 1990.

[52] MILLER, D. G., *Registers in singing: empirical and systematic studies in the theory of the singing voice*. PhD thesis, University of Groningen, the Netherlands, 2000.

[53] MIZUNO, H., ABE, M., and HIROKAWA, T., "Waveform-based speech synthesis approach with a formant frequency modification," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 195–198, 1993.

[54] MOORE, E. and CLEMENTS, M. A., "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information," in *ICASSP 2004*, vol. 1, (Montreal, Canada), pp. 533–536, IEEE, 2004.

[55] MORRIS, R. W. and CLEMENTS, M. A., "Modification of formants in the line spectrum domain," *IEEE Signal Processing Letters*, vol. 24, pp. 515–520, 2002.

[56] MOULINES, E. and CHARPENTIER, F., "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec. 1990.

[57] NECIOGLU, B., *Objectively measured descriptors for perceptual characterization of speakers*. PhD thesis, Georgia Institute of Technology, 1999.

[58] OLIVEIRA, L. C., "Estimation of source parameters by frequency analysis," in *EUROSPEECH 93*, (Berlin, Germany), pp. 99–102, ISCA, 1993.

[59] OPPENHEIM, A. and SCHAFER, R. W., "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-16, pp. 221–228, 1968.

[60] PINTO, N., CHILDERS, D., and LALWANI, A., "Formant speech synthesis: improving production quality," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1870–1887, Dec. 1989.

[61] PRAME, E., "Vibrato extent and intonation in professional western lyric singing," *Journal of the Acoustical Society of America*, vol. 102, pp. 616–621, 1997.

[62] QUATIERI, T. F. and MCAULAY, R. J., "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, pp. 1449–1464, Dec. 1986.

[63] QUATIERI, T. F. and MCAULAY, R. J., "Phase coherence in speech reconstruction for enhancement and coding applications," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 207–210, Apr. 1989.

[64] QUATIERI, T. F. and MCAULAY, R. J., "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, Mar. 1992.

[65] ROBISON, C., BOUNOUS, B., and BAILEY, R., "Vocal beauty: A study proposing its acoustical definition and relevant causes in classical baritones and female belt singers," *The NATS Journal*, pp. 19–30, Sept. 1994.

[66] ROSENBERG, A. E., "Effect of glottal pulse shape on the quality of natural vowels," *JASA*, vol. 49, pp. 583–590, 1971.

[67] ROTHENBERG, M., "A new inverse-filtering technique for deriving the glottal airflow waveform during voicing," *JASA*, vol. 53, pp. 1632–1645, 1973.

[68] ROTHENBERG, M., "The voice source in singing," in *Research aspects of singing*, pp. 15–31, Stockholm: Pub. no. 33 of the Royal Swedish Academy of Music, 1981.

[69] SCHAFER, R. W. and RABINER, L. R., "System for automatic formant anaylsis of voiced speech," *Journal of the Acoustical Society of America*, vol. 47, pp. 634–648, Feb. 1970.

[70] STYLIANOU, Y., LAROCHE, J., and MOULINES, E., "High quality speech modification based on a harmonic + noise model," in *Proceedings of EUROSPEECH*, pp. 451–454, Sept. 1995.

[71] SUNDBERG, J., "The acoustics of the singing voice," *Scientific American*, vol. 236, pp. 82–91, 1977.

[72] SUNDBERG, J., "Perception of singing," in *The psychology of music* (DEUTSCH, D., ed.), pp. 171–214, Academic Press, 1982.

[73] SUNDBERG, J., *The Science of the Singing Voice*. Dekalb, Illinois: Northern Illinois University Press, 1987.

[74] SUNDBERG, J., "Human singing voice," in *Encyclopedia of Acoustics* (CROCKER, M. J., ed.), pp. 1687–1695, John Wiley and Sons, Inc., 1997.

[75] SUNDBERG, J., ANDERSSON, M., and HULTQVIST, C., "Effects of subglottal pressure variation on professional baritone singers' voice sources," *JASA*, vol. 105, no. 3, pp. 1965–1971, 1999.

[76] SUNDBERG, J. and KULLBERG, A., "Voice source studies of register differences in untrained female singing," *Logopedics Phoniatrics Vocology*, vol. 24, pp. 76–83, 1999.

[77] SVEC, J. and PESAK, J., "Vocal breaks from the modal to falsetto register," *Folia Phoniatrica et Logopaedica*, vol. 46, pp. 97–103, 1994.

[78] SVEC, J., SCHUTTE, H. K., and MILLER, D. G., "A subharmonic vibratory pattern in normal vocal folds," *Journal of Speech and Hearing Research*, vol. 39, no. 1, pp. 135–143, 1996.

[79] SVEC, J., SCHUTTE, H. K., and MILLER, D. G., "On pitch jumps between chest and falsetto registers in voice," *Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1523–1531, 1999.

[80] SYRDAL, A., STYLIANOU, Y., GARRISON, L., CONKIE, A., and SCHROETER, J., "TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis," in *Processdings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 273–276, May 1998.

[81] TAKAGI, T. and KUWABARA, H., "Contributions of pitch, formant frequency and bandwidth to the perception of voice-personality," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 889–892, 1986.

[82] TITZE, I., "Voice quality: Part II," *The NATS Journal*, Sept. 1992.

[83] TURAJLIC, E., RENTZOS, D., VASEGHI, S., and HO, C. H., "Evaluation of methods for parametric formant transformation in voice conversion," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 724–727, 2003.

[84] VAN DEN BERG, J. and VENNARD, W., "Toward an objective vocabulary," *NATS Bulletin*, vol. 15, pp. 10–15, 1959.

[85] VELDHUIS, R., "A computationally efficient alternative for the lijencrants-fant model and its perceptual evaluation," *JASA*, vol. 103, no. 1, pp. 566–571, 1998.

[86] VENNARD, W. and HIRANO, M., "The physiological basis for vocal registers," in *Contributions of Voice Research to Singing* (LARGE, J., ed.), pp. 45–58, College-Hill Press, 1973.

[87] `http://www.vocaloid.com`.

[88] WAPNICK, J. and EKHOLM, E., "Expert consensus in solo voice performance evaluation," *Journal of Voice*, vol. 11, no. 4, pp. 429–436, 1997.

VITA

Matthew E. Lee was born on February 5, 1975 in Baraboo, Wisconsin, and was raised in Tallahassee, Florida where he graduated from Lincoln High School in 1993. He attended Georgia Tech in Atlanta, GA, where he obtained a Bachelor of Electrical Engineering degree in December 1997 and a Master of Science in Electrical and Computer Engineering degree in June 1999. He received the Ph.D. Degree from Georgia Tech in 2005.

His interests include sinusoidal modeling, music analysis and synthesis, speech and audio coding, automatic speech recognition, and golf.