TOWARD A HIGH-QUALITY SINGING SYNTHESIZER WITH VOCAL TEXTURE CONTROL

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Hui-Ling Lu
July 2002

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Julius O. Smith III
(Principal Adviser)

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Robert Gray

I certify that I have read this dissertation and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Stephen P. Boyd

Approved for the University Committee on Graduate Studies:

To my family

Abstract

Spectral modeling and physical modeling have been used in the past to achieve high-quality singing synthesis. However, spectral models are known to be difficult to articulate and are therefore relatively limited in expressivity. On the other hand, it is not straightforward to adjust physical model parameters to reproduce a specific recording. In this thesis, a high-quality singing synthesizer is proposed with an associated analysis procedure to retrieve the model parameters automatically from the desired voices. Since over 90% of singing is voiced sound, the focus of this research is to improve naturalness of the vowel tone quality. In addition, an intuitive parametric model is developed to control the vocal textures of the synthetic voices ranging from “pressed,” to “normal,” to “breathy” phonation.

To trade-off between complexity of the model and the corresponding analysis procedure, a source-filter synthesis model is proposed. Based on a simplified human voice production system, the source-filter synthesis model describes human voices as the output of the vocal tract filter excited by a glottal excitation. The vocal tract is modeled as an all-pole filter which has often been used in the past to model non-nasal voiced sound. To accommodate variations in vocal textures, the glottal excitation model employs two elements: a parametric derivative-glottal-wave and modulated aspiration noise. The derivative glottal wave is given by the transformed Liljencrants-Fant (LF) model. The aspiration noise is represented as pitch-synchronous, amplitude-modulated Gaussian noise.

A major contribution of this thesis is the development of an analysis procedure that estimates the parameters of the proposed synthesis model to mimic desired voices. First, a source-filter deconvolution algorithm based on convex optimization techniques is proposed to estimate the vocal tract filter from sound recordings. Second, the inverse-filtered glottal excitation is decomposed into a smoothed derivative glottal wave and a noise residual component using Wavelet Packet Analysis, through which proper parameterizations of the glottal excitation can then be found. Finally, baritone recordings are analyzed to construct a parametric model for controlling vocal textures in synthesized singing.

Acknowledgements

I would like to thank my principal advisor, Professor Julius O. Smith III, for his guidance and support throughout the years. I am grateful to Julius for giving me the academic freedom to work on this unique dissertation topic. Without his encouragement throughout my Ph.D. study, this work could not have been made. In addition to technical advices, he also helps me a lot on my English writing, which makes the publication of this thesis possible.

I would like to thank my associate advisor, Professor Robert M. Gray, for being on my readings and orals committee. I thank him for his wonderful suggestions on my dissertation writing. I would like to thank Professor Stephen P. Boyd for being on my readings committee. One part of this work was initiated from the project of convex optimization course taught by Professor Boyd. I thank him for his valuable advices on convex optimization formulation. I am grateful to Professor Perry Cook at Princeton University for his help throughout the course of my study. In addition to being on my orals committee, he was generous to provide me the C++ version of his singing synthesis program and answer me all sorts of questions regarding singing voice synthesis. I am also grateful to Professor R. Fabian Pease for serving as the chairman of my orals committee.

I appreciate the friendship and help that I have gotten from people at the Center for Computer Research in Music and Acoustics (CCRMA). In particular, I like to thank Tim Stilson, Stefan Bilbao, Craig Stuart Sapp, Harvey Thornburg, Yoon Kim, Fernando Pablo Lopez-Lezcano, Andrew Einaudi, Polly Huang, Ching-Wen Chao, Stefania Serafin, Gary Scavone, Yi-Wen Liu, Maureen Chowning and Christian Herbst. Furthermore, I would like to express my gratitude to my friend, Jeff Hsieh, in the department of Electrical Engineering for helping me edit my dissertation.

Numerous people had given me technical helps while this research was ongoing. I would like to thank Niels Lous at Institute for Perception Research for providing the simulation software of his articulatory voice synthesizer. I would like to express my gratitude to the speech and motor control research group at NTT. They had enlightened me on articulatory motion measurements of human voice production system. I am also grateful to Johan Liljencrants, Jorge C. Lucero, Xavier Pelorson, Mico Hirschberg, Paavo Ilmari Alku, and Johan Sundberg.

I am greatly indebted to my parents, my sister, my brother and my son. Their unconditional love gives me strength to pursue my career. Without my parents' support, I would not even have a chance to start my study at Stanford. I am blessed that Byran joined our family during the final year of my Ph.D. study. His smile pushed me to go on. Finally, I am grateful to my husband, Jen-Chung Chang, for his companionship, love and faith.

Preface

The first impression that people have to the mentioning of singing synthesis is usually score-to-singing synthesis. Score-to-singing synthesis converts a score along with lyrics and specified singing style to singing voices. Similar to text-to-speech (TTS) synthesis [23] [40] [73], score-to-singing synthesis can be decomposed into two subsystems: the rule system and the sound synthesis system. The rule system converts lyrics to phonemes* and generates from the input score and stored musical rules the pitch contour, sound level, duration, and vibratos. The sound synthesis system then generates the singing voice from these control parameters. Each sound synthesis algorithm requires a parametric database to store the information for acoustic realizations of each phoneme and each sound synthesis5 model comes with different co-articulation rules† to connect adjacent phonemes. The sound synthesis system is the focus of this research.

Currently, there are two fundamental approaches to sound synthesis [70]: physical modeling and spectral modeling.‡ Physical modeling constructs a synthesis model from the point of view of how the source generates sound. On the other hand, spectral modeling approaches the problem from the way we perceive sound. There are pros and cons to each of these two approaches.

Physical models are in general more expressive and intuitive to control. For human voice modeling, if the physical model is realistic and the vocal tract configuration is accurate enough, co-articulation is relatively easy. Simple linear interpolation between two target

*More precisely, it is converted to an "allophone" where different acoustic realizations for a specific phoneme under different contexts are considered.

†The extreme is a concatenative synthesis scheme where the co-articulation information is simply stored in a massive database.

‡An example of spectral modeling is Macon's Lyricos singing synthesizer[51] where sinusoidal modeling is used to render singing voices.

phonemes would do most of the job. However, physical models tend to be very complicated and non-linear, which makes the task of analysis/re-synthesis intractable. Human voices in particular are unlike any musical instruments in that measurements of human voice production system are often invasive and inaccurate. Spectral modeling, on the other hand, is easier to analyze and re-synthesize but does not provide expressive control and is difficult to articulate. The motivation of this thesis is to build a synthesis model having tractable analysis/re-synthesis while retaining the expressivity and intuitiveness of a physical model.

A pseudo-physical model for singing voice synthesis is proposed in this thesis. The proposed model is an approximation of the physical model and it retains the ability of analysis/re-synthesis. It is capable of generating high-quality non-nasal singing voices and it is computational affordable. In addition to pitch and volume control, the proposed model also provides flexible control of vocal textures. A major contribution of this thesis is the development of an automatic analysis procedure that estimates the parameters of the proposed synthesis model to mimic the desired voices. Furthermore, a parametric model for vocal texture control is constructed by analyzing baritone singing recordings.

Chapter 1 of this thesis first reviews human voice production system. The proposed synthesis model is illustrated in Chapter 2. The automatic analysis procedure for analysis/re-synthesis is described in Chapter 3. Chapter 4 presents the proposed parametric model for vocal texture control. Finally, Chapter 5 gives conclusions and future directions.

Contents

List of Tables

List of Figures

Chapter 1

The human voice production system

1.1 Introduction

The physiology of the human voice production system [30] is first presented to provide background knowledge for the introduction of a simplified source-filter theory. Figure 1.1 shows a cross-section of the human voice production system. The human voice production system consists of three basic sub-systems: the subglottal area, the larynx and the supraglottal area. The subglottal area, the section below the larynx, is comprised of diaphragm, lungs and trachea. The supraglottal area, the section above the larynx, is comprised of larynx tube, pharyngeal cavity, oral cavity and nasal cavity. The supraglottal area is often called the vocal tract. By controlling the articulators, e.g., the jaw, tongue, lips and velum, one can change the physical configuration of the vocal tract.

According to classical acoustic theory, exciting the vocal tract with pressure from the lungs produces the human voice. The excitation is either periodic pulses as in voiced speech, turbulence flow as in unvoiced speech, or a mix of these two. Figure 1.2 describes the functional components of the human voice production system and gives a brief illustration of voiced sound generation. The lung pressure provides the energy for the whole system. Once the pressure below the vocal folds is large enough and the vocal folds are not too

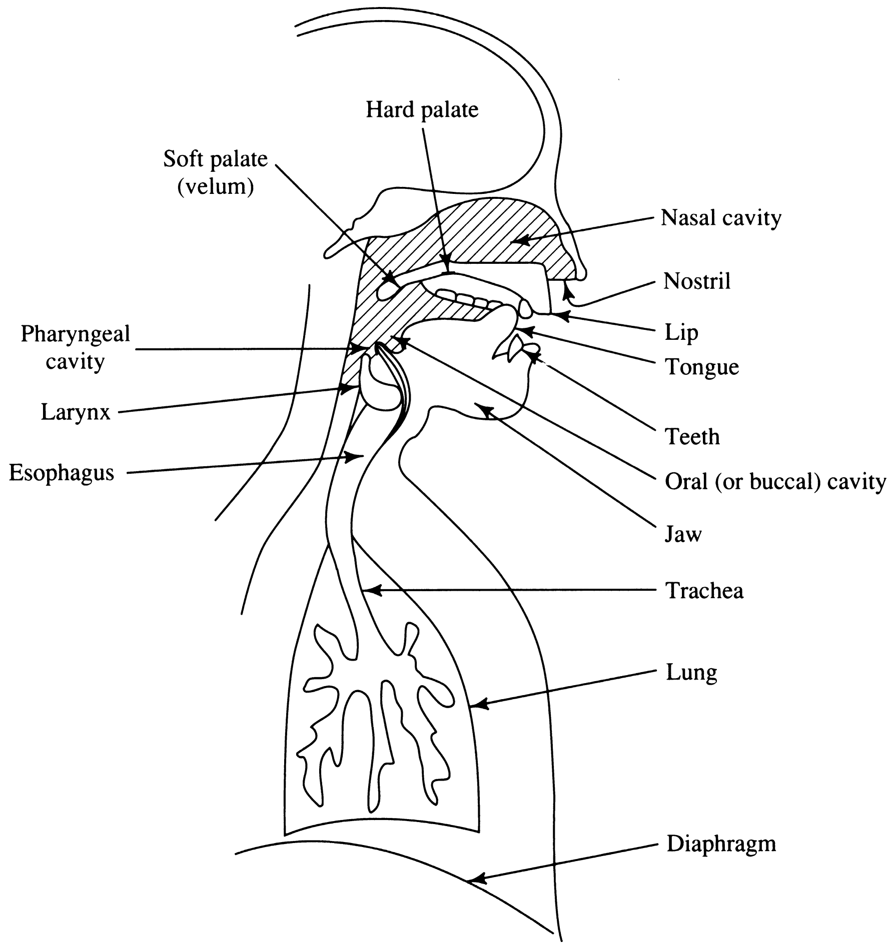

Figure 1.1: A cross-section of the human voice production system [30].

far apart, the vocal folds will start to oscillate. The periodic glottal flow is then generated and excites the vocal tract resonator. The vocal tract is a spectrum-shaping resonator. It consists of two paths. One is the nasal cavity and the other is the oral cavity. The openness of the velum controls the degree of coupling to the nasal tract. If the velum closes, only oral sound output is produced. For western bel canto singing, a style of operatic singing, the velum is often closed and the output sound is non-nasaled.

The quasi-periodic excitation pulses (as in voiced sound) result from self-oscillation of the vocal folds driven by the pressure from the lungs. The glottis is a small slit between the vocal folds and it regulates the air pressure from the lungs. Van der Berg in 1957 [4] gave the first theory on vocal fold oscillation, the aerodynamic-myoelastic theory. This theory

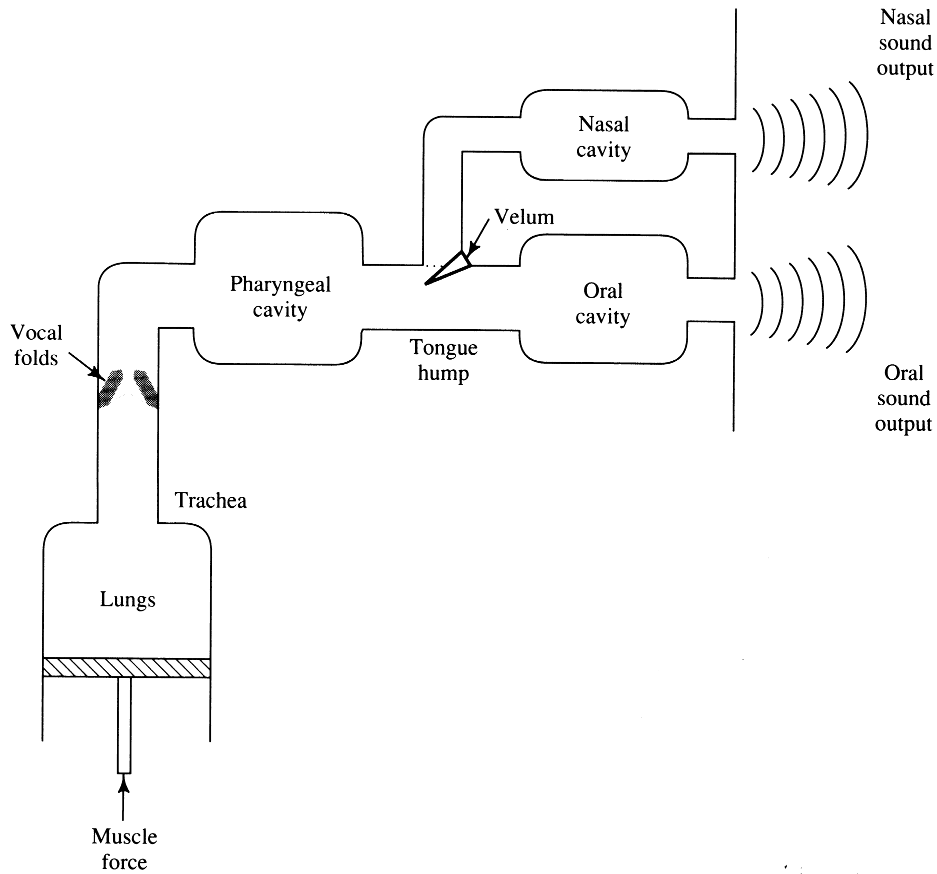

Figure 1.2: Functional components of the human voice production system [19]

offered an approximate qualitative description of the oscillation. It described the oscillation as the result of the balance between the subglottal pressure (the pressure at the inlet of the glottis), the Bernoulli pressure and the elastic restoring force of the vocal fold. The vocal folds are initiated in the prephonatory position,* in which the subglottal pressure builds up due to this obstruction. The glottis is pushed open when the subglottal pressure overcomes the elastic force that holds the vocal folds together. When the glottis is open, the airflow velocity between them increases. By the Bernoulli principle, the pressure in the glottis becomes negative, which in turn sucks the vocal folds together and causes the glottis to close. This cycle continues and produces the quasi-periodic glottal source pulses.

In general, the generation of the glottal source interacts with the vocal tract. Different

*The prephonatory position is the initial configuration of the vocal folds before the beginning of oscillation.

vocal tract configuration produces different glottal source waveform. However, since the glottis is mostly highly impeded, the interaction between the glottal source and the vocal tract is often assumed to be negligible. This is the basis of a source-filter synthesis model where the glottal source is generated irrespective of varying vocal tract configurations and the glottal impedance does not affect the vocal tract resonance structure.

The following sections are devoted to detailed discussions on oscillation patterns of the vocal folds, turbulence noise theory and vocal tract modeling.

1.2 The vocal fold oscillation

1.2.1 The physiology of the vocal folds

The larynx is the organ of phonation (the production of human voice) that houses the vocal folds. The larynx is located at the anterior part of the neck at the top of the trachea. Figure 1.3 shows the lateral view of its basic structure [49].

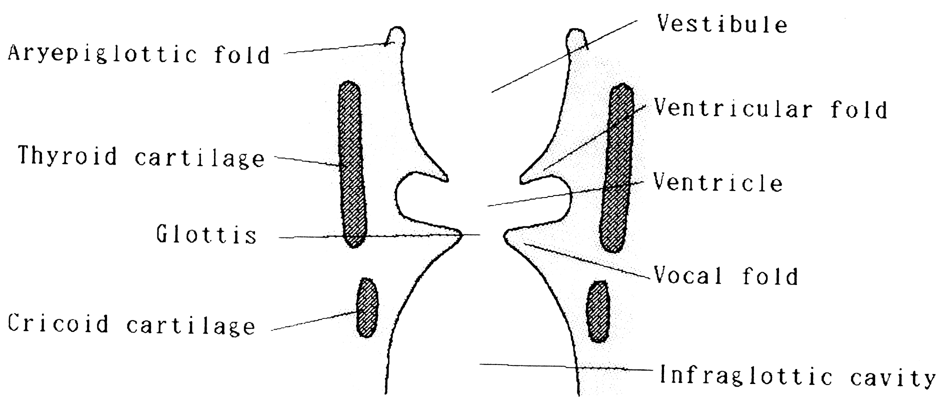

Figure 1.3: The lateral view of the larynx [49]

The vocal folds (also known as vocal cords) are a pair of elastic structures of tendon, muscles and mucous located inside the larynx. In between, they create a slit-like constriction (the glottis), which can regulate the airflow through them. Above the vocal folds, there are two folds called the ventricular or the false vocal cords. A specialized singing style, throat singing (or biphonic, overtone singing), uses both the false vocal cords and the vocal folds. For most of other phonation styles, only the vocal folds are involved. In this thesis, only the sound source generated by the oscillation of vocal folds will be considered.

Figure 1.4 shows the superior view of the larynx. By means of various muscle contractions, the vocal folds can be varied in length, thickness, and tension to form various configurations. During normal breathing, the arytenoid cartilage pushes the vocal folds sufficiently apart to allow free air passage without creating sound. At the beginning of phonation, the aryntenoid cartilage draws the vocal folds together (shown as the dotted lines in Figure 1.4). The vocal folds become tense at the same time. Once the vocal folds are in this prephonatory position, the oscillation of the vocal folds begins when there is sufficient difference between the subglottal (below the glottis) pressure and the supraglottal (above the glottis) pressure. If the vocal folds are adducted sufficiently, airflow may be hindered enough to create a turbulence noise at the glottis.

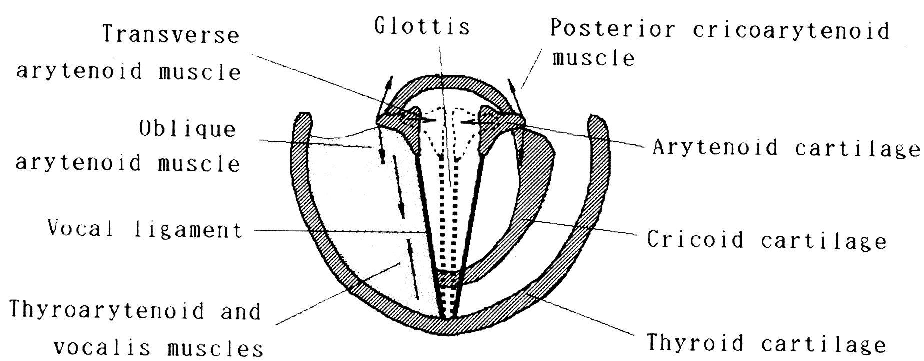

Figure 1.4: The superior view of the larynx

1.2.2 Phonation types

The pitch of the voice is mainly determined by the fundamental frequency of oscillation of the vocal folds. The fundamental frequency primarily depends on the stiffness, mass and the dimension of the vocal folds. By lengthening and tensing the vocal folds, the fundamental frequency is increased.

In addition to pitch control, the vocal folds can control the voice quality by changing its oscillation patterns. The voice quality has a wide range of definitions. Across different pitch ranges, there are three major phonation types: the pulse register, chest (normal) register, and the falsetto register.

During the pulse register phonation (or the vocal fry), the vocal folds are closed most of the time and each glottal pulse may be long enough to be perceived separately. The vocal folds are short, thick, and relaxed. It results in a very low (in the range of 3-50 Hz) and irregular fundamental frequency. The pulse register phonation often occurs at the end of an utterance as lung pressure dies out and the laryngeal muscles relax.

The chest(normal) register phonation has an approximate frequency range of 80-400 Hz for male and 200-500 Hz for female voices. Figure 1.5 shows the sagital section of the vocal folds in one cycle of the oscillation for normal phonation. The oscillation cycle can be divided into two phases, the open and the closed glottis phases. Open quotient is in general defined as the ratio of the open phase to the total oscillation period.† During the normal phonation, the open quotient is in the range of 0.4 to 0.8. As shown in Figure 1.5, the glottis is static during the closed phase. Within the open phase, the glottis has a convergent shape during the opening period and has a divergent shape during the closing period. The closing period is more rapid than the opening period due to the Bernoulli force which adducts the vocal folds.

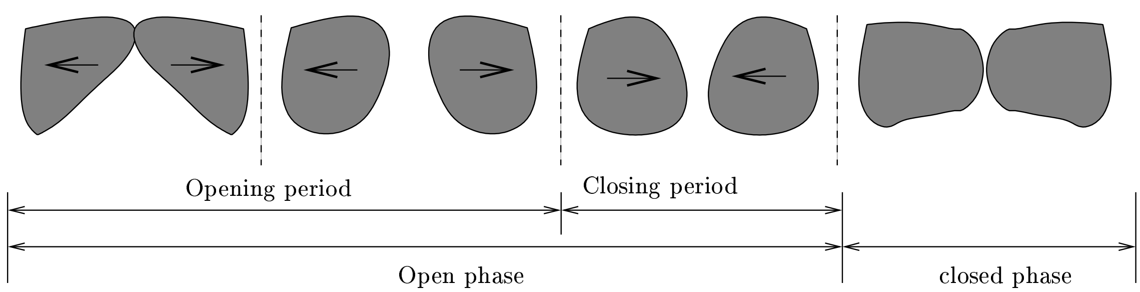

Figure 1.5: The oscillation pattern of the normal phonation

Figure 1.6 shows two cycles of glottal area for the upper and the lower edges of the vocal folds. It also shows the associated glottal flow (volume velocity) and derivative glottal volume velocity. These figures are generated by a realistic articulatory speech synthesizer‡

†The open quotient may be defined slightly differently in some parametric glottal waveform models such as the Liljencrants-Fant model in Chapter 2.

‡An articulatory synthesizer models the human voice production system directly. It synthesizes voices

[45] speaking the vowel $/i/$.[§]

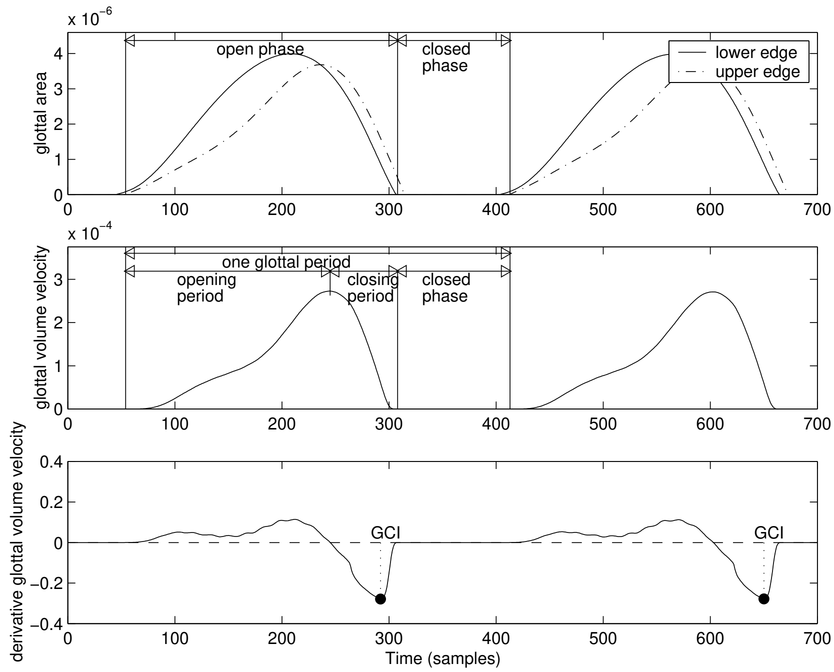

Figure 1.6: (a). glottal areas at the lower and upper edges of the vocal folds. (b). glottal volume velocity. (c). derivative glottal volume velocity.

From Figure 1.6 (a), we can observe that there is a time lag for the upper edge of the vocal fold to pick up the motion of the lower edge of the vocal fold. This is consistent with the oscillation pattern shown in Figure 1.5. Similar to the movement of the glottal area, the glottal flow builds up more gradually during the opening period and decreases more quickly during the closing period. Discontinuities in the slope of glottal volume velocity at the closing period are defined as the GCI's (glottal closure instants) in the derivative glottal volume velocity as shown in Figure 1.6 (c). These GCI's constitute major excitations for the vocal tract resonator.

The falsetto register phonation often occurs with high fundamental frequencies. During the falsetto register, the vocal folds are so thin that only the central portions of the vocal folds vibrate. The lateral lag effects of the normal register disappear. In falsetto, the vocal folds do not close completely, which adds the breathy quality to the voice.

from slowly-varying physiological parameters such as lung pressure, vocal cord tension, shape of tongue, coupling to the nasal cavity, and lip opening [65].

[§]This synthesizer will be described in Appendix A.

This thesis does not model the pulse register phonation and the falsetto register phonation. Instead, the focus is on fine tuning vocal texture control ranging from pressed, to normal, to breathy phonations [41]. That is, we would like to control the degree of breathiness of the voice. "Voice quality" has a wide range of possible meanings in the literature. To avoid confusion , "vocal texture" is used in this thesis to indicate variations among pressed, normal and breathy phonations.

The differences of these phonations come from their prephonatory positions. Figure 1.7 depicts the superior view of the prephonatory positions for these three phonation types and the corresponding typical glottal waves and derivative glottal waves. Different patterns of the glottal wave and of the derivative glottal wave may be observed with respect to varying vocal textures.

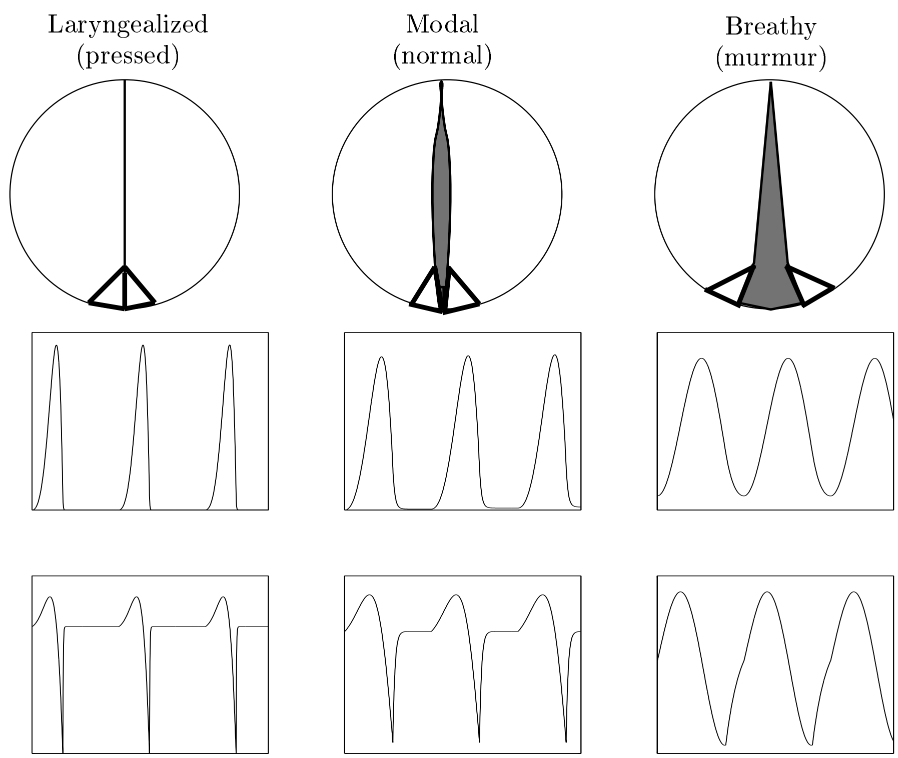

Figure 1.7: (a). Superior view of prephonatory positions. (b). Glottal volume velocity. (c). Derivative glottal volume velocity for three phonation types.

In the beginning of the pressed (laryngealized) phonation, the arytenoid cartilages close off the glottis and the vocal folds may even be compressed against each other. The glottal flow generated by this phonation mode is very narrow with a very small open quotient. The fall-off of the frequency response of the glottal flow is about -12 dB/octave. Moreover, its fundamental frequency is lower than the second harmonic.

In the prephonatory position of the breathy (or murmur) phonation, the arytenoid cartilages are parted but the vocal folds are still sufficiently approximated such that they can vibrate when enough lung pressure is applied. During the breathy phonation, the arytenoid cartilages keep a posterior portion of the glottis open while the anterior parts of the vocal folds vibrate. Therefore, there is considerable DC flow which results in generation of turbulence aspiration noise. Hence, the glottal airflow (volume velocity) has both periodic pulses and aspiration noise component. In practice, aspiration noise is the perceptually most important component to breathy sounded singing voices.

Unlike the pressed and normal phonation, the vocal folds will not close simultaneously during the closing period for the breathy phonation. The anterior parts of the vocal folds will close first followed by the posterior parts. Hence, the glottal flow is more rounded at the closing period and more sinusoidal like. Therefore, the fundamental frequency of breathy phonation is stronger than other higher harmonics.

1.2.3 Existing derivative glottal wave modeling

In speech modeling literatures, there are numerous models for the smoothed derivative glottal wave. Cummings and Clement [17] did an extensive literature survey on various models of the derivative glottal wave. The glottal models are divided into three rough categories: parametric non-interactive glottal volume velocity models, interactive parametric and mechanical glottal models, and three-dimensional physiological and numerical glottal models. These models differ in how closely they approximate the physiology of voice production and how much interaction is assumed among the glottal source, subglottal area, and supraglottal area. Parametric non-interactive glottal volume velocity models are based on source-filter theory where no interaction between the glottal source and vocal tract is assumed. Interactive parametric and mechanical glottal models are loosely based on physical observations of the glottal behavior. They include mechanical models of the vocal cords¶ and parametric

¶The most well-known mechanical model is the two-mass model proposed by Ishizaka and Flanagan [36].

volume velocity or glottal area function models that vary according to the vocal tract configuration. Three-dimensional physiological and numerical glottal models are based directly on the physiological properties of the glottis.

Although a physiologically based model is attractive for control flexibility of physical parameters, a parametric, non-coupled glottal waveform model was chosen for this study to obtain good synthesis quality with the necessary control flexibility. The Liljencrants-Fant model (LF model) [25] and the KLGLOTT88 model [41] appear to be the most widely discussed parametric non-interactive glottal waveform models. The transformed LF model, an extended LF model, is chosen in this study to simplify vocal texture control of the derivative glottal wave synthesis. The KLGLOTT88 model is also used in this study to estimate the vocal tract filter from real recordings. The simplicity of the KLGLOTT88 model allows for the estimation of vocal tract filter to be expressed as a convex optimization problem that can be solved highly efficiently to obtain the model parameters.

1.3 Turbulence noise theory

Turbulence occurs in the flow downstream from a constriction if the flow velocity at the exit is sufficiently large [5]. The Reynolds number, a dimensionless number, indicates the likelihood of turbulence,

$$R_e = \frac{2U}{\nu\sqrt{A_c\pi}} \tag{1.1}$$

where U is the volume velocity, A_c is the effective area of the constriction and ν is the kinematic viscosity of the fluid. Turbulence is likely if the Reynolds number is greater than a critical value.[‖]

Turbulence flow is the excitation source of unvoiced speech. It can exist in the vicinity of the glottis or near a constriction above the glottis. If the principle source of turbulence noise is near the glottis, the noise is called *aspiration* noise.

Depending on the generation mechanism, turbulence noise source can be characterized by three canonical source types: monopole, dipole and quadrupole. All sources of turbulence flow-induced sound are some combinations of these three source types. For example,

[‖] This critical value depends on the geometry of the constriction.

Narayanan has derived a physically motivated parametric hybrid source model which represents fricative consonants as a combination of monople and distributed dipole sources [60].

A monopole source results from a net unsteady mass injection into the fluid region. Random fluctuations in velocity constitute a monopole source and it can be modeled as a volume velocity. For example, the periodic glottal volume flow acts as a monopole source on the vocal tract [33].

Without net mass injection, a dipole source is a net fluctuating force in the medium. A dipole source can be obtained by a rapid flow of air impinging on an obstacle or a surface since the obstacle or surface will generate a fluctuating force on the medium. The dipole source is often represented as a sound pressure source. Based on experimental evidence [74] [3], the magnitude of the dipole pressure source is proportional to volume velocity U and inversely proportional to constriction area A_c.** For breathy phonation, the magnitude of the aspiration noise source is pitch synchronous due to the periodicity of glottal flow and glottal area. Moreover, the spectrum of a dipole source has a broad peak at a frequency proportional to $\frac{v}{d}$, where v is the particle velocity of the airstream and d is the characteristic dimension of the constriction. However, the shape of source spectrum depends on the geometry of the constriction.

A quadrupole is formed by placing two opposite dipoles at a small distance from each other. A quadrupole exists in unbounded medium where there is no net mass injection or net forces. For instance, turbulence in free space induces a quadrupole.

It has been shown that aspiration noise contributes to breathy vocal texture [9] [41]. Aspiration noise is primarily generated by glottal flow impinging on laryngeal surfaces downstream from the glottis. Hence, distributed dipole sources located throughout the vicinity of the false vocal folds and the epiglottis are the dominant sources for aspiration noise. A monopole source is also expected due to random velocity fluctuations in the flow through the glottis. However, the effect of the monopole source is relatively small and can be neglected.

Since the dominant source of aspiration noise is dipole, it is often modeled by a series

**Depending on different experimental conditions, the exact relationship may vary.

pressure source [29] [72] in articulatory speech synthesizers,

$$p = \begin{cases} g \cdot x \cdot (R_e^2 - R_{e,crit}^2) & \text{, if } R_e > R_{e,crit}; \\ 0 & \text{, if } R_e < R_{e,crit}. \end{cases} \tag{1.2}$$

where g is an empirical gain, x is a random number uniformly distributed between -0.5 and 0.5 and $R_{e,crit}$ is the critical Reynolds number. On the other hand, Klatt [41] uses a simple model for the aspiration noise for source-filter type synthesis models with a pitch-synchronous amplitude modulated volume velocity source. The noise residual model in this thesis is based on the same structure as Klatt's model.

1.4 Vocal tract modeling

Sound waves propagating in the vocal tract are often assumed to be planar along the vocal tract axis since soft tissues along the vocal tract prevent radial propagation. This assumption is valid for frequencies with wavelengths that are large compared to the diameter of the vocal tract. At 4k Hz, the wavelength $\lambda = \frac{c}{F} = \frac{340\,\text{m/sec}}{4000} = 8.5$ cm, which is much larger than the average diameter of the vocal tract (≈ 2 cm). Hence, this assumption is valid for frequencies less than 4k Hz.

Moreover, Sondhi [71] has shown that the human vocal tract can be modeled by a straight tube since its curvature only changes the resonance frequency by a few percent from a straightened tube model. A series of concatenated uniform cylindrical acoustic tubes as shown in Figure 1.8 is used to approximate this straightened vocal tract. The acoustic tubes have cross-sectional areas, A_i, and the same length l.

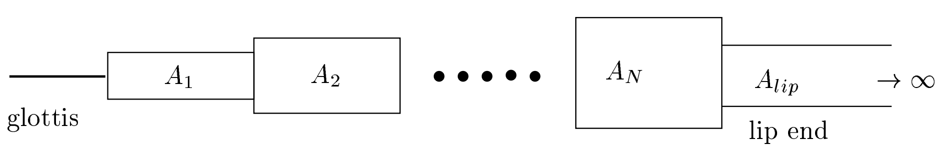

Figure 1.8: Acoustic tubes approximation for human vocal tract.

The vocal tract is assumed to be lossless and hard-walled. The model neglects heat conduction, thermal viscosity and wall vibration. The effects of these losses primarily broaden the bandwidth of the formants.[††] For example, wall vibration broadens the bandwidth of the lower formants since the walls of the vocal tract are rather massive, which is more responsive to low frequencies. Viscosity and thermal losses contribute to broadening the bandwidth of the higher formants [42] [74] [30].

Sound wave propagation follows the laws of conservation. From conservation of momentum and of mass, the wave equation for sound propagation in a lossless uniform (constant cross-sectional area) acoustic tube can be derived:

$$\frac{\partial^2 U(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 U(x,t)}{\partial t^2} \tag{1.3}$$

$$\frac{\partial^2 P(x,t)}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P(x,t)}{\partial t^2} \tag{1.4}$$

where c is the velocity of sound in air, $U(x,t)$ is the volume velocity and $P(x,t)$ is the pressure at point x and time t. The solutions to the above wave equations can be expressed as sums of traveling waves:

$$P(x,t) = P^+\left(t - \frac{x}{c}\right) + P^-\left(t + \frac{x}{c}\right) \tag{1.5}$$

$$U(x,t) = U^+\left(t - \frac{x}{c}\right) + U^-\left(t + \frac{x}{c}\right) \tag{1.6}$$

where $P^+\left(t - \frac{x}{c}\right)$ is the right-going (positive) pressure wave, $P^-\left(t + \frac{x}{c}\right)$ is the left-going (negative) pressure wave, $U^+\left(t - \frac{x}{c}\right)$ is the right-going volume velocity wave and $U^-\left(t + \frac{x}{c}\right)$ is the left-going volume velocity wave.[‡‡]

If adjacent tubes have different cross-sectional areas, scattering occurs. Partial energy will be reflected back and the remaining will get transmitted. Wave impedance (or the characteristic impedance) of a uniform lossless acoustic tube is derived here and the implementation of scattering is described in Chapter 2. Wave impedance is defined as the ratio between the pressure and the volume velocity of a wave propagating in an infinite medium, i.e. there is no reflection and therefore no negative traveling wave component in the medium. The wave impedance of a uniform lossless tube is derived from conservation

[††] Wall vibration is the most significant source of losses [19].

[‡‡] Some researchers [54] [42] [19] have defined $U(x,t) = U^+\left(t - \frac{x}{c}\right) - U^-\left(t + \frac{x}{c}\right)$ to implicitly invert the sign definition of $U^-\left(t + \frac{x}{c}\right)$.

of momentum,

$$A\frac{\partial P(x,t)}{\partial x} = -\rho\frac{\partial U(x,t)}{\partial t} \tag{1.7}$$

where A is the cross-sectional area of the tube and ρ is the density of air. Since there is no negative traveling wave component, $P(x,t) = P^+(t - \frac{x}{c})$ and $U(x,t) = U^+(t - \frac{x}{c})$. Therefore,

$$\begin{aligned} A\frac{\partial P^+(t-\frac{x}{c})}{\partial x} &= -\rho\frac{\partial U^+(t-\frac{x}{c})}{\partial t} && (1.8)\\ \Rightarrow \frac{\partial P^+(t-\frac{x}{c})}{\partial x} &= \frac{\rho c}{A}\frac{\partial U^+(t-\frac{x}{c})}{\partial x}. && (1.9)\end{aligned}$$

By integrating both sides of Equation (1.9) and neglecting the acoustically unimportant DC term, the wave impedance R is derived as

$$R \triangleq \frac{P(x,t)}{U(x,t)} = \frac{\rho c}{A}.$$

In general, in a finite-length acoustic tube with wave impedance R, the following relationships between pressure traveling waves and volume velocity traveling waves may also be derived from conservation of momentum:

$$\begin{aligned} P^+\left(t-\frac{x}{c}\right) + P^-\left(t+\frac{x}{c}\right) &= R\cdot\left[U^+\left(t-\frac{x}{c}\right) - U^-\left(t+\frac{x}{c}\right)\right] && (1.10)\\ P^+\left(t-\frac{x}{c}\right) &= R\cdot U^+\left(t-\frac{x}{c}\right) && (1.11)\\ P^-\left(t+\frac{x}{c}\right) &= -R\cdot U^-\left(t+\frac{x}{c}\right). && (1.12)\end{aligned}$$

Based on the fundamentals presented here, Chapter 2 describes the simulation of sound wave propagation within this simplified vocal tract in discrete time.

Chapter 2

The synthesis model

2.1 The source-filter model

To trade off between complexity of the model and the corresponding analysis procedure, a source-filter type synthesis model [24] is proposed. This model is based on a simplified human voice production system where the source-tract interaction is neglected. As shown in Fig. 2.1, the simplified human voice production system is decomposed into three elements: glottal source, vocal tract, and radiation impedance. The radiation impedance is approximated with a simple differencing filter. Since both the vocal tract filter and the radiation filter are linear and time-invariant (over short time frames), they commute. The glottal source and the radiation impedance are then combined and form the "derivative glottal wave." The result of these manipulations is a source-filter type synthesis model paradigm depicted in Figure 2.2, in which the concept of the source-filter modeling is clearly illustrated: the human voice is modeled as the vocal tract filter excited by a glottal excitation.

The ability to vary vocal textures of singing voices is an essential focus in this thesis and has to be incorporated into the desired glottal excitation model in addition to providing flexible pitch and volume controls. Klatt [41] has summarized important acoustic features for different vocal textures. Two acoustic parameters are considered perceptually important indicators of vocal texture: (1) the degree of aspiration noise intruding at high frequencies in vowels, and (2) the relative strength of the fundamental component of the glottal source wave. Hence, the proposed glottal excitation model consists of two parts, as depicted

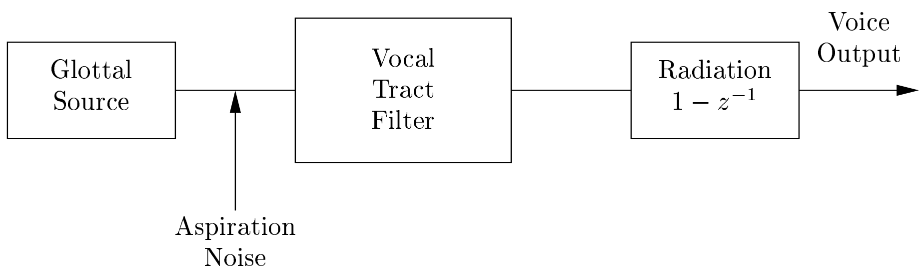

Figure 2.1: Simplified human voice production system

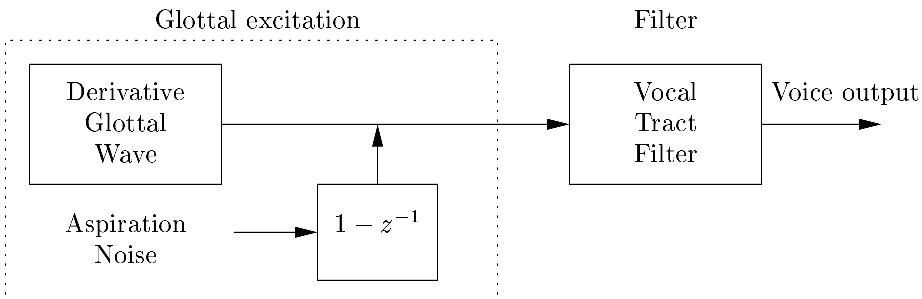

Figure 2.2: Source-filter type synthesis model paradigm

in Figure 2.2 : (1) high-pass aspiration noise; and (2) smooth, quasi-periodic, derivative glottal wave. The shape of the derivative glottal wave controls the relative strength of the fundamental component of the glottal source.

Figure 2.3 gives an overview of the proposed synthesis model. The derivative glottal wave is modeled by the transformed LF model [25] [26], which is an extension of the LF model. The high-pass aspiration noise is modeled by a noise residual model, which is a pitch-synchronous, amplitude-modulated Gaussian noise. The vocal tract filter is modeled as an all-pole filter since only non-nasal voiced sound is considered. The details of the proposed synthesis model are given in the following sections. Section 2.2.1 describes the transformed LF model. The noise residual model and the vocal tract filter implementation are described in Section 2.2.2 and Section 2.3, respectively. Section 2.4 summarizes the implementation of the overall synthesis model.

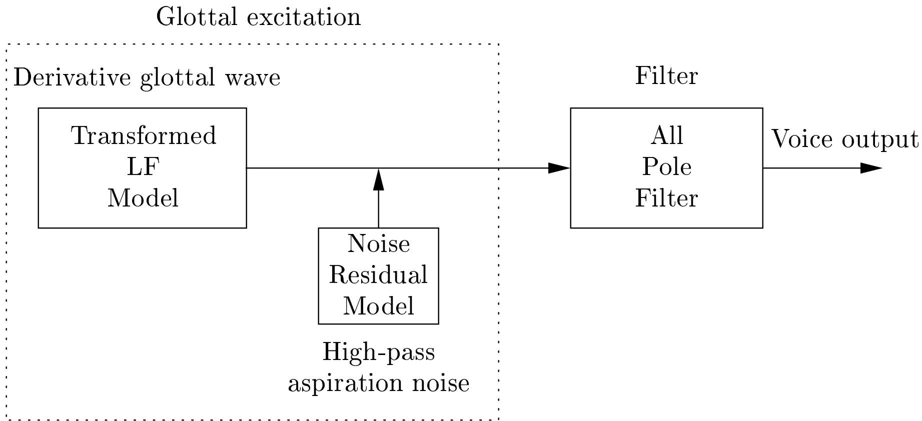

Figure 2.3: Overview of the proposed synthesis model

2.2 Glottal excitation model

2.2.1 Transformed LF Model

The transformed LF model is implemented to generate the smoothed derivative glottal wave. The reason for choosing the transformed LF model is that this model is very easy to control. Only a single parameter, R_d, is needed to control the wave-shape of the derivative glottal wave. For instance, by increasing the wave-shape control parameter R_d, the phonation mode can be changed from pressed to normal to breathy as shown in Figure 2.4.

The transformed LF model is basically an extension of the LF model. It provides a control interface to the LF model to simplify the adjustment of the wave-shape of the derivative glottal wave.

LF model

The LF model [25] is a parameterized time-domain model of one cycle of the derivative glottal wave. This time-domain model characterizes the wave-shape of the derivative glottal wave in the open and the closed phases using only four parameters. The LF model is chosen because it has been shown that the model can accommodate a wide range of natural variations. Its parameters may be estimated from inverse filtering of recorded samples and

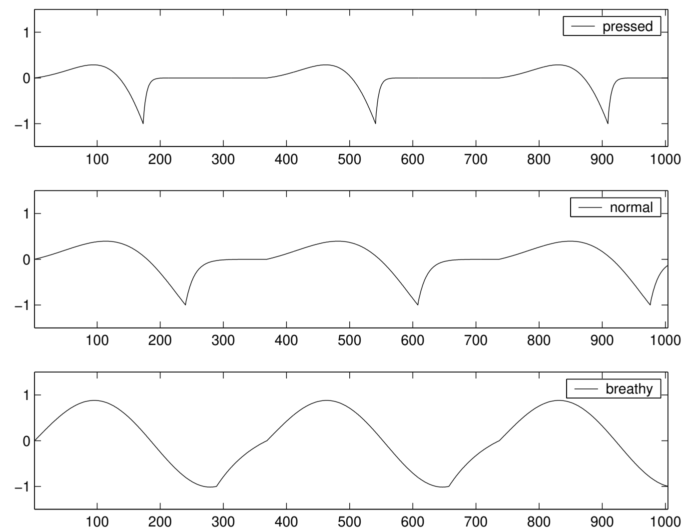

Figure 2.4: Derivative glottal waves for three phonation modes via the transformed LF model.

they vary with diverse voice qualities such as loudness, fundamental frequency and tenseness [38] [12].

The LF model consists of two segments. The first segment characterizes the derivative glottal flow from the glottal opening to the maximum negative peak. The second segment characterizes the closure of the glottis. Figure 2.5 plots two periods of the glottal wave (top) and the derivative glottal wave (bottom) generated from the LF model using a typical set of normal phonation parameters.

The LF model may be parameterized as follows:

$$g(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g t) & , 0 \leq t \leq T_e; \\ -\frac{E_e}{\epsilon T_a} \left[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)} \right] & , T_e \leq t \leq T_c \leq T_0. \end{cases} \tag{2.1}$$

Along with the magnitude of the glottal closure excitation, E_e, the modeled waveform may be specified by two independent sets of parameters: the direct synthesis parameters $(E_0, \alpha, \omega_g, \epsilon)$, and the timing parameters (T_p, T_e, T_a, T_c). The parameter T_p denotes the instant of the maximum glottal flow. The parameter T_0 is the fundamental period. The parameter T_c denotes the ending of the return phase. The parameter T_a is the effective duration of the return phase. As shown in Figure 2.5, T_a is the time constant of $e^{-\epsilon t}$. It

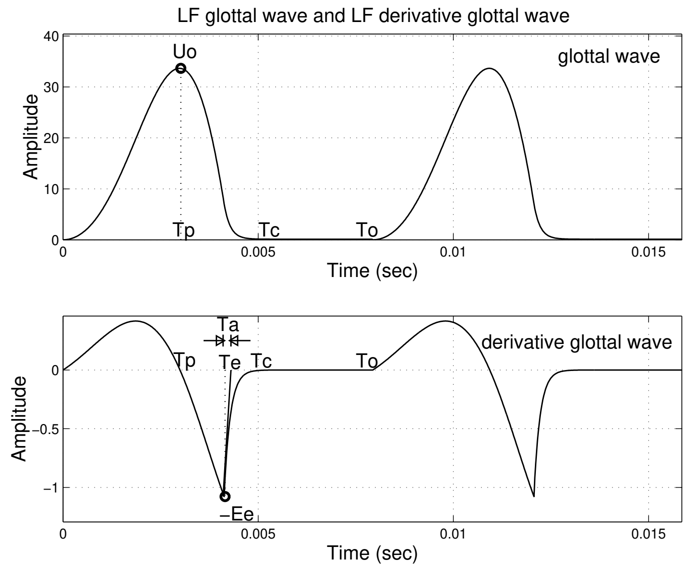

Figure 2.5: The glottal wave and the derivative glottal wave generated from the LF model.

is the time interval between T_e and the point where the tangent to the second segment of the derivative glottal wave hits the time axis. The exponent ϵ in Equation (2.1) is related to T_a and can be uniquely determined from T_a. Since the timing parameters can be easily identified from the estimated derivative glottal wave, one usually obtains the timing parameters first and then derives the direct synthesis parameters from the timing parameters with the following constraints:

$$\int_0^{T_0} g(t)\,dt = 0 \tag{2.2}$$

$$\omega_g = \frac{\pi}{T_p} \tag{2.3}$$

$$\epsilon T_a = 1 - e^{-\epsilon(T_c - T_e)} \tag{2.4}$$

$$E_0 = -\frac{E_e}{e^{\alpha T_e}\sin(\omega_g T_e)}. \tag{2.5}$$

Further details regarding the implementation of the LF model are given in Lin [43].

The return phase, T_a, is a major constituent of the LF model. The effective duration of the return phase is perceptually the most important parameter of the LF model. This parameter has been shown to be inversely proportional to the frequency, $F_a = \frac{1}{2\pi T_a}$, at which the spectrum of the derivative glottal wave attains an extra -6 dB/oct slope. Hence, increasing T_a will lower the cut-off frequency F_a of an equivalent low-pass filter. Therefore, T_a determines the spectral tilt of the glottal source. For instance, the closure of each glottis cycle is less abrupt for the breathy phonation. This gradual closure of the glottis results in a larger T_a and introduces fewer higher harmonics thereby giving a greater spectral tilt.

In addition to the direct synthesis parameters and the timing parameters, one may also describe the model with a set of normalized timing parameters (R_a, R_g, R_k), defined as follows:

$$R_a = T_a/T_0 \tag{2.6}$$

$$R_g = T_0/(2 \cdot T_p) \tag{2.7}$$

$$R_k = (T_e - T_p)/T_p. \tag{2.8}$$

In the transformed LF model [26], a new wave-shape parameter, R_d, is introduced. This parameter is proportional to (U_0/E_e), where U_0 is the peak value of the glottal flow, exclusive of a superimposed constant leakage flow, and E_e is the excitation amplitude at the glottal closure instants. R_d may be estimated from the measured set of LF parameters [27] as

$$R_d = (1/0.11)(0.5 + 1.2 \cdot R_k)(R_k/4R_g + R_a). \tag{2.9}$$

The accuracy of this formula is within 0.5 dB for $R_d < 1.4$. This formula tends to over-estimate with a maximum error of 1.5 dB at $R_d = 2.7$.

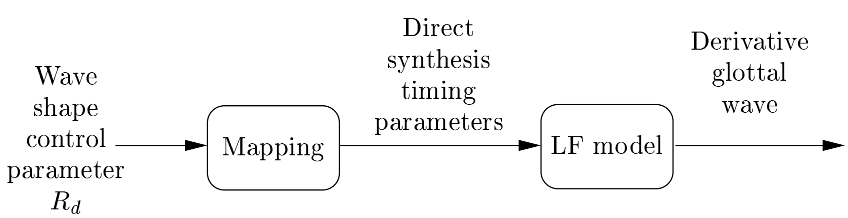

Figure 2.6: Synthesis phase: conversion from the transformed LF model to the LF model.

It has been shown that R_d is one of the most effective parameters for quantifying the quality of the voice source with a single numerical value. Another advantage of using R_d is that R_a, R_g and R_k may be predicted from it. The predicted values are denoted R_{ap}, R_{gp} and R_{kp}, respectively.* The following prediction equations are derived by Fant [26] through

*These notations follow the original paper of the transformed LF model [26].

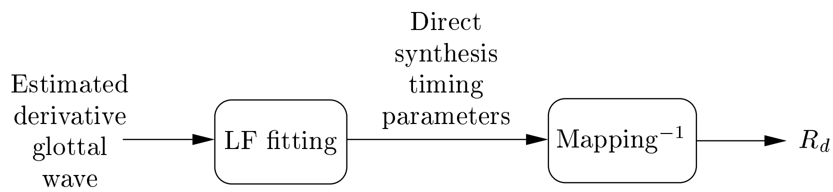

Figure 2.7: Analysis phase: conversion from the LF model to the transformed LF model.

linear regression:

$$R_{ap} = (-1 + 4.8R_d)/100 \tag{2.10}$$

$$R_{kp} = (22.4 + 11.8R_d)/100. \tag{2.11}$$

R_{gp} is then obtained by substituting Equation (2.10) and (2.11) into Equation (2.9). The deviation of the predicted values from the desired values are characterized in the following deviation constants:

$$K_a = R_a/R_{ap}\ ,\ K_g = R_g/R_{gp}\ ,\ K_k = R_k/R_{kp}. \tag{2.12}$$

In Chapter 3, the analysis results of the derivative glottal wave for real recordings are summarized into a single term, R_d. In the synthesis stage, R_a, R_g, and R_k are first predicted from R_d. The timing parameters are then obtained from Equations (2.6), (2.7) and (2.8). Finally, direct synthesis parameters may be retrieved.

In addition to R_d, the "open quotient" parameter is another good indicator of different phonations. The open quotient parameter, OQ, is defined as Te/T_0.[†] Breathy voices tend to have a larger open quotient. The open quotient is related to the relative amplitude of the voice fundamental, H_1, and the second harmonic, H_2. Through regression analysis of data generated by the LF model for different phonation types, Fant [27] found the following relationships:

$$H_1 - H_2 = -6 + 0.27 \cdot \exp(5.5OQ) \tag{2.13}$$

$$H_1 - H_2 = -7.6 + 11.1R_d. \tag{2.14}$$

These relations provide good approximations of R_d's up to 2.7.

[†]This definition is slightly different from the general definition given in Chapter 1.

Figure 2.6 and 2.7 illustrate the relationship between the transformed LF model and the LF model. During the synthesis phase, the user specifies the wave-shape control parameter R_d. The transformed LF model would then map R_d into corresponding timing parameters for the LF model. During the analysis phase, the LF model is used to fit the estimated derivative glottal wave. Timing parameters are then extracted and mapped to the wave-shape control parameter R_d.

2.2.2 Noise residual model

For singing voices to sound "breathy", aspiration noise is perceptually most important [41]. Without the presence of aspiration noise, the increase of the fundamental component may induce the sensation of nasality in a high-pitch voice. Therefore, despite the fact that breathiness can be simulated to some extent by using a sophisticated glottal-source model, a more natural simulation of breathiness requires the addition of aspiration noise.

The aspiration noise (turbulence noise) is pitch synchronous with the smoothed quasi-periodic derivative glottal wave since the likelihood of turbulence flow is proportional to the flow and inversely proportional to the radius of the aperture. The likelihood of the existence of the turbulence noise has been calculated by Cook [14], who concluded that the likelihood of turbulence exists during the entire open phase and maximum sound radiation power is achieved right after the instant the vocal folds begin to close. A high power burst of noise is also likely at the glottal opening instant, corresponding to highly pressurized air rushing through a small slit. Hence, it is expected that there are two pulses of noise per cycle in some cases. One occurs when the vocal folds are closing and the other one occurs when the vocal folds are opening. Pitch-synchronous noise is also found in other musical instruments such as bowed strings [7].

Based on psychoacoustics experiments, Hermes [31] did a study on the perception of synthetic breathy vowels. The synthetic vowels in his experiments are generated based on the source-filter theory, and the derivative glottal wave modeling is simplified as a low-pass filtered pulse train. The results indicated that the noise is segregated from the speech signal and is perceived as a separate stream when stationary noise is used. Hence, adding stationary noise does not contribute to the breathy timbre of the vowel. This is solved by using pitch-synchronous amplitude modulated noise such that the noise has a temporal envelope of the same periodicity as the pulse train.

In Hermes's study, a low-pass filtered pulse train combined with a de-emphasized, high-pass filtered, pitch-synchronous, amplitude-modulated noise is used as the excitation source to the vocal tract filter. This essentially integrates the noise with the pulse train in the sense that there is a perceivable reduction in the loudness of the noise stream and a perceivable timbre change in the breathy vowel. It is found that the reduction of the noise loudness is maximal when the interval between the noise bursts and the pulses is less than 1 ms in a glottal period of 8 ms. The cut-off frequency of the high-pass filter is chosen from 1200 Hz to 2k Hz. A lower cut-off frequency results in a greater degree of breathiness. The de-emphasis is implemented by the low-pass filter $H(z) = 1/(1 - p \cdot z^{-1})$, where the pole p is set to 0.9. Childers [12] also proposed a similar noise model, but without the de-emphasis.

In this thesis, the nature of the double pulses of the aspiration noise is neglected as a first approximation. Therefore, the noise residual before spectral shaping is roughly modeled as a pitch synchronous amplitude modulated Gaussian noise with larger power around the glottal closure instants.

Figure 2.8 shows the noise synthesis model which models the high-pass aspiration noise. The noise residual model consists of two components, an additive Gaussian white noise representing the noise floor of the high-pass aspiration noise and a pitch synchronous amplitude-modulated Gaussian noise. The first block is the Gaussian noise unit generator, which generates a zero mean, unit variance, Gaussian white noise sequence. The second block is a scaled Hanning window centered on the glottal closure instants and it modulates the amplitude of the Gaussian noise. The scale of the Hanning window for each glottal period is specified in the A_n parameter sequence. L indicates the lag of the center of the Hanning window with respect to glottal closure instants. The lag is specified as a percentage relative to the length of the glottal period. The locations of the Hanning window centers are then calculated from the glottal closure instants and the desired lag. After the amplitude modulation, a spectral shaping filter is used to account for the average spectral density of the aspiration noise. The shaping filter also includes the high-pass filtering introduced by commuting the radiation filter with the vocal tract filter as shown in Figure 2.2.

The spectral shaping filter is obtained by matching an autoregressive moving average (ARMA) filter to the average frequency spectrum of the noise residuals extracted from

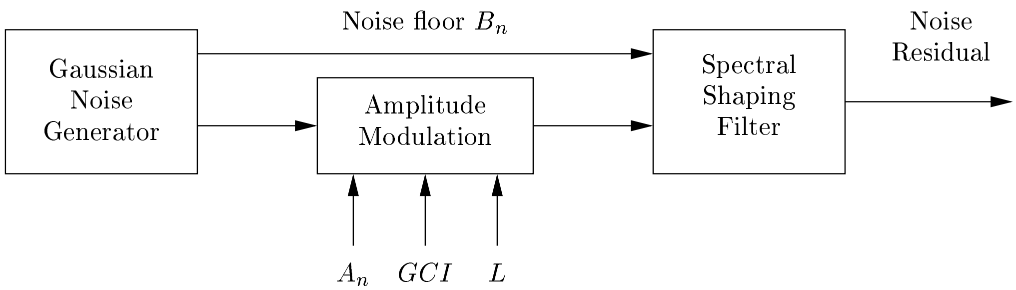

Figure 2.8: Noise residual synthesis model

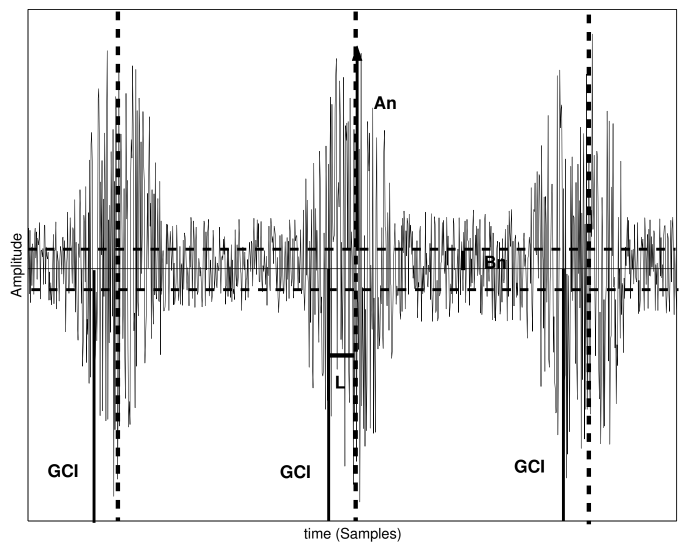

Figure 2.9: Noise residual example.

singing recordings. In the final implementation, the spectral shaping filter is actually neglected since it is not perceptually important according to informal listening tests.

Figure 2.9 shows one example of the noise residual. One can observe a noise floor component with standard deviation B_n. A pitch synchronous hanning-windowed Gaussian noise sits on top of the noise floor.

2.3 Vocal tract filter

There are several methods to simulate sound wave propagation within the vocal tract in discrete time. Flanagan [28] and Ishizaka [36] discretized differential wave equations directly with the finite difference method. This method is much more time-consuming compared to other approaches. Sondhi [72] represented the vocal tract in the frequency domain with a chain matrix. Meyer [55] simulated the vocal tract with a wave digital filter, which was obtained by discretizing the transmission line analog of an acoustic tube using the bilinear transform. Similarly, by sampling the one-dimensional linear wave solution across time and space, Liljencrants [42] and Cook [14] modeled the vocal tract as a digital waveguide filter. The digital waveguide filter[‡] is chosen for this thesis to simulate sound wave propagation within the vocal tract.

There are two advantages of a digital waveguide filter implementation. First, it is intuitive to control this model by varying the area function of acoustic tubes. Second, although the current vocal tract model is assumed to be lossless, acoustic tube losses can be easily incorporated into the digital waveguide model [42].

As discussed in Chapter 1, several assumptions have been made in order to simplify vocal tract modeling. The vocal tract is assumed to be a series of concatenated uniform lossless cylindrical acoustic tubes with specified cross-sectional areas. Plane wave propagation along the vocal tract is also assumed. Moreover, the glottis is assumed to be closed at all times so that a infinite glottal impedance may be assumed. The lip end is assumed to be an infinitely long acoustic tube; therefore, there is no reflected wave at the lip end and the lip impedance is real. Based on these assumptions, the vocal tract filter is implemented by a digital waveguide filter as shown in Figure 2.10 [19], where τ is the propagation time for sound wave to travel across one acoustic tube.

At the junction where adjacent tubes have different cross-sectional areas, scattering occurs due to an impedance mismatch. Scattering follows conservation of energy and mass, which implies that the pressure and volume velocity is continuous at the junction. Markel and Gray [54] have shown several ways to implement the scattering junction. In this work, the scattering junction is modeled by a Kelly-Lochbaum junction [39] as shown in Figure 2.11. The derivation of Kelly-Lochbaum junction is briefly given here. From the continuity of pressure and volume velocity at the junction between the m^{th} section and the $(m+1)^{th}$

[‡]In short, a digital waveguide filter is built by combining unit-length waveguide sections [69].

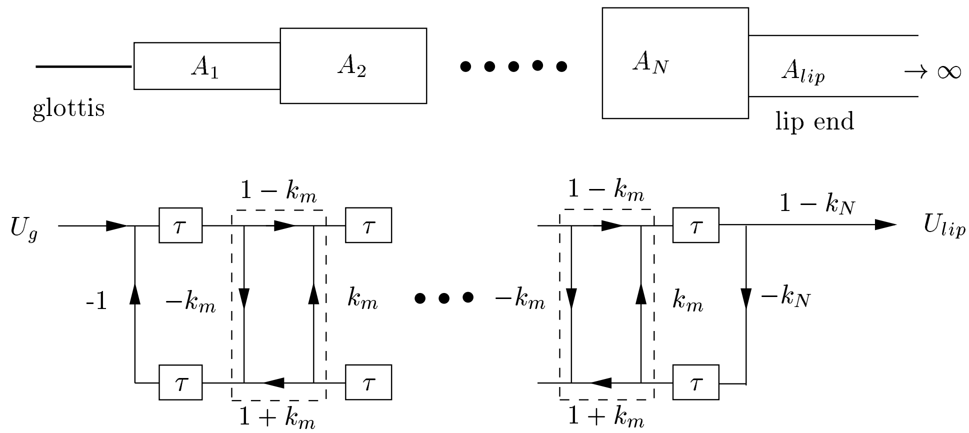

Figure 2.10: Digital waveguide implementation for vocal tract filter.

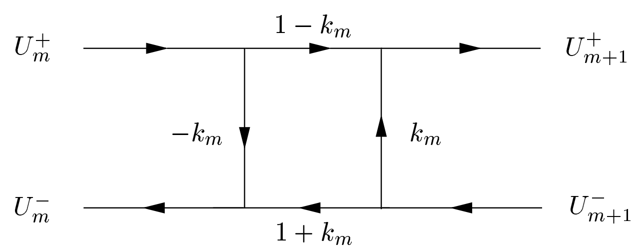

Figure 2.11: The Kelly-Lochbaum junction.

section,

$$P_m^+\left(t-\frac{x}{c}\right)+P_m^-\left(t+\frac{x}{c}\right)=P_{m+1}^+\left(t-\frac{x}{c}\right)+P_{m+1}^-\left(t+\frac{x}{c}\right) \tag{2.15}$$

$$U_m^+\left(t-\frac{x}{c}\right)+U_m^-\left(t+\frac{x}{c}\right)=U_{m+1}^+\left(t-\frac{x}{c}\right)+U_{m+1}^-\left(t+\frac{x}{c}\right). \tag{2.16}$$

Given $P^+(t-\frac{x}{c}) = R\cdot U^+(t-\frac{x}{c})$ and $P^-(t+\frac{x}{c}) = -R\cdot U^-(t+\frac{x}{c})$, one can obtain the following relationships:

$$U_m^-\left(t+\frac{x}{c}\right)=\frac{R_m-R_{m+1}}{R_m+R_{m+1}}U_m^+\left(t-\frac{x}{c}\right)+\frac{2R_{m+1}}{R_m+R_{m+1}}U_{m+1}^-\left(t+\frac{x}{c}\right) \tag{2.17}$$

$$U_{m+1}^+\left(t-\frac{x}{c}\right)=\frac{2R_m}{R_m+R_{m+1}}U_m^+\left(t-\frac{x}{c}\right)+\frac{R_{m+1}-R_m}{R_m+R_{m+1}}U_{m+1}^-\left(t+\frac{x}{c}\right). \tag{2.18}$$

By defining the scattering coefficient as the ratio between the difference and the sum of the adjacent cross-sectional areas,

$$k_m = \frac{A_m - A_{m+1}}{A_m + A_{m+1}} = \frac{R_{m+1} - R_m}{R_m + R_{m+1}}, \tag{2.19}$$

the Kelly-Lochbaum junction is shown to be

$$U_m^-\left(t + \frac{x}{c}\right) = -k_m U_m^+\left(t - \frac{x}{c}\right) + (1 + k_m)U_{m+1}^-\left(t + \frac{x}{c}\right) \tag{2.20}$$

$$U_{m+1}^+\left(t - \frac{x}{c}\right) = (1 - k_m)U_m^+\left(t - \frac{x}{c}\right) + k_m U_{m+1}^-\left(t + \frac{x}{c}\right). \tag{2.21}$$

Recall that τ is the propagation time for a sound wave to travel across one acoustic tube and N is the number of acoustic tubes excluding the glottis end and the lip end. If the sampling period is chosen as 2τ, the transfer function of the digital waveguide filter can be shown to be an Nth order all-pole filter [19].[§] Moreover, N can be determined from the following relationship:

$$\tau = \frac{T}{2} = \frac{1}{2f_s} = \frac{L}{c \cdot N}$$

where T is the sampling period, f_s is the sampling rate, c is the velocity of sound in air and L is the vocal tract length. Since the average vocal tract length for male is about 17 cm, $N = f_s \cdot \frac{0.17\,\mathrm{m}}{340\,\mathrm{m/sec}} \cdot 2 = \frac{f_s}{1000}$. This also explains the empirical rule of thumb for the LPC order selection of $(2 + f_s/1000)$ [54].

Since the energy is totally reflected back at the glottis end and the lip impedance is real, scattering coefficients in this simplified vocal tract model are equivalent to negative reflection (or PARCOR) coefficients defined in LPC analysis. One can obtain reflection coefficients from auto-regressive coefficients via the Levinson-Durbin recursion and vice versa via inverse Levinson-Durbin recursion [35]. In the analysis step, the auto-regressive coefficients are first estimated from real recordings. They are then converted into scattering coefficients for synthesis.

[§]A series of concatenated uniform lossless cylindrical acoustic tubes is an all-pole filter whenever both the glottis impedance and the lip impedance are real.

2.4 Overall synthesis model implementation

Figure 2.12 summarizes the overall implementation of the proposed synthesis model. So far, the transformed LF model, the noise residual model and the digital waveguide vocal tract filter have been introduced. Chapter 4 will describe the vocal texture control model. For synthesis, the user only needs to specify the degree of breathiness, glottal excitation strength,[¶] fundamental frequency contour, and the phoneme. The transformed LF model outputs the derivative glottal wave. The noise residual model generates the high-pass aspiration noise. The sum of these two components forms the glottal excitation. After passing through the vocal tract resonator, the output is a voice having the degree of breathiness specified by the user.

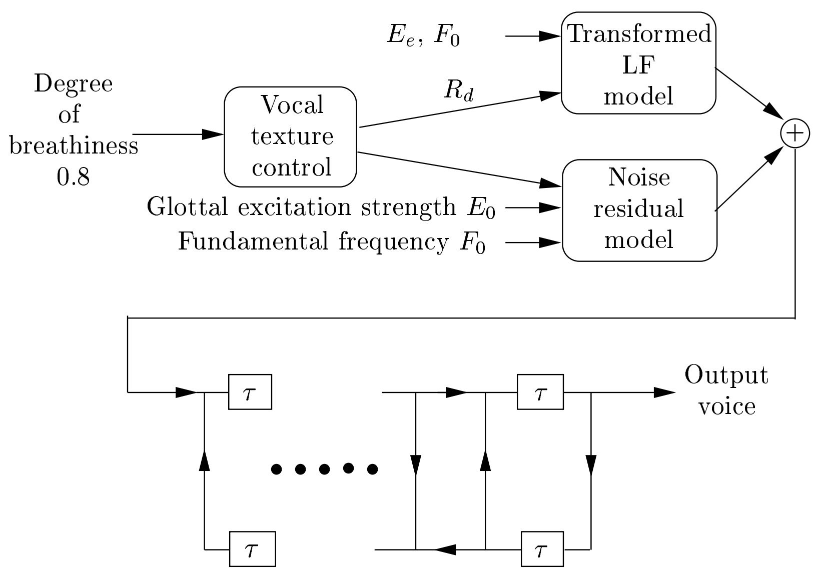

Figure 2.12: Implementation of the proposed synthesis model.

[¶]The glottal excitation strength is positively correlated to output volume of the synthesized voice.

Chapter 3

Analysis algorithms

In order to mimic the original recording, an analysis procedure was developed to estimate, from desired recordings, parameters of the synthesis model. The analysis procedure is illustrated in Figure 3.1.

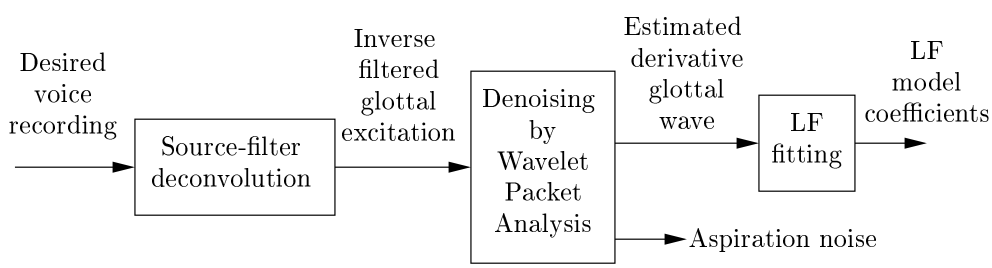

Figure 3.1: A analysis procedure

The first step is to obtain the glottal excitation from sound recordings. A source-filter deconvolution algorithm based on convex optimization is proposed to estimate the glottal excitation. The deconvolution algorithm is described in Section 3.1. The second step is to decompose the inverse filtered glottal excitation into a smoothed derivative glottal wave and a noise residual component. Several wavelet thresholding methods have been studied and compared. The wavelet packet analysis with soft thresholding is found to be the most suitable method for separating the derivative glottal wave and the noise residual. This is described in detail in Section 3.2. After denoising, the smoothed derivative glottal wave is fitted to the LF model via constrained nonlinear optimization, described in Section 3.3. Furthermore, the statistics of the noise residual is measured in order to match the noise residual model proposed in Chapter 2. From the statistics of these noise residual parameters, a parametric model was developed to control the degree of breathiness and this is demonstrated in Chapter 4.

3.1 Source-filter deconvolution

A novelty of the proposed source-filter decomposition [46] algorithm is that it provides a derivative glottal wave constraint when estimating the vocal tract filter. Hence, the resulting inverse-filtered derivative glottal wave is closer to the true glottal excitation, at least when the source-filter interaction is negligible. Although the proposed source-filter deconvolution algorithm is targeted for singing voice synthesis applications, the algorithm itself may potentially be a high-quality, yet low-bit rate speech coding method for voiced sound. A brief discussion of this algorithm is provided here.

The derivative glottal wave component in the proposed algorithm is described by the KLGLOTT88 model [41] as opposed to the transformed LF model since the mathematical description of the KLGLOTT88 model is simple enough to allow the source-filter deconvolution to be formulated as a convex optimization problem. The KLGLOTT88 model consists of the Rosenberg model [17] describing the basic wave-shape of the derivative glottal wave and a spectral tilt implemented by a low-pass filter, $\frac{1}{1-\mu z^{-1}}$ with $\mu > 0$. The synthesis model paradigm in Figure 2.2 is extended in Figure 3.2 to include the KLGLOTT88 model as the glottal excitation.

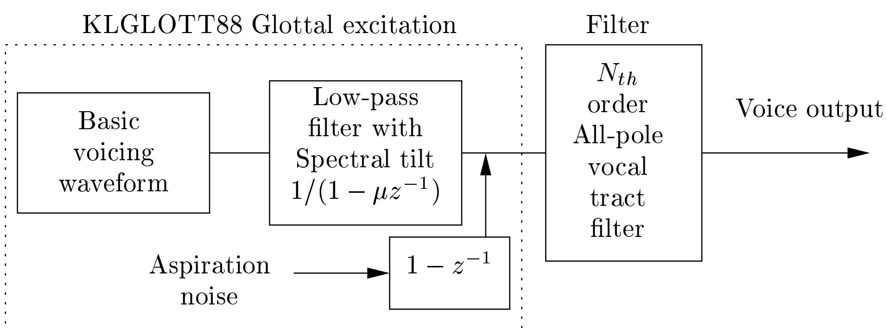

Figure 3.2: Synthesis model in analysis phase

In the KLGLOTT88 model, the basic voicing waveform describes the wave-shape of the derivative glottal wave without the return phase, i.e., assuming the glottal closure is abrupt. In other words, the derivative glottal wave is assumed to be zero in the closed phase. In the open phase, the derivative glottal wave is modeled by a simple second order polynomial,

$$g(n) = \begin{cases} 2 \cdot a \cdot n/F_s - 3 \cdot b(n/F_s)^2 & , 0 \leq n \leq T_0 \cdot OQ \cdot F_s; \\ 0 & , T_0 \cdot OQ \cdot F_s \leq t \leq T_0 \cdot F_s. \end{cases} \tag{3.1}$$

$$a = \frac{27 \cdot AV}{4 \cdot (OQ^2 \cdot T_0)} \tag{3.2}$$

$$b = \frac{27 \cdot AV}{4 \cdot (OQ^3 \cdot T_0^2)} \tag{3.3}$$

where T_0 is the fundamental period of the voice, F_s is the sampling frequency, AV is the amplitude parameter, and OQ is the "open quotient" of the glottal source (0 for always closed and 1 for always open).

Figure 3.3 gives an example of two glottal cycles of the basic voicing waveform with OQ assumed to be 0.6.

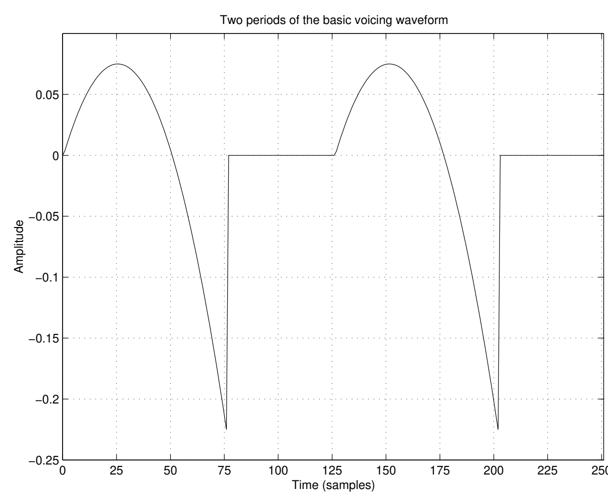

Figure 3.3: Two periods of the basic voicing waveform

Although the basic waveform always has abrupt changes at the glottal closure instants, the low-pass spectral tilt filter gives some smoothing of the return phase. The parameter μ in the low-pass spectral tilt filter adjusts the degree of the abruptness of the glottis closing. This first-order low-pass filter and the vocal tract filter may further be combined to form an order $N+1$ all-pole filter shown in Figure 3.4. The glottal spectral tilt can be associated with a single positive real root of the order $N+1$ filter, where N is even.

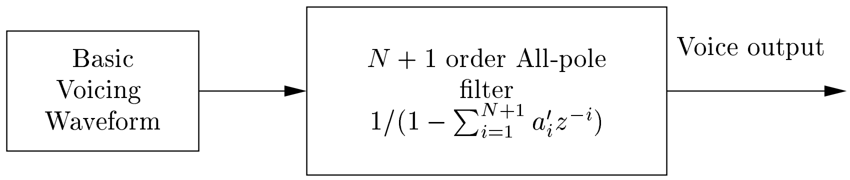

Figure 3.4: Simplified model

Due to the simplicity of the basic waveform, we can formulate the parameter estimation as a convex optimization problem by estimating the parameters of the all-pole vocal tract filter and the glottal source waveform pitch synchronously. More explicitly, if we know the fundamental period T_0 and the open quotient OQ, the task is then to estimate the filter coefficients of the $N+1$ order all-pole filter and the two shaping parameters (a and b) of the basic waveform using one period of the voice pressure data. The merits of using convex optimization methods are computational efficiency and global optimality.

Section 3.1.1 first describes how filter parameter estimation is formulated as a convex optimization problem, followed by a discussion of its implementation. The source-filter deconvolution procedure is then discussed in Section 3.1.2. Since estimation is performed pitch synchronously, it is also necessary to detect glottal closure instants. Different glottal closure instants detection methods are compared in Section 3.1.3. Finally, the performance of the source-filter deconvolution is discussed in Section 3.1.4.

3.1.1 Convex optimization formulation

A good approximation to the original speech pressure wave is achieved by minimizing the error between the estimated speech pressure wave and the true speech pressure wave. However, in our convex optimization formulation, how well a model works is measured by so-called "equation error" [44] in which the squared error between the estimated glottal wave and the true glottal wave signal is minimized. The time domain error criterion can be translated into a weighted spectral domain error criterion using Parseval's theorem. Since the spectrum of the basic voicing waveform in KLGLOTT88 model has a -6dB/oct spectral tilt, a low-pass spectral weighting is imposed to minimize the equation error. This makes the estimated derivative glottal wave more robust to the residual formant ripples and high frequency noise.

Denoting the filter coefficients as $\begin{bmatrix} \hat{a}'_1 & \cdots & \hat{a}'_{N+1} \end{bmatrix}^T$, the known speech signal as $y(n)$, and the estimated derivative glottal waveform as $\hat{g}(n)$, the error between the true glottal wave (that belongs to the KLGLOTT88 model by assumption) and the estimated glottal wave can be formed. For $i = 1, \cdots, m$, where m is the number of sampling points in one glottal period, in the open phase, we have

$$\begin{aligned} g(i) - \hat{g}(i) &= 2a \cdot i - 3b \cdot i^2 - y(i) + \hat{a}'_1 y(i-1) + \ldots + \hat{a}'_{N+1} y(i-N-1) && (3.4) \\ &= \begin{bmatrix} y(i-1) & y(i-2) & \cdots & y(i-N-1) & 2i & -3i^2 \end{bmatrix} X - y(i) && (3.5) \\ &= a_i^T X - b_i && (3.6) \end{aligned}$$

In the closed phase, we have

$$\begin{aligned} g(i) - \hat{g}(i) &= 0 - y(i) + \hat{a}'_1 y(i-1) + \ldots + \hat{a}'_{N+1} y(i-N-1) && (3.7) \\ &= \begin{bmatrix} y(i-1) & y(i-2) & \cdots & y(i-N-1) & 0 & 0 \end{bmatrix} X - y(i) && (3.8) \\ &= a_i^T X - b_i && (3.9) \end{aligned}$$

$X = \begin{bmatrix} \hat{a}'_1 & \cdots & \hat{a}'_{N+1} & a & b \end{bmatrix}^T$ is the parameter vector we wish to estimate. Note that the chosen error is linear in the parameters, as is characteristic of equation error formulations.

Assuming OQ and T_0 are known, we could explicitly write out the equation error for one glottal period in matrix form, given that the close phase starts at sample $i = n + 1$.

$$\begin{aligned}
\begin{bmatrix} g(1) - \hat{g}(1) \\ \vdots \\ g(n) - \hat{g}(n) \\ g(n+1) - \hat{g}(n+1) \\ \vdots \\ g(m) - \hat{g}(m) \end{bmatrix} &= \begin{bmatrix} y(0) & \cdots & y(-N) & 2 & -3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y(n-1) & \cdots & y(n-N-1) & 2n & -3n^2 \\ y(n) & \cdots & y(n-N) & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y(m-1) & \cdots & y(m-N-1) & 0 & 0 \end{bmatrix} X - \begin{bmatrix} y(1) \\ \vdots \\ y(n) \\ y(n+1) \\ \vdots \\ y(m) \end{bmatrix} \\
&= \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} X - \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}
\end{aligned}$$

The minimization of the equation error for a glottal period is, therefore, equivalent to the following convex optimization problem:

$$\begin{array}{ll} \min_X & \|AX - B\| \\ \text{Subject to} & a > 0,\ b > 0 \text{ and } a = T_0 \cdot OQ \cdot b \end{array} \tag{3.10}$$

where $A = \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix}^T$, $B = \begin{bmatrix} b_1^T & \cdots & b_m^T \end{bmatrix}^T$.

The problem can be solved using LP (linear programming) if the L_1 norm or the L_∞ norm is used. If the L_2 norm is used, Equation (3.10) is a quadratic programming (QP) problem. There are several ways to solve a QP. The NEOS guide of optimization software lists a collection of software packages for solving a QP at http://www-fp.mcs.anl.gov/otc/Guide/SoftwareGuide/. In this thesis, the SUMT (sequential unconstrained minimization method) [6] is used. Review of SUMT can be found in Appendix B. For this problem, the L_∞ norm yields the most efficient method, followed by the L_2 norm. However, the L_∞ norm minimization tends to give more unstable filter estimates. To trade-off between robustness and computational efficiency, the L_2 norm minimization was chosen for our application. By uniformly sampling the OQ values and solving the problem at each sampled OQ, the best estimate obtained is the one having the minimum error.

In addition to the convex constraints of the glottal wave-shape parameters (a and b), the last filter coefficient, $\hat{a}'_{N+1}$, is also constrained to a predefined range. By observing that this coefficient is the product of all the poles and the spectral tilt μ, the constraints on $\hat{a}'_{N+1}$ loosely regularize the poles of the all-pole filter.

The order of the vocal tract filter N is chosen to be an even number since the vocal tract is a resonator. Its poles are expected to be complex conjugate pairs, p_i, p_i^* for $i = 1, \cdots, N/2$. Therefore,

$$\begin{aligned} 1 - \hat{a}'_1 z^{-1} - \cdots - \hat{a}'_{N+1} z^{-(N+1)} &= (1 - p_1 z^{-1})(1 - p_1^* z^{-1}) \cdots (1 - p_{\frac{N}{2}} z^{-1})(1 - p^*_{\frac{N}{2}} z^{-1})(1 - \mu z^{-1}) \\ \therefore \hat{a}'_{N+1} &= \mu \cdot \prod_{i=1}^{\frac{N}{2}} |p_i|^2 > 0. \end{aligned}$$

The coefficient $\hat{a}'_{N+1}$ is positive since the spectral tilt μ has to be positive for low-pass filtering. In addition, a stable filter estimate requires that the magnitude of poles are less than 1. Assuming $/mu$ is less than 0.9 and the magnitude of poles are less than 0.985, $\hat{a}'_{N+1}$ is required to be less than $0.9 \cdot 0.985^N$ in our implementation. Hence, two additional convex inequality constraints ($\hat{a}'_{N+1} > 0$ and $\hat{a}'_{N+1} < 0.9 \cdot 0.985^N$) are introduced to the original optimization problem in Equation (3.10).

Since the estimated all-pole filter is not guaranteed to be stable, a Steiglitz-McBride iteration [44] with stability projection is applied in the convex optimization loop to overcome this problem. This stabilization procedure is stopped when it can not converge in a few iterations and the unstable poles are simply reflected around the unit circle.

An interesting observation is that the proposed method is equivalent to closed-phase LPC if only the speech signal during the closed phase of the glottal cycle is used to estimate filter coefficients. Hence, the proposed method extends the closed-phase LPC by including the open phase of the derivative glottal wave and fitting a constrained KLGLOTT88 model there.

Implementation

The actual implementation of the formulation minimizes the square of the L_2 norm and is expressed in the following form:

$$\begin{array}{ll} \min_X & \|AX - B\|_2^2 \\ \text{Subject to} & a > 0,\ b > 0,\ \hat{a}'_{N+1} > 0,\ \hat{a}'_{N+1} < 0.9 \cdot 0.985^N \text{ and } a = T_0 \cdot OQ \cdot b \end{array} \tag{3.11}$$

where $X = \begin{bmatrix} \hat{a}'_1 & \cdots & \hat{a}'_{N+1} & a & b \end{bmatrix}^T$ is a vector of length $N+3$.*

The solution X has to satisfy the equality constraint, $a = T_0 \cdot OQ \cdot b$,

$$\begin{bmatrix} 0 & \cdots & 0 & 1 & -T_0 \cdot OQ \end{bmatrix}^T \cdot \begin{bmatrix} \hat{a}'_1 \\ \vdots \\ \hat{a}'_{N+1} \\ a \\ b \end{bmatrix} = A_E \cdot X = 0, \tag{3.12}$$

where A_E is a rank 1 matrix.

X is chosen to be $F \cdot Z + X_0$, where F is the null space of A_E,† $Z \in R^{N+2}$ and X_0 is any solution which satisfies Equation (3.12). In the current implementation, X_0 is chosen to be $\begin{bmatrix} 0 & \cdots & 0 & T_0 \cdot OQ & 1 \end{bmatrix}^T$. In addition to the equality constraint, there are four inequality constaints. All of them can be re-written in terms of the variable Z.

- $a > 0$
  $\Rightarrow \begin{bmatrix} 0 & \cdots & 0 & -1 & 0 \end{bmatrix} \cdot X < 0$
  $\Rightarrow f_1(Z) \triangleq d_1^T \cdot (F \cdot Z + X_0) < 0,$
  where $d_1 = \begin{bmatrix} 0 & \cdots & 0 & -1 & 0 \end{bmatrix}^T$
- $b > 0$
  $\Rightarrow \begin{bmatrix} 0 & \cdots & 0 & 0 & -1 \end{bmatrix} \cdot X < 0$
  $\Rightarrow f_2(Z) \triangleq d_2^T \cdot (F \cdot Z + X_0) < 0,$
  where $d_2 = \begin{bmatrix} 0 & \cdots & 0 & 0 & -1 \end{bmatrix}^T$

*N is the order of the vocal tract filter.
†Hence, F has rank $N + 2$

- $\hat{a}'_{N+1} > 0$
  $\Rightarrow \begin{bmatrix} 0 & \cdots & 0 & -1 & 0 & 0 \end{bmatrix} \cdot X < 0$
  $\Rightarrow f_3(Z) \triangleq d_3^T \cdot (F \cdot Z + X_0) < 0,$
  where $d_3 = \begin{bmatrix} 0 & \cdots & 0 & -1 & 0 & 0 \end{bmatrix}^T$

- $\hat{a}'_{N+1} < 0.9 \cdot 0.985^N$
  $\Rightarrow \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & 0 \end{bmatrix} \cdot X - 0.9 \cdot 0.985^N < 0$
  $\Rightarrow f_4(Z) \triangleq d_4^T \cdot (F \cdot Z + X_0) - 0.9 \cdot 0.985^N < 0,$
  where $d_4 = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 \end{bmatrix}^T$

The objective function can also be written in terms of variable Z.

$$\begin{aligned} f_0(Z) &= \|AX - B\|_2^2|_{X=F\cdot Z+X_0} \\ &= Z^T F^T A^T A F Z + 2(X_0^T A^T A F - B^T A F)Z + (B^T B + X_0^T A^T A X_0 - 2B^T A X_0) \end{aligned}$$

Therefore, solving Equation (3.11) is equivalent to solving the following problem.

$$\begin{array}{ll} \min_Z & f_o(Z) \\ \text{Subject to} & f_i(Z) < 0,\, i = 1, \cdots, 4. \end{array} \tag{3.13}$$

SUMT is applied to solve Equation (3.13). The basic concept behind of SUMT is to solve a sequence of smooth unconstrained problems by transforming the constraints into a smooth barrier function. SUMT is outlined here.
Given strictly feasible Z, $t = t^{(0)} > 0$, $\mu > 1$, $\epsilon > 0$.
Loop

1. Centering step: Starting from Z, compute the optimal point $Z^\star(t)$ which minimizes $tf_0(Z) + \phi(Z)$. The Newton's method with backtracking line search‡ [6] is applied to the current implementation of source-filter deconvolution.

2. $Z = Z^\star(t)$.

3. $t = \mu t$.

‡Backtracking line search parameters are fixed. $\beta = 0.5$, $\alpha = 0.2$.

until $m/t < \epsilon$.[§]

For each outer iteration of SUMT, the problem is to reduced to solving the following unconstrained convex optimization problem.

$$\min_{Z}\ t \cdot f_0(Z) + \phi(Z) \tag{3.14}$$

where the log barrier function $\phi(Z) = -\sum_{i=1}^{4} \log(-f_i(Z))$.

The computational complexity of SUMT (Appendix B) depends on the choice of $t^{(0)}$ and μ. Following the rule of thumb in Section B.3.5, $t^{(0)}$ is chosen to be $4/\|AX_0 - B\|_2^2$ since $4/\|AX_0 - B\|_2^2$ is the estimated initial duality gap. In order to find the best μ to minimize computation, the total number of Newton iterations is calculated for a wide range of μ. Figure 3.5 plots the average total number of Newton iterations with respect to μ for three sung pitches with μ ranging from 10 to 180. The averaged results are based on baritone sung vowels /a/, /i/, and /e/. Since the sung pitch affects the size of matrices A and B, the optimal value of μ depends on the sung pitch. It turns out that the optimal μ tends to be larger when the sung pitch is higher.

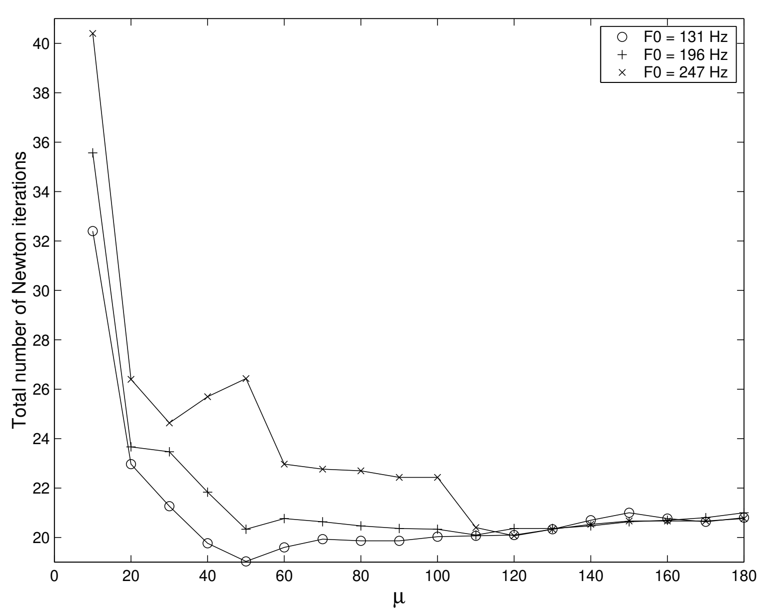

Figure 3.5: Total number of Newton iterations with respect to μ.

[§]The desired accuracy ϵ is chosen to be 0.001 in the current implementation.

3.1.2 Estimation Procedure

The overall source-filter deconvolution procedure for analysis/re-synthesis is illustrated in Figure 3.6. In phase I, the vocal tract filter and the glottal wave parameters are estimated using a sequential unconstrained minimization technique with linear search on the open quotient value. The objective of phase II is to smooth the estimated model parameter sequence obtained from the SUMT procedure.

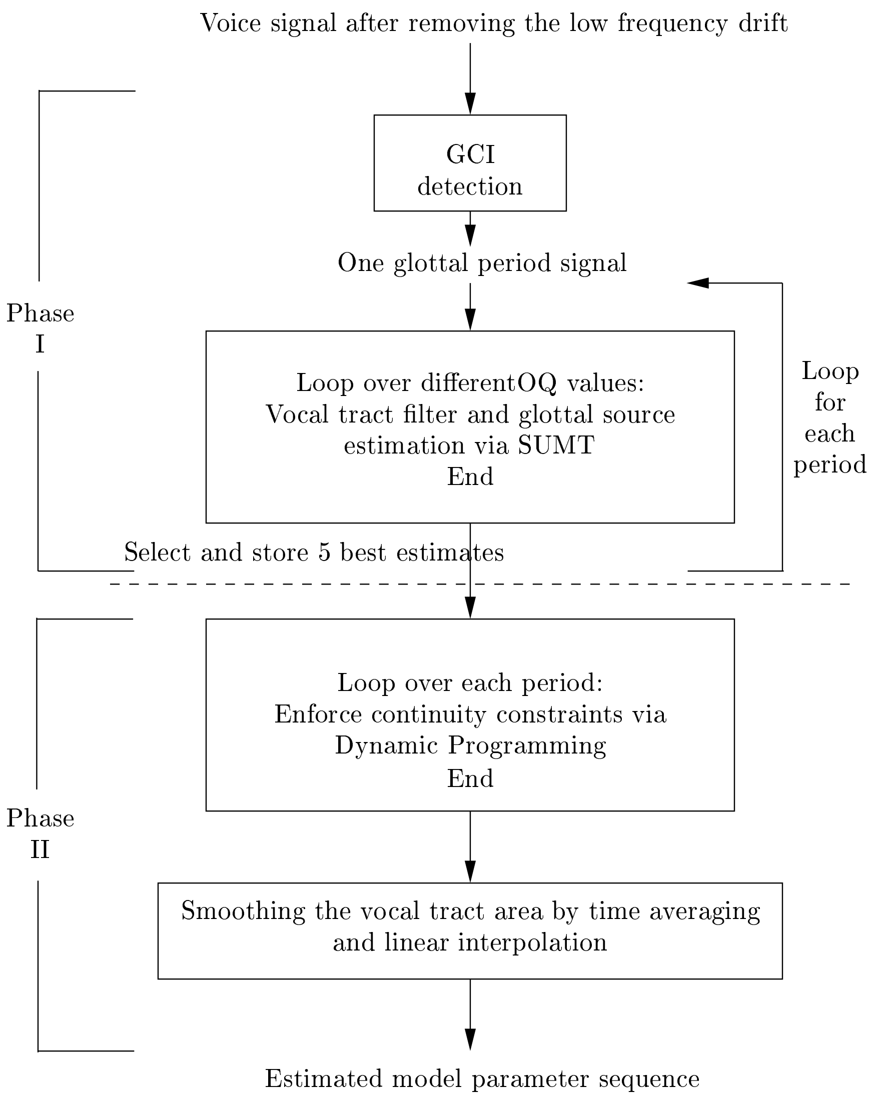

Figure 3.6: Analysis procedure for source-filter decomposition

The voice signal is first high-pass filtered by the filter $\frac{1-z^{-1}}{1-0.99z^{-1}}$ to remove the low frequency drift near DC. Since the SUMT estimation is performed pitch synchronously, it is necessary to estimate the fundamental period T_0 [1] and retrieve the pitch-synchronized speech pressure data that start at the glottal closed phase prior to SUMT estimation. Several methods for GCI detection from speech signals have been discussed in the literature [78] [50]. Most are based on either the short-time energy of the signal or the linear prediction residual signal. These methods operate on block data; since voice is rarely precisely stationary, which implies ambiguity in the instants of the detected glottal closure. To overcome this problem, Smits [67] developed another type of method to extract the instants of significant excitation (glottal closures) for speech signals. This method assumes that the excitation signal within a pitch period, starting from the significant excitation, is minimum phase. The glottal closure instants are then estimated as the sum of the frame start time and the average group-delay of the LP residual within an analysis frame.

The group-delay method has been shown to be robust against noise and distortion, since the average phase characteristics of the signal are determined mainly by the strength of the excitation [57]. Hence, this algorithm was chosen as the main algorithm for GCI detection. The estimates from this method tend to have a lag bias. This is resolved by implementing the Frobenius norm approach [50] in conjunction with the group-delay method to determine the final GCI locations. The details of the group-delay method and the Frobenius norm approach is described in Section 3.1.3.

From the study of synthetic data, it was found that the normalized estimation error of the inverse filtered glottal excitation defined in Equation (3.21) has an upper bound of 7% even when the detected GCIs are off by 15% from the correct locations. Hence, GCI detection is a byproduct of the SUMT estimation, where the GCIs are determined by locating the local minimum of the inverse-filtered glottal excitation. In experiments for real recordings of sustained vowels, the GCIs are actually further corrected using this method.

After the GCI detection, the SUMT estimation is applied, for each glottal period, to the speech pressure wave data, in which linear search is performed over different OQ values. For the vicinity of the optimal OQ value, the error surfaces with respect to OQ for synthetic data simulations are found to be almost quasi-convex. Hence, the search range of the optimal OQ can be reduced to a small line segment around the estimated OQ value of the previous glottal period to speed up the computation. The five best estimates are stored for each glottal period.

Since the estimates are obtained for each period, there is no inherent continuity constraint between adjacent periods. The dynamic programming method is applied to enforce the continuity constraints, in which a discontinuity penalty is added to the error at each possible path:

$$w_{OQ} \cdot |OQ_{i+1} - OQ_i| + w_{AV} \cdot |AV_{i+1} - AV_i| + w_{VT} \cdot \|\underline{K}_{i+1} - \underline{K}_i\|, \tag{3.15}$$

where $\underline{K}_i$ is the reflection coefficient vector at period i; OQ_i and AV_i are the OQ value and AV value at period i. Reflection coefficients are closely related to the vocal-tract cross-sectional areas if the vocal tracts are assumed to be concatenated lossless cylindrical tubes. Alternatively, one could apply the continuity constraints for AV and the all-pole filter coefficients directly in the SUMT procedure. For further smoothing of the vocal tract filter, the vocal tract areas are down-sampled by averaging over several periods and then up-sampled by linear interpolation. The above flow chart provides the general procedure for acquiring the model parameters for the purpose of analysis and re-synthesis. The last two blocks are optional for the purpose of obtaining a set of model parameters for sustained vowels.

3.1.3 Glottal closure instants detection

Most of the existing GCI detection algorithms are based on the source-filter human voice production model. The voiced sound is generated by exciting the vocal tract filter with a quasi-periodic derivative glottal volume velocity wave (Figure 2.2). The glottal wave in general is gradually building up and has rapid transition during the closing period. This causes the main excitations to occur at glottal closure points. Most GCI detection methods rely on the assumption that glottal closures are strong and abrupt.

The GCI detection plays an important role in pitch synchronous voice processing such as closed-phase LPC [81] and PSOLA for prosody manipulations of the speech sounds [8]. For the closed-phase LPC application, the GCI detection is used to estimate the closed phase of the glottal cycle. Linear prediction is then applied to this segment of the data to obtain precise vocal tract characteristics since the closed phase is where the vocal tract and the glottal source have no interaction.

In the following two sections, the *Frobenius norm algorithm* and the *group delay algorithm* for GCI detection are introduced. Our initial GCI detection method is a hybrid combination of these two methods.

Frobenius norm algorithm

The Frobenius norm algorithm operates on block data and is used in conjunction with the main GCI detection method, the group-delay method. This algorithm [50] is based on the idea that linear prediction fits better when processing a segment of speech signal that does not contain the glottal closure points (main excitation). Otherwise, large prediction errors are introduced if glottal closures are included.

Consider a segment of pre-emphasized speech signal, $s(i)$, for $i = 1, 2, \cdots, p+m$, where $m \geq p+1$. The following data matrix is formed,

$$\mathbf{S} = \begin{bmatrix} s_{p+1} & s_p & s_{p-1} & \cdots & s_1 \\ s_{p+2} & s_{p+1} & s_p & \cdots & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p+m} & s_{p+m-1} & s_{p+m-2} & \cdots & s_m \end{bmatrix} = \begin{bmatrix} \overline{s}_{p+1} & \overline{s}_p & \cdots & \overline{s}_1 \end{bmatrix}.$$

Assume the data matrix has full column rank so that it can be decomposed as the sum of $(p+1)$ rank one matrices by SVD:

$$\mathbf{S} = \sum_{i=1}^{p+1} \sigma_i \overline{u}_i \overline{v}_i^t \tag{3.16}$$

where $U = [\overline{u}_1 \overline{u}_2 \cdots \overline{u}_m]$ and $V = [\overline{v}_1 \overline{v}_2 \cdots \overline{v}_{p+1}]$ are orthogonal matrices with $U^t U = V^t V = I_{p+1}$. $\sigma_i' s$ are singular values where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{p+1} > 0$.

The Frobenius norm (or Hilbert-Schmidt norm, Euclidean norm) of a matrix $\mathbf{S} = \{s_{ij} : 1 \leq i \leq m, 1 \leq j \leq p+1\}$ is defined as

$$\|\mathbf{S}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p+1} s_{ij}^2} = \sqrt{\mathrm{Trace}\{\mathbf{S}^{\mathbf{T}}\mathbf{S}\}}. \tag{3.17}$$

Since $\mathbf{S}^{\mathbf{T}}\mathbf{S} = \left(\sum_{i=1}^{p+1}\sigma_i\overline{u}_i\overline{v}_i^t\right)^T\left(\sum_{j=1}^{p+1}\sigma_j\overline{u}_j\overline{v}_j^t\right) = \sum_{i=1}^{p+1}\sigma_i^2$, the Frobenius norm is also the square root of the sum of the squared singular values:

$$\|\mathbf{S}\|_F = \sqrt{\sum_{i=1}^{p+1}\sigma_i^2}. \tag{3.18}$$

From the above equation, we have the following inequality,

$$\sigma_{p+1}^2 \leq \frac{1}{p+1}\|\mathbf{S}\|_F^2 \leq \sigma_1^2. \tag{3.19}$$

Assuming the linear prediction order is p, s_{p+i} can then be predicted as $-a_1 \cdot s_{p+i-1} - a_2 \cdot s_{p+i-2} - \cdots - a_p \cdot s_i$. The linear prediction error vector may be expressed as,

$$\mathbf{e} = \mathbf{Sa} \tag{3.20}$$

where $\mathbf{a} = \begin{bmatrix} 1 & a_1 & a_2 & \cdots & a_p \end{bmatrix}^T$. When the data matrix does not contain the excitation point, $\mathbf{e}$ goes to zero (without considering the noise) such that the vectors are linearly dependent and their singular values are small. On the other hand, when the data matrix contains the excitation point, the norm of $\mathbf{e}$ and singular values $\sigma_i's$ become large [50]. Hence, the Frobenius norm of the data matrix $\mathbf{S}$ reaches its maximum when s_{p+1} is excited by the main excitation. In order to find the main excitation (GCI), a sequence of matrix $\mathbf{S}$ is first constructed by windowing the preemphasized input speech signal sample by sample. The Frobenius norm function is then calculated and the local maxima of the Frobenius norm function are extracted as GCI's.

Group delay algorithm

The group delay algorithm assumes that the excitation signal starting from the glottal closure instants is a minimum phase signal within one pitch period [67]. It determines the glottal closure instants by using the properties of minimum phase signals and group delay function, i.e., the average slope of the phase spectrum (average group delay) of a minimum phase signal is zero.[¶] Since the physical vocal tract is assumed to be a minimum phase

[¶]The group delay is defined as $\triangleq -\frac{d\phi(w)}{dw} \Rightarrow$. Hence, the average group delay $= \frac{1}{2\pi}\int_0^{2\pi}\frac{d\phi(w)}{dw}dw = \frac{1}{2\pi}\int_0^{2\pi}d\phi(w) = \frac{1}{2\pi}(\phi(2\pi)-\phi(0)) \triangleq$ which is the winding number. For rational functions, the winding system, the speech pressure waveform following the glottal closure instant is a minimum phase signal within one pitch period. The analysis procedure of this algorithm is described below [57].

- Calculate the linear prediction residual signal from the speech pressure input data using the LPC autocorrelation method. Since the filter estimated with the LPC is minimum phase, the linear prediction residual signal following the glottal closure instant is still a minimum phase signal within one pitch period.

- Smooth the residual signal and compute its short-time energy function for every sample. The weighted residual signal is obtained by multiplying the residual signal with this short-energy function, which is an estimation of the strength of glottal excitation. The robustness of this group delay method depends mainly on the strength of the glottal excitation [57].

- For each sampling instant of the (weighted) residual signal, select a frame of data centered at the sampling instant and compute the group delay function.

- Smooth the group delay function using a 3-point median filter.

- Compute the average of the group delay function. The group delay is non-zero whenever the (weighted) residual signal does not start from the glottal closure instant.

- Determine the phase slope function, which is the negative of the average of the group delay function for each sampling instant of the (weighted) residual signal.

- Smooth the phase slope function and identify its positive zero-crossings as the glottal closure points.

Comparison

Figure 3.7 compares the GCI detection error rate for the Frobenius norm method, the group delay methods and the hybrid method. The error rate is the absolute error of GCI estimation, in the number of samples, divided by the number of samples in the glottal

number is the number of zeros outside the unit circle minus the number of poles outside the unit circle. Since a minimum system has no pole or zero outside the unit circle, its average group delay is zero.

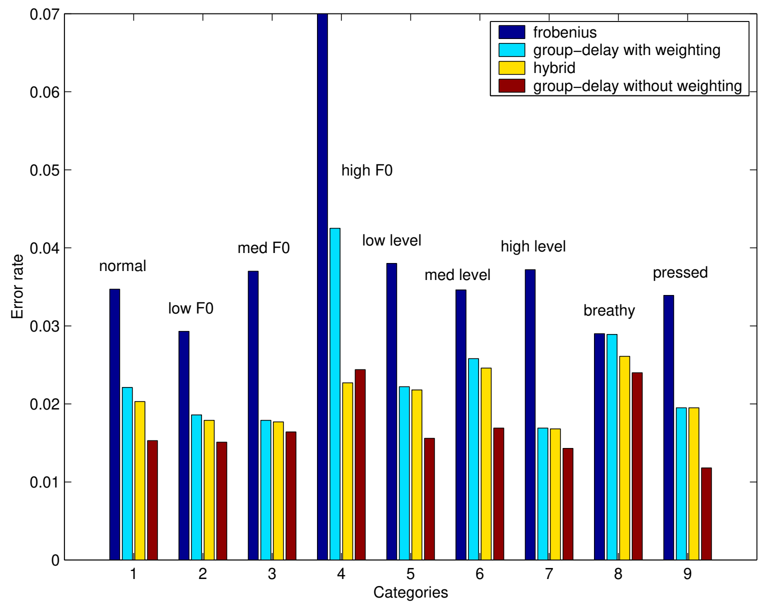

Figure 3.7: Error rate comparisons of GCI detection algorithms.

period. The group delay method could either apply to the weighted residual signal or the non-weighted residual signal. The hybrid method is a combination of the weighted group delay method and the Frobenius norm method. Since the weighted group delay method tends to have a lag bias, the Frobenius norm method is used whenever the detected GCI via the Frobenius norm leads the detected GCI via the group delay method. The weighted group delay method is used in the exception that the Frobenius norm results in a larger absolute error between the estimated period length and the average pitch period. The speech pressure data used are synthesized by filtering synthetic derivative glottal waves generated with the LF model with an all-pole vocal tract filter. Nine different scenarios are presented in Figure 3.7 that take into consideration of different levels of loudness, fundamental frequencies and breathiness. The model parameters for these scenarios are obtained by fitting the LF model to the inversed filtered derivative glottal wave [38]. The error rates shown in Figure 3.7 are averaged over vowel configurations of /a/, /e/, and /i/. From Figure 3.7, we see that the group delay methods are more accurate than the Frobenius norm method. It also shows that the hybrid method improves over the weighted group delay method. Although the hybrid method has a larger error rate compared to the non-weighted group delay method in this synthetic data set, the hybrid method is used in our GCI initialization since the weighted group delay method is shown to be more robust to noisy input [57].

3.1.4 Performance evaluation of the source-filter deconvolution

In this section, the proposed source-filter deconvolution estimation procedure is applied to both synthetic data and baritone real recordings to evaluate its performance.

Simple synthetic data

First, ignoring inherent model mismatches, the input synthetic data were generated using the analysis model depicted in Figure 3.4. The sampling rate of the synthetic data is 48k Hz and the order of the filter is 40. The fundamental period of the synthetic voice, T_0, is chosen as an integer multiple of the sampling period. There exist interpolated estimation methods which can be incorporated to extend to the non-integer period case [68].

Error surface with respect to OQ To verify that linear search over OQ is valid, error surfaces with respect to OQ under different possible conditions are examined here. The desired value of OQ is 0.6. The error is defined as the square of the normalized glottal wave estimation error in a given glottal cycle,

$$\frac{\sum_{i=1}^{m}(g(i) - \hat{g}(i))^2}{\sum_{i=1}^{m}(y(i) - \hat{y}(i))^2}. \tag{3.21}$$

Figure 3.8 shows the effect of increasing the pitch of the analyzed. This is achieved by shortening the glottal period and decreasing the number of data points. Figures 3.9 and 3.10 illustrate the effect of inaccurate GCI detection, where the GCI errors are given in percentage per glottal cycle. Figure 3.11 shows the effect of insufficient filter order. Figure 3.13 shows the effect of noisy pressure wave input.

From Figure 3.8, we can observe that the optimal point of the error surface has a narrower notch at lower fundamental frequencies while the OQ estimates are all accurate.

Figure 3.9 shows that as the estimated GCI's are advanced ahead of the true GCI's, the optimal OQ estimates are further deviated. Larger GCI estimation error leads to a more flattened error surface and under estimates the OQ.

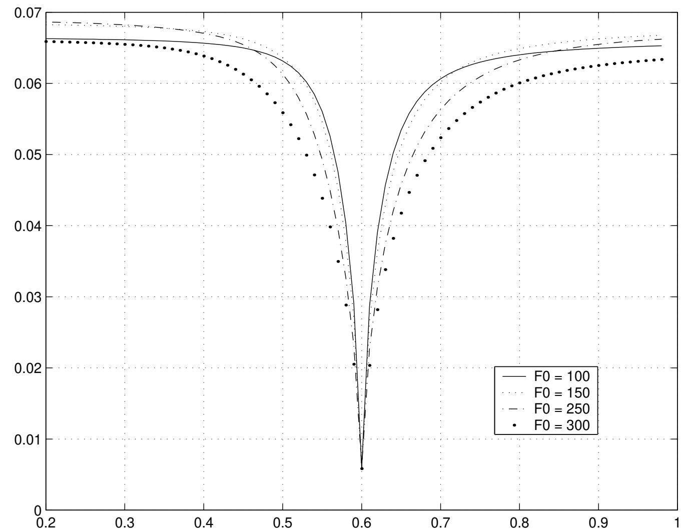

Figure 3.8: Error surface at various fundamental frequencies.

Similarly, Figure 3.10 shows that with increasing GCI lag errors, i.e., the estimated GCI's are delayed versions of the true GCI's, the error surface becomes more flattened and the OQ's are over estimated.

Figure 3.11 shows that when the order of the filter is reduced below the order of the filter that generates the synthetic data ($N = 40$, the true order), the OQ estimate is biased slightly above the true value, 0.6, but more importantly, the error surface is flattened.

Figure 3.12 plots the error computed by Equation (3.21) with respect to the order of the estimated filter and shows that the error remains at minimum as long as the order of the filter is larger than the desired order (40). Hence, given unknown filter order, the order of the filter can be determined as the smallest value which minimizes the error.

Figure 3.13 depicts the error surface with noisy input data. As the SNR decreases, the minimum error increases and the error surface is more rounded near the optimal OQ value. However, the estimated $OQ's$ are still very close to the desired value.

In summary, the error surfaces with respect to OQ are almost convex near the optimal OQ values. By the inherent physical continuity constraint of the OQ value, the linear search range of the OQ values can be limited to the neighborhood of the previous optimal OQ estimate. Therefore, a linear search over OQ is valid most of the time.

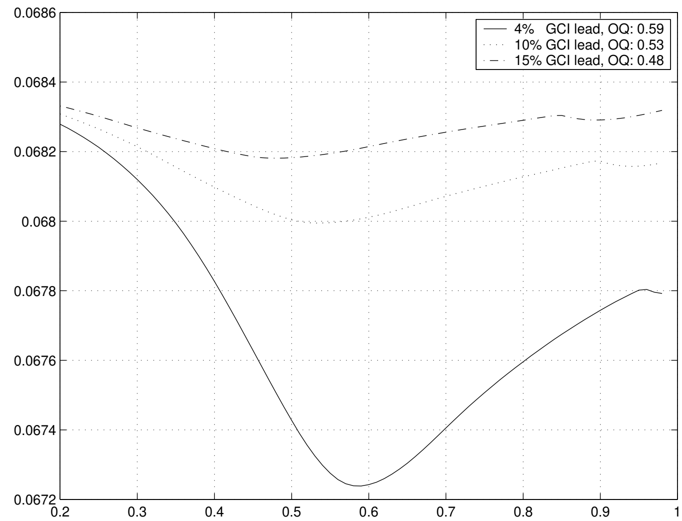

Figure 3.9: Error surface with incorrect GCI detection (GCI lead error).

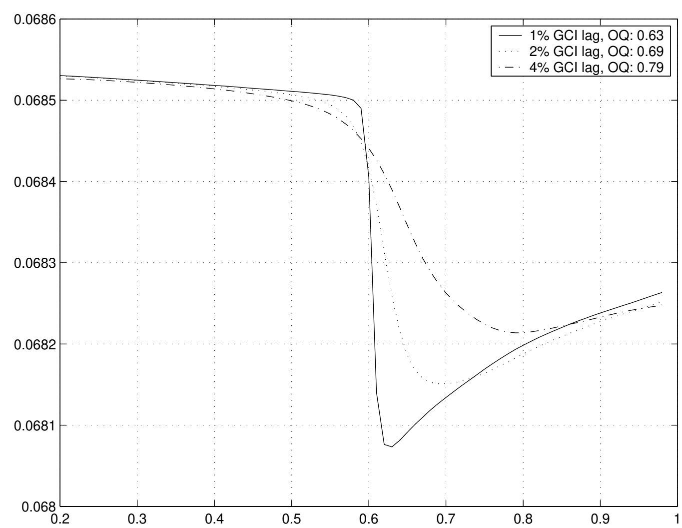

Figure 3.10: Error surface with incorrect GCI detection (GCI lag error).

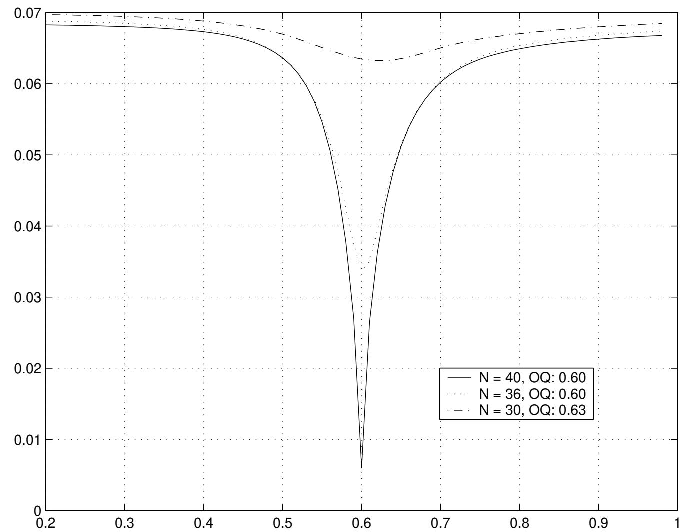

Figure 3.11: Error surface with insufficient filter order.

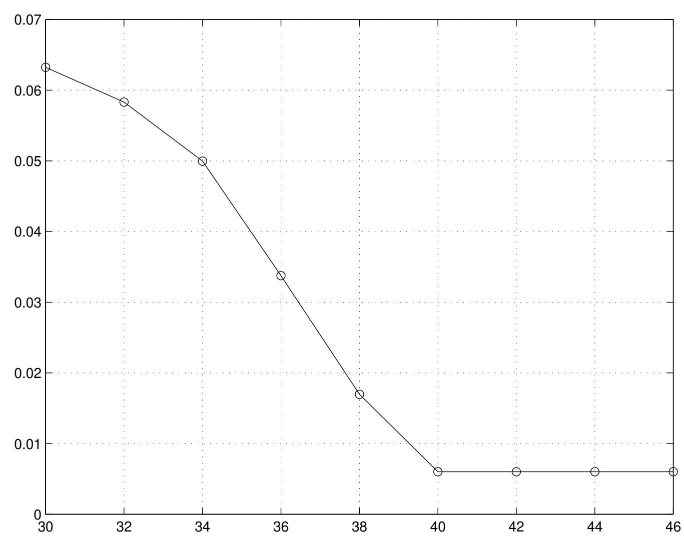

Figure 3.12: Normalized error with respect to the order of the estimated filter .

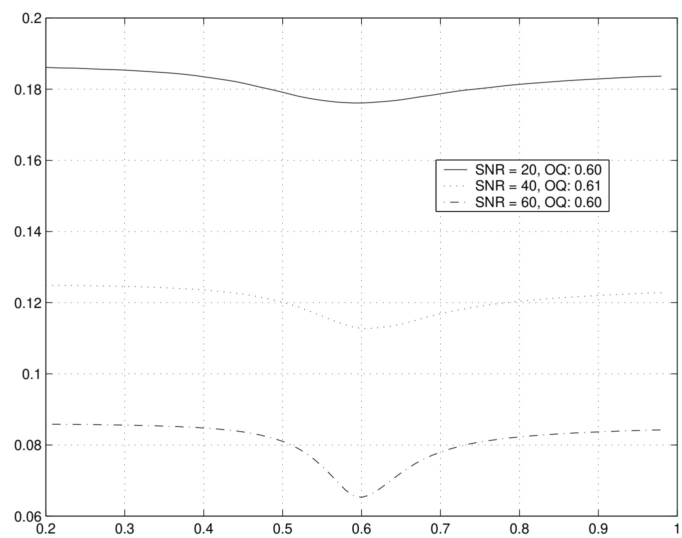

Figure 3.13: Error surface with noisy data.

Vocal tract estimation comparisons The following figures show the desired and estimated frequency responses of the vocal tract filter. The vocal tract filter estimates computed by the autocorrelation method of LPC analysis [54] are compared to ones obtained by the proposed SUMT method. The input data in the LPC analysis are pre-emphasized by a high-pass filter $H(z) = (1 - 0.95z^{-1})$ to account for a typical -6 dB/oct spectral tilt of the derivative glottal wave in speech. Therefore, the periodic excitation source for the estimated filter in the LPC analysis is assumed as an impulse train, which is an approximation of a double derivative of the glottal wave. In addition to the frequency response comparison, the pseudo-area distance between the desired and the estimated vocal tracts is also compared. The pseudo-area is derived from the auto-regressive coefficients of the vocal tract filter by assuming that the vocal tract area at the lip end is 1 and the area of the glottis is zero. The area distance measure is a normalized L_2 distance,

$$\sqrt{\sum_{i=1}^{N+1} (a(i) - \hat{a}(i))^2/(N+1)},$$

where N is the order of the filter, $a(i)$ is the area of the ith section of the desire vocal tract and $\hat{a}(i)$ is the area of the ith section of the estimated vocal tract.

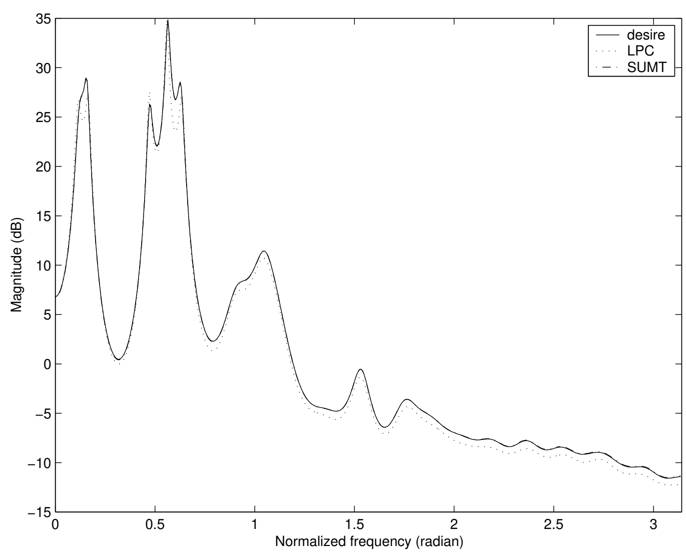

Figure 3.14: Frequency response of the vocal tract filter with fundamental frequency $f_0 =$ 300 Hz.

Figure 3.14 shows the frequency response of the vocal tract filter for synthetic input data having a fundamental frequency of 300 Hz. The SUMT estimation recovers the vocal tract filter perfectly. The bandwidth of the formants of the LPC estimate tends to be narrower than the desired bandwidth. The spectral tilt of the LPC estimate is larger than the desired one, hence, its inversed glottal waveform estimate is noisier. Table 3.1 summarizes the area distance measures for both the LPC and the SUMT estimates. It is apparent that the area distance of the LPC estimates are larger.

Table 3.1: Comparison of the area distance measure for LPC and SUMT estimation at various fundamental frequencies.

F_0	100	150	250	300
SUMT	$9 \cdot 10^{-4}$	$9 \cdot 10^{-4}$	$8.7 \cdot 10^{-4}$	$8.8 \cdot 10^{-4}$
LPC	0.2787	0.1907	0.0577	0.0447

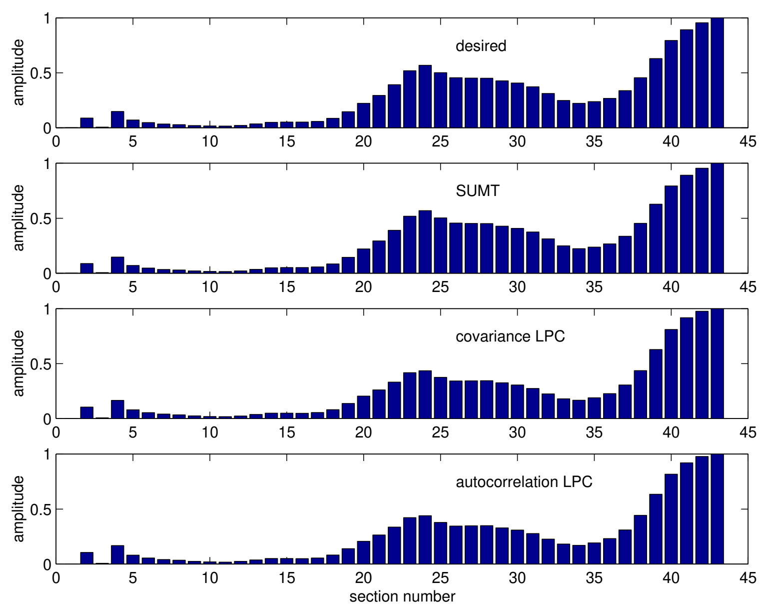

Figure 3.15: Pseudo-area functions for simple synthetic data.

The desired and the estimated area functions are compared in Figure 3.15. The area function of the SUMT estimate matches the desired almost exactly. In addition to the area function via the autocorrelation method of LPC, Figure 3.15 also shows the area function via the covariance method of LPC since it and the autocorrelation method are often used for speech modeling in practice. Both area functions via the autocorrelation method and the covariance method are very close to the desired one.

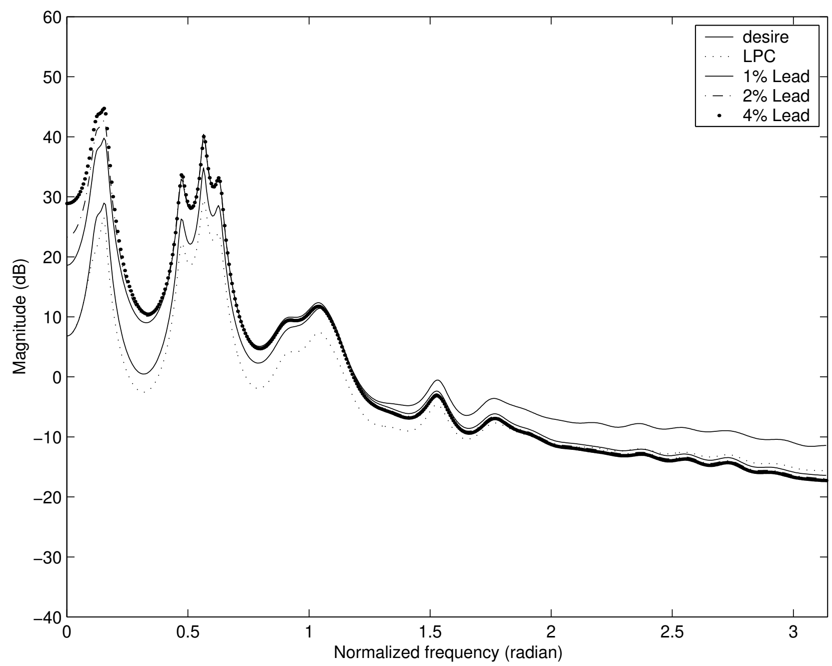

Figure 3.16: The frequency response of the vocal tract filter with GCI lead error.

Table 3.2: Comparison of the area distance measure for LPC and SUMT estimation with GCI lead errors.

GCI lead error	1%	2%	4%
SUMT	0.2670	0.4495	0.7713
LPC	0.1907	0.1907	0.1907

Figure 3.16 shows the frequency response estimates when there are lead errors in the GCI detection. The fundamental frequency of the input synthetic data is 150 Hz. Although the first few formant estimates of the SUMT method are very close to the desired, their strengths and bandwidths are deviated from the desired response. All of the area distances of the SUMT estimates with GCI lead errors $(1\%, 2\%, 4\%)$ are larger than the area distance of the LPC estimate (tabulated in Table 3.2).

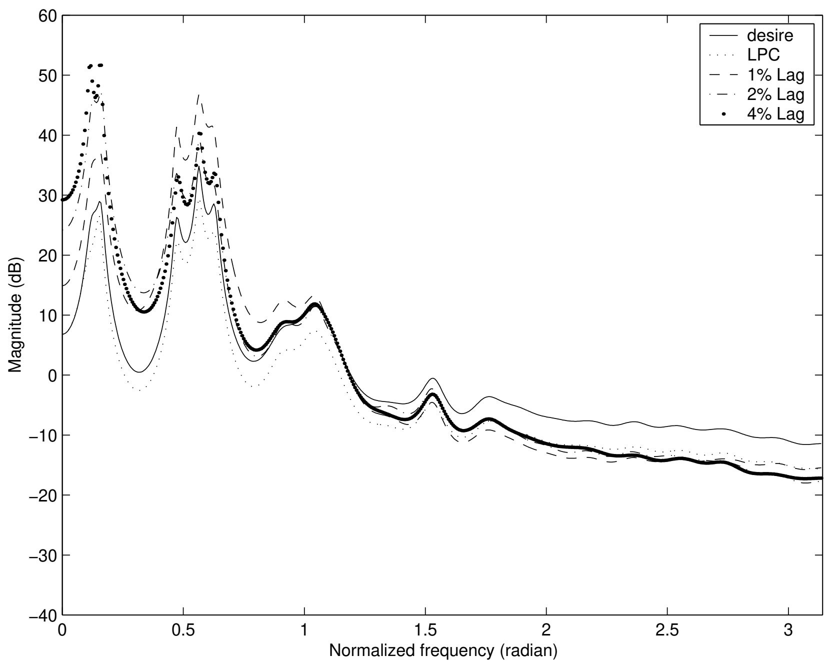

Figure 3.17: Frequency response of the vocal tract filter with GCI lag error.

Table 3.3: Comparison of the area distance measure for LPC and SUMT estimation with GCI lag errors.

GCI lag error	1%	2%	4%
SUMT	0.0762	0.3843	0.7025
LPC	0.1907	0.1907	0.1907

Figure 3.17 shows the frequency response estimates when there are lag errors in the GCI detection. The fundamental frequency of the input synthetic data is 150 Hz. Again, the

first few formants of the SUMT estimation have more deviations from the desired response compared to the LPC estimate. Table 3.3 shows that only the SUMT estimate with a 1% GCI lag error has a smaller area distance than the LPC estimate. Figures 3.16 and 3.17 illustrate the importance of the accuracy of the GCI detection to our pitch synchronous estimation method.

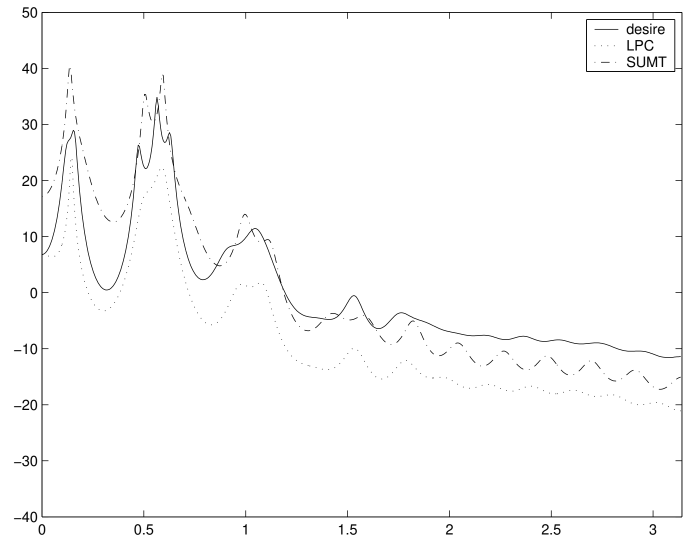

Figure 3.18: Frequency response of the vocal tract filter with a insufficient filter order, 30.

Figures 3.18 and 3.19 illustrate the effect of filter order insufficiency on vocal tract filter estimation. When the filter order is 30, some formants in both the SUMT and the LPC estimation are combined with nearby formants since the order is too small. The LPC tends to under-estimate the formant strength, on the other hand, the SUMT estimate tends to over-estimate the strength of the first few formants. When the filter order approaches to the desired order, the SUMT estimate is better than the LPC estimate.

Figures 3.20 through 3.22 illustrate the frequency response at various SNR conditions. The LPC estimates are highly deviated at high frequency range, especially under low SNR conditions. For SNR $= 60$, the SUMT estimate is capable of matching the desired frequency response of the first few formants. Table 3.4 summarizes the area distance for both the LPC estimates and SUMT estimates. The SUMT estimates are more robust to noisy data in terms of the area distance measure.

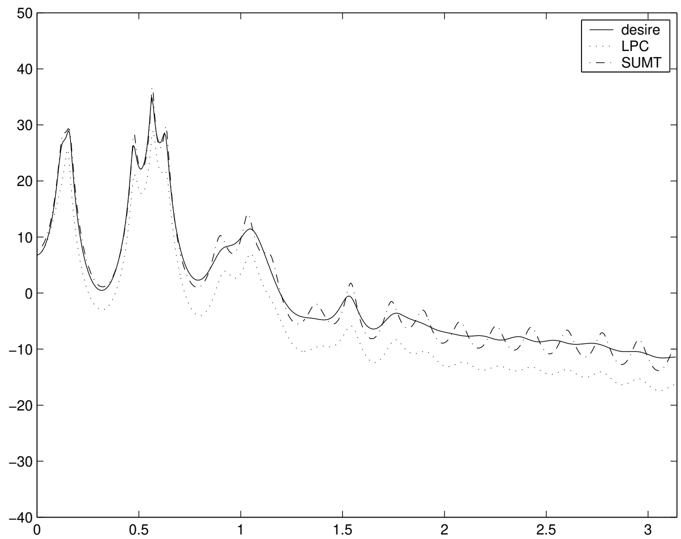

Figure 3.19: Frequency response of the vocal tract filter with a insufficient filter order, 36.

Table 3.4: Comparison of the area distance measure for LPC and SUMT estimation at various SNR conditions.

SNR	20	40	60
SUMT	0.1197	0.0670	0.0344
LPC	0.4306	0.2561	0.1598

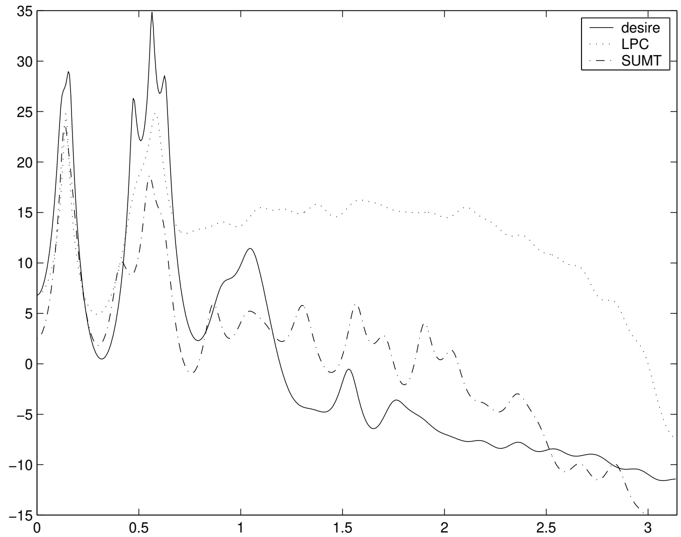

Figure 3.20: Frequency response of the vocal tract filter with SNR = 20 dB.

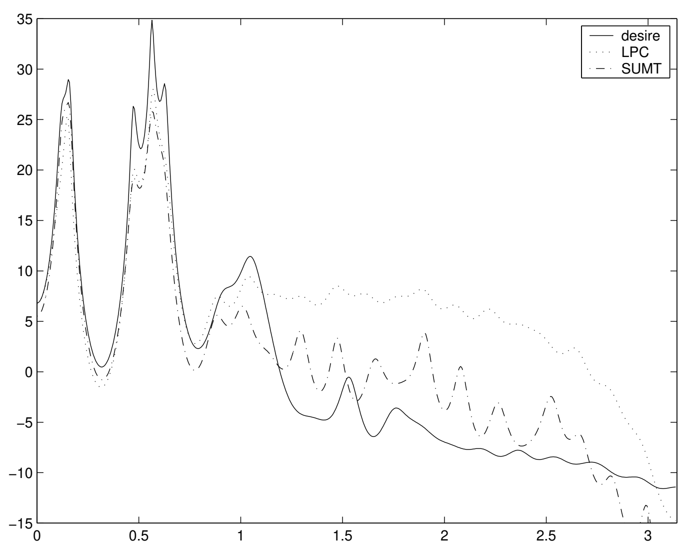

Figure 3.21: Frequency response of the vocal tract filter with SNR = 40 dB.

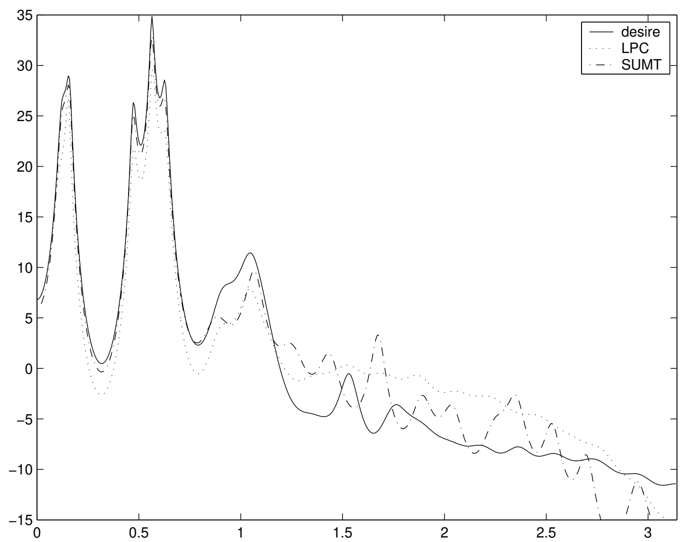

Figure 3.22: Frequency response of the vocal tract filter with SNR = 60 dB.

Synthetic data generated from two-mass model

In this section, the input synthetic data are generated using an up-to-date articulatory speech synthesizer with a two-mass vocal fold model [45] [13]. The objective is to examine the effect of source-tract filter interaction on SUMT estimations. The complexity of the articulatory synthesizer is gradually increased to generate three different input synthetic data sets. The first data set has a real lip reflection and the lip radiation is implemented with a differencing filter as is assumed in the SUMT estimation model. The second data set considers more realistic lip reflection and lip radiation. The third data input includes the viscosity and inertia of the air passing through the glottis. No trachea is attached to the articulatory synthesizer for generating these three data sets. Moreover, all of the input synthetic data are based on sustained vowel /i/ since it contains more source-filter interaction due to its constricted vocal tract configuration.

Three vocal tract filter estimation methods are compared: the autocorrelation method of LPC, the closed-phase LPC covariance method [12] and the SUMT estimation. The closed-phase LPC estimation only uses the speech pressure data during the closed phase of the glottal period where the vocal tract and the glottal wave have no interaction at all. Hence, it is expected that the closed-phase LPC achieves the best vocal tract filter estimate as long as the length of the data during the closed phase is at least twice the order of the vocal tract filter. Therefore, it is interesting to compare the SUMT estimation to the closed-phase LPC to see how close they are. Unlike the autocorrelation method, the vocal tract filter estimated from the LPC covariance method is not guaranteed to be stable. This is stablized by reflecting the unstable poles around the unit circle.

Figure 3.23 to Figure 3.25 gives the results of the first data set where the lip reflection is assumed to be real and the lip radiation is a first order differencing filter. From top to bottom, Figure 3.23 shows the desired area function used in the articulatory synthesizer, the pseudo-area functions from the SUMT estimation, the covariance LPC estimation and the autocorrelation LPC estimation. Since accurate information of closed-phase region is provided for the closed-phase LPC estimation in our simulations, the closed-phase LPC can perfectly recover the desired area function for this data set. Since it is trivial to show the area function for the closed-phase LPC estimate, the covariance method of LPC estimation is compared instead. The area of the glottis section is not included here since the glottis

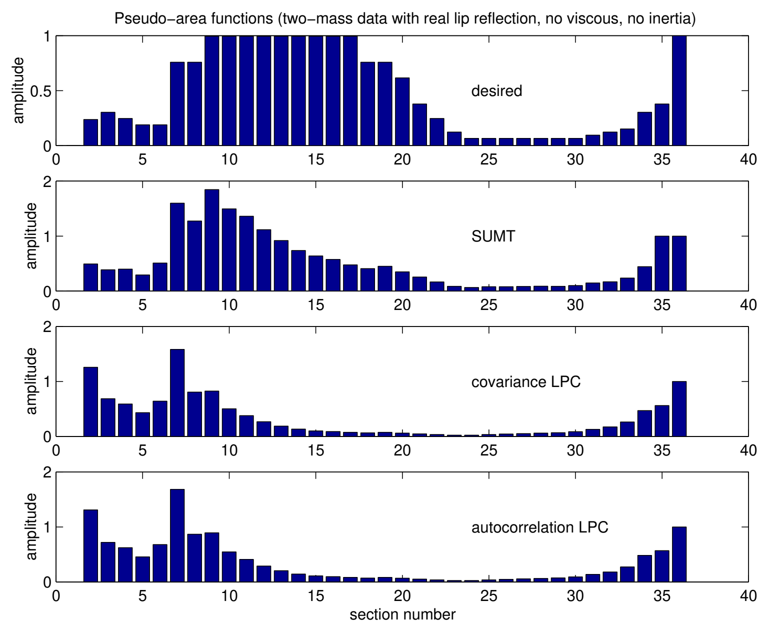

Figure 3.23: Area functions for synthetic two-mass data set I.

area is actually time varying. Moreover, the area of the lip section is normalized to 1 in this figure. The area functions for the SUMT and LPC estimates are converted from the auto-regressive coefficients under the assumption that the impedance of the glottis is infinite, i.e., the area of the glottis is zero. Figure 3.23 shows that the SUMT estimate captures the rough shape of desired vocal tract. But due to source-tract interaction, the SUMT estimate cannot recover the desired vocal tract perfectly. However, this is better than both the covariance and the autocorrelation LPC's, which are not reliable at all.

Figure 3.24 compares frequency responses of the vocal tract filter. The desired frequency response is obtained from the true area function with the area of the glottis assumed to be zero. Hence, this is the frequency response of the closed-phase glottal period. Figure 3.24 shows that the desired frequency response of the closed-phase region coincides with the frequency response obtained from the closed-phase LPC estimation. The autocorrelation LPC method has the worst match to the desired closed-phase frequency response. The SUMT estimate has a good match up to the fifth formant. It is interesting to note that the frequency response of the SUMT estimation is more damped than the desired closed-phase frequency response. The reason is that the SUMT estimation averages out the open phase and the closed-phase frequency responses and that the open phase response is normally more

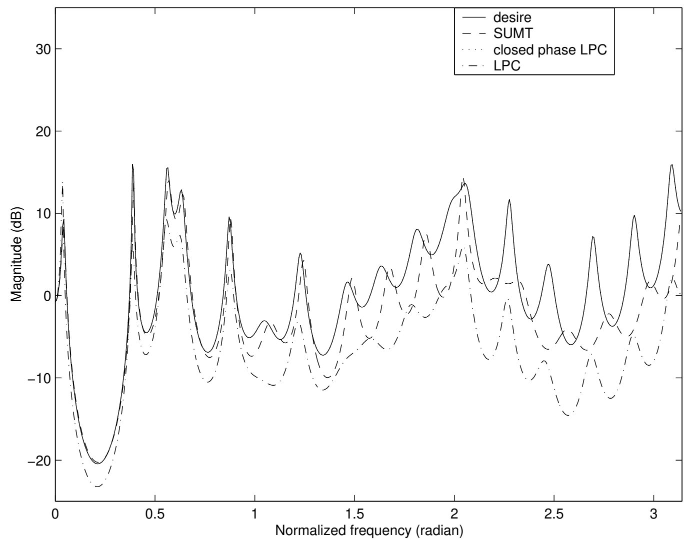

Figure 3.24: Frequency responses of the vocal tract filter for synthetic two-mass data set I.

damped. Actually, if the glottal area increases, the formant bandwidths will increase and the formant frequency will stay the same if the lip reflection impedance is real. Hence, one way to simulate the source-tract interaction is to increase the formant bandwidth during the open phase of the glottal period [11] [43]. However, this implementation is not considered in our synthesis model since the resulting effect is hardly audible from informal listening tests.

Figure 3.25 compares the glottal waves and the derivative glottal waves. The estimated glottal wave is obtained by inverse filtering the speech pressure input with the estimated filter. It is observed that the closed-phase LPC recovers the glottal wave exactly. The derivative glottal wave obtained from the autocorrelation LPC has large deviations from the desired waveform. Although the SUMT estimation can not recover the derivative glottal wave as perfectly as the closed-phase LPC, the error is small and it is still very close to the desired derivative glottal wave.

Results shown in Figure 3.26 are based on the second data set where lip reflection and radiation are more realistic. Again, the closed-phase LPC produces the best estimate of the derivative glottal wave. The derivative glottal wave estimated from SUMT has spurious ripples at the closed phase glottal period, but it closely follows the formant ripples of the true derivative glottal wave at the open-phase glottal period. Formant ripples are

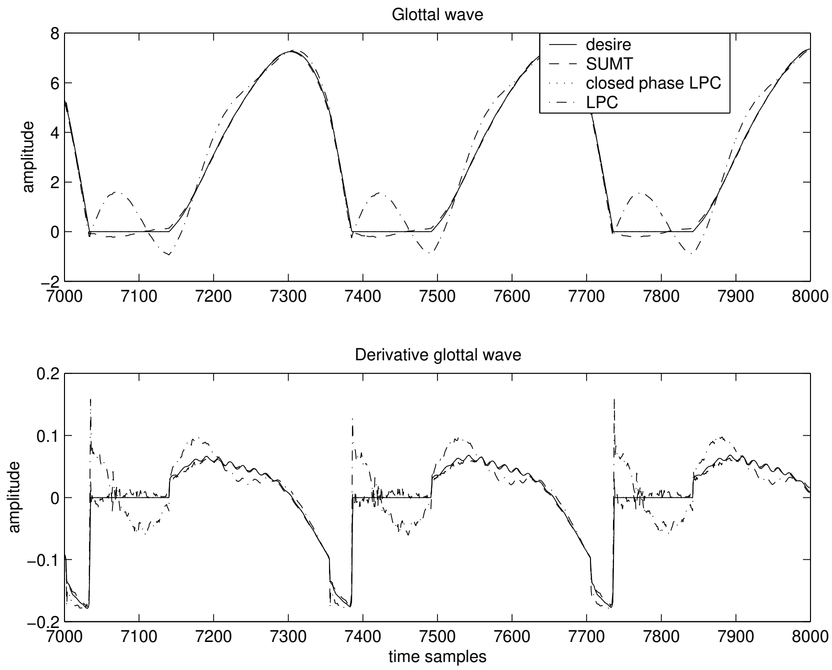

Figure 3.25: Glottal waves and derivative glottal waves for synthetic two-mass data set I.

one of the by-product of source-tract interaction. The autocorrelation LPC estimate has larger deviations and it tends to have spiky errors near the glottal closure instants since the underlying assumption of the linear prediction model breaks down around main excitations.

Results in Figure 3.27 are produced with the third data set as the input synthetic data where viscosity and inertia are taken into consideration. It turns out that the results are quite similar in comparison to the second data set. The estimation errors for the derivative glottal wave and the glottal wave do not increase due to the consideration of flow viscosity and inertia. In summary, the SUMT estimation has better performance than the autocorrelation LPC estimation in estimating the glottal wave excitation. The SUMT estimation can faithfully reproduce the formant ripples generated by source-tract interaction. However, the derivative glottal wave estimated by SUMT has inevitable spurious ripples in the closed phase glottal period since the vocal tract filter estimate averages in the open phase response.

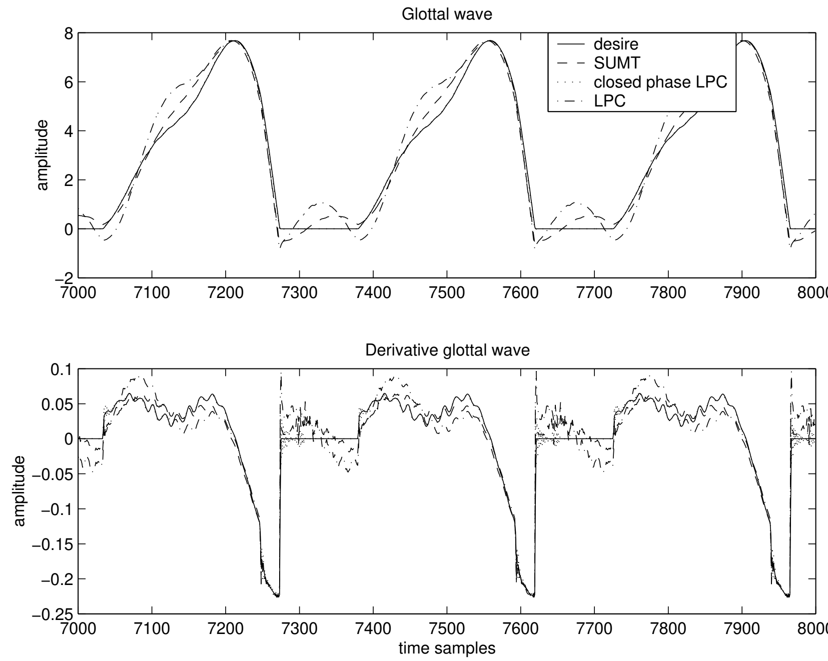

Figure 3.26: Glottal waves and derivative glottal waves for synthetic two-mass data set II.

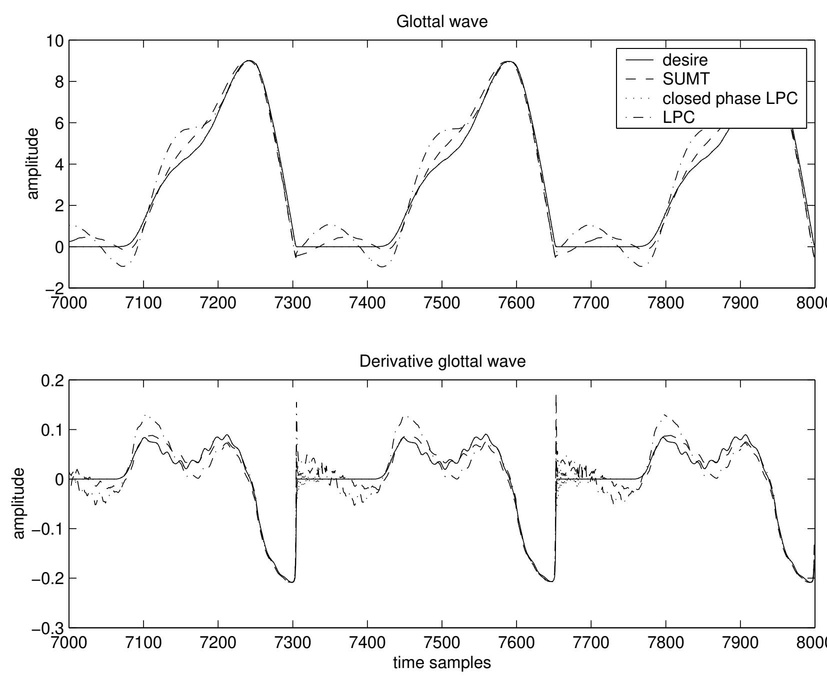

Figure 3.27: Glottal waves and derivative glottal waves for more realistic synthetic two-mass data.

Real recordings

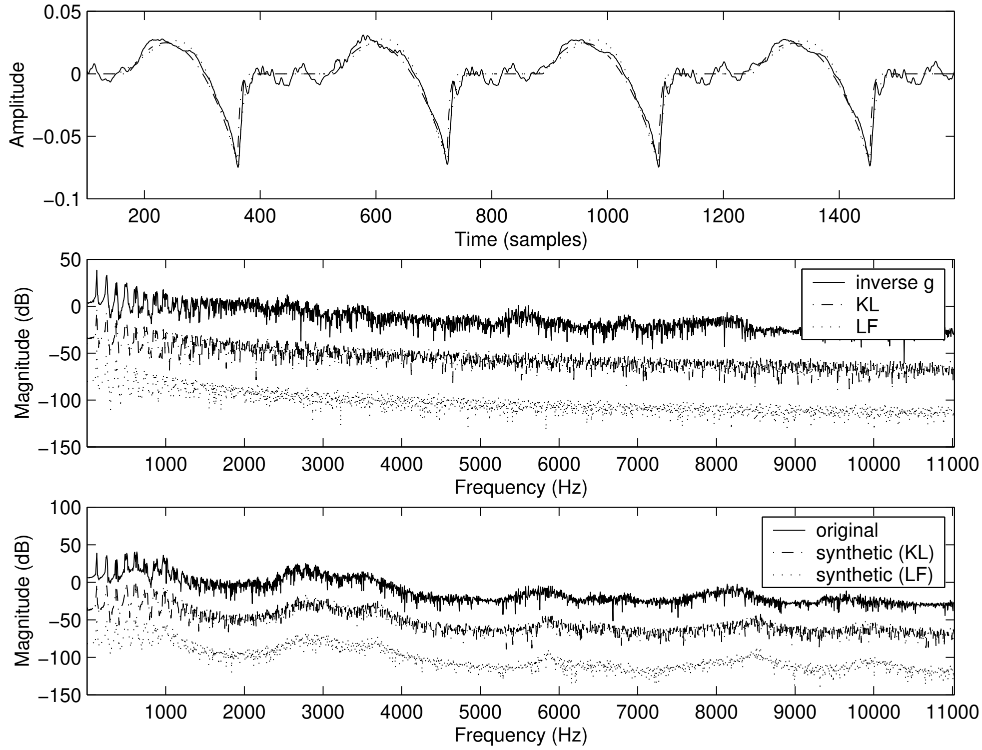

Figure 3.28: Source-filter deconvolution results for a low-pitched baritone sung vowel /a/.

Figure 3.28 illustrates the effectiveness of the source-filter deconvolution algorithm. These results are estimated from a baritone sung vowel /a/ at pitch B2. The first plot overlays the inverse filtered glottal excitation, the KLGLOTT88 and the LF synthetic derivative glottal waves. The parameters for the KLGLOTT88 model are obtained at the source-filter decomposition step. The LF model parameters are obtained by fitting the inverse filtered signal to the LF model. The LF fitting method is illustrated in Section 3.3. Similar to the two-mass synthetic data case, some ripples occur during the closed phase. These inevitable ripples appear to result from the source-tract interaction that is beyond the assumptions of the source-filter model. The second plot shows a snapshot of the spectra of the inverse-filtered glottal excitation and two fitted synthetic derivative glottal waves. The third plot shows the spectra of the original recording, the synthetic sound generated by exciting the estimated vocal tract filter with the LF synthetic derivative glottal wave and the synthetic sound generated by the KLGLOTT88 synthetic derivative glottal wave excitation. Both the

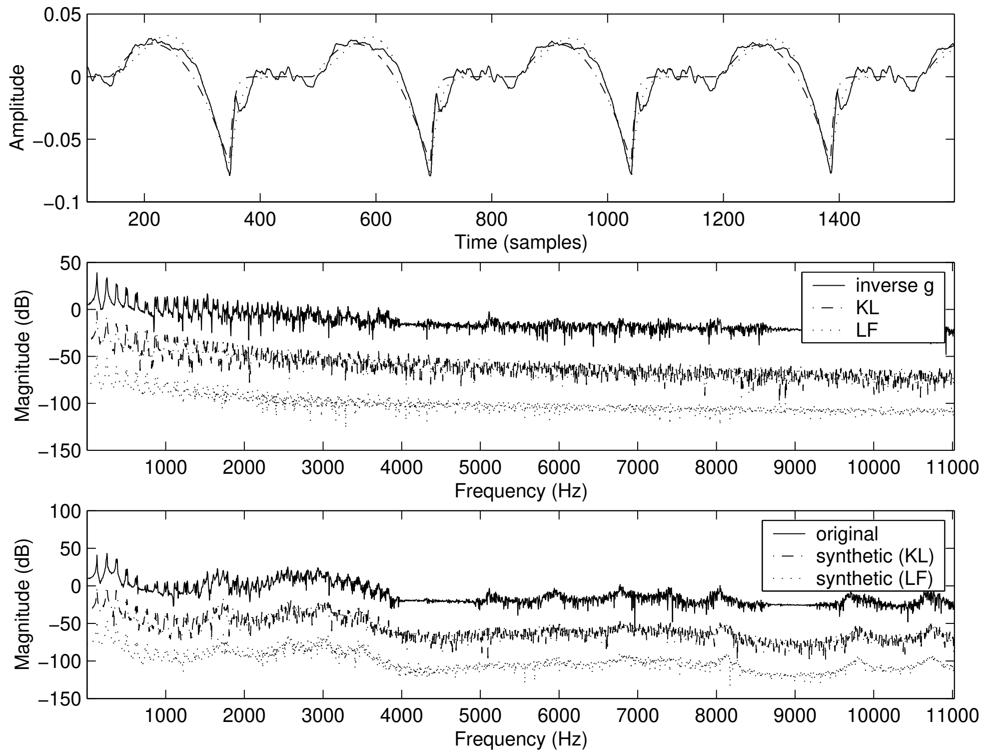

Figure 3.29: Source-filter deconvolution results for a low-pitched baritone sung vowel /i/.

frequency responses of the synthetic vowels match the frequency response of the original recording quite well. In Figures 3.28 and 3.29, we have separated these spectra by scaling their magnitude for better viewing. Figure 3.29 shows the estimated results from a baritone sung vowel /i/ at pitch B2. Vowel /i/ is supposed to have larger source-tract interaction. However, the results shown here are quite similar to the results of vowel /a/. There is no indication of serious degradation on the performance of the SUMT estimation.

To validate that the error surfaces with respect to OQ are almost convex around the optimal value of OQ, Figure 3.30 plots a sequence of error surfaces for a low-pitched baritone sung vowel /i/. The estimated OQ's (diamonds) for each period are shown in the figure along with a glottal period index number. The minimum values of these error surfaces are slightly adjusted for better viewing of the estimated OQ sequence. One can observe that the error surfaces are well-behaved and the estimated OQ sequence is also smoothed.

Although the function of the source-filter deconvolution algorithm is to extract the vocal

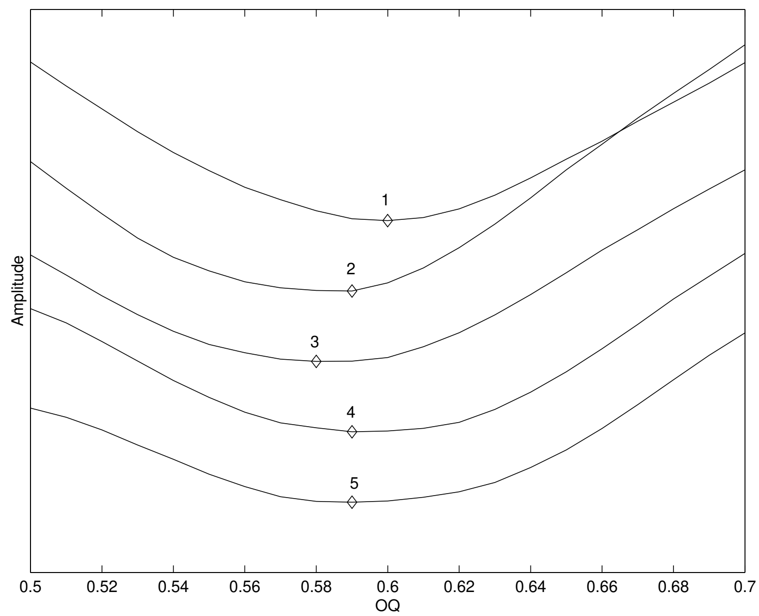

Figure 3.30: A sequence of error surfaces with respect to OQ for a low-pitched baritone sung vowel /i/.

tract filter in the proposed analysis procedure, the algorithm by itself can be an effective analysis/re-synthesis method for non-breathy voices. If a synthetic vowel is generated from the assumed model as in Figure 3.2 with parameters extracted from the source-filter deconvolution algorithm, the synthetic vowel sounds almost like the original singing based on informal subjective listening tests. The subjective listening tests use the method of triple-stimulus with hidden reference. For each test, three sound files (R-A-B) are played, where R is the reference signal (original recording) and the hidden reference that is identical to R is randomly assigned to either A or B. The unassigned one is the synthetic sound file. The listener is asked to judge the impairments of A compared to R and B compared to R according to the ITU-R 5-point impairment scale [37], in which a score 5 indicates that the subject cannot perceive the difference between two testing sound files, and a score 4 indicates the difference between two sound files are perceptible but not annoying. There are 10 untrained subjects in these listening tests. The maximum difference between the the impairments of A compared to R and B compared to R is no larger than 0.2. Therefore, the synthetic sounds are perceived almost identical to the original ones. Analysis/re-synthesis sound examples can be found at http://www-ccrma.stanford.edu/˜vickylu/thesis/.

3.2 Estimating glottal aspiration noise

Glottal source modeling has been shown to be an important factor for improving the naturalness of speech synthesis [11]. To accommodate variations of vocal textures, the glottal excitation is modeled as the sum of the derivative glottal wave and high-pass aspiration noise. The transformed LF model [26] is proposed for the derivative glottal wave (DGW). Moreover, the high-pass glottal aspiration noise is represented as pitch-synchronous, amplitude-modulated Gaussian noise in the noise residual model.

In Section 3.1, a source-filter deconvolution procedure is developed to obtain glottal excitation estimates from real recordings [46]. The focus of this section is to extract the aspiration noise component from the glottal excitation estimates such that we can parameterize the noise residual model properly in our synthesis model. Since the glottal excitation is modeled as the sum of the DGW, which is comparatively smooth, and the high-pass aspiration noise, which is stochastic in nature, we can approach their separation as a denoising problem.

Several wavelet thresholding and best-basis thresholding methods are studied and compared in order to obtain reliable estimates of the glottal aspiration noise. The comparison results conclude that the best-basis (wavelet packet analysis) soft-thresholding method is the most effective way to extract the glottal aspiration noise [48].

3.2.1 Denoising via wavelet or best basis thresholding

This section summarizes the basic concepts of denoising via wavelet and best-basis thresholding. The thresholding methods compared in the next section are also described here.

Assume we have a noisy data record $X = f + W$ in the time domain, where f has a certain degree of regularity[‖] [53] and W is a zero-mean Gaussian white noise sequence with variance σ^2. The noisy data, X, can also be represented in another basis $B = \{g_m\}_{m=0}^{N-1}$ using new coordinates as

$$X_B[m] = f_B[m] + W_B[m], \tag{3.22}$$

‖If the degree of the regularity is an integer, it is the order of the differentiability of f.

where $X_B[m] = \langle X, g_m \rangle$, $f_B[m] = \langle f, g_m \rangle$ and $W_B[m] = \langle W, g_m \rangle$. Here, only orthonormal bases are considered.

Since W is a zero-mean white noise sequence with variance σ^2, $W_B[m]$ is also a white noise sequence with variance σ^2, for any choice of orthonormal basis. If the new basis is chosen such that f can be compactly represented by a few large coefficients $f_B[m]$, thresholding out the smaller coefficients of $X_B[m]$ corresponding to $W_B[m]$ will reduce the effect of the noise. This is the basic idea of denoising via best-basis thresholding.

The thresholding estimator of f given $X = f + W$ can be expressed as

$$\tilde{F} = \sum_{m=0}^{N-1} \sigma_T(X_B[m]) g_m. \tag{3.23}$$

where $\tilde{F}$ denotes the "denoised" vector in terms of the best basis, and $\sigma_T(x)$ is the thresholding function.

Thresholding method

There are two thresholding alternatives: hard and soft thresholdings. A "hard thresholding" can be implemented as

$$\sigma_T(x) = \begin{cases} x & \text{if } |x| > T \\ 0 & \text{if } |x| \leq T \end{cases} \tag{3.24}$$

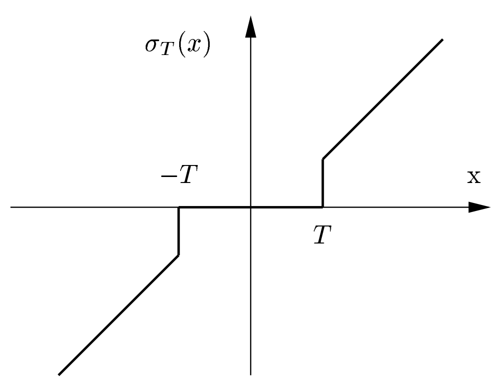

Figure 3.31: Hard thresholding

A "soft thresholding" can be defined [53] as, e.g.,

$$\sigma_T(x) = \begin{cases} x - T & \text{if } x \geq T \\ x + T & \text{if } x \leq -T \\ 0 & \text{if } |x| \leq T \end{cases} \tag{3.25}$$

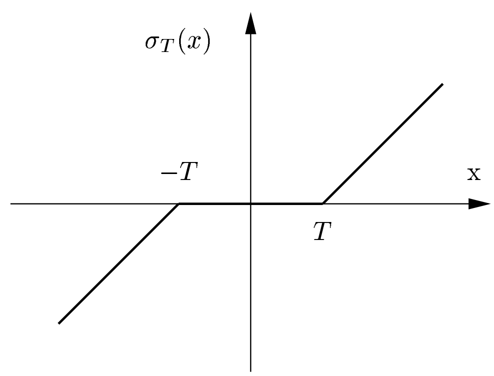

Figure 3.32: Soft thresholding

This definition of soft thresholding minimizes a quadratic distance to the data, penalized by an L_1 norm, i.e.,

$$\sigma_T(X) = arg \min_Y \left(\|Y - X\|_2 + 2T\|Y\|_1 \right). \tag{3.26}$$

Since the thresholded coefficients $|\sigma_T(X_B[m])|$ are reduced by T in the soft-thresholding case, and since the noise coefficients $|W_B[m]|$ are rarely above the threshold T, we have that $|\sigma_T(X_B[m])| < |f_B[m]|$ with very high probability. Therefore, the soft-thresholding estimate reconstructed from large amplitude coefficient consists primarily of attenuated signal components from f, with very little contribution from the noise W.

The *risk* of the thresholding estimator is defined as

$$r_t(f) = E\left\{ \|f - \tilde{F}\|^2 \right\}. \tag{3.27}$$

The risk is therefore the expected value of the squared norm of the eliminated coefficients in the best-basis expansion for orthonormal bases.

Threshold Evaluation

There are several ways to determine the threshold value T. In general, the threshold is chosen so that there is a high probability it will be just above the maximum level of the noise coefficients $|W_B[m]|$. For comparisons in the next section, three major criteria for determining the threshold value are considered: the universal threshold, SURE threshold, and multiscale SURE thresholds.

The *universal threshold* is set to be [20]

$$T_{\text{u}} \triangleq \sigma\sqrt{2\ln N}$$

It can be shown [53] that the maximum amplitude of the noise coefficient W_B has a very high probability of being just below this threshold as the data size N increases, assuming W_B is a vector of N independent Gaussian random variables of variance σ^2.

The SURE threshold is the threshold value that minimizes the Stein unbiased risk estimator [22]

$$\tilde{r}_t(f,T) = \sum_{m=0}^{N-1} \Phi(|X_B[m]|^2) \tag{3.28}$$

where

$$\Phi(u) = \begin{cases} u - \sigma^2 & \text{if } u \leq T^2 \\ \sigma^2 + T^2 & \text{if } u > T^2. \end{cases} \tag{3.29}$$

Let $X_B^r[k] = X_B[m_k]$ denote the coefficients $X_B[m]$ sorted in order of decreasing magnitude, i.e., such that $|X_B^r[k]| \geq |X_B^r[k+1]|$ for $1 \leq k < N$. If $|X_B^r[l-1]| > T \geq |X_B^r[l]|$, Eq. (3.28) can be rewritten as

$$\tilde{r}_t(f,T) = \sum_{k=l}^{N} |X_B^r[k]|^2 - (N-l+1)\sigma^2 + (l-1)(\sigma^2 + T^2). \tag{3.30}$$

To minimize the above equation, T has to be $|X_B^r[l]|$. Therefore, the SURE threshold

$$T_s = arg\min_{|X_B^r[l]|, l=1,\ldots,N} \tilde{r}_t(f, |X_B^r[l]|) \tag{3.31}$$

where $\tilde{r}_t(f, |X_B^r[l]|) = \sum_{k=l}^{N} |X_B^r[k]|^2 - (N-l+1)\sigma^2 + (l-1)(\sigma^2 + |X_B^r[k]|^2)$.

The original SURE threshold scheme uses the same threshold for every scale of the wavelet decomposition and its thresholding value is calculated from the finest-scale wavelet coefficients. In the multiscale scheme, a different SURE threshold is calculated by minimizing the risk at each scale of the wavelet decomposition. By using a different threshold for each scale, one can use a smaller threshold at coarser scales where the noise coefficients $|W_B|$ are smaller.

The best-basis thresholding is essentially the wavelet packet analysis which minimizes the Stein unbiased risk estimate. That is, the best basis B is chosen by minimizing Eq. (3.28). A hard or soft thresholding is then applied to threshold the resulting wavelet packet coefficients by the threshold value, $\sigma\sqrt{2\ln P}$, where P is the number of distinct vectors occurring among all bases in the library. For a wavelet packet, $P = N\log_2(N)$.

By using different threshold evaluation criteria with the choice of either hard or soft thresholding, the following thresholding methods will be compared in Section 3.2.3: Universal hard (or soft) thresholding, SURE soft thresholding, multiscale SURE soft thresholding and best-basis hard (or soft) thresholding. SMS (sinusoidal modeling) decomposition [66], a Short-time Fourier analysis method, is also compared.

3.2.2 Implementation issues

For our application, the objective is to extract the aspiration noise component W from the inversed glottal excitation estimates X, where X consists of the aspiration noise component and the DGW [47] component f, i.e., $X = f + W$.

To adapt the denoising methods of the previous section to our application, we need to consider the following issues.

- Select the wavelet filter: One needs to find a good basis family that can represent the derivative glottal wave component, f, as compactly as possible.
- Estimate the variance of noise component: The variance estimate is necessary in order to evaluate the threshold value.
- Select the wavelet decomposition level.

- Select the thresholding method: Hard and soft thresholdings are two thresholding alternatives.
- Select the threshold evaluation criteria: Universal, Stein Unbiased Risk Estimate (SURE), and multiscale SURE thresholds are three major criteria for threshold evaluation in this study.

In the previous section, thresholding methods and threshold evaluation criteria have been discussed. This section will describe how to resolve the remaining implementation issues.

First, we need to find a good basis family that can compactly represent the derivative glottal wave component. An energy compactness criterion is used to select the wavelet filter. The energy compactness criterion compares the sparseness of the representation using different wavelet families. It is defined as one over the number of coefficients needed to accumulated 0.9 of the total energy of the signal. The smaller the number is, the more compact it is. In this study, only orthonormal wavelets are considered so that we can easily obtain the signal energy from the wavelet coefficients without scaling.

Since the true f is not available, to determine which wavelet filter to use for our application, the energy compactness is tested on a synthesized DGW. The synthesized DGW is obtained from the LF model [26]. This synthetic DGW is chosen to have the same degree of breathiness and the same average pitch as the glottal excitation obtained from real data via inverse filtering. It is synthesized using a typical breathy phonation wave-shape parameter ($Rd = 2.9$) which was obtained in a previous study [47]. The length of the wavelet filter depends on the pitch period of the derivative glottal wave. The shorter the pitch period is, the shorter the wavelet filter is. For the data using in this study, the 7^{th} order Daubechies wavelet was chosen via this energy compactness criterion.

In order to find the threshold value, we also need to estimate the variance of the noise sequence. In the thresholding denoising theory, the noise sequence is assumed to be a zero-mean, i.i.d. Gaussian white noise process with variance σ^2. Hence, the wavelet coefficients obtained via orthonormal wavelet expansion are also a zero-mean, Gaussian white noise sequence having variance σ^2. The standard deviation σ is typically estimated as [61]

$$\tilde{\sigma} = \frac{M_x}{0.6745}, \tag{3.32}$$

where M_x is the median of the absolute value of the wavelet coefficients of the noise sequence at the finest scale. Equation (3.32) is derived from the fact that the expected median magnitude of a zero mean Gaussian white noise sequence of variance σ^2 is about 0.6745σ.[**] Since the likelihood of the aspiration noise is proportional to the flow and inversely proportional to the radius of the aperture [14], the aspiration noise is pitch synchronous with the derivative glottal wave. Because of the pitch-synchronous nature of the aspiration noise, its wavelet coefficients in the finest-scale decomposition are also pitch-synchronous. We will under-estimate the threshold evaluated from the variance via Eq. (3.32). In order to extract large noise components around glottal closure instants, the variance of the noise is actually estimated via the unbiased sample variance[††] of noise components with magnitude larger than 0.1 of the maximum noise magnitude. This results in a larger estimate.

In this thesis, the wavelet decomposition is implemented as a two-channel subband coder [52] [75]. In our implementation, the approximation component of the wavelet decomposition is kept without thresholding. At the maximum wavelet decomposition level, the signal is decomposed to only one point. The larger the decomposition level is, the more the denoised signal gets low-passed. Hence, sharp corners of the glottal closure instants of the derivative glottal wave will not be well modeled. This in turn contributes the low frequency components of the estimated noise residual. On the other hand, if the decomposition level is too small, the estimated noise will be under-estimated since too much noise is left out in the denoised signal. Therefore, a criterion for selecting decomposition level is needed.

An empirical criterion for selecting the wavelet decomposition level is derived from synthetic data simulations. The criterion is to choose the largest level such that the magnitude of normalized low-pass filtered noise residual[‡‡] is small enough. The normalized low-pass filtered noise residual is normalized over the magnitude of the estimated noise residual. This criterion will be quantified for the best-basis soft thresholding method and it will be illustrated in the next section.

[**] Let x be a zero mean Gaussian white noise sequence of variance σ^2. Since M_x is the median of the absolute value of $x's$, probability $P(|x| < M_x) = 0.5$, i.e., $P(-\frac{M_x}{\sigma} < \frac{x}{\sigma} < \frac{M_x}{\sigma}) = 0.5$. For Gaussian distribution, $\frac{M_x}{\sigma} = 0.6745$.

[††] Unbiased sample variance is defined as $\frac{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2}{N-1}$, where $\overline{x}$ is the sample mean $\frac{\sum_{i=1}^{N} x_i}{N}$.

[‡‡] The noise residual is the difference between the original noisy input and the denoised result.

3.2.3 Results and Discussion

To compare performances of different denoising methods, denoising procedures are applied to synthetic input data. The input noisy data is the synthetic DGW (in breathy phonation) plus the pitch synchronously windowed Gaussian white noise (using a Hanning window with 0.5 duty cycle). The denoising results for one particular realization are shown in Figures 3.33 to 3.38. From the denoising results, it can be observed that universal thresholding methods tend to under-estimate the threshold value. Hence, large noise components around glottal closure instants are not extracted and the pitch-synchronous nature of the noise residual is less prominent.

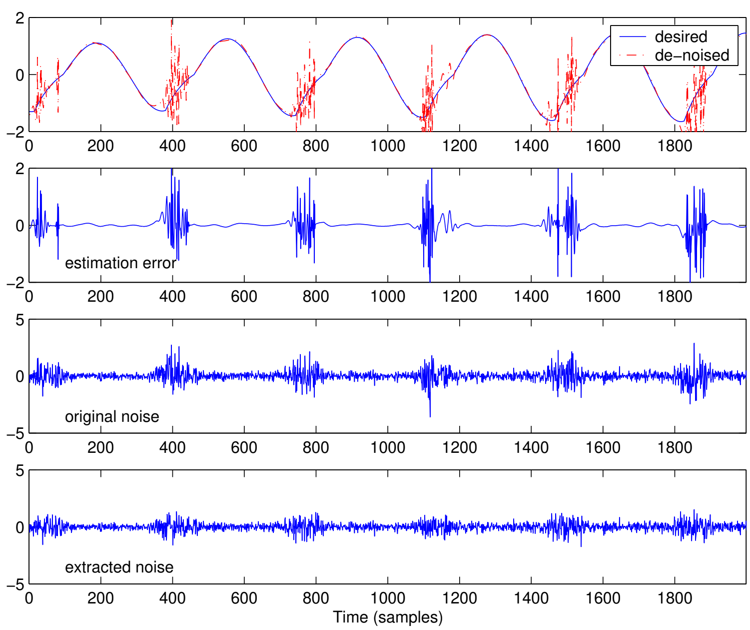

Figure 3.33: Universal hard thresholding

The same denoising procedures were applied to 100 runs of synthetic input data and the average signal to noise ratio (SNR) is calculated. The SNR is measured as energy ratio between the clean synthetic DGW and the estimation error of denoised DGW. The original SNR before denoising is 5.83(dB). Table 3.5 shows the resulting average SNR for different denoising methods in the best scenario. In this case, wavelet and best-basis decompositions are performed at a level which gives the best resulting signal to noise ratio. Moreover, the best SNR for SMS was achieved by changing the number of sinusoidals.

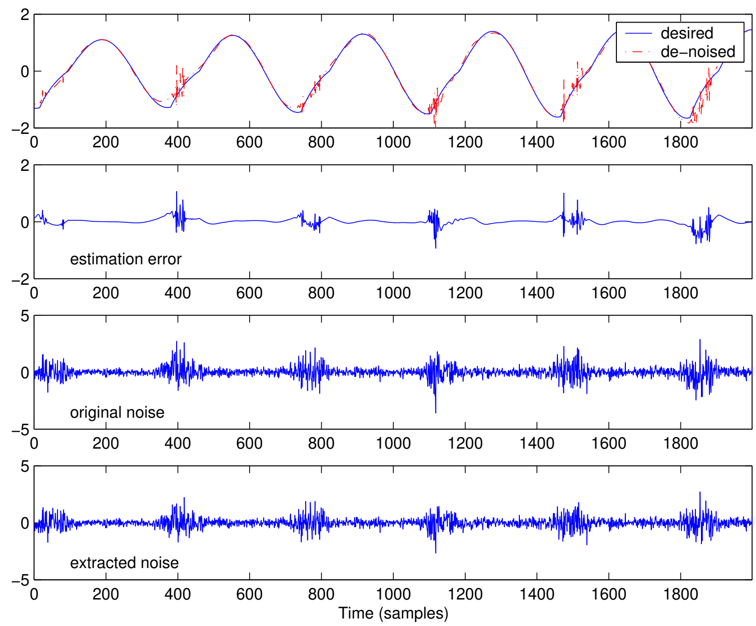

Figure 3.34: Universal soft thresholding

Comparing the average SNR of the denoised signal (the estimated DGW), except SMS, the best-basis soft-thresholding method achieves the highest SNR. The performance of the best-basis hard-thresholding method is very close to the best-basis soft-thresholding method. SURE soft thresholding comes out third. Multiscale SURE and SURE soft thresholding methods have comparable SNR values. Universal hard thresholding and universal soft thresholding left in too much noise in the denoised signal.

Among the thresholding methods, the best-basis soft-thresholding method emerges as the best option we have found thus far for extracting the glottal aspiration noise since it yields the best SNR. However, it is impossible to choose the decomposition level by selecting the one which gives the best SNR for real data. Therefore, a criterion for selecting decomposition level is needed. As described in the previous section, the selection criterion uses the magnitude of the low-pass filtered estimated noise residual to select the decomposition level which results in the best SNR. Denote the normalized low-pass filtered estimated noise residual as R. For the best-basis soft thresholding, the selected level is the largest level that satisfies

$$R < (\max(R) - \min(R)) \cdot 0.04 + \min(R).$$

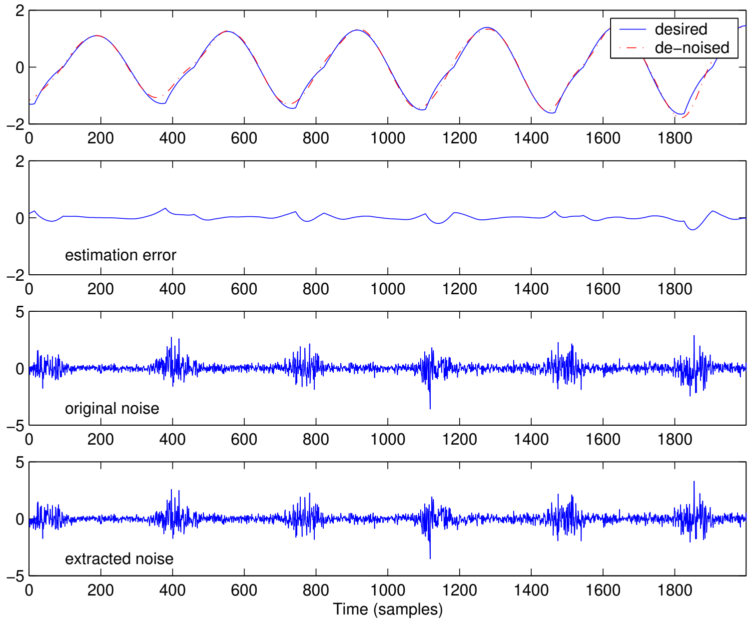

Figure 3.35: SURE soft thresholding

Figure 3.39 illustrates the reason for choosing this criterion. The decomposition level is labeled along the x-axis. The y-axis shows the normalized low-pass filtered estimated noise residual, R. These results are generated from 100 runs of synthetic data simulations via best-basis soft thresholding. The original SNR is 5.83 dB. The desired level is 5 which can result in the best SNR after denoising. From this figure, you can see that the normalized low-pass filtered estimated noise residual starts to grow faster at decomposition level 5. By using the proposed empirical level selection criterion, 95 out of 100 runs are correctly estimated. Others are estimated as level 4.

To show that the best-basis soft thresholding method can faithfully extract glottal aspiration noise, it has been applied to real breathy sung vowels. Figure 3.40 shows the extracted noises for various decomposition levels for a breathy sung vowel /a/. The decomposition level is labeled along the y-axis. One can observe that the low-pass filtered estimated noise residual increases with the decomposition level. On the other hand, the aspiration noise will be under-estimated if the decomposition level is too small. Figure 3.41 shows the denoising result with the decomposition level selected from the proposed criterion. The upper plot shows the inverse-filtered glottal excitation and the denoised signal. The bottom plot is the extracted aspiration noise. The glottal closure instants (GCI) and

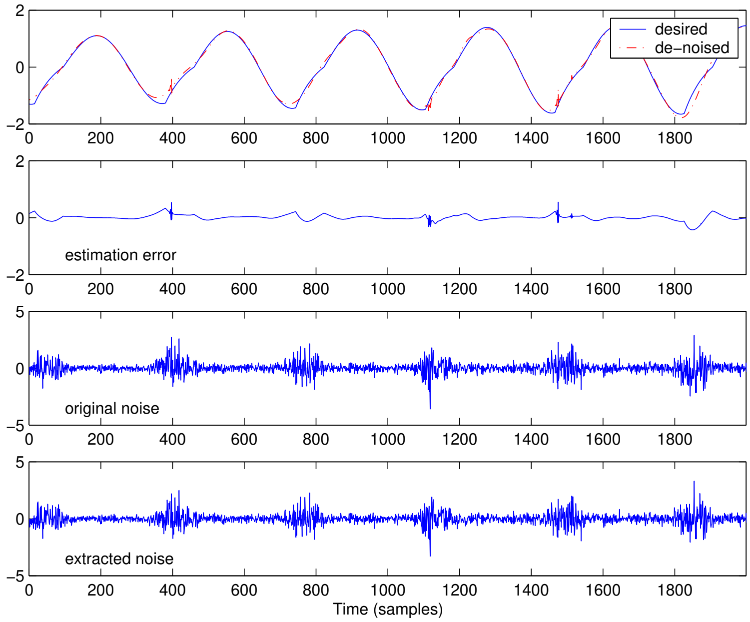

Figure 3.36: Multiscale SURE soft thresholding

the glottal open instants (GOI) are also indicated in the bottom plot. From this figure, it is clearly that noise bursts occur right after the glottal closure instants and approximately after the glottal open instants. This result is consistent with the study by Cook [14].

Similarly, it is necessary to find a criterion to choose the number of sinusoidal components for SMS method. If the SMS model consists of more sinusoidal components, less noise will be extracted. On the other hand, if the number of sinusoidal components is too small, the sharp edges around glottal closure instants (GCI) will not be accurately modeled. An empirical criterion is also established from synthetic data simulations. However, it does not work well when it is applied to real data as shown in Figure 3.42. Again, the upper plot shows the inverse-filtered glottal excitation and the denoised signal. The bottom plot is the extracted aspiration noise. More low frequency drift is observed in the extracted aspiration noise. It indicates that the proposed criterion for SMS under-estimate the number of sinusoidal components. Although SMS gives the best SNR in synthetic data simulations. it is not straightforward to apply SMS to real data. Therefore, it concludes that best-basis (wavelet packet analysis) soft thresholding method is the best option to extract the noise residual.

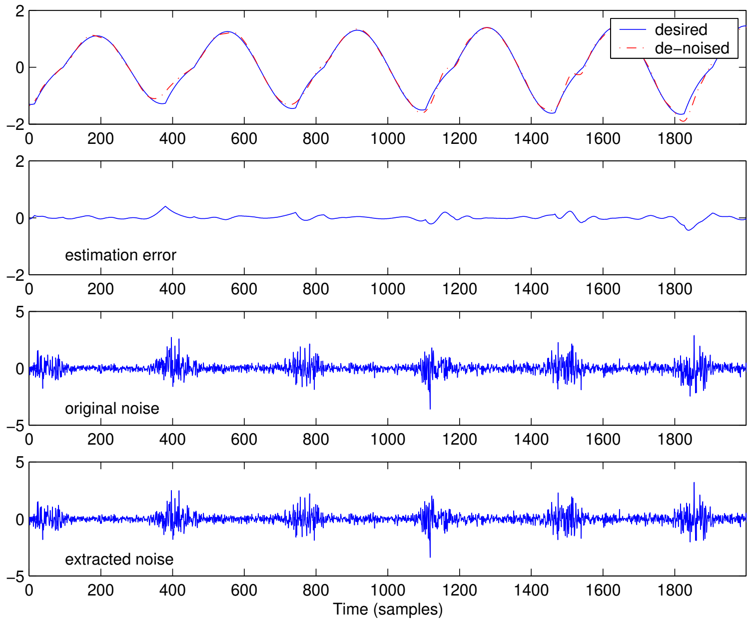

Figure 3.37: Best basis soft thresholding

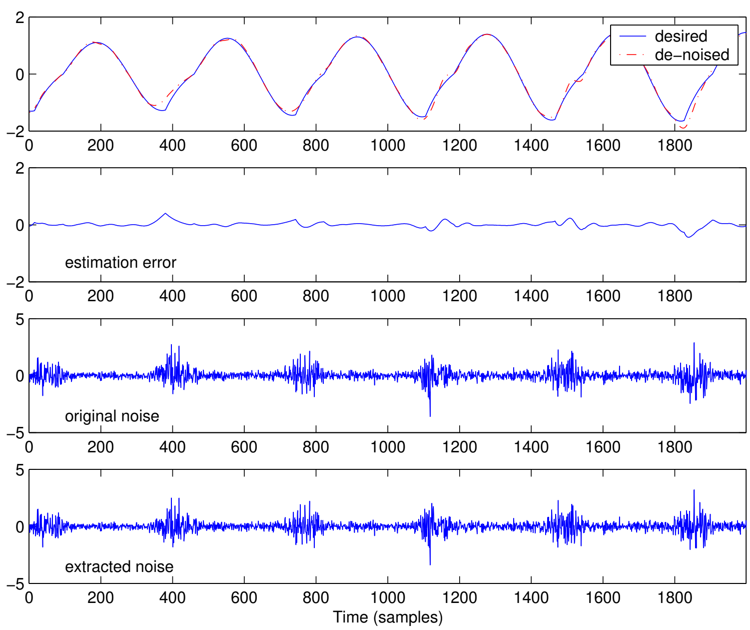

Figure 3.38: Best basis hard thresholding

Table 3.5: Comparisons of various denoising methods for synthetic data.

Method	Denoised SNR (dB)
Universal Hard	9.56
Universal Soft	16.95
Multiscale SURE Soft	17.88
SURE Soft	18.33
Best-basis Hard	20.93
Best-basis Soft	20.96
SMS	24.65

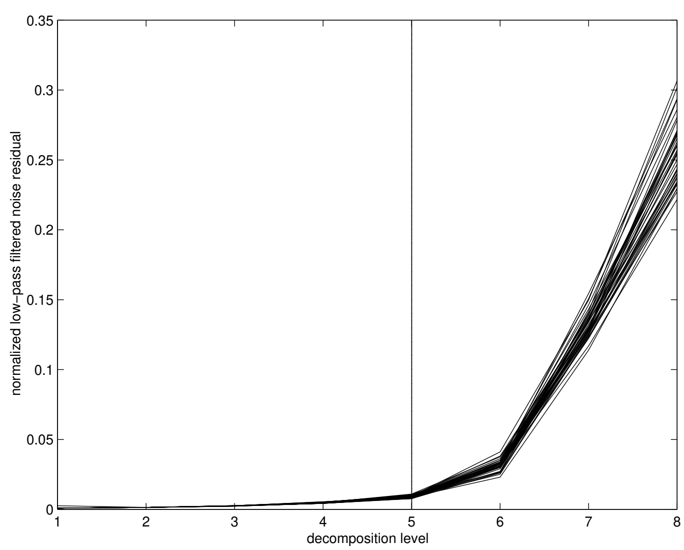

Figure 3.39: Normalized low-pass filtered estimated noise residual.

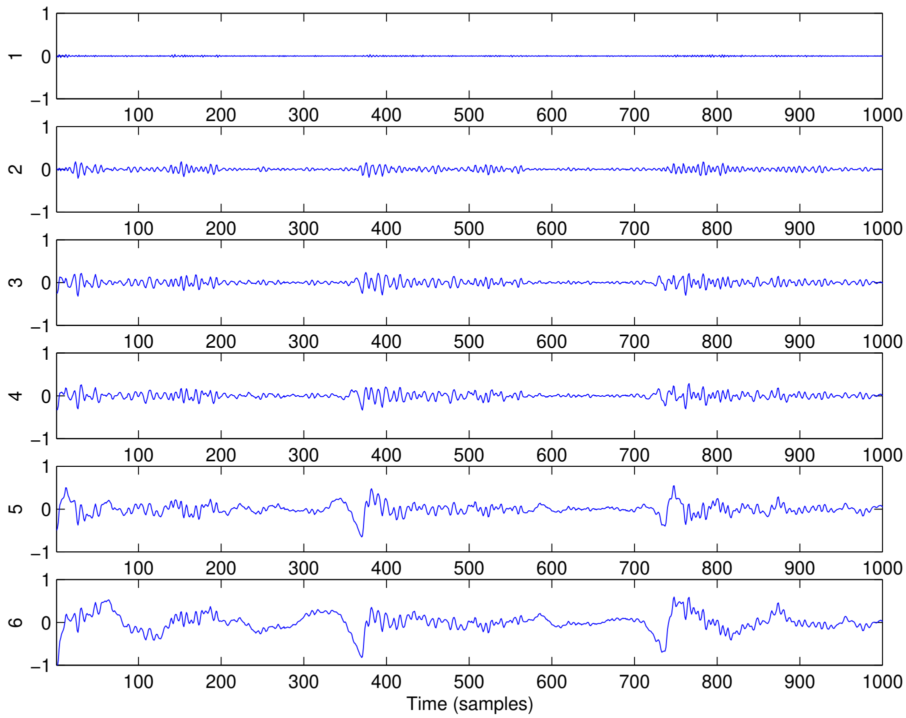

Figure 3.40: Extracted glottal aspiration noise for a breathy sung /a/ with various decomposition levels.

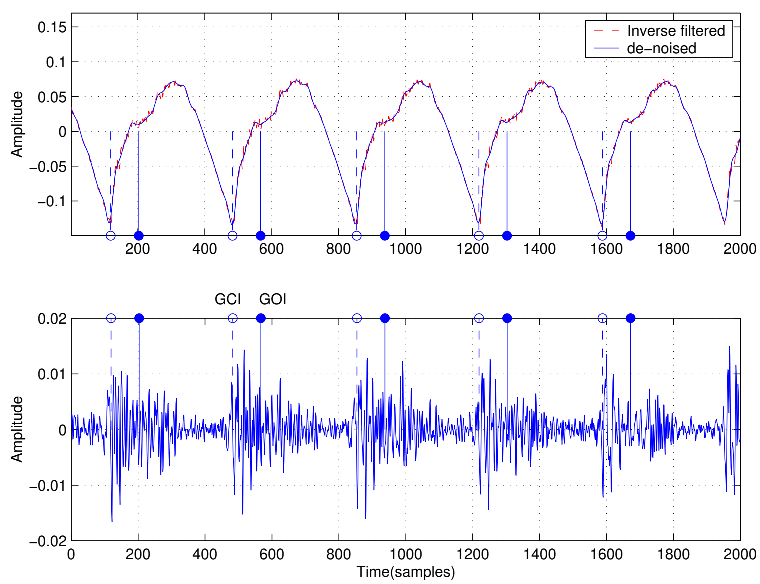

Figure 3.41: Extracted glottal aspiration noise for a breathy sung /a/ via best-basis soft-thresholding.

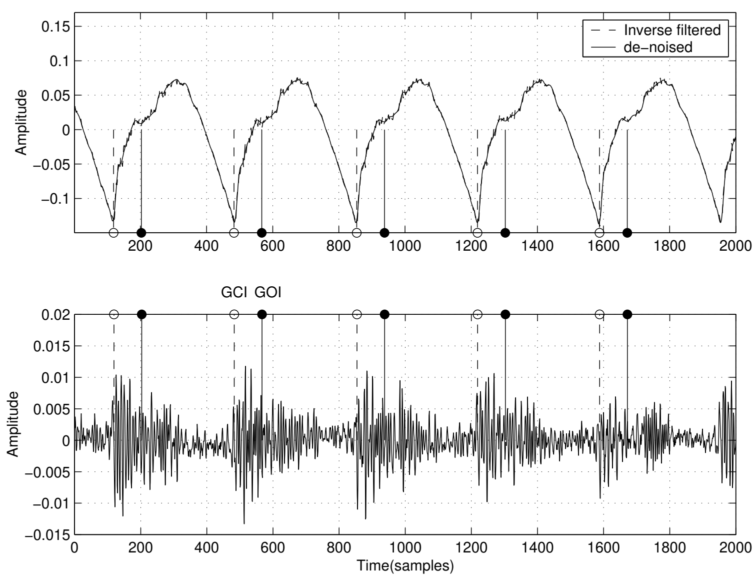

Figure 3.42: Extracted glottal aspiration noise for a breathy sung /a/ via SMS.

A statistical analysis of the noise residual for breathy vowels was carried out. First, the degree of pitch synchrony was evaluated by measuring the intervals between adjacent peaks in the smoothed noise time-domain magnitude envelope. The intervals were normalized such that an interval of 1 is one period. A histogram of these normalized amplitude-envelope peak intervals (see Figure 3.43) shows that the noise residual is highly pitch synchronous. We also found that the average noise peaks occurs at a lag of 10% of the period after the glottal closure instants. The corresponding interval between the major excitation and the noise bursts is less than 1 ms. This is consistent with D.J. Hermes's experiments [31] with synthetic noise.

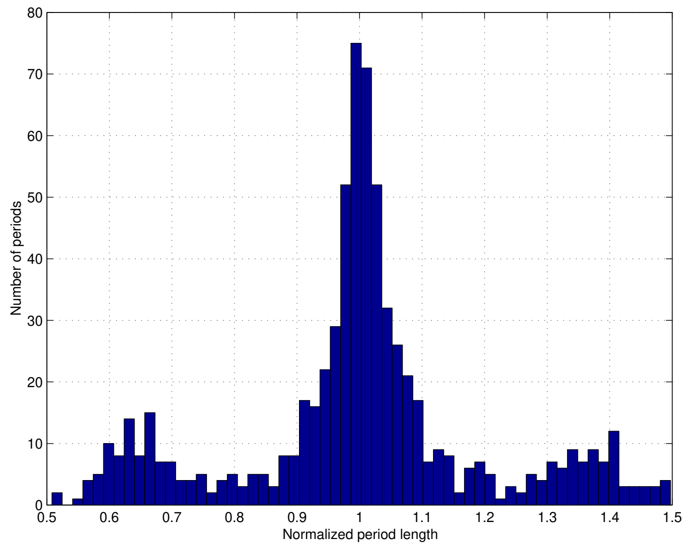

Figure 3.43: Histogram of normalized noise period length.

In order to determine the variance of the Gaussian noise around the glottal closure instants and the duty cycle of the amplitude modulation, the strength and the duty cycle of the noise bursts were also evaluated. The strength of the noise burst is measured as the normalized amplitude-envelope peak for each glottal period. The amplitude is normalized by the strength of the glottal excitation, E_e. It was found that the normalized strength could be approximated as a constant except at the onset and the offset of the vowel. The average strength was 0.04 in our experiments. The duty cycle is measured as the normalized time duration for the noise bursts decay to 10% above the noise floor. The duty cycle is less consistent across different sung tones. The average duty cycle measured was 42%. In addition to the strength and the duty cycle, the noise floor is also measured from the extracted noise residual.

For the purpose of controlling breathiness of synthetic voices, a parametric model has established from the statistics of these noise residual. Chapter 4 is devoted to discuss this vocal texture control model.

3.3 LF model fitting

Since the LF model has a more complicated mathematical description, it can not easily be estimated directly in the source-filter deconvolution step. Instead, the KLGLOTT88 model is used in the source-filter deconvolution step to achieve the convex formulation. Once the estimated derivative glottal wave is obtained, the LF model is used to re-parameterize the derivative glottal wave period-by-period. Two steps are involved in LF model fitting. For each period of the derivative glottal wave, the LF model timing parameters and glottal excitation, E_e, are first retrieved by direct estimation methods [77] [2] which measure these parameters directly from the time-domain waveform. Since the estimated derivative glottal wave is usually noisy due to model mismatch, the direct estimation method may yield unreliable results. The LF model parameters are further refined via constrained nonlinear optimization using the sequential quadratic programming method [34].

The overall procedure of LF model fitting is illustrated here. In the following discussion, for each glottal cycle, t_e is the time index where the derivative glottal wave reaches its local minimum and t_p is the time index where the glottal wave has local maximum. Moreover, t_o is the time index where each glottal cycle starts, and t_c is the time index where the return phase of derivative glottal wave ends.

- A good initial estimate is essential to the nonlinear optimization step to achieve global minimum. Since the estimated derivative glottal wave often contains ripples and noise, it is low-pass filtered before performing the direct estimation. The low-pass filter is implemented by convolution with a 7-point Blackman window.
- Find the excitation strength E_e and the time index t_e by locating local minimum of the estimated derivative glottal wave.
- Search for t_p as the first zero-crossing from the left of t_e.

- t_o is the time point to the left of t_p where the estimated derivative glottal wave first becomes lower than a small threshold. t_o is also constrained by the estimated value of the open quotient OQ from Equation (2.14).

- t_c is determined as the first sample after t_e that is smaller than some threshold.

- LF parameter T_a is estimated as $(t_c - t_e) \cdot 2/3$.

- LF parameter T_p is estimated as $(t_p - t_o)$.

- LF parameter T_e is estimated as $(t_e - t_o)$.

- LF parameter T_c is estimated as $(t_c - t_o)$.

- From previous steps, one obtains the first approximation of the LF timing parameters (T_p, T_e, T_a, T_c) and excitation strength E_e. To further refine these estimates, constrained nonlinear optimization is applied. In the constrained nonlinear optimization step, E_e, T_e and T_c are kept unchanged. T_p is constrained to be within $\pm 20\%$ of the estimated value from the above direct estimation method. Moreover, T_a is confined between 0 and $T_c - T_e$. The error criterion is the time domain root mean square error of one glottal pulse period, starting at t_o and ending at t_c.

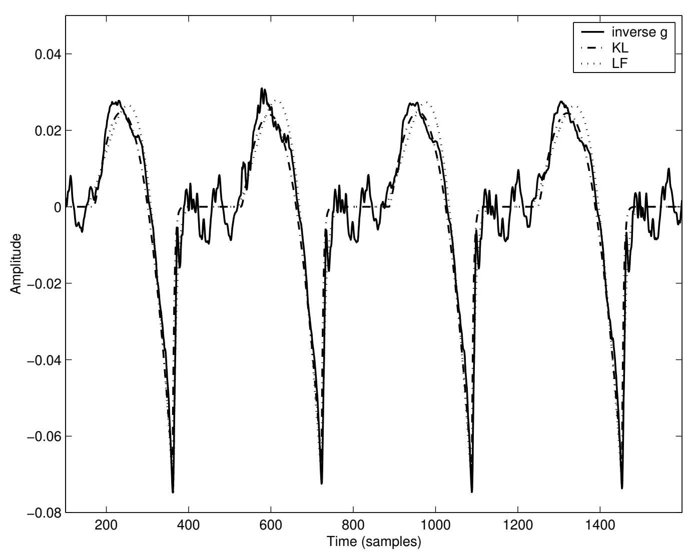

Figure 3.44: Fitting the derivative glottal wave by LF model.

Figure 3.44 shows one example of LF model fitting. The input data is a baritone sung vowel /a/ at pitch B2. This figure overlays the inverse filtered glottal excitation and the fitted derivative glottal waves for both the KLGLOTT88 model and the LF model. Both models yield reasonable fitting results.

Table 3.6: The summary of the LF model parameterizations.

Condition	OQ	F_a	K_k	K_a	R_d
Pressed	0.49	675	1.07	1.23	0.84
Normal	0.64	515.3	1.04	0.95	1.19
Breathy	0.78	161	0.89	1.21	2.90
Low F_o	0.61	349	1.05	0.97	1.30
Medium F_o	0.64	326	1.02	1.05	1.63
High F_o	0.69	300.7	0.90	1.35	2.26
Vowel /a/	0.63	349	0.99	1.10	1.57
Vowel /e/	0.66	318.9	1.01	1.10	1.79
Vowel /i/	0.65	298.3	1.00	1.12	1.81

Table 3.6 summarizes the LF model parameters from the analysis of baritone sung vowels. The results are averaged across different singing phonations, pitch scales, and vowel types. The low pitch is $B2$, the medium pitch is $G3$ and the high pitch is $B3$. Comparing to Karlsson's results for speech data [38], we have a larger R_d parameter and a lower cut-off frequency F_a, i.e., a larger spectral tilt. From the summary, we also see that the deviation constants are close to 1 in most cases; hence, one could predict the LF parameters quite well from a single R_d parameter. Furthermore, we see that increasing the degree of breathiness or the phonation frequency will result in a larger R_d.

In order to show that the proposed analysis procedure in this chapter can perform the task of analysis/re-synthesis, the original recordings and the re-synthesized sound files have been compared via informal listening tests as illustrated in Section 3.1.4. These sound files can be found at http://www-ccrma.stanford.edu/˜vickylu/thesis/. The synthetic voices are generated from the estimated model parameters of the proposed synthesis model in Figure 2.3. It has been shown that synthetic sounds mimic the original recordings quite well.

Chapter 4

Vocal texture control

In previous chapters, a high-quality singing voice synthesis model is proposed with an associated analysis procedure to retrieve the model parameters automatically from the desired voices. This synthesis model is capable of varying vocal textures from "laryngealized" (or "pressed"), to normal, to "breathy" phonation. From the study of model parameterizations by analyzing the baritone recordings with varying vocal textures, it suggests that it is not straightforward nor intuitive to control the vocal texture by direct interpolation between these parameterizations. In this chapter, a simple parametric vocal texture control model is described. This model is used to determine the parameterization of the glottal excitation to achieve the desired vocal texture.

To accommodate variations of vocal textures, the glottal excitation model consists of two elements: the deterministic derivative glottal wave and the aspiration noise.

The derivative glottal wave is modeled by the transformed LF model. The transformed LF model is a parameterized time-domain model of one cycle of the derivative glottal wave. To determine the derivative glottal wave for each glottal cycle, only two parameters need to be specified: (1) the glottal excitation strength, denoted E_e, and (2) a single wave-shape parameter, denoted R_d. The glottal excitation strength provides the amplitude envelope of output voice, while the wave-shape parameter controls the shape of one normalized cycle of the oscillation.

As shown in Figure 2.8, the aspiration noise source consists of two components.* One is the pitch-synchronous, amplitude-modulated Gaussian noise. The other one is the additive Gaussian noise floor. The pitch-synchronous noise component of the aspiration noise source is generated by scaling the amplitude of a zero mean, unit variance, Gaussian white noise sequence with Hanning window centered around the glottal closure instants. In summary, there are four model parameters per glottal cycle for the aspiration noise model. They are: (1) the Gaussian noise floor level B_n, (2) the amplitude of the Hanning window A_n, (3) the duration of the Hanning window, and (4) the lag of the Hanning window midpoint relative to the glottal closure instants L.

Recall that Klatt [41] has summarized two acoustic parameters as important indicators of vocal texture. They are: (1) the degree of aspiration noise at high frequencies in vowels, and (2) the relative strength of the fundamental component of the glottal source wave. In the transformed LF model, the wave-shape parameter, R_d, can effectively control the relative strength of the fundamental component of the glottal source as shown in Equation (2.14). By changing the aspiration noise model parameters, we can also change the degree of aspiration noise at high frequencies. However, it is not intuitive to control vocal textures by these parameters directly. A proper mapping between the degree of breathiness and these direct model parameters is necessary.

Figure 4.1 shows the basic idea of vocal texture control. The idea is that users only need to specify the desired vocal texture and the glottal excitation strength. If the desired vocal texture is non-breathy, the vocal texture model will determine the wave-shape control parameter R_d for the transformed LF model. If the desired vocal texture is breathy, in addition to the wave-shape parameter, the vocal texture model will also determine the corresponding parameters for the noise residual model.

4.1 Non-breathy mode

By analyzing baritone recordings of sustained vowels with varying vocal textures, correlations among the model parameters of the glottal excitation were explored. It was found that

*The spectral shaping filter has been neglected since it is not sufficiently perceptually important.

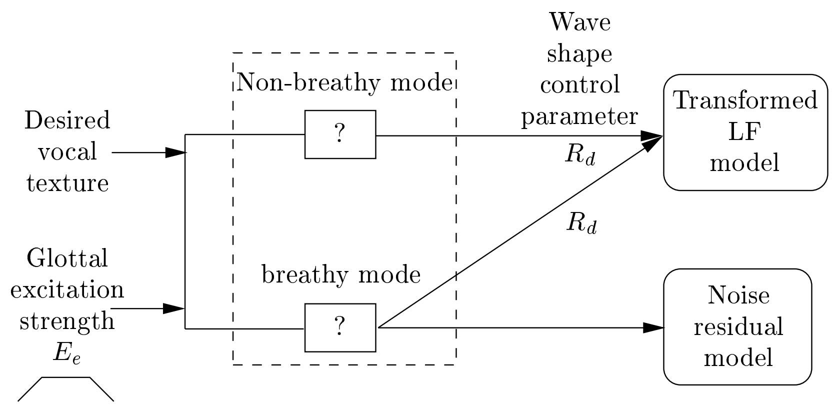

Figure 4.1: Implementation diagram of vocal texture control.

the normalized glottal excitation strength $\tilde{E}_e$ † and the wave-shape parameter R_d are highly correlated for non-breathy phonation modes (normal and pressed mode). By observing the scatter plots of R_d and $\tilde{E}_e$, R_d is chosen to be approximated as an exponential function of $\tilde{E}_e$:

$$a + b \cdot e^{-c \cdot \tilde{E}_e} \tag{4.1}$$

where a, b, and c are scalars. The value of a controls R_d during the sustained portion of singing for each note; b controls R_d when E_e approaches zero (at the beginning of the attacks and at the end of releases), and c controls the decay rate of the exponential model.

Figure 4.2 plots R_d with respect to $\tilde{E}_e$. The upper plot is for normal phonation mode and the bottom plot is for pressed phonation mode. Both raw data points and the fitted exponential model curve are shown in the figure. All of the data points are obtained from baritone singing recordings of sustained vowels (/a/, /e/, /i/) across different pitch scales. From Figure 4.2, it is clear that R_d and $\tilde{E}_e$ are highly correlated for both the normal and pressed phonation modes and they can be well modeled by the exponential model in Equation (4.1).

The data and fitted models for normal and pressed modes are shown together in Figure 4.3 to illustrate the differences between the fitted parameters of the exponential model for different phonations. Pressed phonation corresponds to a smaller R_d value, resulting in a glottal wave shape with abrupt closure and a longer closure time. By fitting the model to

† $\tilde{E}_e \triangleq \frac{1}{g} \cdot E_e$, where g is a scalar gain and $\tilde{E}_e$ is normalized to be within 0 and 1 for each sustained vowel recording.

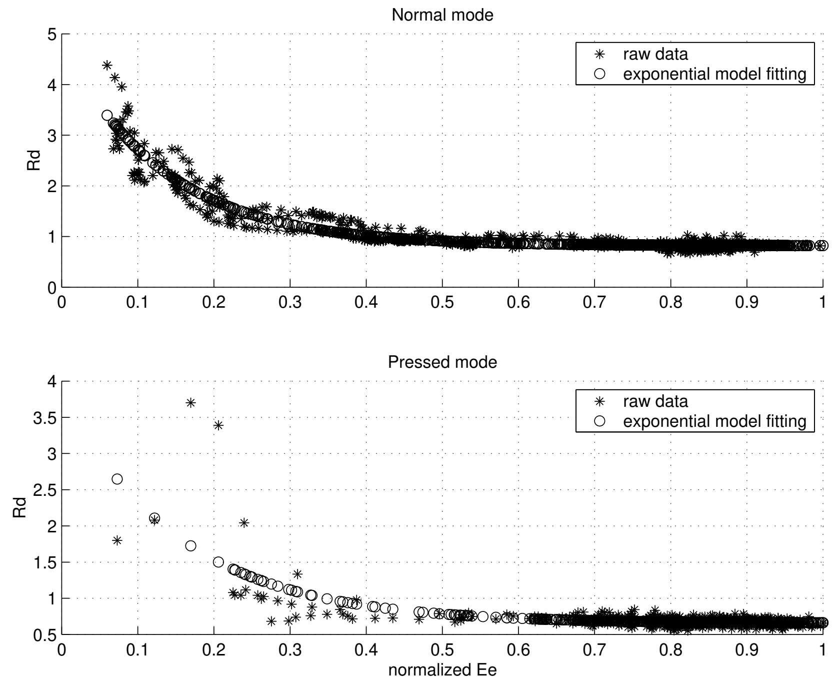

Figure 4.2: The correlation between R_d and the normalized glottal excitation.

real recordings, it was found that the values of a, b, and c are smaller for pressed phonation.

Figure 4.4 shows the method to control vocal textures for non-breathy mode. Two sets of parameters of the exponential model $((a_{\text{pressed}}, b_{\text{pressed}}, c_{\text{pressed}})$ and $(a_{\text{normal}}, b_{\text{normal}}, c_{\text{normal}}))$ for pressed and normal phonations were estimated from real recordings. By simple linear interpolation between these two set of parameters, a new set of parameters of the exponential model for the desired vocal texture is obtained. The value of the wave-shape parameter, R_d, is then predicted from the normalized glottal excitation strength, $\tilde{E}_e$, which is already known from the specified amplitude envelope when synthesizing. Finally, the glottal excitation with the specified R_d is generated from the transformed LF model.

4.2 Breathy mode

For the breathy mode vocal texture control, the noise to harmonic ratio, NHR, is introduced. The NHR is calculated per glottal cycle. It is defined as the high-pass[‡] noise residual

[‡]The cut-off frequency is 2k Hz.

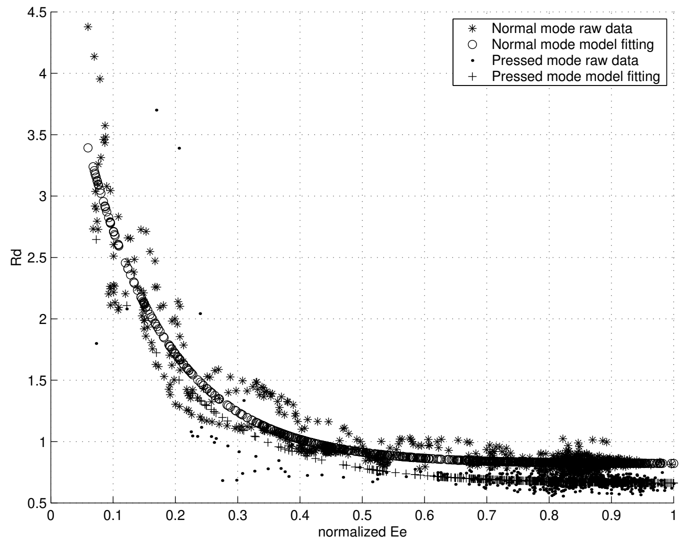

Figure 4.3: The exponential models for both of the pressed and normal phonation modes.

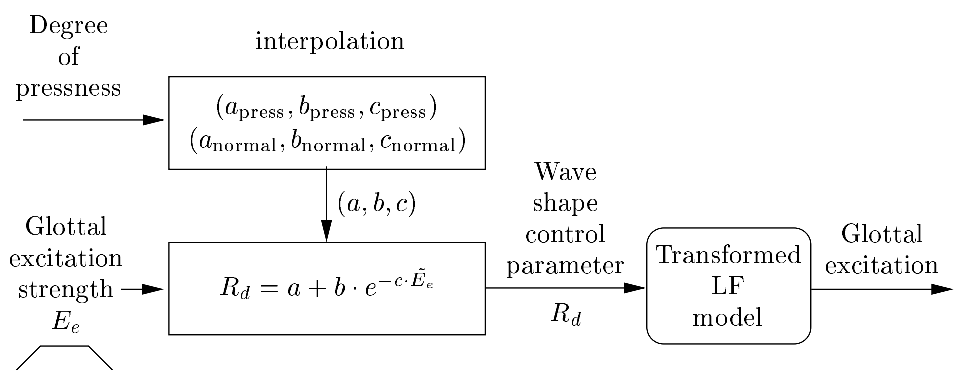

Figure 4.4: Vocal texture control for non-breathy mode.

energy over the high-pass derivative glottal wave energy, where the noise residual and the derivative glottal wave are obtained from the analysis procedure described in Chapter 3. The reason for introducing this new parameter is that it is a good indicator for degree of breathiness.

By analyzing baritone recordings for the breathy phonation, the wave-shape parameter R_d is found to be roughly positively correlated to NHR as shown in Figure 4.5. As a result,

R_d is approximately modeled as a simple log function of the NHR,

$$x = \ln((\sqrt{NHR} - 1.4847/0.5016)/0.7396. \tag{4.2}$$

$$R_d = \begin{cases} 1.5 & \text{, if } x < 1.5; \\ x & \text{, if } 1.5 \le x \le 4; \\ 4 & \text{, } x > 4. \end{cases} \tag{4.3}$$

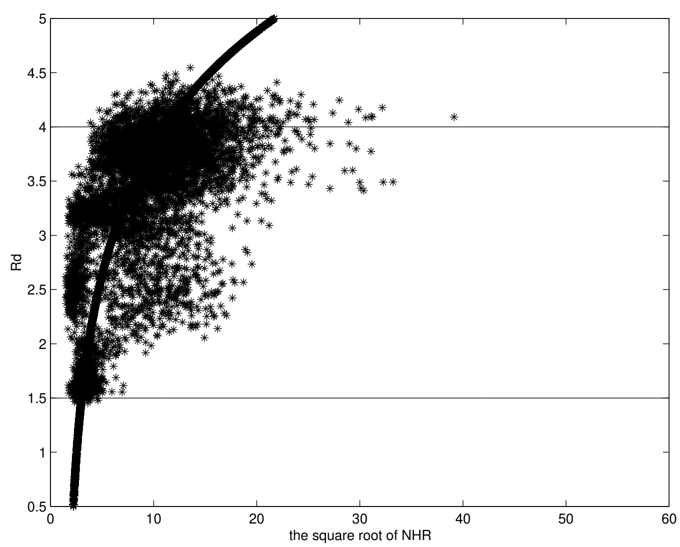

Figure 4.5: The scatter plot for R_d and $\sqrt{NHR}$.

Figure 4.5 shows the scatter plot of R_d and $\sqrt{NHR}$. The fitted model is also plotted. The scattering of the Figure 4.5 is large since the value of R_d depends not only on NHR, but also on the singing pitch. Since R_d is only the secondary parameter to control the degree of breathiness, this approximate model is sufficient. Some sound examples are shown at http://www-ccrma.stanford.edu/˜vickylu/thesis/ to illustrate that inaccuracy of R_d estimate has only minor impact on the sensation of breathiness.

To reduce the computation of NHR, another parameter $NHR2$ is introduced to approximate NHR. For $NHR2$, the high-pass derivative glottal wave energy is replaced by the glottal excitation strength E_e. Figure 4.6 shows the scatter plot of $\sqrt{NHR2}$ and $\sqrt{NHR}$. From this figure, one can observe the linear dependency between $NHR2$ and NHR. Hence, $NHR2$ can be a fine approximation of NHR.

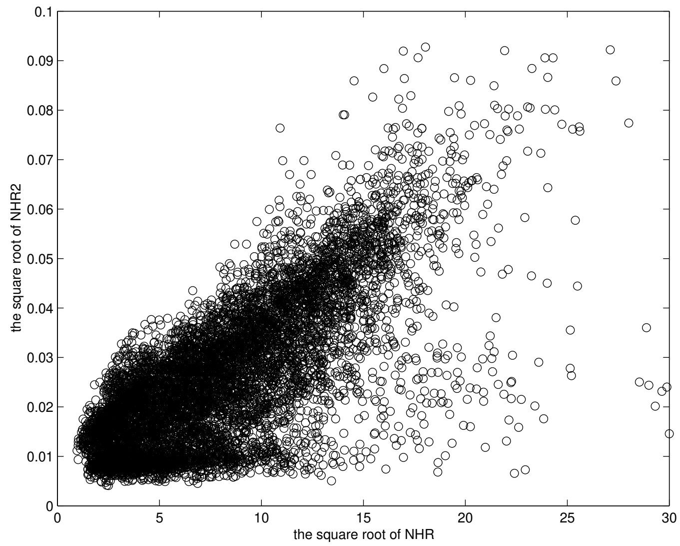

Figure 4.6: The scatter plot for $\sqrt{NHR2}$ and $\sqrt{NHR}$.

In addition to R_d, one needs to determine parameters of the noise residual model in the breathy phonation. It was found that the scale of the pitch synchronous noise component (A_n) is linearly correlated to the variance of the noise floor (B_n). Hence, these two parameters can be reduced to one. Moreover, the duty cycle and window lag for the pitch synchronous noise component have a relatively minor perceptual impact compared to NHR. These two parameters can be fixed without affecting the vocal texture very much, as long as (1) the Hanning window width is between 40% and 80% of the glottal period, and (2) the lag of the center of the Hanning window after glottal opening is between 0% and 15% of the period.

Figure 4.7 shows the block diagram for vocal texture control in the breathy mode. Similar to the non-breathy mode, users only need to specify the degree of breathiness and the glottal excitation strength. The NHR is adjusted proportionally to the degree of breathiness. The wave-shape parameter is determined from the adjusted NHR. A derivative glottal wave is then generated. For the noise residual component, the variance of the noise floor, B_n, is set to one. The amplitude of the Hanning window, A_n, is calculated from B_n via a linear equation. With the fixed duty cycle and window lag, the noise component is generated by the noise residual model as described in Chapter 2. The strength of the output noise residual is then scaled by the NHR. The glottal excitation is finally obtained

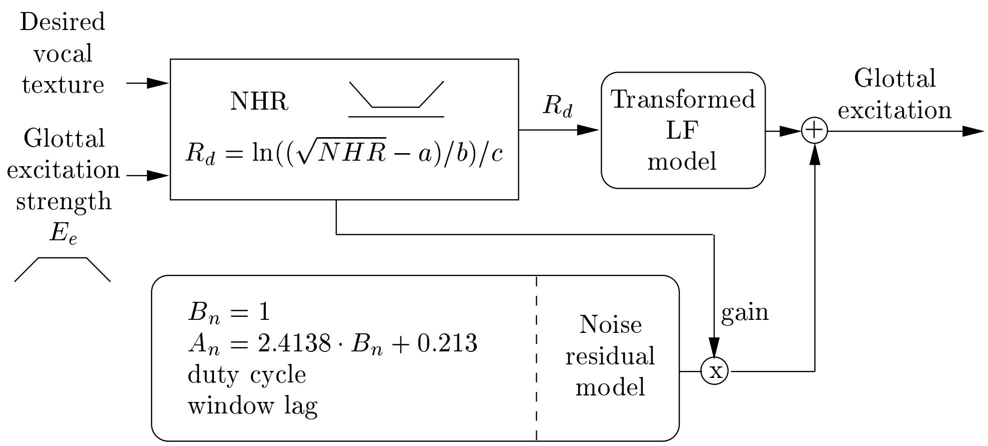

Figure 4.7: Vocal texture control for breathy mode.

by summing the derivative glottal wave and the noise residual.

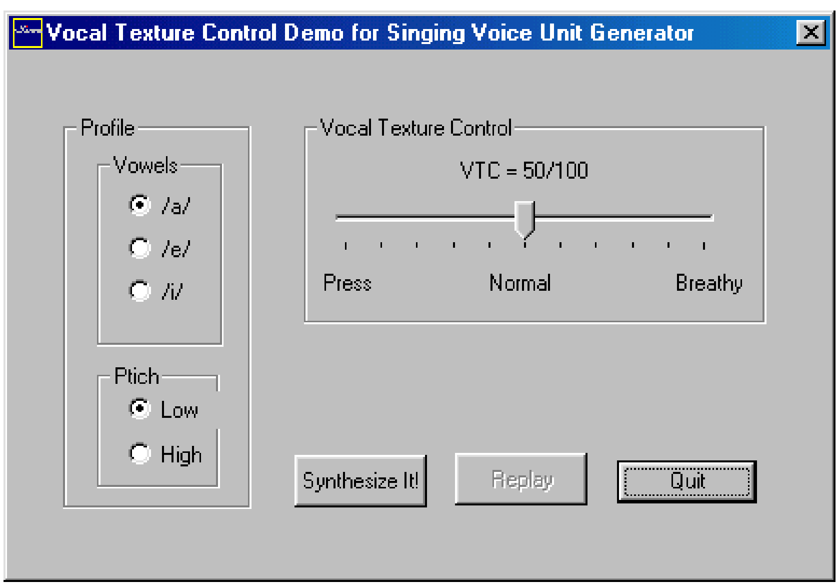

Figure 4.8: Vocal texture control interface.

In summary, control of the vocal texture is made very simple by exploiting correlations among the synthesis model parameters, and by choice of the intuitive NHR parameter. An interactive vocal texture control interface is implemented as shown in Figure 4.8 to demonstrate the effectiveness of the proposed simple parametric model.

Chapter 5

Conclusions and future directions

To achieve high-quality singing synthesis, spectral modeling and physical modeling have been implemented in the past. However, spectral models are known to be articulation difficult and expressivity limited at low algorithm complexities. On the other hand, it is not always straightforward to adjust physical model parameters to reproduce a specific recording. The objective of this research is to build a synthesis model which is as close to a physical model as possible without losing the ability of analysis/re-synthesis. In this thesis, a pseudo-physical model for singing voice synthesis was proposed. An automatic analysis procedure for analysis/re-synthesis was also developed for the proposed synthesis model. This model is capable of generating high-quality non-nasal singing voices and is computationally affordable. In addition to pitch and volume control, this synthesis model also provides flexible control of vocal textures ranging from "pressed", to "normal", to "breathy" phonation. By analyzing baritone singing recordings, a parametric model was constructed to simplify vocal texture control.

Source-filter deconvolution is the first component of the proposed analysis/re-synthesis procedure. This pitch synchronous source-filter deconvolution scheme based on the SUMT algorithm was developed to estimate the vocal tract filter and the glottal excitation simultaneously. In the speech literature, the closed-phase LPC is considered to be the most accurate method of estimating the vocal tract filter since there is no source-tract interaction during the closed phase. However, its usage can be limited since the closed phase may be very short or non-existent in practice, making estimation via the closed-phase LPC impractical. For example, the derivative glottal wave in high-pitched singing voices tends to have a very short closed phase. Moreover, the closed-phase LPC estimation is sensitive to the detection accuracy of closed-phase glottal excitation intervals. The proposed source-filter deconvolution scheme in this thesis is an extension of the closed-phased LPC in the sense that an open quotient is estimated but a complete glottal cycle is used to estimate the vocal tract filter. Under the assumption that one can neglect source-tract interaction, the resultant vocal tract estimation can be considered an average of the vocal tract filter during the closed-phase region and the open-phase region. Nevertheless, it has been shown in Section 3.1.4 that the proposed scheme can recover the glottal excitation quite well even when source-tract interaction exists.

Another interesting component of the analysis/re-synthesis procedure is denoising. The best-basis (wavelet packet analysis) soft-thresholding method has been adapted to denoise pitch synchronous noise such as aspiration noise in this thesis. This method is chosen since it is successful at separating the aspiration noise and the derivative glottal. With the introduction of an automatic decomposition level selection criterion, the best-basis soft-thresholding method is shown to outperform SMS [66].

A list of possible directions for future research are proposed below based on current research:

- For the purpose of score-to-singing synthesis, a complete sound synthesis component requires the construction of a parametric database of acoustic realizations for each phoneme and its associated co-articulation rules. It is possible to use the proposed analysis procedure to perform such task. Since the proposed model is pseudo-physical, it is expected that the co-articulation rules will be simplified compared to other schemes such as linear prediction and sinusoidal modeling.

- Subjective informal listening tests have been used in this thesis to evaluate the quality of analysis/re-synthesis. However, the method for objective measurement of perceived audio quality is available as an ITU standard (PEAQ) [79]. Objective measurements may be used in future development of co-articulation rules where massive listening tests are required.

- The proposed source-filter deconvolution algorithm has the potential for low-bit rate high quality speech coding. Such development requires a comparison of its sound quality with other existing low-bit rate coding methods.

- The source-filter deconvolution algorithm developed in this thesis is capable of improving several existing applications. For example, the fine structures of the inverse filtered glottal waves have been used in speaker identification [63]. Another example is speaking style modification through glottal excitation adjustments [15] [16].

- In this study, the proposed synthesis model neglects the effects of source-tract interaction. Several researchers have simulated source-tract interaction by increasing the bandwidths of the formants during the open phase of the glottal cycle [11] [43]; since both the vocal folds and subglottal area contribute damping to the human voice production system. Based on this idea, one can extend the current source-filter deconvolution method by considering a time-varying all pole filter where the auto-regressive coefficients are updated at the transitions between closed-phase glottal cycle and open-phase glottal cycle.

- Aperiodicity features [32] and noise features [59] [64] [58] are two important categories for describing pathological voices. Michaelis [56] has shown that glottal to noise excitation ratio is the best candidate for noise feature. Along with other aperiodicity features, such as jitter and shimmer, the glottal to noise excitation ratio gives the best description of pathological voices since it correlates less with other aperiodicity features. Therefore, the analysis procedure for separating the glottal aspiration noise and the derivative glottal wave provides the means of studying pathological voices.

- In this thesis, sound transformation of vocal textures has been constructed for non-nasal vowels. It would be interesting to extend this research to continuous speech or singing phrases.

Appendix A

An articulatory voice synthesizer

An articulatory synthesizer models the human voice production system directly. It synthesizes voices from slowly-varying physiological parameters such as lung pressure, vocal cord tension, shape of tongue, coupling to the nasal cavity, and lip opening. Lous at IPO (Institute for Perception Research) [45] has proposed an articulatory speech synthesizer with symmetrical two-mass vocal folds for prosthesis design. The vocal fold model is an extension of two-mass model of Ishizaka and Flanagan [36] which can capture essential features of human voice production by suitable parameterization. The vocal tract is modeled by a series of lossless cylindrical acoustic tubes and simulated by a digital waveguide filter. Radiation at the lip end is assumed as a flanged pipe in free space. Before details of the two-mass model proposed by Lous are illustrated, a general structure of a two-mass type model is first reviewed in Section A.1.

A.1 General structure of physical glottal modeling

In general, a physical model of glottal source includes a mechanical model which describes the biomechanical behavior of the vocal fold tissue, a geometry model which models geometry of the vocal folds, a collision model which describes how the collision of vocal cords changes the geometry and mechanical behavior of the vocal folds, and a aerodynamics model which computes the glottal flow and fluid-dynamic forces acting on the masses.

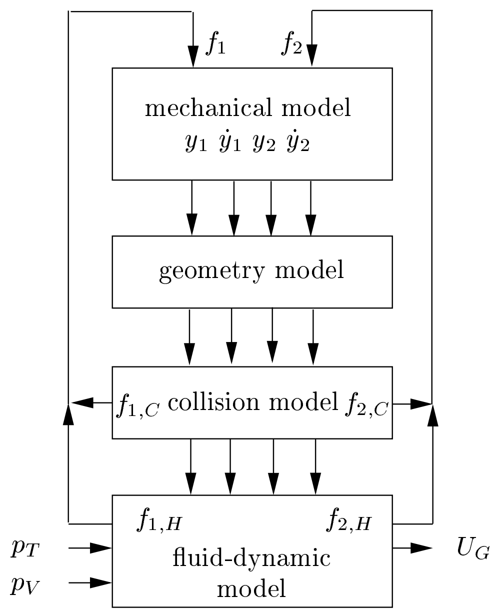

Figure A.1: General signal flow chart of a two-mass model.

Figure A.1 shows a general signal flow chart of a two-mass model [80]. Two-mass model simplifies a vocal cord as two masses coupled with a spring. These two masses are also connected to the larynx by a spring and a dashpot. These masses are assumed to move only in vertical direction with respect to glottal flow. In the diagram, y_1 and y_2 denote the position of two masses. $\dot{y}_1$ and $\dot{y}_2$ denote the velocity of two masses. P_T is the air pressure upstream and P_V is the air pressure downstream of the glottis. Moreover, U_G is the glottal volume velocity.

A.1.1 Geometry model

Figure A.2 shows the geometry of the two-mass model proposed by Lous. This two-mass model assumes a two-dimensional structure of the vocal folds. The geometry is assumed to be constant in the z-direction over the vocal fold length L_g. The glottis is also assumed to be symmetric with respect to the flow channel axis. At one side of the glottis, it consists two equal weighted point masses and three mass-less plates that support masses. Aerodynamic force is then acting on these three plates.

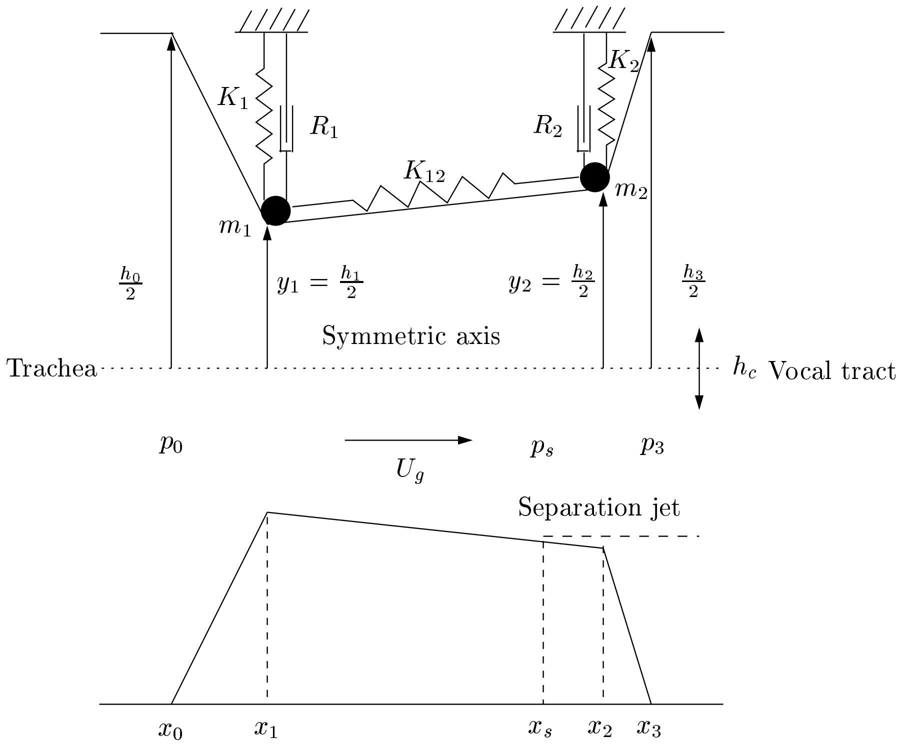

Figure A.2: Symmetrical two-mass model for the vocal folds.

A.1.2 Mechanical and collision model

The mechanical behavior of vocal fold tissues is modeled by lump inertia, elasticity and damping. Two point masses (m_1, m_2)* are used to model the inertia in each fold. Each mass is assumed to attach to an imaginary rigid wall by a spring and a dashpot. Springs, K_1, K_2, model the elasticity and dashpots, R_1, R_2, model the damping of the vocal folds. When collision between two vocal cords occurs, the spring coefficients (K_1, K_2) and dashpot coefficients (R_1, R_2) are increased to model the change of vocal fold tissues. Between two masses, there is also a spring K_{12} which induces a coupling force. The following equations give a summary of the mechanical behavior of the two-mass model:

$$m_1\ddot{y}_1 + R_1\dot{y}_1 + K_1 y_1 + K_{12}(y_1 - y_2) = F_{h1} \tag{A.1}$$

$$m_2\ddot{y}_2 + R_2\dot{y}_2 + K_2 y_2 + K_{12}(y_2 - y_1) = F_{h2} \tag{A.2}$$

*More specifically, they are line masses with length L_g.

where F_{h1} is the force acting on mass m_1 and F_{h2} is the force acting on mass m_2.

A.1.3 Aerodynamics model

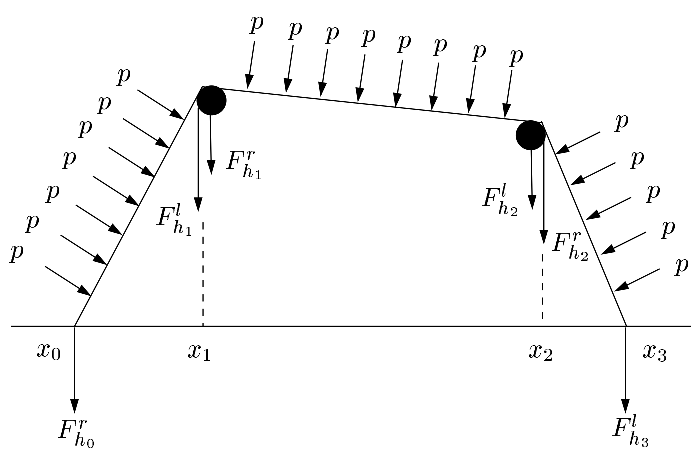

Figure A.3: Aerodynamic reaction forces resulting from a pressure distribution $p(x)$ in the glottis.

Since the dimension of the glottis is small compared to the acoustic wave length, the glottal flow is commonly assumed to be locally incompressible, frictionless and quasi-stationary. Under these conditions, the conservation of mechanical energy of a fluid particle follows the Bernoulli's equation† [33],

$$\frac{1}{2}\rho_0 v^2 + p = \text{constant}$$

where v is the particle velocity and p is the pressure. This model also assumes a free jet bounded by two shear layers‡ of opposite vorticity is formed at a separation point $x_s(t)$. That is, the one-dimensional glottal flow traveling from trachea separates from the wall to form a free jet at a point $x_s(t)$. The flow separation point depends on the geometry of the glottis. Pelorson [62] has shown that the moving separation point is important in aerodynamics modeling. In this model, the separation point $x_s(t)$ is at the position where the vertical height of the flow channel $h_s(t)$ satisfies $h_s(t)/h_1(t) = s$§ as shown in Figure A.2. If $h_2(t) > sh_1(t)$, $x_s(t) = x_2$.

†If flow inertia and viscosity are considered, the Bernoulli's equation is modified to include the pressure drop due to the flow inertia and viscosity.

‡A shear layer is a plane which separates the low velocity region from the mean flow.

§Separation constant s is a pre-determined constant.

Assume turbulence in the free jet dissipates all the kinetic energy with a negligible recovery of pressure. Hence, the pressure in the glottis for $x > x_s(t)$ is equal to the pressure at the entrance of the vocal tract. Before the separation point, pressure can be obtained by the Bernoulli's equation. For this model, the pressure is described by

$$p(x,t) + \frac{\rho_0}{2}\left(\frac{U_g(t)}{h(x,t)L_g}\right)^2 = p_0(t) + \left(\frac{U_g(t)}{h_0 L_g}\right)^2,$$

where ρ_0 is the mean air density and $U_g(t)$ is the volume flux through the glottis.

Aerodynamic reaction forces result from the pressure distribution $p(x)$ in the glottis. As shown in Figure A.3, the force components F^r_{h1} at x_1, and F^l_{h2} at x_2, result from the pressure distribution on the mass-less plate connecting masses m_1 and m_2. Similar force components can be determined from the aerodynamic pressure on the plates modeling the glottis inlet (F^r_{h0} and F^l_{h1}) and outlet (F^r_{h2} and F^l_{h3}). By considering the equilibrium of force and moments, the force acting on masses can be calculated by

$$F_{h_i}(t) = \int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{x_i - x_{i-1}} p(x,t)\, dx.$$

Appendix B

Convex optimization

The materials in this appendix is taken from [6], which the reader may consult for more details.

B.1 Convex Set

A set C is convex if the line segment between any two points in C belongs to C, that is,

$$x, y \in C, \theta \in [0,1] \Rightarrow \theta x + (1-\theta)y \in C. \tag{B.1}$$

B.2 Convex function

A function $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is convex if the domain of f is a convex set and if for all $x, y \in$ domain of f, the line segment between $(x, f(x))$ and $(y, f(y))$ are above f as shown in Figure B.1. That is,

$$x, y \in \mathbf{dom} f, \theta \in [0,1] \Rightarrow f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y). \tag{B.2}$$

If a convex function f is differentiable, the first order Taylor approximation of f is a global

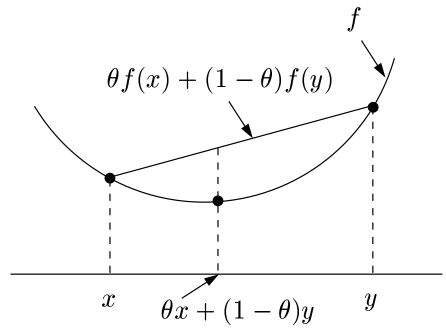

Figure B.1: Illustration of a convex function, f.

under-estimator of f as shown in Figure B.2 . That is,

$$x, x_0 \in \mathbf{dom} f, f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0). \tag{B.3}$$

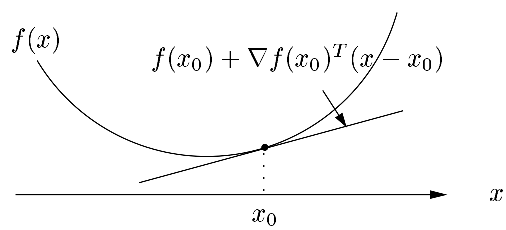

Figure B.2: First order Taylor approximation of f is a global under-estimator of f.

B.3 Convex optimization problem

An optimization problem is a problem of finding an x which minimizes the objective function $f_0(x)$ and satisfies all required inequality and equality constraints. That is,

$$\begin{array}{ll} \text{Minimize} & f_0(x) \\ \text{Subject to} & f_i(x) \leq 0, i = 1, \cdots, m \\ & h_i(x) = 0, i = 1, \cdots, p \end{array} \tag{B.4}$$

where $f_i, h_i : \mathbf{R}^n \to \mathbf{R}$ and x is the optimization variable. A point x is feasible if it satisfies all constraints. The feasible set C is defined as the set of all feasible points. An optimization problem is infeasible if the feasible set is empty. If all of the inequality constraints are convex functions and the equality constraints are affine, the feasible set is convex. A problem becomes a convex optimization problem if the objective function is also convex. A convex optimization problem can be represented as follows:

$$\begin{array}{ll} \text{Minimize} & f_0(x) \\ \text{Subject to} & f_i(x) \leq 0, i = 1, \cdots, m \\ & a_i^T x - b_i = 0, i = 1, \cdots, p \end{array} \tag{B.5}$$

Convex optimization problems are often tractable both in theory and in practice. They may be solved numerically with great efficiency and they grow gracefully with the problem size. A convex optimization problem has one global solution, since any local solution is also global. Moreover, it can handle non-differentiable as well as smooth problems. Duality theory exists for proving infeasibility or for providing lower bounds on objective functions. As such, convex optimizaion problems have non-heuristic stopping criteria.

B.3.1 An optimality criterion for differentiable f_0

If the objective function f_0 of a convex probelm is differentiable, $x \in C$ is optimal if and only if

$$y \in C \Rightarrow \nabla f_0(x)^T (y - x) \geq 0. \tag{B.6}$$

Figure B.3 shows a geometrical interpretation of this optimality condition. The shaded region is the convex feasible set C. The contour lines of f_0 are also illustrated. This figure shows that $-\nabla f_0(x)$ defines a supporting hyperplane for C at x. From this criterion, $\nabla f_0(x) = 0$ for $x \in C$ implies that x is optimal. For unconstrained problems, x is optimal iff $\nabla f_0(x) = 0$ since the feasible set C becomes $\mathbf{R}^n$.

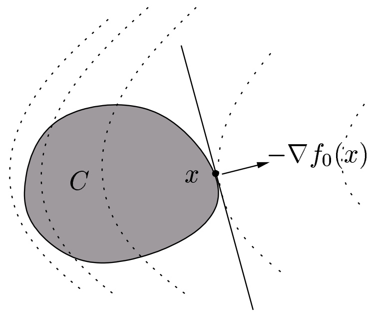

Figure B.3: Geometrical interpretation of the optimality criterion (B.6).

B.3.2 Examples of convex optimization problem

Quadratic program (QP)

A quadratic program has a quadratic objective function, linear inequalities and equalities. That is,

$$\begin{array}{ll} \text{Minimize} & x^T P x + 2q^T x + r \\ \text{Subject to} & Ax \leq b,\ Fx = g \end{array} \tag{B.7}$$

A quadratic program is a convex optimization problem if P is semi-positive definite.* The source-filter deconvolution problem in this thesis is an example of convex quadratic program.

B.3.3 Smooth unconstrained minimization methods

If the convex optimization problem in Equation (B.5) is unconstrained† and its objective function f is differenetiable, $\nabla f(x^\star) = 0$ is a necessary and sufficient condition for $x^\star$ to be optimal. Denote $f^\star$ as the optimal value $f(x^\star)$. Unconstrained optimization problems usually have to be solved iteratively, from which a minimization sequence $x^{(k)}$ is generated such that $f(x^{(k)}) \to f^\star$ as $k \to \infty$.‡ The following section will describe one general iterative algorithm for solving unconstrained optimization problems.

*$P \geq 0$ iff $x^T P x \geq 0$, for $\forall x$.

†There is no inequality or equality constraint.

‡Here k denotes the iteration number.

Descent methods

A descent method produces a minimization sequence $x^{(k)}$, $k = 1, \cdots$, from the following iteration

$$x^{(k+1)} = x^{(k)} + t^{(k)} v^{(k)}$$

where the vector $v^{(k)}$ is called the search direction and the positive scalar $t^{(k)}$ is called the step size at iteration k. The minimization sequence generated by the descent method guarantees that $f(x^{(k+1)}) < f(x^{(k))}$, except when $x^{(k)}$ is optimal. Since f is convex, $\nabla f(x^{(k)})^T (y - x^{(k)}) \geq 0$ implies $f(y) \geq f(x^{(k)})$ from Equation (B.6). Hence, the search direction $v^{(k)}$ in the descent method must satisfy

$$\nabla f(x^{(k)})^T v^{(k)} \leq 0.$$

The general structure of the descent method is given below:
Given a starting point $x \in \mathbf{dom} f$.
Loop

1. Compute the search direction $v^{(k)}$.
2. Line search: choose a step size $t^{(k)} > 0$.
3. Update the optimization variable. $x^{(k+1)} = x^{(k)} + t^{(k)} v^{(k)}$.

until stopping criterion is satisfied

Various descent methods exist and they differ in the way the search direction is computed or how the line search for choosing the step size is performed. For example, the *Gradient algorithm*, or the *gradient descent method*, uses the negative gradient $-\nabla f(x)$ as its search direction. The *steepest descent method* calculates the search direction as the normalized steepest descent direction that gives the largest decrease in the linear approximation of f,

$$v_{nsd} = arg \min_{z, \|z\|=1} \nabla f(x)^T z.$$

The *Newton's method* uses the following search direction,

$$v = -\nabla^2 f(x)^{-1} \nabla f(x)$$

which minimizes the second-order expansion of $f^{\S}$ at x. The *Newton's method* is described in more detail later. Three common types of line search are also described here.

Line search

The purpose of line search is to find a step size t for updating the estimate x along the search direction. There are three common types of line search.

- Fixed step size: $t > 0$ is a specified fixed value.
- Exact line search: t is chosen to minimize f along the search direction, i.e.,

$$t = t^{\star} = arg \min_{s} f(x + sv).$$

- Backtracking line search: To reduce computation, most line search in practice are inexact in the sense that the step size is chosen to approximately minimize f along the search direction. The backtracking line search method is a very simple and effective inexact line search method. This method requires two constant parameters α and β with $0 < \alpha < 0.5$ and $0 < \beta < 1$. The outline of the backtracking line search method is given here.
  $t = 1$.
  while $(f(x + tv) > f(x) + \alpha t \nabla f(x)^T v)$
  $\quad t = \beta t$.
  end

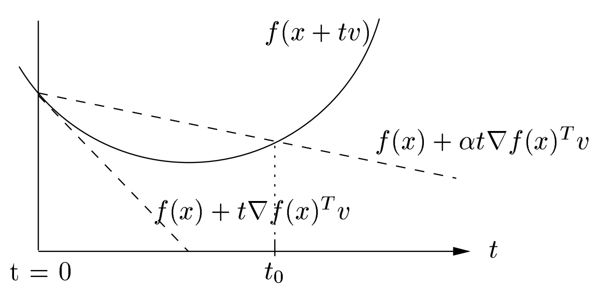

Figure B.4: Backtracking line search. The step size estimate t is the first value when $t < t_0$.

[§]The second order expansion of f at x is given as $f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)$.

Figure B.4 illustrates the idea of backtracking. $f(x) + t\nabla f(x)^T v$ is the linear approximation of $f(x + tv)$ at $t = 0$. It is also the under-estimator of $f(x + tv)$.[¶] Since $0 < \alpha < 0.5$ and $t\nabla f(x)^T v < 0$, $f(x) + \alpha t\nabla f(x)^T v$ has a smaller slope than $f(x) + t\nabla f(x)^T v$ as shown in Figure B.4. Assume that the line $f(x) + \alpha t\nabla f(x)^T v$ intersects with $f(x+tv)$ at $t = t_0$. The interval between $t = 0$ and $t = t_0$ is very likely to include the optimal value $t^\star$. Backtracking line search gives the approximated estimate of $t^\star$. It starts with unit step size ($t = 1$) and terminates whenever the updated t is smaller than t_0.

Newton's method

Assume that f is a twice differentiable convex function. The search direction of the so called *Newton Step* is defined as $v = -\nabla^2 f(x)^{-1}\nabla f(x)$. The search direction is obtained by minimizing the second order approximation of f, i.e., $f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T\nabla^2 f(x)(y - x)$. That is, $x + v$ minimizes f if f is truly quadratic. If the estimate x is updated by $x + v$, this method is also referred to as the *pure* Newton's method. The idea of *pure* Newton's method is that $x + v$ gives a better approximation of $x^\star$ than x when x is near the optimum $x^\star$, since f is nearly quadratic in the vicinity of the optimum solution.

When x estimate is updated as $x + tv$ ($t \leq 1$) instead, it is called the Newton's method, also referred to as the *damped* or the *guarded* Newton's method. An outline of the Newton's method is given below.

Given a starting point $x \in \mathbf{dom} f$.

Loop

1. Compute the Newton step $v^{(k)} = -\nabla^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$.
2. Line search: compute the step size t via backtracking line search.
3. Update the optimization variable: $x^{(k+1)} = x^{(k)} + t^{(k)}v^{(k)}$.

until stopping criterion is satisfied.

Moreover, since $\nabla f(x)^T v = -\nabla f(x)^T\nabla^2 f(x)^{-1}\nabla f(x) < 0$ (unless $\nabla f(x) = 0$), the Newton's method is a form of the descent methods.

[¶] A function is convex iff it is convex on all lines. Hence, $f(x_0 + tv)$ is convex in t for all x_0 and v.

B.3.4 Duality

Consider the original convex optimization problem in Equation (B.4) as a primal problem. The *Lagrangian* associated with a primal problem is defined as

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

where λ_i is the Lagrange multiplier with respect to the ith inequality constraint $f_i(x)$ and ν_i is the Lagrange multiplier with respect to the ith equality constraint $h_i(x)$. The vectors λ and ν are called the dual variables associated with the primal problem. By augmenting the objective function with a weighted sum of the constraint functions, the constraints are formulated into the Lagrangian.

Lagrange dual function

Assume that the primal problem has a non-empty domain $D = \cap_{i=0}^{m} \mathbf{dom} f_i \bigcap \cap_{i=0}^{p} \mathbf{dom} h_i.$, and that the optimal value of the primal problem is denoted as $p\star$. The *Lagrange dual function* is defined as

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)). \tag{B.8}$$

The Lagrange dual function is concave in (λ, ν) since it is a pointwise minimum of affine functions of (λ, ν). The dual function may become $-\infty$ for some values of λ.

Duality gap

It can be shown that the dual function yields lower bounds on the optimal value $p^\star$ for the primal problem. That is,

$$g(\lambda, \nu) \leq p^\star \leq f_0(x), \forall \lambda \succeq 0^{\|}, \forall \nu. \tag{B.9}$$

$^{\|}$ $\lambda \succeq 0$ means that every element in the vector λ is greater than 0.

Duality gap is defined as $f_0(x) - g(\lambda)$. If $\lambda \succeq 0$ and $g(\lambda, \nu) > -\infty$, (λ, ν) are dual feasible. In this case, $g(\lambda, \nu)$ yields a meaningful lower bound on the optimal value.

Lagrange dual problem

Associated with the primal problem, the Lagrange dual problem is defined as

$$\begin{array}{ll} \text{Maximize} & g(\lambda, \nu) \\ \text{Subject to} & \lambda \succeq 0. \end{array} \tag{B.10}$$

The Lagrange dual problem is commonly formulated to find the best lower bound on $p^\star$. If the optimal value of the dual problem is denoted as $d^\star$, the optimal duality gap is therefore $p^\star - d^\star$. For convex primal problems, strong duality, i.e. $p^\star = d^\star$, usually holds. Hence, the dual optimal point $\lambda^\star$ is the certificate of optimality for primal optimal point $x^\star$.

B.3.5 Sequential unconstrained minimization methods

To solve the convex optimization problem in Equation (B.4), one family of interior-point methods is reviewed in this section. It is called the *sequential unconstrained minimization*, *barrier method*, or the *path-finding method*. The basic idea behind this method is to solve or approximately solve a sequence of smooth unconstrained problems** after transforming the constraints into a smooth barrier function.

A *barrier function* is a convex and smooth function which grows without bound when x approaches the boundary of the feasible set. The most popular barrier function is the *logarithmic barrier*. It is defined as

$$\phi(x) = \begin{cases} -\sum_{i=1}^{m} \log(-f_i(x)) & \text{if } f_i(x) < 0,\, i = 1, \cdots, m \\ +\infty & \text{otherwise.} \end{cases} \tag{B.11}$$

When the objective function for minimization is augmented with the barrier function, the optimal solution would be pushed away from the boundary of the feasible set. Consider

**The following discussion assumes that equality constraints have been eliminated.

the modified problem as

$$x^\star(t) = arg\,\min(tf_0(x) + \phi(x)),\, t > 0. \tag{B.12}$$

This modified problem can be solved by smooth unconstrained minimization techniques. The solution curve $x^\star(t)$ is called the *central path.* The barrier limits the search of $x^\star(t)$ in the feasible set and t gives the relative weight between the objective and the barrier function. Hence, $x^\star(t)$ converges to the optimal solution when $t \to \infty$.

Since (B.12) is an unconstrained convex optimization problem, solution $x^\star(t)$ satisfies $\nabla(tf_0(x^\star(t)) + \phi(x^\star(t))) = 0$. That is,

$$\nabla f_0(x^\star(t)) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x^\star(t)) = 0,\, \lambda_i = \frac{1}{-f_i(x^\star(t))t}. \tag{B.13}$$

Hence, $x^\star(t)$ also minimizes the Lagrangian $L(x,\lambda) = f_0(x) + \sum_i \lambda_i f_i(x)$. This implies that a point on central path yields a dual feasible point and a lower bound, $g(\lambda)$,

$$g(\lambda) = f_0(x^\star(t)) + \sum_i \lambda_i f_i(x^\star(t)) = f_0(x^\star(t)) - m/t. \tag{B.14}$$

If the optimal value of the original convex optimization problem ((B.5)) is denoted as $p^\star$, the following inequality holds,

$$f_0(x^\star(t)) \geq p^\star \geq f_0(x^\star(t)) - m/t. \tag{B.15}$$

This inequality condition also shows that $x^\star(t)$ would become optimal as $t \to \infty$. From this relationship, it is attempting to solve the convex optimization problem by solving the unconstrained minimization problem (Equation (B.12)) with a very large $t = m/\epsilon$ that satisfies an accuracy of ϵ. However, this method has a very slow convergence in practice.

SUMT

Instead of solving one unconstrained minimization problem with a very large t, a sequence of unconstrained minimization problem is solved. $x^\star(t)$ is computed for a sequence of increasing values of t, until $t \geq m/\epsilon$. When $t \geq m/\epsilon$, it is guaranteed that the ϵ-suboptimal solution of the original problem is obtained.[††] This method is called the *sequential unconstrained minimization technique* (SUMT). Its outline is given below.

Given strictly feasible x, $t = t^{(0)} > 0$, $\mu > 1$, $\epsilon > 0$.

Loop

1. Centering step: Starting from x, compute the optimal point $x^\star(t)$ which minimizes $tf_0 + \phi$. In the current implementation of source-filter deconvolution, the Newton's method is applied to minimize $tf_0 + \phi$.

2. $x = x^\star(t)$.

3. $t = \mu t$.

until $m/t < \epsilon$.

Computation complexity of SUMT

Each execution of step 1 is a centering step. The iterations executed during the centering step are referred to as inner iterations. In addition to the initial centering step which computes $x^\star(t^{(0)})$, the desired accuracy ϵ can be achieved by $\lceil \frac{\log(\epsilon t^{(0)}/m)}{\log \mu} \rceil$ number of centering steps since the duality gap is $m/(\mu^k t^{(0)})$ after $k+1$ centering steps. The total number of Newton steps depends on the choice of μ and $t^{(0)}$.

For a small μ, the number of centering steps is larger since t is only updated by a small factor. However, the number of inner iterations for each centering step is smaller since the previous iterations provide very good starting points for x. In practice, SUMT works well for a wide range of μ.

If $t^{(0)}$ is too large, too many iterations in the initial centering step will be required. However, extra centering steps are necessary if $t^{(0)}$ is chosen too small. The rule of thumb is to choose $t^{(0)}$ such that $m/t^{(0)}$ is approximately the same order as $f_0(x^{(0)}) - f^\star$ since $m/t^{(0)}$ is equal to the duality gap for the first centering step. If a dual feasible point λ is known, $t^{(0)} = m/\eta$ with a duality gap of $\eta = f_0(x^{(0)}) - g(\lambda)$ is a reasonable choice.

[††]This can be shown from Equation (B.15)

Bibliography

[1] S. Ahmadi and Andreas S. Spanias. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. *IEEE Trans. Speech, Audio Processing*, 7(3):333–338, 1999.

[2] P. Alku and E. Vilkman. Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering. In *J. Acoust. Soc. Am.*, 98(2), Pt. 1, pp. 763–767, 1995.

[3] P. Badin, C. Shadle, Y.P.T. Ngoc, J.N. Carter, W.S.C. Chiu, C. Scully, and K. Stromberg. Frication and aspiration noise sources: Contribution of experimental data to articulatory synthesis. In *Int. Conf. Spoken Language Processing*, pp. 163–166, 1994.

[4] J. Van den Berg. Myoelastic-aerodynamics theory of voice production. *J. Speech Hear. Res.*, 1, pp. 227–244, 1957.

[5] W.K. Blake. *Mechanics of flow-induced sound and vibration, volume 1: General concepts and elementary sources.* Academic Press, New York, 1986.

[6] S. Boyd and L. Vandenberghe. *Course Reader for EE364: Introduction to Convex Optimization with Engineering Applications.* Stanford University, 1999.

[7] C. Chafe. Pulse noise in self-sustained oscillations of musical instruments. *ICASSP*, 2:1157–1160, 1990.

[8] F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Proc. EUROSPEECH-89*, 2:13–19, 1989.

[9] D. G. Childers and C. K. Lee. Vocal quality factors: analysis, synthesis and perception. *J. Acoust. Soc. Am.*, 90(5):2394–2411, 1991.

[10] D. G. Childers and C. F. Wong. Measuring and modeling vocal-tract interaction. *IEEE Trans. Biomedical Engineering*, 41(7):663–671, 1994.

[11] D. G. Childers and H. T. Hu. Speech synthesis by glottal excited linear prediction. *J. Acoust. Soc. Am.*, 96(4):2026–2036, 1994.

[12] D. G. Childers and C. Ahn. Modeling the glottal volume-velocity waveform for three voice types. *J. Acoust. Soc. Am.*, 97(1):505–519, 1995.

[13] D. G. Childers. *Speech processing and synthesis toolboxes.* John Wiley & Sons, 2000.

[14] P.R. Cook. *Identification of Control Parameters in an Articulatory Vocal Tract Model with Applications to the Synthesis of Singing.* PhD thesis, Stanford University, 1990.

[15] K. E. Cummings and M. A. Clements. Analysis of glottal waveforms across stress styles. *Proc. 1990 Int. Conf. Acoust. Speech Signal Process*, (1):369–372, Apr. 1990.

[16] K. E. Cummings and M. A. Clements. Application of the analysis of glottal excitation of stressed speech to speaking style modification. *Proc. 1993 Int. Conf. Acoust. Speech Signal Process*, (2):207–210, Apr. 1993.

[17] K. E. Cummings and M. A. Clements. Glottal models for digital speech processing: a historical survey and new results. *Digital signal processing*, 5:21–42, 1995.

[18] Christophe d'Alessandro and Boris Doval. Voice quality modification using periodic-aperiodic decomposition and spectral processing of the voice source signal. *3rd International Workshop on Speech Synthesis*, 277–282, 1998.

[19] J. R. Deller, J. G. Proakis, and J. H. Hansen. *Discrete-time processing of speech signals.* Macmillan, New York, 1993.

[20] Donoho, D.L. and Johnstone, I.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[21] Donoho, D.L. De-noising by soft thresholding. *IEEE Trans on Inf. Theory*, 41(3):613–627, 1995.

[22] Donoho, D.L. and Johnstone, I.M. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

[23] Thierry Dutoit. *An introduction to text-to-speech synthesis.* Kluwer Academic Publishers, 1997.

[24] G. Fant. *Acoustic theory of speech production.* Mouton, The Hague, 1970.

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

[26] G. Fant. The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR*, 2-3:119–156, 1995.

[27] G. Fant. The voice source in connected speech. *Speech Communication*, 22:125–139, 1997.

[28] J.L. Flanagan and L. Landgraf. Self-oscillating source for vocal tract synthesizers. *IEEE Trans. on Audio and Electroacoustics*, AU-16(1):57–64, 1968.

[29] J.L. Flanagan, K. Ishizaka, and K.L. Shipley. synthesis of speech from a dynamic model of the vocal cord and vocal tract. *Bell Syst. Tech. J.*, 54:485–506, 1975.

[30] J.L. Flanagan. *Speech Analysis, Synthesis, and Perception*, page 1–85. Springer Verlag, New York, 1972.

[31] D. J. Hermes. Synthesis of breathy vowels: some research methods. *Speech communication*, 10:497–502, 1991.

[32] James Hillenbrand. Perception of aperiodicities in synthetically generated voices. *J. Acoust. Soc. Am.*, 83(6):2361–2371, 1988.

[33] A. Hirschberg. Some fluid dynamic aspects of speech. *Bulletin de la Communication Parlée*, 2:7–30, 1992.

[34] P. E. Gill, W. Murray, and M. H. Wright. *Numerical Linear Algebra and Optimization.* Addison Wesley, 1991.

[35] M. H. Hayes. *Statistical digital signal processing and modeling.* John Wiley & Sons, 1996.

[36] K. Ishizaka and J. Flanagan. Synthesis of voiced sounds from a two-mass model of vocal cords. *Bell Syst. Tech. J.*, 51:1233-1269, 1972.

[37] ITU-R Rec. BS.562-3. Subjective assessment of sound quality. *International Telecommunications union*, Geneva, Switzerland, 1990.

[38] I. Karlsson and J. Liljencrants. Diverse voice qualities: models and data. *STL-QPSR.*, 2:143–146, 1996.

[39] J.L. Kelly. and C.C. Lochbaum. Speech synthesis. *Proceedings of the Fourth International Congress on Acoustics.*, September:1–4, Paper G42, 1962.

[40] Dennis. H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, 82(3):737–793, 1987.

[41] Dennis. H. Klatt and Laura C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857, 1990.

[42] J. Liljencrants. *Speech synthesis with a reflection-type line analog.* DS dissertation, Department of Speech Communication & Music Acoustics, Royal Inst. of Technology (KTH), Stockholm,1985.

[43] Qiguang Lin. *Speech Production Theory and Articulatory Speech Synthesis.* PhD dissertation, Dept. of Speech Communication & Music Acoustics, Royal Inst. of Technology (KTH), Stockholm, 1990.

[44] L. Ljung and T. Soderstrom. *Theory and Practice of Recursive Identification.* the MIT Press, 1987.

[45] N. J. C. Lous, G. C. J. Hofmans, R. N. J. Veldhuis, and A. Hirschberg. A symmetrical two-mass vocal-fold model coupled to vocal tract and trachea, with application to prothesis design. *Acta Acustica*, 84:1135–1150, 1998.

[46] Hui-Ling Lu and J. O. Smith. Joint estimation of vocal tract filter and glottal source waveform via convex optimization. *Proc. IEEE Workshop on application of signal processing to audio and acoustics*, pp. 79–82, 1999.

[47] Hui-Ling Lu and J. O. Smith. Glottal source modeling for singing voice. *Proceedings of the International Computer Music Conference (ICMC-00, Berlin)*, pp. 90–97, 2000.

[48] Hui-Ling Lu and J. O. Smith. Estimating glottal aspiration noise via wavelet thresholding and best basis thresholding. *Proc. IEEE Workshop on application of signal processing to audio and acoustics*, 2001.

[49] J. C. Lucero. *The dynamics of the vocal fold oscillation.* PhD thesis, Shizuoka University, 1993.

[50] C. Ma, Y. Kamp, and L.F. Willems. Frobenius norm approach to glottal closure detection from the speech signal. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 2(2):258–264, 1994.

[51] M.W. Macon, L. Jensen-Link, J. Oliverio, M.A. Clements and E.B. George. A system for singing voice synthesis based on sinusoidal modeling. *Proceeding of International Conf. on Acoustics, Speech, and Signal Processing*, 1:435–438, 1997.

[52] Mallat, S.G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Pattern Anal. and Machine Intell.*, 11(7):674-693, 1989.

[53] Mallat, S.G. *A wavelet tour of signal processing.* Academic Press, San Diego, 1999.

[54] J. D. Markel and A. H. Gray, Jr. *Linear prediction of speech.* Springer-Verlag, New York, 1976.

[55] P. Meyer, R. Wilhelms, and H.W. Strube. A quasiarticulatory speech synthesizer for German language running in real time. *J. Acoust. Soc. Am.*, 86(2):523–539, 1989.

[56] D. Michaelis, M. Fröhlich, and H.W. Strube. Selection and combination of acoustic features for the description of pathologic voices. *J. Acoust. Soc. Am.*, 103(3):1628–1639, 1998.

[57] P. Satyanarayana Murthy and B. Yegnanarayana. Robustness of group-delay-based method for excitation of significant instants of excitation from speech signals. *IEEE Trans. Speech, Audio Processing*, 7(6):609–619, 1999.

[58] P.J. Murphy. Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *J. Acoust. Soc. Am.*, 105(5):2866–2881, 1999.

[59] Hiroshi Muta and Thomas Baer. A pitch-synchronous analysis of hoarseness in running speech. *J. Acoust. Soc. Am.*, 84(4):1292–1301, 1988.

[60] S. Narayanan and A. Alwan. Noise source models for fricative consonants. *IEEE Trans. Speech, Audio Processing*, 8(2):328–344, 2000.

[61] Ogden, R.T. *Essential wavelets for statistical applications and data analysis.* Birkhäuser, Boston, 1997.

[62] X. Pelorson, A. Hirschberg, R. R. van Hassel, and A.P.J. Wijnands. Theoretical and experimental study of quasisteady flow separation within the glottis during phonation. *J. Acoust. Soc. Am.*, 96(6):3416–3431, 1994.

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

[64] Yingyong Qi, R. E. Hillman and C. Milstein. The estimation of signal-to-noise ratio in continuous speech for disordered voices. *J. Acoust. Soc. Am.*, 105(4):2532–2535, 1999.

[65] J. Schroeter and M.M. Sondhi. Techniques for estimating vocal-tract shape from the speech signal. *IEEE Trans. speech and audio processing*, 2(1):133–150, 1994.

[66] Xavier Serra. Musical Sound Modeling with sinusoids plus noise, published in C. Roads, editors, *Musical Signal Processing*, Swets & Zeitlinger Publishers, 1997.

[67] R. Smits and B. Yegnanayarana. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. Speech Audio Processing*, (3):325–333, 1995.

[68] J. O. Smith and B. Friedlander. Adaptive interpolated time-delay estimation. *IEEE Trans. Aerospace and Electronic Systems*, 21(2):180–199, Mar. 1985.

[69] J. O. Smith. Waveguide filter tutorial. *Proceedings of the 1987 International Computer Music Conference*, 9–16, 1987.

[70] J. O. Smith. Viewpoints on the history of digital synthesis. *Proceedings of the 1991 International Computer Music Conference*, 1–10, 1991.

[71] M.M. Sondhi. Resonances of a bent vocal tract. *J. Acoust. Soc. Am.*, 79:1113–1116, Apr. 1986.

[72] M.M. Sondhi and J. Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. Acoustic, speech and signal processing*, ASSP-35(7):955–967, Jul. 1987.

[73] R. Sproat and J. Olive. An approach to text-to-speech synthesis. *Speech coding and synthesis*, Elsevier press, 1995.

[74] K.N. Stevens. *Acoustic phonetics.* The MIT Press, Cambridge, 2000.

[75] G. Strang and T. Nguyen. *Wavelet and filter banks.* Wellesley-Cambridge Press, 1996.

[76] H. Strik, B. Cranen and L. Boves. Fitting a LF-model to inverse filter signals. *Proceedings of the "3rd European conference on speech communication and technology" (EUROSPEECH-93)*, 1:103–106, 1993.

[77] H. Strik. Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *J. Acoust. Soc. Am.*, 103(5):2659–2669, 1998.

[78] H.W. Strube. Determination of the instant of glottal closure. *J. Acoust. Soc. Am.*, 56:1625–1628, 1974.

[79] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. PEAQ-The ITU standard for objective measurement of perceived audio quality. *J. Audio Eng. Soc.*, 48(1/2):3-29, 2000.

[80] R.N.J. Veldhuis, I.J.M. Bogaert and N.J.C. Lous. Two-mass models for speech synthesis. *Proceedings of 4th European Conf. on Speech Communication Technology*, 1853-1856, 1995.

[81] B. Yegnanarayana and R. N. J. Veldhuis. Extraction of the vocal tract system characteristics from speech signals. *IEEE Trans. Speech, Audio Processing*, 6(4):313–327, 1998.