A NOVEL FRAMEWORK FOR HIGH-QUALITY VOICE SOURCE ANALYSIS AND SYNTHESIS

A thesis submitted for the degree of
Doctor of Philosophy

by

EMIR TURAJLIC

DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING,
BRUNEL UNIVERSITY

September 2006

Abstract

The analysis, parameterization and modeling of voice source estimates obtained via inverse filtering of recorded speech are some of the most challenging areas of speech processing owing to the fact humans produce a wide range of voice source realizations and that the voice source estimates commonly contain artifacts due to the non-linear time-varying source-filter coupling. Currently, the most widely adopted representation of voice source signal is Liljencrants-Fant's (LF) model which was developed in late 1985. Due to the overly simplistic interpretation of voice source dynamics, LF model can not represent the fine temporal structure of glottal flow derivative realizations nor can it carry the sufficient spectral richness to facilitate a truly natural sounding speech synthesis. In this thesis we have introduced Characteristic Glottal Pulse Waveform Parameterization and Modeling (CGPWPM) which constitutes an entirely novel framework for voice source analysis, parameterization and reconstruction. In comparative evaluation of CGPWPM and LF model we have demonstrated that the proposed method is able to preserve higher levels of speaker-dependant information from the voice source estimates and realize a more natural sounding speech synthesis. In general, we have shown that CGPWPM-based speech synthesis rates highly on the scale of absolute perceptual acceptability and that speech signals are faithfully reconstructed on consistent basis, across speakers, gender. We have applied CGPWPM to voice quality profiling and text-independent voice quality conversion method. The proposed voice conversion method is able to achieve the desired perceptual effects and the modified speech remained as natural sounding and intelligible as natural speech. In this thesis, we have also developed an optimal wavelet thresholding strategy for voice source signals which is able to suppress aspiration noise and still retain both the slow and the rapid variations in the voice source estimate.

Statement of copyright

Declaration

The work described in this thesis has not been previously submitted for a degree in this or any other university, and unless otherwise referenced it is the author's own work.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AV GD	Average Group Delay
BAMS	Bayesian Adaptive Multiresolution Smoother
CGPW	Characteristic Glottal Pulse Waveform
CGPWPM	Characteristic Glottal Pulse Waveform Parameterization and Modeling
DE	Direct Estimation
DMOS	Degradation Mean Opinion Score
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
EGG	Electroglottographic signal
EM	Expectation Maximization
EW GD	Energy Weighted Group Delay
EWI	Enhanced Waveform Interpolative coding
FE	Fit Estimation
FEM	Finite Element Modeling
GCI	Glottal Closure Instant
GM	Glottal Matrix
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LF	Liljencrants-Fant's model
LPC	Linear Predictive Coding
LRT	Likelihood Ratio Test
ML	Maximum Likelihood
MOS	Mean Opinion Score
MXDV	Maximum Deviation
NZC	Negative Zero Crossing

PSOLA	Pitch Synchronous Overlap and Add
SNR	Signal to Noise Ratio
SNR_seg	Segmental Signal to Noise Ratio
TI-H	Translation Invariant Hard thresholding
TI-S	Translation Invariant Soft thresholding
WLC	Window Length Coefficient

Acknowledgments

I would like to thank my supervisor and my friend Prof. Saeed Vaseghi for his assistance throughout the course of my PhD research. I would like to extend my sincere gratitude to my parents and family for all their love, support, encouragement and confidence. I am also thankful to my fellow researches of the Brunel University speech processing lab. Our discussions played an important role in the development of this thesis. I would also like to extend my gratitude to my friends for their companionship, encouragement, and invaluable distractions.

I am sincerely thankful to all the people who have directly or indirectly contributed to this thesis.

Emir Turajlic

Chapter 1

Introduction

Although, a range of techniques have been developed to enable the study of laryngeal dynamics, e.g. electroglottography, electromagnetic glottography, transillumination and high-speed imaging, in this thesis we have opted for the closed-phase pitch synchronous inverse filtering of recorded speech as it is a non-invasive, does not cause any discomfort to the speaker and it does not require bulky or expensive equipment. However the analysis, parameterization and modeling of inverse filtering results are some of the most challenging areas of speech processing owing to the fact humans produce a wide range of voice source realizations and that the voice source estimates commonly contain artifacts due to the nonlinear time varying source filter coupling. Some examples of these distortions include first formant ripple that describes a sinusoidal like perturbation that overlays the glottal flow derivative waveform, the skewness to the right due to the inertive loading of subglottal and supraglottal systems and the nonlinear increase in glottal excitation strength when the first formant frequency appears near the multiple of pitch frequency. Over the years, two distinct approaches to voice source signal modeling have emerged, physical and analytical modeling. Physical models attempt to describe the laryngeal level of speech production system in terms of physiological quantities, whereby the glottal airflow is viewed as a product of viscoelastic-aerodynamic interactions in the glottis. While being a valuable tool for understanding the physiology of voice production, physically informed models have a very limited range of applications as they tend to have a large number of parameters and a high computational cost. On the other hand, the analytical approach to voice source modeling is less concerned with the underlying physical processes behind the vocal fold oscillations, but rather attempts to directly describe the glottal airflow waveforms with an opportune combination of mathematical functions. As a result they benefit from a reduced number of control parameters and improved computational efficiency.

Current glottal pulse models, including the popular Liljencrants-Fant's model have adopted overly simplistic interpretations of voice source dynamics, and yet, they do not offer the desired levels of modeling accuracy and parameterization robustness. Liljencrants-Fant's model does not have enough degrees of freedom to represent the fine temporal structure of glottal flow derivative realizations and in the best of circumstance it can only model their general shape. The deficiencies of the current analytical voice source models are perhaps the best revealed in the quality of synthetic speech where inadequate voice source modeling would produce distinct perceptual effects. We have to stress that analytical representation of voice source signal does not carry the sufficient spectral richness to facilitate a truly natural sounding speech synthesis. Clearly, there is a need for a more sophisticated voice source model that can enable a high quality voice source analysis and source-filter-based speech synthesis. With these motivations, we have developed Characteristic Glottal Pulse Waveform Parameterization and Modeling (CGPWPM) which is an entirely novel framework for voice source analysis, parameterization and reconstruction and a more accurate alternative to Liljencrants-Fant's model. CGPWPM is, to our knowledge, the only method that facilitates adaptive voice source modeling, where the form and structure of the model is directly dependant on the observed voice source signal. It is a fully automatic method, but we have described how CGPWPM can be employed very effectively semi-automatic manner as well. Another important feature of CGPWPM framework is that it provides the means to accurately estimate the statistical properties of non-stationary turbulent components related to aspiration noise. Unlike the vast majority of voice source parameterization techniques CGPWPM is able to produce accurate parameterization results over entire natural read speech sentences, and not just on sustained vowels and constrained segments of well behaved voice source signal. In comparison to Liljencrants-Fant's model, CGPWPM enables by far superior source filter based speech synthesis. The quality of CGPWPM-based synthetic speech is often perceptually indistinguishable from natural speech. However, the quality of the proposed method is best demonstrated by the fact that it is able to accurately parameterize and synthesize pathological voices where the corresponding voice source waveforms exhibit a range of complex and multidimensional dysphonic manifestations.

The above mentioned characteristics of CGPWPM, enable the proposed method to be applied to almost all areas of speech processing, but in particular to speech synthesis, speech coding, clinical research, voice quality profiling, voice and voice quality conversion, speaker identification and verification, speech enhancement.

We have to say that the proposed method has some small resemblance to Waveform Interpolative (WI) coding which is currently a state of the art speech synthesis and coding technique. However, the similarities only lie in the fact that both techniques make use of some type of characteristic waveform to perform reconstruction of acoustic signals. It is important to stress that that Waveform Interpolative coding is performed on either linear predictive coding (LPC) residue or speech waveform and not on any credible estimates of voice source signal. In WI coding the reconstruction process is based on simple linear interpolation of consecutive characteristic waveforms that were periodically estimated from the observed waveform. On the other hand, CGPWPM uses closed-phased pitch-synchronous closed phase inverse filtering to obtain high-quality voice source estimates. Voice source reconstruction is based on a single characteristic waveform and more importantly, the proposed method is able to parametrically represent the non-linear evolution of characteristic waveform in time. In this thesis, we have also attempted to develop the optimal wavelet thresholding strategy for voice source signals. Our principal aim was to preserve the shape of a non-stationary signal that is observed in additive noise for further glottal excitation analysis, e.g. voice source parameterization. We were in particular concerned with preserving the glottal pulse shape in the regions of glottal closure instants as we were aware that even a small degree of over-smoothing could considerably compromise the authenticity of parametric voice quality description. Having considered voice source analysis, parameterization, reconstruction and voice source denoising together with voice quality profiling and voice quality conversion, we can say that this thesis deals with all important aspects of voice source processing. The thesis is organized as follows:

In Chapter 1, an overview of speech production mechanisms is presented in order to provide background knowledge for Chapter 2 where we examine the principal approaches to speech modeling within the framework of source-filter theory. A discussion on the anatomy and physiology of the organ groups involved in speech production is provided. We also examine various phonation types and relate the temporal characteristics of the glottal airflow to the Laver's framework of voice quality. Furthermore, we describe the most common laryngeal disorders and their effect on voice quality. We present an overview of turbulence noise theory and describe the relationship between the voice quality and the aspiration noise levels along the auditory continuum.

In Chapter 2, we examine the principal approaches to speech modeling within the source-filter framework. The main focus is placed on voice source modeling. Physically informed models of voice source such as the one-mass, two-mass, body cover, and mucosal wave models are discussed in order to provide a platform for understanding the underlying principles behind the vocal fold dynamics. A review of the analytical voice source models is also presented. A simplified source-filter model, Childers' aspiration noise model, and Liljencrants-Fant model of glottal airflow are described in detail as they are used to synthesize tests signals throughout the thesis. We gain intuition for sound in the vocal tract by using a concatenated acoustic tube model and the traveling wave equations. Under the assumption of linearity and time-invariance, a transfer function of air flow velocity from the glottis to the lips is developed.

In Chapter 3, we present a group delay approach to glottal closure instant (GCI) estimation. Specifically, *average group delay* and *energy weighted group delay* measures are discussed in detail. Their properties are studied on synthetic and natural speech datasets. We have proposed a GCI estimation method that is based on a group delay algorithm and a translation-invariant hard-thresholding of the linear predictive coding residue. The performances of the two group delay measures and the proposed method are evaluated for a range of fixed and pitch-synchronous analysis window lengths. The optimal GCI estimation strategy is reported. In this thesis, we have adopted the closed-phase pitch synchronous inverse filtering of recorded speech as the means for obtaining the estimates of the glottal flow derivative waveforms. Nevertheless, we have also presented a brief overview of other approaches and methods that enable the examination of laryngeal dynamics during *voiced* phonation. Subsequently, we have described a formant modulation analysis technique that can be used to ascertain the extent of non-linear source-filter coupling. Both, the inverse filtering method and the formant modulation analysis are employed on a range of voice qualities to enable a qualitative evaluation of the temporal glottal excitation structure.

In Chapter 4, an overview of wavelet thresholding is presented including a range of commonly used thresholding methods such as Universal thresholding, SureShrink thresholding, Hybrid-Sure thresholding, Translation-Invariant thresholding, Hypothesis-

Testing-based thresholding, Block thresholding, and Bayesian Adaptive Multi-resolution Smoother. We have introduced a set of experiments designed to obtain the optimal denoising strategy for voice source signals. Note that the optimal denoising strategy for voice source signals is first developed on simulated signals and is eventually evaluated on the natural acoustic data

In Chapter 5 and Chapter 6, a novel framework for voice source analysis, parameterization and reconstruction is described. The framework is denoted as Characteristic Glottal Pulse Waveform Parameterization and Modeling. The novelty of this approach requires an extensive elaboration on a range of issues. As such, the problems of voice source parameterization and voice source reconstruction are treated separately. Chapter 5 describes the CGPWMP approach to voice source parameterization. The parameterization performance is evaluated on synthetic and natural speech datasets. The CGPWPM is applied to voice quality profiling where we aim to develop a parametric voice quality description for a range of voice quality types.

In Chapter 6, we will describe the voice source reconstruction aspect of the Characteristic Glottal Pulse Waveform Parameterization and Modeling (CGPWPM) system. Subsequently, CGPWPM is applied under the source-filter model of speech production to develop a speech synthesis and a voice conversion method. A comparative assessment between the Characteristic Glottal Pulse Waveform model and the Liljencrants-Fant model of the glottal flow derivative waveform is presented. Subjective A/B listening test are used to establish which of the two models provides a perceptually more acceptable synthetic speech. These two models are also evaluated in the context of speaker identification. The aim of the speaker identification experiment is to determine whether the fine structural elements of the glottal flow derivative waveform (which can be modeled by CGPWPM system) contain speaker identity related information. The quality of CGPWPM-based speech synthesis is formally evaluated via Mean Opinion Scores and Degradation Mean Opinion Scores. CGPWPM-based voice quality conversion is evaluated via subjective triadic comparison listening tests.

Chapter 2

The Human Voice Production System

ABSTRACT

In this chapter, we present an overview of speech production mechanisms. A discussion on the anatomy and physiology of the organ groups involved in speech production is provided. We examine the principal approaches to speech modeling within the source-filter framework. The main focus is placed on voice source modeling. Physically informed models of voice source such as the one-mass, two-mass, body cover, and mucosal wave models are discussed in order to provide a platform for understanding the underlying principles behind the vocal fold dynamics. A review of the analytical voice source models is also presented. We examine various phonation types and relate the temporal characteristics of the glottal airflow to the Laver's framework of voice quality. Furthermore, we describe the most common laryngeal disorders and their effect on voice quality. A simplified source-filter model, Childers' aspiration noise model, and Liljencrants-Fant model of glottal airflow are described in detail as they are used to synthesize tests signals throughout the thesis.

2.1 Source-filter theory of human speech production

From the physiological point of view, the voice production consists of three sub-components: the *larynx*, the *subglottal* area and the *supraglottal* area. The *subglottal* area consists of diaphragm, lungs and trachea. The *supraglottal* area refers to the speech production organs above the larynx; namely, vocal tract and lips. Figure 2.1 shows a cross-section of laryngeal and supralaryngeal area of speech production system

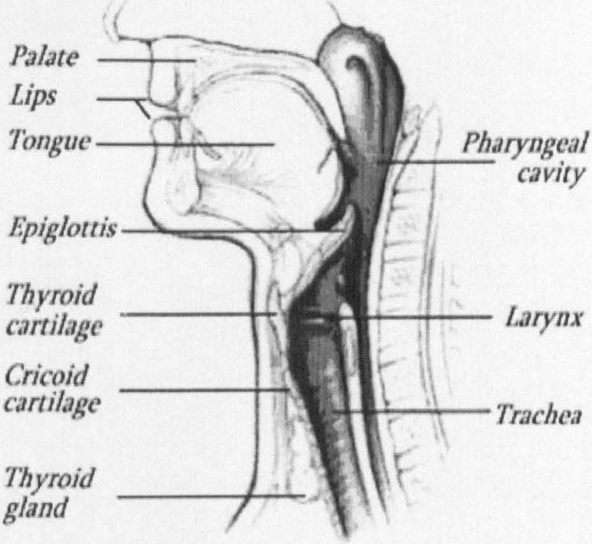


Figure 2.1: A cross-section of laryngeal and supralaryngeal area of speech production system. The image is obtained from [145].

The lungs act as a power supply and provide airflow to the laryngeal stage of speech production mechanism. The expending and contracting of lungs is referred to as *inspiration* and *expiration,* respectively. During *inspiration,* the air flows in the lungs through a relatively open glottis with an average area of 3-4 cm^2. The direction of the airflow is reversed during the *expiration*. The area of glottis depends on the type of *expiration*. During breathing, the glottal area is around 1 cm^2, while during phonation it varies around 0.05 to 0.1 cm^2. The glottal valve serves to control the airflow and the extent of utilized lung capacity. For the short utterances at normal loudness, only 25% of lung capacity is exploited. The louder and longer speech requires a greater use of the lung capacity.

In the larynx, the airflow from the lungs is modulated by the action of the vocal folds, to yield a *periodic* and a *turbulent* airflow source. The production of sound is characterized by the vibratory pattern of vocal folds and the configuration of laryngeal muscles and cartilages. The vocal tract spectrally shapes the voice source and gives the sound information related to the linguistic layer of speech communication. The vocal tract comprises of larynx tube, pharyngeal cavity, oral and nasal cavity. The position of velum controls the degree of coupling to the nasal tract. When velum closes, only oral sound is produced. The configuration of the articulators, i.e. jaw, tongue, lips and velum controls the vocal tract geometry and the spectral shape of the sound. The resonant frequencies of the vocal tract are called formant frequencies and they appear in spectrum as regions of concentrated acoustic energy. Following the sound initiation and the vocal tract spectral shaping, the variation of air pressure at the lips gives rise to a propagating sound wave that listeners perceive as speech.

In 1960, Fant [42] introduced the source-filter theory of human speech production. The theory postulates that the speech production can be viewed as a two stage process, where the first stage corresponds to the initiation of sound, and the second stage is responsible for the acoustic filtering. The voice source and the acoustic filter are considered independent of each other. The primary sources of voicing are attributed to the vibration of vocal folds and aspiration noise. The linear filter describes the spectral characteristics of the acoustic tube formed by the pharynx, oral cavity, and lips.

The source-filter model is an over-simplification of the speech production process. Fricative sounds which are created at the front of the oral cavity are not modulated by the resonances of the vocal tract to the same extent as the *voiced* and *aspirated* sounds. As such, the source filter model is not very accurate for fricative sounds. Under the source-filter speech production model, it is assumed that the glottal impedance is infinite and that the glottal airflow source is not affected by the vocal tract. In reality, the pressure in the vocal tract cavity just above the glottis "backs up" against the glottal flow and interacts non-linearly with the flow [45], [44], [118]. A range of manifestations of source-filter coupling have been identified. First formant ripple is often observed in the glottal flow derivative signal obtained via inverse filtering [21]. The glottal flow derivative waveforms may be skewed to the right

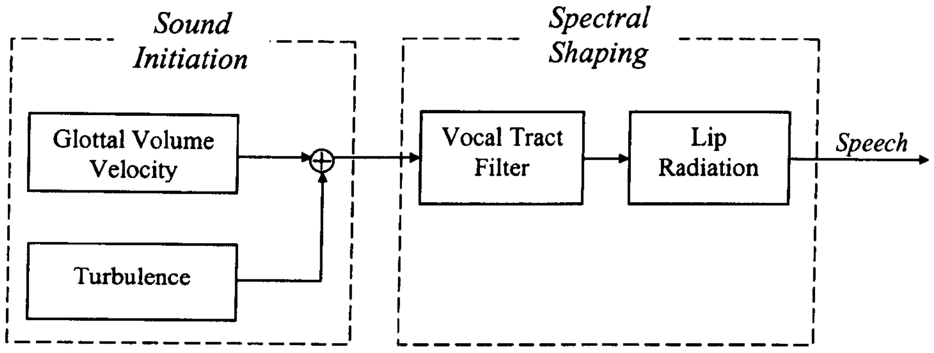


Figure 2.2: Source-filter model of the speech production system.

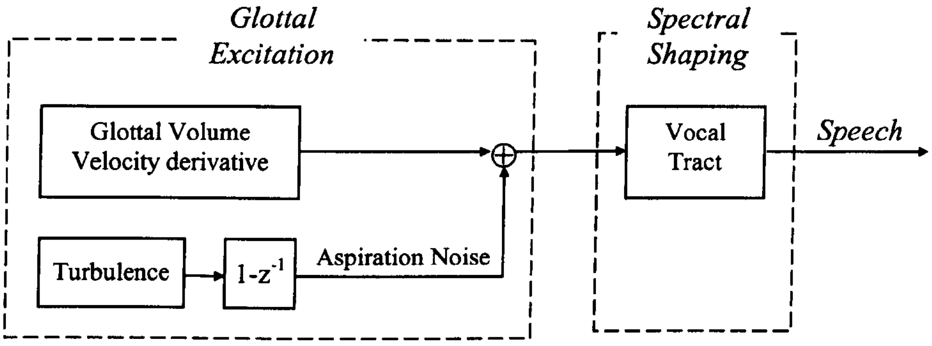


Figure 2.3: Simplified speech production model.

due to the inertive loading by the subglottal and supraglottal acoustic systems [119]. A nonlinear increase in the voice source strength can occur when the frequency of the first vocal tract resonance is near an integral multiple of glottal cycle frequency [4]. The temporally variant glottal impedance can also affect the vocal tract frequency response, especially in the region of first formant [4].

Nevertheless, due to the simplifications in speech analysis and processing, the source-filter theory is a widely adopted framework for speech modeling. For a wide range of speech realizations, especially for *voiced* speech, these secondary effects are negligible and the source filter model is perfectly adequate. Figure 2.2 shows the speech production system as a connection of three separate and independent processes: voice source, vocal tract and lip radiation. The voice source is spectrally shaped by the vocal tract and is thereafter radiated by the lips. The lip radiation is thought to have spectral properties similar to a differencing filter [49], [78]. If the two are equated, the lip radiation block can be replaced with the filter $R(z) = 1 - z^{-1}$. Linearity and short term time-invariance of the source-filter model allow the *vocal tract* and the *lip* radiation to be interchanged. Thus the speech production model can be simplified as shown in the Figure 2.3. The simplified speech production describes glottal excitation being filtered by the vocal tract to produce speech. The transfer function for the source-filter model of speech production is developed in Chapter 3, Subsection 3.3.1.

This chapter is organized as follows. In Section 2.2, we describe the anatomy and physiology of larynx. We also examine various phonation types and relate the temporal characteristics of glottal airflow to the Laver's framework of voice quality. We describe the most common laryngeal disorders and their effect on voice quality. Furthermore, we present an overview of physically informed and analytical approaches to voice source modeling. In Section 2.3, we discuss the physiology and anatomy of vocal tract and describe the concatenated lossless tube vocal tract model and its digital equivalent. Section 2.4 concludes the chapter with a discussion on turbulence noise production theory and with a brief overview of the prominent aspiration noise models.

2.2 Voice source

2.2.1 Anatomy and physiology of larynx

During the process of phonation, the larynx transforms the potential energy of the compressed air, below the larynx, into the kinetic energy of regressive airflow. If the speed of transformation is sufficiently high, the changes in the air pressure generate acoustic waves that propagate into the surrounding air. Figure 2.4 and Figure 2.5 illustrate the anatomy of larynx. The larynx is a musculo-cartilaginous structure supported by the muscles from the hyoid bone. The hyoid bone is a part of the laryngeal system but also provides a support for the tongue. The cartilage framework of the larynx consists of the thyroid, circoid, epiglottis and three paired cartilages, arytenoid, corniculate and cuneiform cartilages. The thyroid is the largest laryngeal cartilage. It is attached to the hyoid bone and circoid cartilage. The circoid cartilage is shaped like a signet ring with the anterior arch and a narrow convex ring. The epiglottis is a leaf like structure that is attached via ligaments to the base of the tongue, thyroid cartilage and the walls of the pharyngeal cavity. The epiglottis functions as a protection for larynx and prevents food from entering through glottis. The arytenoid cartilages approximate a pyramidal shape and are situated at the superior border of the circoid lamina. The corniculate cartilages are positioned at the top of each arytenoid and serve as a protection for the two arytenoid cartilages. The cuneiform elastic cartilages provide a support to the membrane of the aryepiglottic folds.

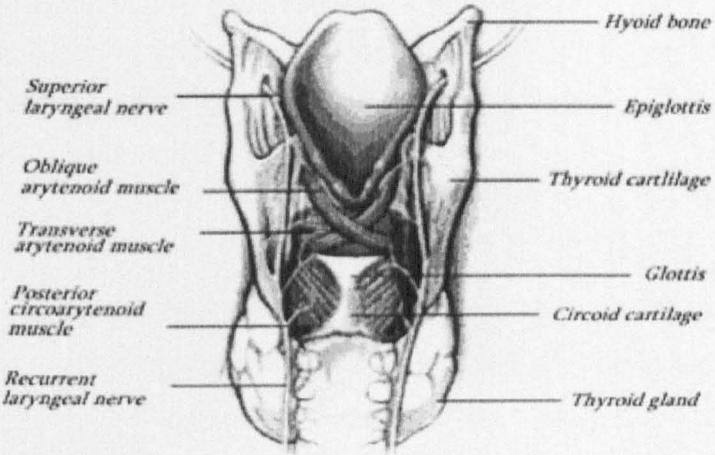


Figure 2.4: Anatomy of larynx - posterior view. The image is obtained from [145].

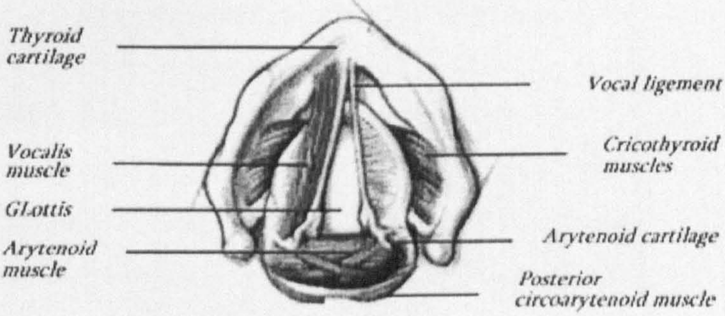


Figure 2.5: A cross- section of larynx as viewed from above. The image is obtained from [145].

The muscular structure of larynx is divided into two groups, the extrinsic and the intrinsic muscles. The extrinsic muscles, commonly referred to as strap muscles, provide a support to the larynx and their function is to move the laryngeal system as a whole. When swallowing the larynx is moved upward and the epiglottis folds down over it, covering the entrance of trachea and preventing the food from entering into the lungs. During yawning, the larynx is lowered in order to widen the airway. The intrinsic muscles control the movement of laryngeal cartilages during phonation. The posterior cricoarytenoid muscle is responsible for vocal fold opening. It is positioned on the posterior surface of the circoid cartilage and enables lateral displacement of vocal folds through rotation of arytenoid cartilage. The cricothyroid muscle runs from circoid to thyroid cartilage. It causes lengthening of vocal folds by elevating the circoid and lowering the thyroid cartilage. The change in vocal fold length affects the stiffness of vocal folds, which in turn affects the duration of the vocal fold oscillations. The lateral cricoarytenoid muscles start at the arytenoid muscle and run to the sides of the circoid cartilage. Their contraction causes the posterior part of the vocal folds to approximate. The interarytenoid muscles run horizontally between the two arytenoids and support the cricoarytenoid muscle.

The glottis is a slit-like orifice situated between the two vocal folds. The size of the glottis is controlled by the arytenoid cartilages and muscles within the vocal folds. The vocal folds are located at the narrowest portion of the airway and stretch between the front and back of the larynx, as illustrated in Figure 2.5. The vocal folds are made up of layered structure and they are approximately 17-24 mm long for male adults, and 13-17 mm long for female adults [72]. The outermost layer is a 0.05-0.1 mm thin skin of stratified squamous epithelium, below which are lamina propria and thyroarytenoid muscle. The lamina propria is itself a layered structure consisting of superficial, intermediate and deep layer. The superficial layer, approximately 0.5 mm thick, consists of loosely organized elastic elastin fibers wrapped in interstitial fluids. The intermediate layer is largely made up of the uniformly orientated (in anterior-posterior direction) elastin fibers, but it also contains some collagen fibers. On the other hand, the deep layer is primarily made up of collagen fibers. The collagen fibers are almost inextensible and serve to constrain the vocal fold elongation. The combined width of the intermediate and deep layer is around 1- 2 mm. The bulk of the vocal fold structure is made up of the 7-8 mm thick thyroarytenoid muscle fibers running longitudinally along the folds and laterally to the sides of lamina propria. This specific structure enables changes in the shape, thickness and elasticity of the vocal folds. At the front of larynx, the vocal folds are fixed to the thyroid cartilage. At the back and side of the larynx, the vocal folds are attached to the two arytenoid cartilages that can rotate along the circoid cartilages. The vocal folds are also free to move at the back and side of the larynx.

Helmholtz and Muller, and later Berg developed the myoelastic-aerodynamic theory to explain the dynamics of the vocal folds during phonation [11]. The theory states that the vibration of vocal folds is an interaction of two forces, the subglottal pressure that pushes the vocal folds apart, and the Bernoulli effect that causes the vocal folds to approximate each other and eventually close. Let us considered a case when the vocal folds are open, and the tension in the vocal folds is low. Let us also suppose that the lungs are contracting and causing an air flow through the larynx. The velocity of air will increase as it flows through the narrow glottis. This will result in a pressure drop along the margins of the vocal folds. When the pressure at the glottal margin drops below the pressure exerted by the tension of the vocal folds, the vocal folds will abruptly approximate each other. This phenomenon can be explained by Bernoulli's Principle (2.1). The principle states that as the velocity of the fluid increases across a plane, there is a pressure drop along the plane and consequently, less pressure is exerted perpendicular to the flow. In case of vocal fold oscillations, the Bernoulli principle predicts that continual effect of narrowing the glottis and increasing the air velocity will eventually force the vocal folds to close. The complete closure of glottis is only achieved if the surface of the vocal folds is soft and smooth.

$$\rho\frac{1}{2}(v^2 p) = const \qquad (2.1)$$

The parameters correspond to: *ρ - density of the fluid, p - pressure, v - velocity of the flow.*

As the lungs continue to contract, the closure of the vocal folds causes the build-up in the subglottal pressure. The pressure difference across glottis causes the separation of vocal folds. The folds continue to stretch outwards until the elastic forces cause it to contract back together for the next cycle of the vocal fold dynamics. The movement of vocal folds in the process of separation and closure is referred to as *abduction* and *adduction* respectively.

This is essentially a very simple view of vocal fold vibrations. More sophisticated models, such as the Isaka and Flanagan's *two-mass model* [75] or Titze's *body cover model* [138] are taking into account the more intricate aspects of vocal fold vibrations, e.g. a delay in the non uniform movement between the top and bottom edge of the vocal folds and the non-uniform pressure distribution at the glottis. Further elaboration on these two models will be presented in the next chapter.

2.2.2 Phonation

The definition of phonation, according to Laver, is the use of the laryngeal system to produce an audible source of acoustic energy [93]. The laryngeal configuration and the geometry of the glottis control the phonation and the voice texture. The adjustments to the laryngeal settings are made through changes in stiffness and thickness of vocal folds, the level of elevation/lowering of larynx, amount of adductive and abductive tension, as well as by the changes in the geometry of supraglottal structures. The tension and adjustment forces acting on vocal folds are illustrated in Figure 2.6.

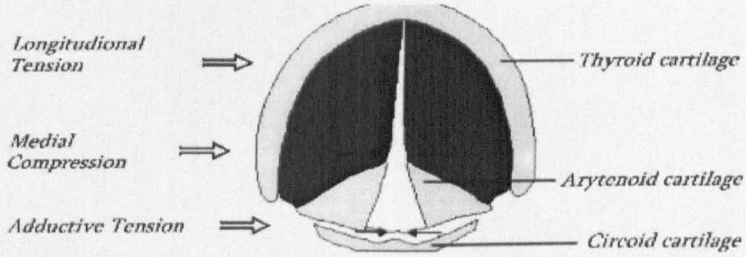


Figure 2.6: The tension and adjustment forces acting on vocal folds

Active and passive longitudinal tension in vocal folds arises from contractions of vocalis muscle and cricothyroid muscle, respectively. The contraction of the lateral thyroarytenoid muscle causes medial tension, while the contraction of the interarytenoid and the lateral cricoarytenoid muscle cause adductive tension. The tension and adjustment forces acting on the vocal folds affect the phonation process, dynamics of glottal airflow and the perceived voice quality.

The phonation can be characterized as *voiced*, *unvoiced*, or *mixed* depending on the extent vocal fold vibration. In *voiced* phonation, the vocal folds vibrate to produce a string of quasi-periodic glottal airflow waveforms. *Nil phonation* and *aspirated sounds* are examples of *unvoiced* phonation. *Nil phonation* is characterized by the lack of acoustic energy generated by the larynx, and it is associated with fricative sounds. The sound source is generated by the constrictions in the vocal tract In *aspirated sounds*, such as '*h*' in the word '*hello*', the turbulent airflow occurs at the glottis when the vocal folds are held partially open. Whispered speech is principally characterized by the *aspirated phonation*. *Mixed* phonation is a combination of *voiced* and *unvoiced* phonation. An example of this type of phonation is a phoneme [z] (as in <u>z</u>ebra), where the vocal folds vibrate and in the same time when the turbulent airflow is produced at the tip of the tongue near the teeth.

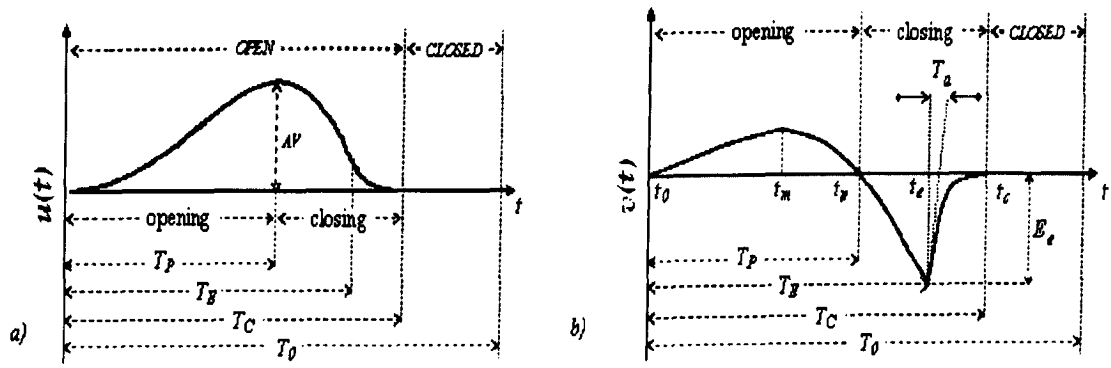


Figure 2.7: (a) Glottal volume velocity, (b) Glottal volume velocity derivative; The graphs are synthesized using the Liljencrants-Fant glottal flow derivative model. Note that in panel b) the lower case parameters indicate the significant instants in the glottal waveform irrespective of time origin. On the other hand the upper case parameters are referenced to the glottal opening instant, t_0.

Figure 2.7 shows an example of glottal volume velocity airflow and its derivative during a single cycle of *voiced* phonation. The important events in the glottal airflow dynamics are the instants of vocal fold abduction (t_0), the instants of maximum positive glottal flow derivative value (t_m), the instants of maximum glottal airflow (t_p), the instants of vocal fold closure onset (t_e), and the instants of complete glottal closure (t_c). The two graphs in Figure 2.7 illustrate the gradual build up of the glottal airflow during the opening phase of the glottal cycle and the rapid airflow decrease during the closing phase of the glottal cycle. During the opening phase, the vocal folds disconnect inferiorly and the opening travels upward with a wave-like motion in the mucous membrane. Often the opening occurs first on the superior surface as a small "chink" that spreads wide open in a zipper like manner [8]. The closing phase of the glottal cycle is initiated with the contact between the lower edges of the vocal folds. The closure subsequently proceeds along the length of the lower edge and eventually the mucosal layers of the folds come together. The highest rate of change of glottal airflow occurs at the glottal closure instant. Since the intensity of the produced acoustic wave is directly related to the intensity of glottal airflow derivative and not to the glottal airflow itself, the glottal closure instant corresponds to the instance of the strongest vocal tract excitation. The length of the glottal cycle T_0 is determined primarily by the subglottal pressure (pressure exerted by the lungs), vocal fold tension, their mass, length and elasticity.

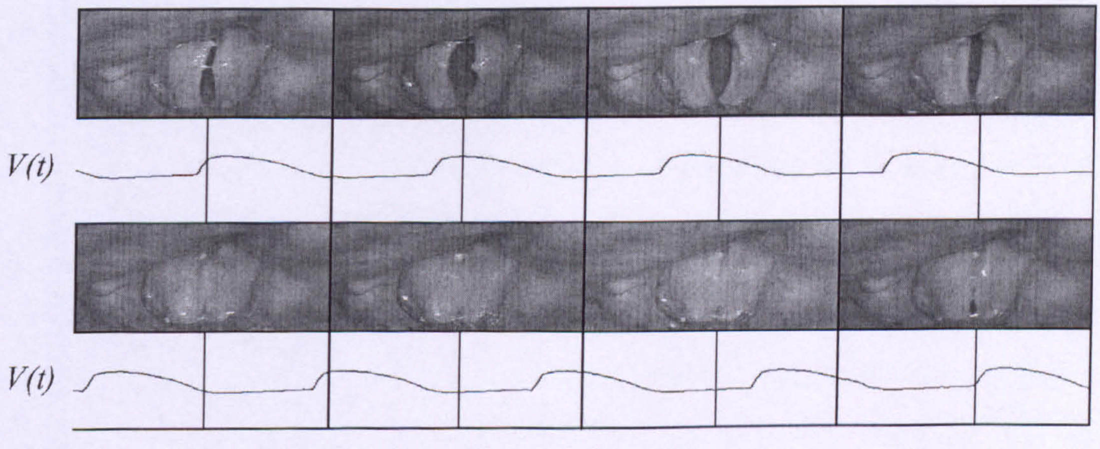


Figure 2.8: Glottal volume velocity graph with the stroboscopically derived images of the vocal folds at regularly spaced intervals during the glottal cycle for a modal voice of an adult male with an average pitch of 120 Hz. The image is obtained from [50].*

Laryngeal adjustments control the temporal characteristics of the glottal airflow and perceived voice quality during the *voiced* phonation. The *voiced* phonation can be broadly classified into five categories: *modal, creaky, harsh, breathy,* and *falsetto phonation.* These phonation types are not mutually exclusive and some of them combine to form compound phonations. When all muscular adjustments are on a moderate level and the pitch and loudness are at the normal conversational level, the resulting phonation is described as *modal voiced* phonation. The *modal* phonation is distinguished by periodic and complete closure of the glottis. While the vocal folds are separated, the glottis has a triangular form and it is the widest at the arytenoids. The vocal folds open and close with a slight vertical phase difference whereby the lower edges of the vocal folds move before the upper edges. The *modal* phonation occurs on average at 120 Hz for female adults and 220 Hz for male adults. Figure 2.8 illustrates the relationship between glottal flow and vocal fold dynamics during *modal* phonation. Note that the right arytenoid cartilage is visible in the upper left corner of the stroboscopically[*] derived images. The upper and lower sides of each frame correspond to the posterior and anterior side of vocal folds, respectively.

Creaky phonation, or *vocal fry,* occurs for weak longitudinal and strong adductive forces. At such muscular adjustments, the vocal folds thicken beyond the modal level. The heavy mass and low tensions produce slow and irregular vibrations. The typical frequency of vibrations for *creaky* voice is in the range of 25-50 Hz. Also, the glottal airflow rate (12-20 cc/s) is significantly lowered from the *modal* phonation (100-350 cc/s).

Harsh phonation is a result of very strong tension and adjustment forces acting on the vocal folds. The upper larynx is constricted and the ventricular folds are pressed against the vocal folds. This causes irregular amplitude and irregular frequency of vocal fold vibration.

Breathy phonation occurs for moderate longitudinal tension, weak medial compression, and very weak adductive tension. The main characteristics of *breathy* voice are incomplete

[*] Stroboscopy is a widely used technique for obtaining a video sequence of the vocal fold vibration. A flashing light source illuminates the glottis. The frequency of flashing is adjusted slightly below the frequency of vocal fold oscillations; each flash occurs at a slightly later phase of glottal cycle than the previous one.

glottal closure and elevated aspiration noise levels. The glottal airflow is higher than for the *modal* voice, while the frequency of vocal fold vibrations is only slightly lower.

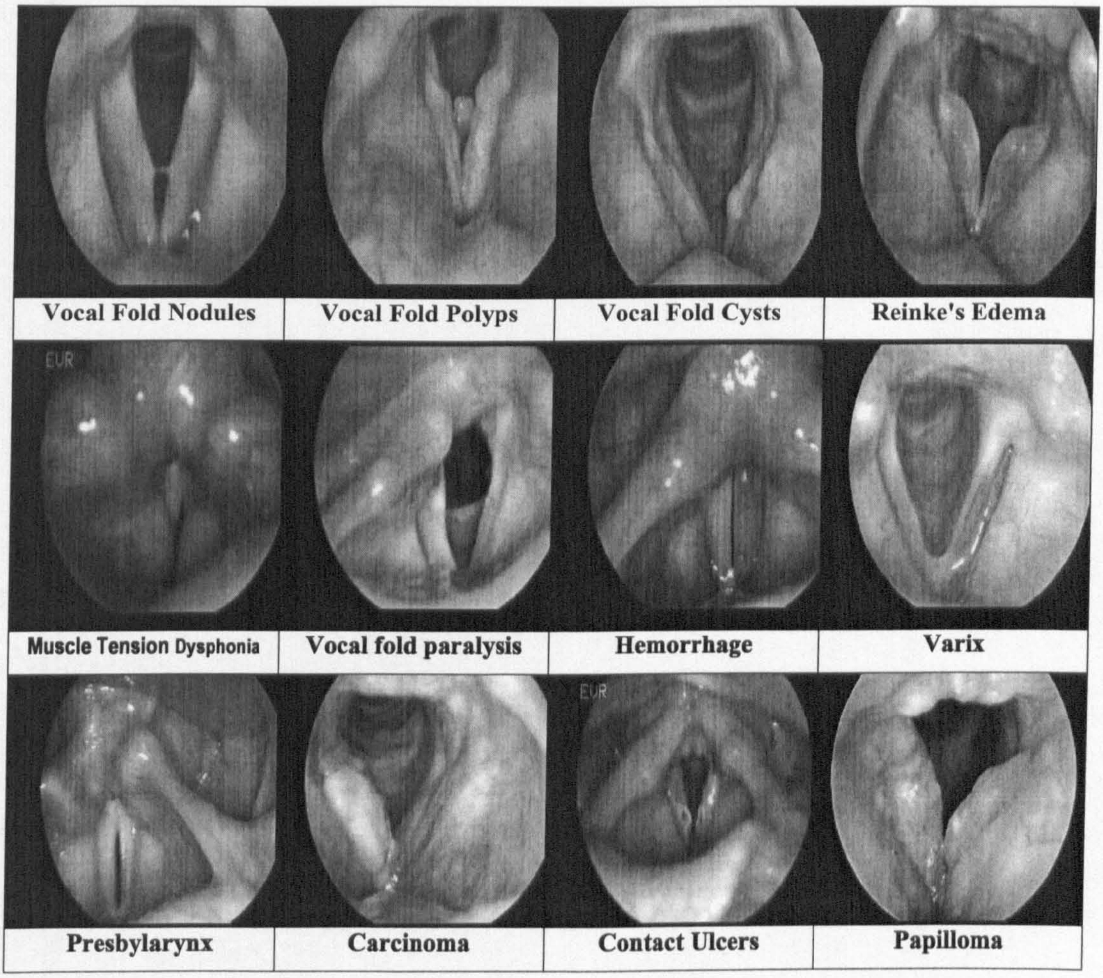
Falsetto phonation is associated with high frequencies of vocal folds vibrations. The adductive tension and the medial compression are strong. The strong longitudinal forces cause the stretching and thinning of the vocal folds. The closure of the glottis is often incomplete and a certain amount of audible turbulence noise is present. Also, the vocal folds tend to move without the vertical phase difference.

Although, the laryngeal settings are the most important factor affecting the voice quality perception, the supralaryngeal settings can also exert some influence on voice quality. Depending on the overall muscular tension settings, both laryngeal and supralaryngeal, the deviation from neural voice quality can be characterized as *lax* or *tense.*

2.2.3 Dysphonia

Dysphonia is defined as an abnormal voice quality. The perception of dysphonia is an important indicator of a wide range of structural, medical, neurological, or behavioral conditions. Voice pathologies are relatively common affecting about 5% of the population [10]. In spite of the fact that there are relatively large number of existent methods for laryngeal and vocal tract diagnosis (laryngoscopy, glottography electromyography, videostroboscopy, videokymography), the researchers are becoming increasingly interested in acoustic analysis of pathological voices [30], [69], [109], [101], [104]. The trend is due to the inherent non-invasive nature of acoustic analysis and its potential to provide quantitive information for objective diagnoses, with less strain on personnel resources and time.

Parametric description of pathological voices is a challenging problem due to the high complexity and multidimensionality of the dysphonic manifestations in the acoustic signal. Most of the work has been focused on perturbation analysis measures. "*Hoarseness*" (in this context it does not refer explicitly to the amount of noise in speech) is a loosely used term describing the perceived voice pathology due to the abnormality at the laryngeal level of voice production. "*hoarseness*" can be quantified in terms of *breathiness* (a degree of non-modulated turbulence noise in produced sound), *hoarseness* (amount of noise in produced sound), and *roughness* (extent of irregular fluctuation in the duration of vocal fold cycle) [71], [92]. In this section, we will describe some of the most common laryngeal disorders and their effect on voice quality. The voice disorders are grouped according their causes: vocal overuse and misuse, nervous system disturbance, disease or trauma. Figure 2.9 shows the stroboscopically derived images of the vocal folds corresponding to the considered voice pathologies.



Vocal Fold Nodules

Vocal Fold Polyps

Vocal Fold Cysts

Reinke's Edema

Muscle Tension Dysphonia

Vocal fold paralysis

Hemorrhage

Varix

Presbylarynx

Carcinoma

Contact Ulcers

Papilloma

Figure 2.9: Stroboscopically derived images of common voice pathologies (obtained from [107]).

Voice Disorders Related to Vocal Overuse and Misuse

Vocal fold nodules are common benign vocal fold lesions. They are usually bilateral and occur at the junction of the anterior 1/3 and posterior 2/3 of the vocal folds. They may vary significantly in size. Voice quality characteristics include hoarseness, breathiness, and lowered pitch. *A vocal fold polyp* is a fluid-filled lesion that occurs either unilaterally or bilaterally. It is often manifested through hoarseness, breathiness, diplophonia (audible perception of two distinct pitches). *Vocal fold cyst* is a fluid-filled growth. A vocal cord cyst can cause hoarseness, breathiness, voice and pitch breaks. *Reinke's Edema* or Polypoid *Degeneration* describes a condition when a membranous portion of the vocal folds is filled with fluid. It is caused by long-term smoking and chronic vocal overuse or misuse. Voice quality characteristics associated with Reinke's Edema include lowered pitch and severe hoarseness. Excessive and unnecessary tension of laryngeal muscles during vocal fold vibration is referred to *Muscle Tension Dysphonia* (MTD). It is thought to be a compensatory mechanism in the presence of an underlying laryngeal pathology.

Voice Disorders Related to Nervous System Disturbance

Vocal fold paralysis[†] / *paresis*[‡] may result from a viral infection, cerebral vascular accident (stroke), trauma to the head, recurrent laryngeal nerve damage following surgery to the head, neck, or chest region, or may be idiopathic (cause unknown). A tumor may also cause immobility of the vocal fold(s). Vocal characteristics consistent with vocal fold paresis/paralysis include breathiness, hoarseness, diplophonia (audible perception of two distinct pitches), decreased pitch range, and an inability to increase loudness.

Voice Disorders Related to Disease or Trauma

A *hemorrhage* of the vocal fold occurs when a blood vessel bursts and bleeds into the submucosal vocal fold layer. It is caused by voice overuse in combination with the intake of the anticoagulants such as aspirin or persistent usage of steroidal inhalants. Dysphonia is manifested through hoarseness, loss of pitch range, and vocal fatigue. A *varix* is a long,

[†] complete absence of movement
[‡] weakness in the movement

defined blood vessel on the vocal fold surface. It is thought to be caused by vocal overuse or misuse in a single traumatic episode. Vocal quality may be hoarse. *Presbylarynx* is a condition that is caused through normal aging of the larynx and is manifested by the loss of vocal fold tone and elasticity, hoarseness, breathiness, decreased loudness, and pitch instability. The appearance of "bowed" vocal folds secondary to the vocal fold atrophy is a symptom of presbylarynx. *Cancer* can affect the oral, pharyngeal, or laryngeal cavities. Laryngeal cancers are generally caused by a chronic irritation due to cigarette smoke. Laryngeal cancers are often life-threatening (if not detected early), and have severe affects on voice quality as well as breathing, and swallowing. A contact ulcer is a small lesion that typically develops on the medial portion of the vocal folds. Contact ulcers are most often caused by vocal overuse and misuse, laryngopharyngeal reflux, smoking, and excessive alcohol consumption. Voice often exhibits hoarseness, breathiness, lowered pitch, as well as the decreased pitch range. *Papillomas* are lesions that can run deep into vocal fold tissue. Papillomas are thought to be caused by viruses and can form throughout the larynx and upper airway. The symptoms are severe hoarseness and breathiness, and in some cases, respiratory problems.

2.2.4 Voice Source Modeling

In general, the glottal excitation models can be classified into two broad groups, the physically informed and analytical models. Physical models attempt to describe the laryngeal level of speech production system in terms of physiological quantities. The glottal airflow is viewed as a product of viscoelastic-aerodynamic interactions in the glottis. On the other hand, the analytical models ignore the vocal fold physiology and are primarily concerned with the representation of the glottal airflow or the glottal airflow derivative waveforms.

2.2.4.1 Physically informed models

In the early 1950's, myoelastic-aerodynamic theory was the principal description of the vocal fold oscillations. Although, the theory explains the basic principals behind the vocal fold behavior, it lacks sophistication to account for more intricate features, such as the non-linear viscoelasticity of vocal fold tissue, non-linear interaction between the glottal airflow and glottal area, collision of the opposite vocal folds, vertical phase difference of vocal folds, etc...

Over the years new theories have been developed that incorporate a wider range of physiological details of voice production system. *Finite-element* models provide a very accurate description of vocal fold physiology and explain a range of oscillation patterns. However they are computationally very intensive. *Lumped-element* models, such as the *mass-spring* and the *body-cover* model attempt to reduce the modeling complexity and computational intensity, while retaining the representation of the most significant characteristics of vocal fold oscillations. The physical voice source models provide a platform for understanding the underlying principles behind the laryngeal physiology. They can also be used in voice synthesis and in the study of voice pathology. However, they typically involve a large number of parameters that are difficult to estimate and control in a manner that is required by most practical applications.

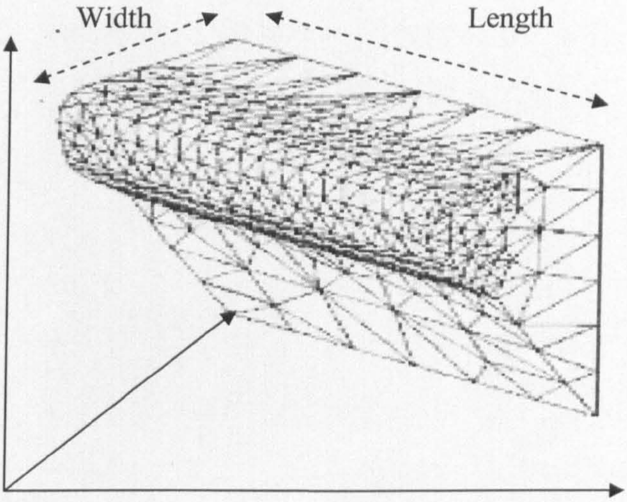


Figure 2.10: Finite element model

Finite elements and continuum models of the vocal folds

Finite-element [144] and *continuum* models [12], [13] were developed in order to provide a highly informative physiological description of vocal folds. The models are based on the simulations of the distributed mechanical displacements and aero-dynamical forces in the glottis.

Finite-element modeling (FEM) uses a large number of mass-like elements to obtain an accurate 3-dimensional representation of vocal fold structure, see Figure 2.10. The FEM models are able to describe the propagation of oscillations within the vocal-fold tissues through the viscoelastic and aerodynamic equations. The model describes a wide range of features of vocal fold dynamics including distributive effects, boundary conditions and a variety of vibration modes. However, finite element modeling is computationally highly expensive and its use is almost completely restricted to research applications.

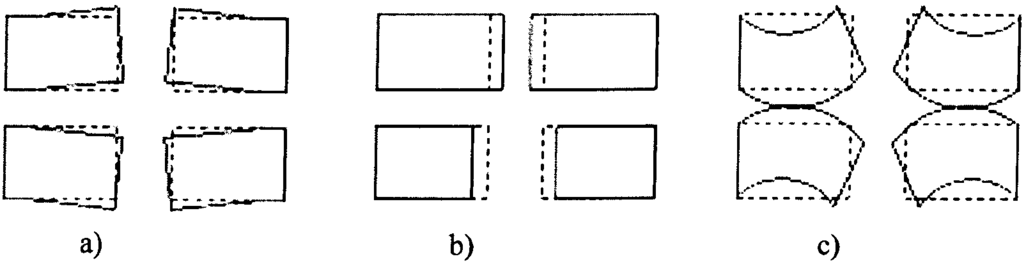


Figure 2.11: Continuum model - Three principle eigenmodes of vocal fold vibration. Dashed boxes describe vocal folds in the rest position.

In continuum models, the vocal folds are assumed to take a restricted number of simple shapes such as a rectangular parallelepiped, see Figure 2.11. The tissue properties are uniform in the plane normal to the longitudinal direction [137]. The basic concept behind the continuum models is to represent complex vibration patterns in terms of a small number of orthogonal modes and thus reduce the modeling complexity and computational intensity. In comparison to the finite-element model, the continuum model offers a less accurate representation of the interactions between the aerodynamic flow and the elastic vocal fold tissue.

On the other hand, the lamped elements models attempt to capture the essential traits of vocal fold vibrations with as few control parameters as possible, such that their use would not be restricted by computational intensity. This is achieved by lamping the large portions of vocal folds anatomy into discrete mass elements.

The one-mass models

In 1968, Flanagan & Landgraf proposed the first lumped element model of the vocal fold oscillations [48]. The *one-mass* model is illustrated in the Figure 2.12a). The model is symmetrical and each vocal fold is represented by identical mass-dumper-spring system. The motion of the vocal folds is restricted to horizontal direction.

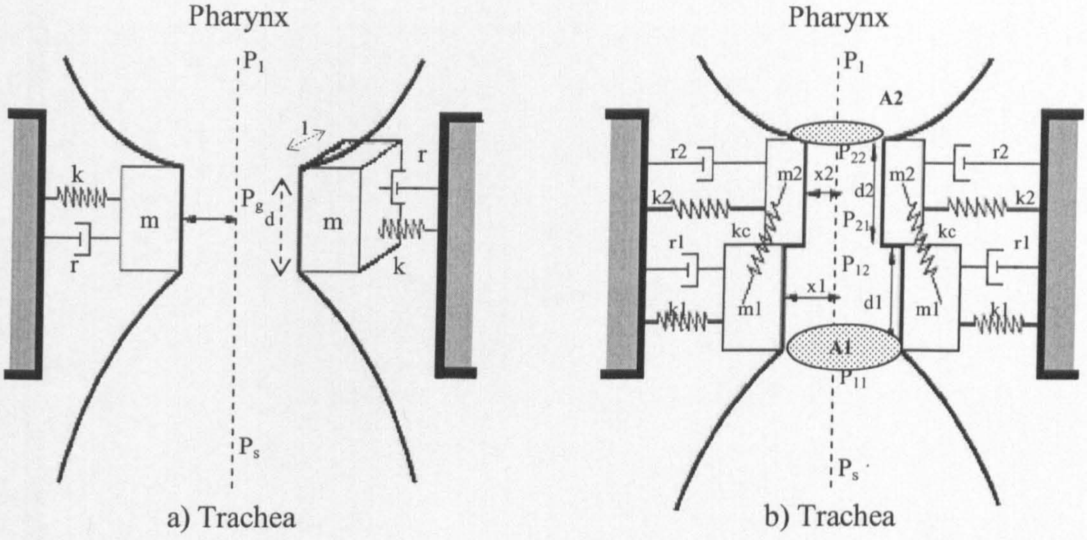


Figure 2.12: Mass–spring models of vocal folds. a) one-mass model; b) two-mass model

In this vocal fold model, the effect of vocal fold load on larynx is ignored. It is also assumed that the supraglottal pressure is 1 atmosphere and the subglottal pressure P_s is of constant value. Under such conditions, vocal fold dynamics is represented by following equation:

$$m\ddot{x} \;+\; r\dot{x} \;+\; kx \;=\; d\,l\,P_g(x) \tag{2.2}$$

The positive constants: *m*, *r* and *k* represent mass, stiffness and viscous damping of vocal folds, respectively. The length and width of vocal folds is represented by *d* and *l*, respectively. The glottal airflow dynamics may be described by Bernoulli's equation. The glottal air-pressure, P_g depends on the geometry of the glottis, and as such it is a function of vocal fold displacement, *x*. The glottal air pressure decreases with the increasing area of the glottis and it reaches its maximum value for a fully closed glottis. The product of $d\,l\,P_g$ describes the aerodynamic force that acts perpendicularly to the vocal fold tissue surface.

$$\frac{d}{dt}(E_k + E_p) = -\,r\,\dot{x}^2 < 0 \tag{2.3}$$

Equation (2.3) shows that the derivative of the sum of the potential and kinetic energy of the system is always negative. The total energy of the system decreases along trajectories and thus, the model can not produce self-sustaining oscillations. Furthermore the system has no limit cycles due to the lack of the necessary degrees of freedom in the oscillatory system. However, self-sustaining oscillations do occur when the inertive load of vocal tract is added to the model [138], [140]. The *one-mass model* allows only lateral movements of oscillating mass and is not able to incorporate vertical phase difference of vocal folds. The observations of the vocal fold movement demonstrate that the vocal fold tissue displacement is non-uniform and that the upper and lower portions of the vocal folds move out of phase. Hence, a more complex *two-mass model* was developed to incorporate the finer features of vocal fold dynamics.

The two-mass models

In 1972, Ishizaka & Flanagan proposed the *two-mass* model of the vocal fold oscillations [75]. According to many researchers, the *two-mass* model adequately describes the most relevant features of vocal fold oscillations: self-sustaining oscillations and the phase difference in motion of upper and lower edge of vocal folds.

The *two-mass model* is illustrated in the Figure 2.12b). The vocal folds are assumed to be bilaterally symmetric and are both modeled by identical pair of masses. As in the *one-mass* model, the masses are allowed only lateral motion. Each mass is subjected to elastic and dissipative forces, and as such this mechanical system can be described as a second order oscillator. The source of oscillation damping is provided by the viscous resistance of the vocal folds and larynx tissues. Further damping is also induced by the adhesiveness of the soft and moist contact surface of vocal folds during their contact. The properties of the vocal fold tissue primarily determined by the levels of muscular tension and vocal fold elongation. For accurate representation of the vocal folds, the stiffness parameters, k_1 and k_2, are modeled as non-linear quadratic functions of their corresponding displacement. The stiffness of the coupling spring, k_c describes the vocal fold stiffness in a direction perpendicular to the direction of the vocal fold oscillations. The coupling spring stiffness is non-linear. Its value is governed by the vocal fold tension and width, and can vary systemically. The collision between the folds is modeled by a restoring contact force. The overall mechanical system can be described by (2.4). The equations relate pressures p_{m1} and p_{m2} to the driving surfaces of the masses, $d_1 l$ and $d_2 l$.

$$\begin{aligned} m_1 \ddot{x}_1 + r_1 \dot{x}_1 + k_1 x_1 + k_C (x_1 - x_2) &= d_1 l p_{m1} \\ m_2 \ddot{x}_2 + r_2 \dot{x}_2 + k_2 x_2 + k_C (x_1 - x_2) &= d_2 l p_{m2} \end{aligned} \tag{2.4}$$

Ishizaka and Flanagan described the pressure distributions inside the glottis via successive discrete steps p_{ij}, see Figure 2.12b). The have assumed that the dimensions of the glottis are small compared to the wavelengths of the produced sound. Furthermore, the glottal flow is viewed as quasi-stationary. This follows the assumption that the glottal flow velocity is significantly higher than the period of the vocal fold oscillations.

$$\begin{aligned} p_s - p_{11} &= 0.69 \, \rho_{air} \frac{U_g^2}{A_1^2} \\ p_{11} - p_{12} &= 12 \, v d_1 \frac{l^2 U_g}{A_1^3} \\ p_{12} - p_{21} &= \frac{1}{2} \rho_{air} U_g^2 \left(\frac{1}{A_2^2} - \frac{1}{A_1^2}\right) \\ p_{21} - p_{22} &= 12 \, v d_2 \frac{l^2 U_g}{A_2^3} \\ p_{22} - p_1 &= \frac{1}{2} \rho_{air} \frac{U_g^2}{A_2^2} \left[2 \frac{A_2}{S} \left(1 - \frac{A_2}{S}\right) \right] \end{aligned} \tag{2.5}$$

The pressure distribution along the glottis is described by the set of equations (2.5). The pressure drop at the inlet of the glottis $p_s - p_{11}$ is obtained from the Bernoulli law for an ideal fluid in a static system. The pressure drops along the masses, $p_{11} - p_{12}$ and $p_{21} - p_{22}$, are directly proportional to the shear air viscosity, v. At the transition point between the masses m_1 and m_2, the glottal volume flow is of constant value. However the velocity of the air particles is different due to the sudden change in the glottis area. The pressure change

$p_{12} - p_{21}$ is equal to the change in kinetic energy per unit volume of the fluid. At the upper edge of the glottis, the pressure approximates to the atmospheric value. The pressure drop $p_{22} - p_1$ is dependant on vocal tract input area, S.

In order to complete the description of the mechanical model, the driving pressures of the two masses, p_{m1} and p_{m2} are related to the pressure distribution along the glottis, as in (2.6). The driving pressures are obtained by computing the mean pressure along the masses.

$$\begin{aligned} p_{m1} &= \frac{1}{2}\left[p_{11} + p_{12}\right] = p_s - 0.69\,\rho_{air}\,\frac{U_g^2}{A_1^2} + 6\,\nu d_1\,\frac{l^2 U_g}{A_1^3} \\ p_{m1} &= \frac{1}{2}\left[p_{21} + p_{22}\right] = p_1 + \frac{1}{2}\,\rho_{air}\,\frac{U_g^2}{A_2^2}\left[2\frac{A_2}{S}\left(1-\frac{A_2}{S}\right)\right] + 6\,\nu d_2\,\frac{l^2 U_g}{A_2^3} \end{aligned} \tag{2.6}$$

The model can be implemented and coupled to the vocal tract model numerically using the Euler method. The *two-mass* model captures the two eigenmodes of vocal fold oscillation. The two eigenmodes are equivalent to those of distributed model. The eigenmodes shown in Figure 2.11b) *and* Figure 2.11c) correspond to the two masses moving in phase, and 180 degrees out of phase, respectively. The model is able to incorporate subtle features of speech production, such as the acoustic interaction between the glottis and vocal tract. As a result of this interaction the model is able to synthesize a more natural speech where the changes in the glottal cycle period and the open quotient due the vocal tract load are adequately represented.

Since the introduction of the *two-mass* model, a great deal of research effort has been done to overcome some of its limitations [87], [94]. For example, the *two-mass* model makes an assumption that the elastic structure of the vocal folds is fixed to a rigid wall. This is not strictly true as repeated experiments have confirmed the existence of surface waves radiation from the throat during the *voiced* phonation. In the *two-mass* model the glottal areas are described as squares and this approximation usually leads to over-abrupt glottal closures and more intense glottal peaks. The model requires estimation of up to 19 parameters and as such it is still computationally intensive. Furthermore, the accuracy of the model is compromised by the abrupt pressure drop at the junction of the two masses.

In contrast, the *body-cover* model of vocal folds obtains an improved pressure distribution in the glottis and is also able to describes the surface wave propagating ahead of the vocal folds

Body-cover and mucosal wave models

Based on the examination of mechanical properties of the vocal fold tissue, Hirano [70] suggested that the vocal fold structure can be split into two distinct components with specific mechanical characteristics. The two components correspond to the outermost *cover* layer, consisting of a pliable tissue (the epithelium and the superficial layer) and the inner or *body* layer, consisting of muscle and ligament fibers. The results of further investigation of the vocal fold dynamics suggest that there is a wavelike motion in the superficial mucosal tissues during the *voiced* phonation. On the basis of these results, Titze developed a *body-cover model* of vocal folds [138]. The *body-cover* model is illustrated in the Figure 2.13a). In addition to enabling representation of the surface wave propagation, the *body-cover model* also has physiologically more realistic control parameters.

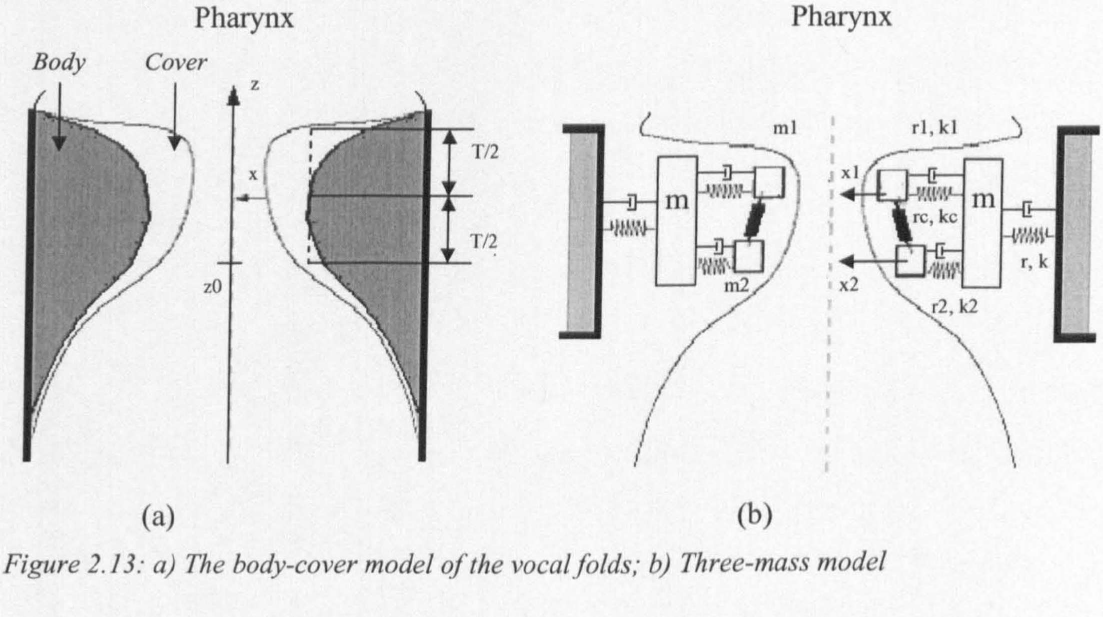


Figure 2.13: a) The body-cover model of the vocal folds; b) Three-mass model

The mechanical aspect of the *body-cover* model can be described as a second order oscillator.

$$m\ddot{x} \; + \; r\dot{x} \; + \; kx \; = \; f(\dot{x},x) \tag{2.7}$$

In this representation, the aerodynamic driving force is a function of velocity as well as of displacement. This implies that the driving force has a component in phase with tissue velocity, and as such it allows the energy transfer from the glottal airflow to the vocal fold tissue. The motion equations for the *body-cover* system are coupled to the aerodynamic driving force via the glottal area. The driving pressure exerts a force on the *cover* masses to produce oscillations. The average driving pressure, P_m is a function of trans-glottal pressure $(P_s\text{-}P_1)$ as in (2.8).

$$P_m \;=\; P_1 + \frac{(P_s - P_1)\,(1 - a_2 / a_1 - k_e)}{k_t} \tag{2.8}$$

, where P_s and P_1 denote the subglottal pressure and the vocal tract input pressure, respectively. The parameters, k_t and k_e describe the pressure loss and recovery coefficients. The two coefficients take into an account kinetic losses due to the turbulent behavior of the airflow at the glottal expansion region. The magnitude of the driving pressure is to a large extent governed by the glottis areas in the regions of the upper lower *cover* mass, a_1, and the lower *cover* mass, a_2. The cross-sectional areas at glottal entry and exit are approximated to:

$$\begin{aligned} a_1 &= 2l_g\,(x_{01} + x + \tau\dot{x}) \\ a_2 &= 2l_g\,(x_{02} + x - \tau\dot{x}) \end{aligned} \tag{2.9}$$

, where l_g corresponds to the length of the glottal folds in antero-posterior direction and τ denotes the time interval for the mucosal wave to travel half the glottal width (the length of the glottis in the direction of the airflow). Since, the driving aerodynamic force of the mechanical system is proportional to the average driving pressure and the driving surface area, the vocal fold motion equation is expressed in terms of subglottal pressure, as in (2.10)

$$m\ddot{x} \ + \ r\dot{x} \ + \ kx \ = \ d\,l_g \frac{P_s\,(x_{01} - x_{02} + 2\tau\,\dot{x})}{k_t\,(x_{01} + x + \tau\,\dot{x})} \tag{2.10}$$

In order to analyze the behavior of vocal folds at the equilibrium position $x = 0$, equation (2.10) is linearized around $x = 0$:

$$m\,x + (r - \frac{2\tau\,d\,l_g P_s}{x_0})\dot{x} + k\,x \ = \ 0 \tag{2.11}$$

It is clear that the glottal airflow induces negative damping on vocal folds. For small values of subglottal pressure, P_s, the overall damping is positive and the equilibrium state is stable. However, for large values of P_s, the overall damping is negative due to the net energy transfer from the glottal airflow to the vocal folds. This energy is used to overcome the energy dissipated in the tissue and maintain the vocal fold oscillations. The amplitude of vocal fold oscillations is limited by the collisions of the opposite vocal folds and other nonlinear effects.

The fact that the *one-mass* and *two–mass* models do not capture the layered structure of vocal folds, has led Story and Titze to propose a lumped-element approximation of the *body-cover* model [127]. In order to maintain the low-dimensionality, they have simply added a third mass to the *two-mass* model to describe the "*body*" of vocal folds. The resulting *three-mass* model, Figure 2.13b) is able to replicate the *body-cover* coupled oscillations. It also provides physiologically realistic parameters. Unlike the *two-mass* model, it is able to adequately describe a naturally common situation when a contraction of thyroarytenoid muscle stiffens the *body*, without any affect on the stiffness of the *cover*.

The complex larynx physiology presents a continuous challenge to researchers. There is a need to retain a relative degree of simplicity and a reduced number of control parameters so as to improve the computational efficiency of the models. On the other hand, there is a need to improve the faithfulness of the voice source model and also a desire to include the complex non-linear phenomena in the vocal fold dynamics, such as chaotic behavior [68], [79], limit cycles [68], [97] and bifurcations [140], [98].

2.2.4.2 Analytical models

Analytical models are less concerned with the underlying physical processes behind the vocal fold oscillations, but rather attempt to directly describe the glottal airflow waveforms with an opportune combination of mathematical functions. The advantage of this approach to voice source modeling is reduced number control parameters. Generally, the parameters of the analytical glottal airflow models can be related to speech perception, voice quality and to a certain extent they can also describe some aspects of the voice source physiology. Many state of the art speech analysis, speech synthesis, speech coding and speech morphing systems employ the analytical models of the glottal airflow. Over the past decades of research, a number of analytical glottal flow models have been proposed in the literature: Hedelin [65], Fant, Liljencrants and Lin [43], Fujisaki and Ljungqvist [53], Klatt & Klatt (RK model) [86], Milenkovic [105], Childers and Hu [22], Veldhuis [141]. The Liljencrants-Fant (LF) model [43] is the most widely accepted voice source model as it is able to adequately represent a wide range of natural glottal airflow variations.

Liljencrants-Fant (LF) model

The LF-model is a parametric description of glottal airflow derivative cycle in the time domain. The LF-model is defined by the following equations:

$$
\begin{aligned}
v(t) &= E_0\, e^{\alpha t} \sin(\omega_g\, t) && ,0 \le t < T_e \\
v(t) &= \frac{E_e}{\varepsilon T_a}\left[e^{-\varepsilon(t - T_e)} - e^{-\varepsilon(T_c - T_e)}\right] && ,T_e \le t < T_c \\
v(t) &= 0 && ,T_c \le t < T_0
\end{aligned}
\tag{2.12}
$$

The constraints in (2.13) are placed on the LF-model in order to satisfy the continuity of the model. Furthermore, the constraints take into account that no air flow takes place during the closure of glottis and that there is no net gain in the airflow over a glottal cycle.

$$\begin{aligned} \int_0^{To} v(t)dt &= 0 \\ \omega_g &= \frac{\pi}{T_p} \\ \varepsilon T_a &= 1 - e^{-\varepsilon(T_c - T_e)} \\ E_0 &= -\frac{E_e}{e^{\alpha T_e}\sin(\omega_g T_e)} \end{aligned} \tag{2.13}$$

The glottal excitation is characterized by the glottal cycle duration, amplitude and glottal shape. An example of the glottal airflow derivative obtained via LF-model is shown in the Figure 2.7. The parameters, E_e and T_0 denote the maximum glottal airflow declination rate and the glottal cycle duration, respectively. The shape of the glottal flow derivative is determined by the following parameters $T_{p,}$ $T_{e,}$ T_c and T_a. They are often referred to as *timing parameters* of the LF-model. They correspond to the significant events in the glottal airflow cycle in relation to the glottal opening instant. The parameters T_p and T_e denote the instants of maximum glottal airflow and vocal fold closure onset, respectively. T_c marks the instant of complete glottal closure and T_a describes the effective duration of the return phase. T_a is defined as time interval between t_e and the point where the tangent to the second segment of the glottal volume velocity waveform intersects the time axis, as illustrated in Figure 2.7b). If T_a is small, then the abruptness of closure is fast. The exponent ε is directly related to T_a and it can be uniquely estimated from the *timing parameters*. In fact, the *timing parameters* are a sufficient representation of the LF-model, and all the other parameters, such as the dummy parameter α, can be derived by solving the equations (2.12) and (2.13).

The integral of LF pulse, Figure 2.7a), is a bell-shaped representation of the glottal airflow and it is characterized by having only positive or null values. The glottal airflow increases during the opening phase of the glottal cycle. It reaches its maximum value of *AV* at the instant T_p. Thereafter, the glottal airflow decreases until it finally reaches zero value at the instant of complete glottal closure.

The ratios, of the timing parameters of the LF-model, such as *open quotient* and *speed quotient* are commonly used to describe the glottal waveform shape and to quantify *voice quality*. The *open quotient*, O_q describes the time interval in which the vocal folds are open, with respect to the duration of glottal cycle. On the other hand, the *speed quotient* S_q reflects the asymmetry of the glottal flow derivative waveform. It is a ratio of rise and fall time in the glottal airflow. In natural speech, the glottal opening phase is longer than the glottal closure phase and thus the *speed quotient* values are found to be above 0.5.

$$O_q = \frac{T_c}{T_0}, \tag{2.14}$$

$$S_q = \frac{T_p}{T_c - T_p} \tag{2.15}$$

Another common representation of the glottal shape is a set of normalized time parameters, defined in (2.16).

$$\begin{aligned} R_a &= \frac{T_a}{T_0} \\ R_g &= \frac{T_0}{2T_p} \\ R_k &= \frac{(T_e - T_p)}{T_p} \\ R_0 &= \frac{T_e}{T_0} \end{aligned} \tag{2.16}$$

The parameter, R_a controls the spectral tilt of the glottal flow derivative by adding an extra 6dB/oct attenuation above the cutoff frequency, f_c.. Doval and d'Alessandro have analytically derived the spectrum of the LF-model and they have shown that the cut off frequency f_c is largely determined by the parameter R_a and much less affected by the values of R_g and R_k [38]. They have also suggested that for most glottal pulse shapes, f_c can be adequately approximated by: $F_a = 1/2\pi T_a$. The parameter R_g controls the frequency scaling whereby an increase in R_g value results in the energy shift from the low frequency harmonics towards the medium and high frequency harmonics. The remaining parameter, R_k is mostly responsible for determining the first harmonics amplitude. The R-parameters can be related to *open quotient* as $O_q = (1+R_k)/(2R_g)$.

Transformed LF-model

The transformed LF-model is developed by Fant in an attempt to describe glottal derivative waveform and quantify the voice source quality by a single parameter, R_d [46]. He argued that the R_a, R_g and R_k parameters of the LF-model are inter-independent, and highly correlated. He demonstrated that that optimal value for the glottal pulse shaper parameter, R_d, is obtained by the following equation:

$$R_d = \frac{1}{0.11}(0.5 + 1.2R_k)(\frac{R_k}{4R_g} + R_a) \qquad (2.17)$$

Furthermore, he demonstrated that the R_a, R_g and R_k parameters may be predicted from the R_d value using the following equations in conjunction with (2.17)

$$\begin{aligned} \widetilde{R}_a &= 0.01(-1 + 4.8R_d) \\ \widetilde{R}_k &= 0.01(22.4 + 11.8R_d) \end{aligned} \qquad (2.18)$$

This relationship between the R-parameters R_a, R_g and R_k ensures that the accuracy of R_d is in the range of 0.5 dB for the values of R_d less than 1.4. The least reliable R_d value is 2.7, where the estimation error is round 1.5dB. In Chapter 5, the R-parameters of the LF-glottal flow derivative model, as well as the *speed quotient* and *open quotient* are used in the development of a parametric voice quality profile. Furthermore, in Chapter 6, the glottal shape parameter

R_d is used as a perceptual distance measure in the source-to-target voice quality conversion experiment.

2.3 Vocal tract

2.3.1 Anatomy and physiology of vocal tract

The vocal tract is comprised of the larynx tube, pharyngeal cavity, oral cavity and the nasal cavity that is coupled to the oral tract via velum. The geometry of the oral tract is governed by the position of the primary articulators, i.e. the velum, jaw, tongue and lips. The oral tract of a typical male speaker has an average length of 17 cm and a spatially varying cross-section of 20 cm^2. These dimensions are slightly smaller for female speakers.

The primary purpose of vocal tract is to spectrally shape the voice source and provide information to the linguistic layer of the speech communication. The vocal tract is essentially an acoustic resonator that enhances some frequencies and attenuates the others. The spectral shaping characteristics of vocal tract are controlled by supralaryngeal settings, i.e. the configuration of articulators. The primary articulator is the tongue, which divides the vocal tract into two resonant cavities and thereby strongly affects the transmission characteristics of the vocal tract. Velum controls the extent of coupling into the nasal tract. When the opening area is less than 20 mm^2 nasality is generally not perceived. Wider openings generally result in nasal resonances, and when the opening area reaches close to 50mm^2 the speech is perceived as nasal. In speech science, the resonant frequencies of the vocal tract are called formant frequencies. The spectral contribution of a resonance is described through formant bandwidth and formant amplitude. It has been shown that the first three formants are sufficient for the perceptual characterization of English vowels and consonants. Higher formants have a less significant linguistic role, but they contribute to the "naturalness" of speech. An example of the temporal evolution of the vocal tract resonances is shown in Appendix A, Figure: A.1. The figure shows the formant trajectories of a male speaker over the utterance "We were away a year ago". The formant trajectories were obtained using the closed-phase, pitch synchronous, 14th order covariance based linear prediction analysis and a Viterbi search algorithm.

The secondary purpose of vocal tract is to generate new sources for sound production, namely *impulse* and *noise* sources. With a complete constriction of vocal tract and the continual *expiration* (breathing out), the air pressure is build-up behind the closure. If the constriction is removed, the abrupt release of pressure produces an *impulse* sound. Phoneticians refer to these sounds as stops (focusing on the closure) or plosives (focusing on the release). On the other hand, when air is forced to flow passed the vocal tract constriction that is just short of being complete, the turbulent sound is produced. For instance, when the lower teeth are pressed against the upper lip and the air is forced through, a sound associated with IPA[§] symbol [f] is formed. This type of turbulent flow is commonly referred to as friction, and the sounds associated with *friction* are called *fricatives*. Generally, constrictions can be made anywhere along the vocal tract, from the larynx to the lips. It is difficult to make a complete constriction in the pharyngeal region, but a narrow fricative constriction in pharynx is possible. Voice source can also arise from the interaction of vortices with the vocal tract boundaries, i.e. the occlusions in the oral tract, teeth and the false vocal folds [9]. This type of sound source is much less understood than either impulse or noise source. However, there is some evidence that the sound sources due to vortices have a notable effect on the temporal, spectral and perceptual characteristics of speech [9]. The supralaryngeal settings also have some influence on phonation and voice quality; mostly by affecting the extent of source-vocal tract coupling. Nevertheless, in voice source research, particularly from the signal processing perspective, the role of vocal tract is commonly neglected at the expense of laryngeal level of sound production.

[§] International Phonetic Alphabet - official kept by International Phonetic Association; founded in Paris, 1886

2.3.2 Vocal tract modeling

The object of the vocal tract modeling is to represent the anatomical, phonetic, and acoustic features of produced sound. Over the years, a number of vocal tract models have been developed (Flanagen [49]; Kröger [91]; Meyer [102]; Maeda [100]; Nowakowska & Zarnecki [112]) and most of them are based on numerical methods and computer simulations. In this section, we will present the lossless tube vocal model and its digital equivalent. The model is based on the assumption that the sound propagates as a plane wave, and the losses due to the heat conduction and the viscous friction at the vocal tract walls are ignored.

Lossless Tube Concatenation Model

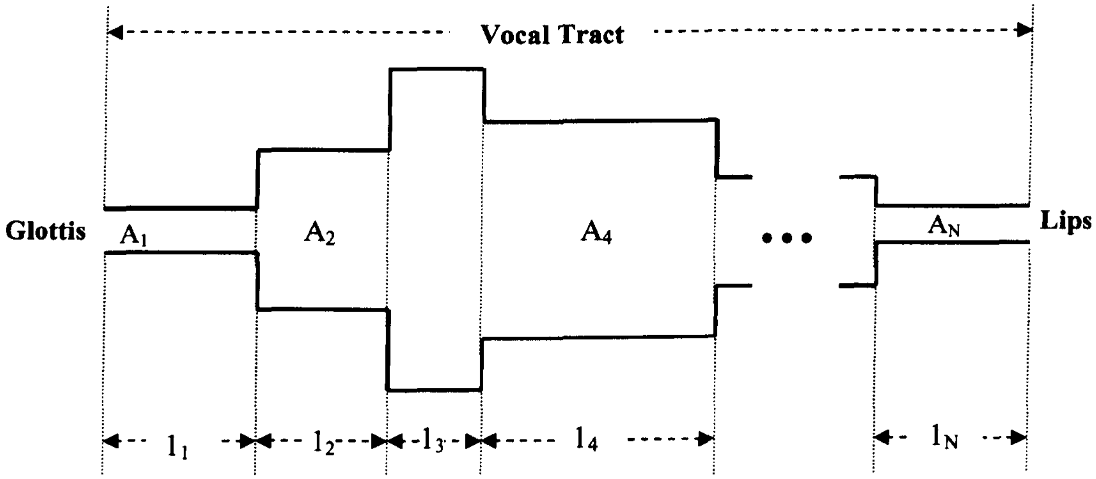


Figure 2.14: The lossless tube model of vocal tract

The diagram of lossless tube concatenated model of vocal tract is shown in Figure 2.14. The sound waves propagating through a lossless tube satisfy the following set of equations:

$$\begin{aligned} -\frac{\partial p}{\partial x} &= \rho\frac{\partial(u/A)}{\partial t} \\ -\frac{\partial u}{\partial x} &= \frac{1}{\rho c^2}\frac{\partial(pA)}{\partial t}+\frac{\partial A}{\partial t} \end{aligned} \qquad (2.19)$$

where $p = p(x,t)$ - sound pressure in the tube; $u = u(x,t)$ - volume velocity; $A = A(x,t)$ -

tube cross-sectional area; ρ - density of the air in the tube; c – velocity of sound.

Assuming the tube is uniform such that $A(x,t) = A$, where A is a constant, the solution of sound propagation is a traveling wave

$$u(x,t) = u^{+}(t-\frac{x}{c}) - u^{-}(t+\frac{x}{c}) \qquad (2.20a)$$

$$p(x,t) = \frac{\rho c}{A}[u^{+}(t-\frac{x}{c}) + u^{-}(t+\frac{x}{c})] \qquad (2.20b)$$

The total volume flow in the tube at any given instant is a superposition of two waves, one going in a forward and the other in a reverse direction, as in Figure 2.15. The total acoustic pressure in a tube is related to the sum of forward and backward volume flow and the acoustic impedance of the tube ($\frac{\rho c}{A}$). The equations (2.20a) and (2.20b) are the same as for the transmission line with $u \approx$ current, $p \approx$ voltage and $\rho c / A \approx$ impedance. At boundaries, the volume flow continuity and the pressure flow continuity must be satisfied. Therefore, at k^{th} boundary:

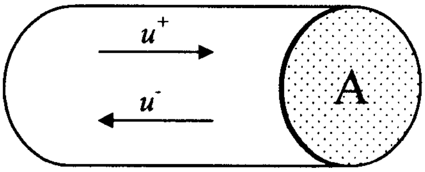


Figure 2.15: Forward and backward wave propagation

$$\begin{aligned} u_k(l_k,t) &= u_{k+1}(0,t) \\ \rho_k(l_k,t) &= \rho_{k+1}(0,t) \end{aligned} \qquad (2.21)$$

Combining the equations (2.20) and (2.21) we obtain the expression for the forward and backward wave propagation:

$$\begin{aligned} u_{k+1}^{+}(t) &= (1+r_k)\,u_k^{+}(t-\tau_k) + r_k\,u_{k+1}^{-}(t) \\ u_k^{-}(t+\tau_k) &= (-r_k)\,u_k^{+}(t-\tau_k) + (1-r_k)\,u_{k+1}^{+}(t) \end{aligned} \qquad (2.22)$$

where $\tau_k = \frac{l_k}{c}$. The reflection coefficient, r_k, bounded by $-1 \le r_k \le 1$ is defined in terms of the cross-sectional areas of the tubes surrounding the junction.

$$r_k = \frac{u^+_{k+1}(t)}{u^-_{k+1}(t)} = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{2.23}$$

Having modeled the propagation of acoustic wave through a set of concatenated lossless tubes we can derive the flow diagram of the vocal tract, see Figure 2.16. In order to obtain the vocal tract transfer function, (2.24), it is assumed that all the concatenated tubes have the length of half the sampling period $l_k = 0.5\,c\,T_{samp}$.

$$V(z) = \frac{0.5(1+r_G)\prod_{k=1}^{N}(1+r_k)\,z^{-N/2}}{D(z)} \tag{2.24}$$

where,

$$D(z) = [1 - r_G]\begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_N \\ -r_N z^{-N} & z^{-1} \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Furthermore, if we approximate glottis area to zero, and assume that the lips do not reflect sound back into the vocal tract, we can write transfer function of the vocal tract in terms of G - gain, $z^{-\frac{N}{2}}$ - acoustic time delay along the vocal tract, and the N^{th} order all pole filter as

$$V(z) = G\frac{z^{-\frac{N}{2}}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2.25}$$

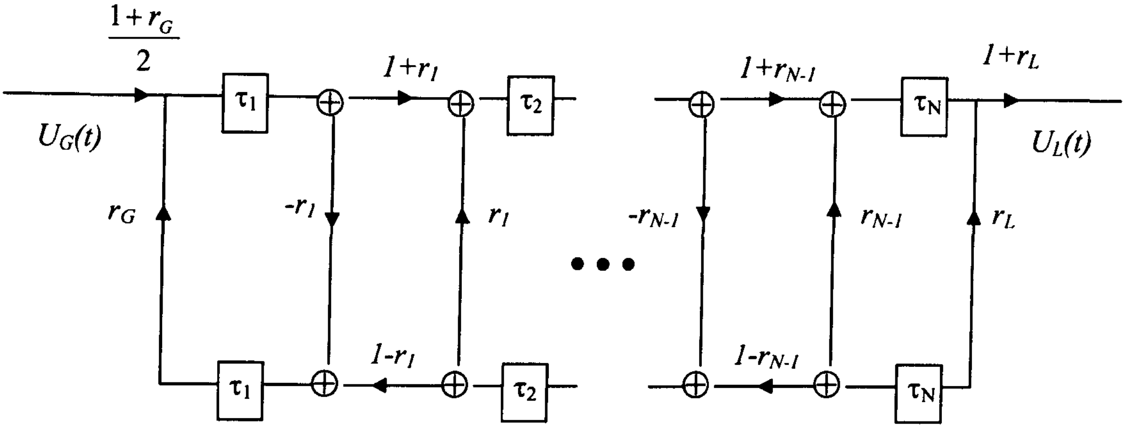


Figure 2.16: Flow diagram of the vocal tract

These assumptions establish the relationship between the number of tube segments in the vocal tract model and the sampling frequency as $N = f_S \frac{2L}{c}$. Since the average vocal tract length is typically around 15-17 cm in adults, $N = f_S/1000$. However, in LPC modeling a rule of thumb is to select a filter of $(2 + f_S/1000)$ th order. The poles of the all pole filter occur in form of complex conjugate pairs and correspond to the resonances or *formants* in the vocal tract frequency response. Experiments have shown that the vocal tract transfer function (of a male adult) has about 1 formant per 1 kHz.

From the transmission line analogues, the lips are treated as a radiation impedance load on the vocal tract. Morse models the lips as the radiation from a spherical baffle [108]. Stevens' model involves a resistive load and three other frequency dependant components [125]. Fant represents the lip radiation with a resistive part, to account for acoustic energy loss, and a reactance part that describes the mass inertia of air at the lips [42]. The most commonly adopted lip radiation model approximates the transfer function between the airflow at the lips and the pressure at the microphone to a 1st-order high-pass response with a corner frequency of $f_c = c/\sqrt{4A} \approx 5kHz$, where A corresponds to lip-opening area. For the sampling frequencies below 20 kHz the lip radiation transfer function can be further simplified:

$$R(z) = 1 - \alpha z^{-1} \tag{2.26}$$

Although the expression for $R(z)$ is obtained as a single zero on the unit circle, it is more realistic to represent the lip radiation with a zero moved slightly inside the unit circle such that $\alpha < 1$. The near field measurements at the lips do not show the 6dB/octave rolloff as predicted by the zero on the unit circle [49]. An extensive overview and comparison of various lip radiation models is provided by Lin [95].

2.4 Aspiration noise

2.4.1 Overview of aspiration noise production theory

Turbulence occurs at the exit of the constriction if the flow velocity is sufficiently high and the area of constriction is small enough. The Reynolds number indicates the likelihood of the turbulence, as in (2.27). The turbulence occurs if the Reynolds number exceeds the threshold value determined by the geometry of constriction.

$$R_e = \frac{2U}{v\sqrt{A_c \pi}} \tag{2.27}$$

The parameters correspond to: U - volume velocity, A_c - effective area of constriction and v - kinematics viscosity of the fluid.

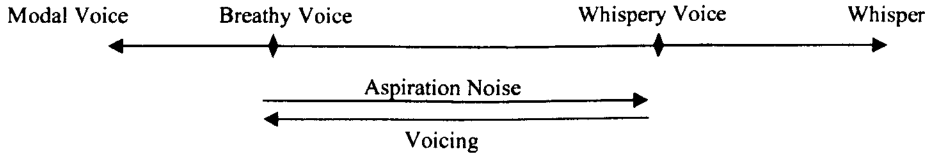


Figure 2.17: The relationship between the voice quality and the aspiration noise levels along the auditory continuum.

Depending on the generation mechanism, turbulence noise source can be differentiated as *monopole*, *dipole* or *quadruple*. A *monopole* source arises from a net unsteady mass injection into the fluid region. A glottal volume flow can be considered as a *monopole* source acting on the vocal tract. The monopole source is often represented through volume velocity. A *dipole* source is a net fluctuating force in the medium without the net mass injection. It can arise from a rapid flow of air hitting an obstacle or a surface. The spectrum of the dipole source has a peak at frequency proportional the flow velocity and inversely proportional to the characteristic dimension of the constriction. The overall shape of the spectrum depends on the constriction geometry. The *dipole* source is often modeled as a sound pressure source. A *quadruple* turbulence noise source occurs when two opposite dipoles are positioned at close proximity to each other. It exists in an unbounded medium without the net mass injection and without the net force.

The turbulence noise that is produced in vicinity of glottis is referred to as *aspiration noise.* The main cause of aspiration noise is a dipole source due to the glottal airflow impinging on laryngeal surface in proximity of false vocal folds and the epiglottis. The extent of aspiration noise during *voiced* phonation is an important correlate of voice quality. The aspiration noise, essentially, adds a degree of "breathiness" to voice texture [34]. In some minor languages contrasting between the *breathy* and *modal* phonation carries linguistic information [59] as well. The role of the voice quality in conveying the paralinguistic information, such as emotions, mood, and attitudes is explored and reported in [85], [58]. *Whispery* voice is found to be an important attribute of acoustic perception of "fear", while "sadness" is found to be commonly associated with the *breathy* voice. *Breathiness* is generally treated as a continuum, without a clear threshold for separating *modal* or *breathy voice* on perceptual, acoustic or physiological basis. The same applies for the transitions between the *breathy* and *whispery* voice. Figure 2.17 illustrates the relationship between the perceptions of voice quality and the relative aspiration noise levels along the auditory continuum.

2.4.2 Aspiration noise modeling

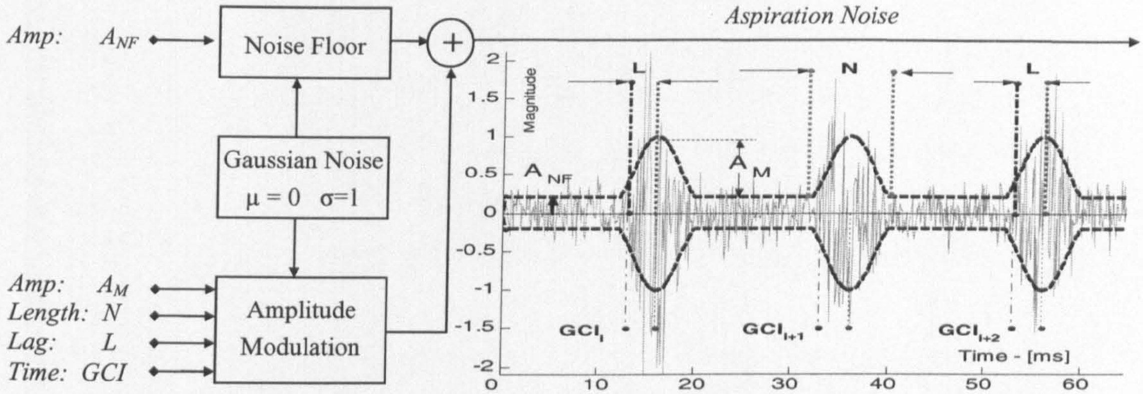


Figure 2.18: The schematic diagram of the aspiration noise model and an example of the synthesized aspiration noise signal. The following parameter values are used in the synthesis of this aspiration noise example: T_0 =20 ms, A_{NF} = 0.21, A_M =0.79, N=50%. L=15%.

The turbulence noise is initiated at the center or just downstream from the constriction region or it can be spatially distributed along the constriction region [126]. Cook argued that the turbulent flow is likely during the entire open phase of the glottal cycle, but the maximum power of the turbulent flow is generated at the instant of glottal closure while slightly smaller bursts of turbulence may occur at the instances of vocal fold abduction onset [26]. The notion that the aspiration noise might be non-stationary was further confirmed by the results of psychoacoustic experiments conducted by Hermes [66]. Using the source-filter synthesis model, he demonstrated that the stationary noise does not adequately represent the aspiration turbulence as the noise is perceived as being acoustically separate from the rest of speech. Subsequently, Hermes developed an aspiration noise model that was able to achieve the desired perceptual effect. This model describes the aspiration noise as de-emphasized, high-pass filtered, amplitude modulated, pitch synchronized white noise. De-emphasis is achieved with the low pass filter $H(z) = 1/(1-0.9z^{-1})$. The high-pass filter with the cut-off frequency in the range of 1.2-2 kHz controls the level of breathiness. The cut-off frequency of the high-pass filter is thought to be inversely proportional to level of perceived breathiness.

Figure 2.18 shows a schematic diagram of the aspiration noise model that is used in this thesis. It is conceptually similar to the model used by Childers [23]. The spectral shaping component includes the de-emphasis filter and the high pass filter. Prior to spectral shaping, the turbulence noise consists of two components; the noise floor that is present throughout the glottal cycle and the pitch synchronous amplitude modulated component. The first block of the aspiration noise model generates the white Gaussian noise of unit variance and zero mean. The noise floor is obtained by scaling the white Gaussian noise by noise floor amplitude parameter, A_{NF}. On the other hand, the pitch-synchronous amplitude modulation is achieved by scaling the white Gaussian noise with the Hamming window. The modulation is controlled by the following parameters, A_M, *N, L, GCI.* The parameters, A_M and *N* describe the amplitude and the duration of the Hamming window, respectively. The parameter, *L* indicates the lag between the midpoint of the Hamming window and the glottal closure instant, *GCI.* The length of the Hamming window, *N* and the lag, *L* are both measured in terms of percentages relative to the glottal cycle duration.

Effectively, the noise floor scaling and the amplitude modulation combine to form the aspiration noise envelope. A particular example of the aspiration noise envelope is shown in Figure 2.19. Childers has shown that the aspiration noise can be adequately modeled as amplitude modulated pitch synchronous Gaussian noise without the spectral shaping components [23]. Our informal listening tests confirm that perceptual effects of the spectral shaping component are not significant, and as such, the spectral shaping component will also be omitted in the final implementation of the aspiration noise model. Figure 2.20 shows an example of the glottal excitation signal for SNR = 20 dB. The synthesis is carried out by directly adding the aspiration noise to the Liljencrants-Fant's representation of the glottal flow derivative signal. This type of synthesis will be used throughout the thesis.

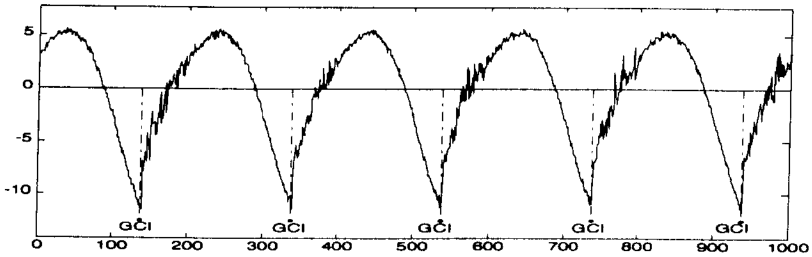


Figure 2.19: Synthesized glottal excitation signal example. The glottal volume velocity derivative is synthesized using the following LF parameters T_p = 58%; T_e = 84%; T_c = 100%; T_a = 12%; The following parameter values are used synthesize of turbulence noise: T_0 =200, A_{NF} = 0.21, A_M =0.79, N=50%, L=15%. SNR=20dB

Chapter 3

Voice Source Estimation

ABSTRACT

This chapter presents a comparative study of the temporal structure of the glottal flow derivative estimates in relation to an idealized view of voice source realizations as defined by Liljencrants-Fant (LF) model. Specifically, we endeavor to ascertain the extent by which LF model can be used to represent the voice source estimates obtained via closed-phase pitch synchronous inverse filtering of recorded speech. The study includes several phonation types and two examples of voice pathology. The study has established the following. Due to the limited degrees of freedom, Liljencrants-Fant's model is only capable of adequately representing the "coarse" glottal pulse structure. We have presented evidence that the unrepresented elements or the "fine" glottal flow derivative structure contains information related to voice source individuality. In addition, we have shown that LF-parameters do not always accurately portray significant events in the vocal fold dynamics which might have a effect on the accuracy and robustness of LF based voice source parameterization techniques. In this chapter, we have also considered a group delay approach to GCI estimation. Specifically, *average group delay* and *energy weighted group delay* measures are discussed in detail. We have proposed a GCI estimation method that is based on a group delay algorithm and the translation-invariant hard-thresholding of the LPC residue. The performances of the two group delay measures and the proposed method are evaluated for a range of fixed and pitch-synchronous group delay window lengths. We have found that the pitch synchronous *energy weighted group delay* measure with the wavelet–denoised LPC residue provides by far the best GCI estimation performance.

3.1 Introduction

Analysis of the vocal fold vibrations is a challenging problem as the larynx is not easily accessible. Nevertheless a number of techniques have been developed to enable the study of laryngeal dynamics. Electroglottography is widely used for clinical and research purposes [139]. The electroglottographic signal is obtained by measuring the impedance changes across the speaker's neck. The change in impedance values is related to the changes in the contact area between the vocal folds. As such, the electroglottographic signal is comparable to the volume velocity in the air stream through the glottis. On the other hand, the electromagnetic glottography uses high frequency electromagnetic waves to measure tissue motion [139]. Optical methods, such as transillumination and high-speed imaging are also useful in obtaining information about the larynx and the nature of vocal fold oscillations [84]. Transillumination uses a fiberscope inserted through the nose to illuminate the glottis. Subsequently, an estimate of the glottis area can be obtained by measuring the light intensity passing through the glottis via a photo sensor that is externally attached to the larynx. High-speed imaging is performed through a speaker's mouth using a mirror located near the uvula or through the fiberscope inserted through the nose. In 1959, Miller introduced the first inverse filtering technique [106]. He applied analogue electronic filters to cancel the effects of two lowest formants and the lip radiation from recorded speech. In 1973, Rothenberg developed an alternative form of inverse filtering, whereby an estimate of volume-velocity of air at the mouth is obtained using a pneumotachographic mask that measures the pressure differential across the mask's screen [117]. The corresponding signal is then inverse-filtered to yield estimates of the glottal volume velocity waveforms. This method is very accurate and robust to the low frequency noise. However, its useful frequency range is constrained to below 1.6 kHz.

The inverse filtering of recorded speech is proving to be the most popular approach to estimate the glottal excitation signals. It is non-invasive, does not cause any discomfort to the speaker and does not require bulky or expensive equipment. It is based on the theoretical framework of the source-filter theory of speech production, and as such, it allows independent studying of voice source and vocal tract. In principle, provided that the vocal tract transfer function is known, the glottal excitation signal can be obtained by feeding the speech signal through the inverse of the vocal tract filter.

Sophisticated inverse filtering techniques relay on accurate estimates of glottal closure instants (GCIs). In Section 3.2, we have presented a group delay approach to GCI estimation. Specifically, *average group delay* and *energy weighted group delay* measures are discussed in detail. Their properties are studied on the synthetic and natural speech datasets. We have proposed a GCI estimation method that is based on a group delay algorithm and a translation-invariant hard-thresholding of the LPC residue. The performances of the two group delay measures and the proposed method are evaluated for a range of fixed and pitch-synchronous group delay window lengths. The optimal GCI estimation strategy is evaluated and reported. In Section 3.3, an overview of the closed-phase pitch-synchronous inverse filtering method for obtaining the estimates of glottal flow derivative waveforms is presented. Subsequently, we describe a formant modulation analysis technique. The formant modulation analysis is used to determine the extent of nonlinear source-filter coupling. Both techniques are employed on a range of voice qualities, including two examples of pathological voices, to enable a comparative study of the temporal structure of the glottal flow derivative estimates in relation to an idealized view of voice source realizations as defined by Liljencrants-Fant model. Section 3.4 concludes the chapter.

3.2 Glottal Closure Instant Detection

Automatic, robust and accurate identification of glottal closure instants can be beneficial to a range of speech processing applications. In PSOLA-based concatenative synthesis and in some voice conversion techniques, the glottal closure instants are indispensable in preserving coherence across segment boundaries [62], [132]. Blind deconvolution of the voice source and vocal tract via pitch synchronous closed-phase analysis relies on the glottal closure instants to segment the glottal cycle into closed and open phases [27], [148]. The use of

GCIs can also be found in speech coding, voice quality analysis, pitch tracking, voice conversion, and speaker verification systems. Over the years, numerous methods and approaches to GCI identification have been proposed. For research purposes, glottal closure instants can be obtained from the electroglottographic (EGG) impedance signals, but in most other applications, signals are not available [90].

Strube uses the peaks of the log-determinant of a sliding autocovariance window to relate the instances of glottal excitation to the discontinuities in the linear model of speech production [131]. McKenna developed a similar method based on Kalman filtering [103]. Navaro-Messa *et al.* estimate the GCIs by examining the features in the time-frequency representation of speech [111]. Cheng and O'Shaughnessy use a maximum likelihood method based on the Hilbert transform to obtain GCI approximations [19]. Ma *et al.* proposed a method whereby the glottal closure instants are identified as the maxima of the Frobenius norm of the signal matrix [99]. Smits and Yegnanarayana were the first researcher to propose the use of a group delay measure to determine the instants of acoustic excitations [124]. Each of these methods has some shortcomings which may result in the GCI identification errors. The main disadvantage of the group delay based GCI estimation methods is their intrinsic sensitivity to the presence of noise.

3.2.1 Group delay measures

The group delay based GCI estimation exploits the properties of the minimum phase signals and the group delay function, i.e. the *average group delay* of a minimum phase signal is zero. On the assumption that the vocal tract is a minimum phase system (all pole vocal tract model), the speech pressure waveform following the glottal closure instant is a minimum phase signal within one glottal pulse cycle. As such, a window of speech or LPC residue will change a sign in the phase slope as it crosses the moment of excitation (zero-phase). For a given input signal $u(r)$, we consider an N-sample windowed segment beginning at a sample r

$$x_r(n) = w(n)\,u(n+r) \qquad n = 0,\ldots, N\text{-}1 \tag{3.1}$$

and its corresponding group delay function

$$\tau_r(k) = \frac{-d\arg(X_r(k))}{d\omega} \tag{3.2}$$

where $X_r(k)$ is a Fourier transform of $x_r(n)$ at frequency k; ω = 2kπ/N; $w(n)$ denotes the analysis window function used in the short term group delay analysis. Since group delay function is highly sensitive to noise, averaging over frequency is required to increase the level of robustness. We will consider two types of group delay measures, *average group delay* d_{AV} , and *energy weighted group delay* d_{EW} :

$$d_{AV}(r) = \frac{1}{N}\sum_{k=0}^{N-1}\tau_r(k) = \frac{1}{N}\sum_{k=0}^{N-1}\frac{\tilde{X}_r(k)}{X_r(k)} \tag{3.3}$$

$$d_{EW} = \frac{\sum_{k=0}^{N-1}\left|X_{r(k)}\right|^2\tau_r(k)}{\sum_{k=0}^{N-1}\left|X_{r(k)}\right|^2} = \frac{\sum_{n=0}^{N-1}nx_r^2(n)}{\sum_{n=0}^{N-1}x_r^2(n)} \tag{3.4}$$

Note that in the *average group delay* measure d_{AV} , the conjugate symmetry of $\tilde{X}$ and X ensures that the corresponding summation is real. Prior to applying a group delay measure, the speech signal is first passed through a 1^{st} order pre-emphasis filter with a 50 Hz corner frequency. The pre-emphasized speech is subsequently inverse filtered using the 22^{nd} order autocorrelation-based linear prediction coefficients and a sliding 20 ms analysis window (Hamming) with 50% overlap. The high frequency noise is removed from the LPC residual signal with a 4 kHz, 2^{nd} order Butterworth low-pass filter to obtain a signal *u(r)*. The group delay measures are applied on *u(r)* using a sliding Hamming window analysis. The energy weighting, 3rd order median filter and a 1.5 kHz low pass filter are applied to the d_{AV} measure to remove the occasional extreme values [110].

In our implementation of the group delay measures, the time origin is shifted to the central point of the group delay window, $w(n)$ in (3.1), so that $d_i'(r) = d_i(r - N/2 - 0.5) - N/2 - 0.5$, where $i \in \{AV, EW\}$.

3.2.2 Performance evaluation on synthetic data

The Effect of Window Length

The effect of the group delay window length on the performance of GCI estimation is investigated using an idealized version of the LPC residual. The synthesized signal consists of an impulse train with additive white Gaussian noise at SNR = 20 dB. The successive impulses are delayed by periods of 60, 50, 40, 30, 20, 10, 10 and 6 samples. The last impulse has twice the intensity of the other impulses. The test signal is shown in Figure 3.1 a).

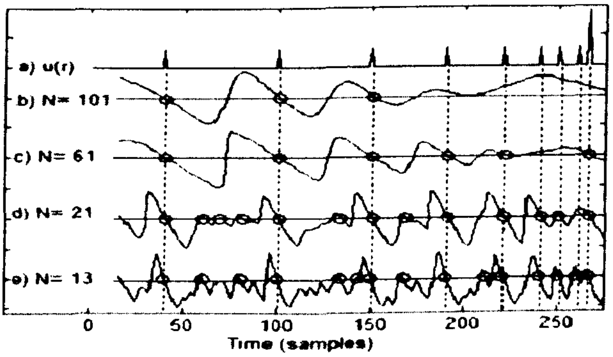


Figure 3.1 a) A train of impulses with periods of 60, 50, 40, 30, 20, 10, 10 and 6 samples with additive noise at SNR=20 dB. In panels b)-e) energy weighted group delay functions are displayed for a range of group delay window lengths.

The energy waited group delay function is obtained for four group delay window lengths, N=101, N=61, N =13, N=21. The results are shown in Figures 3.1 b)-1e), respectively. On the first glance we can establish that the shape of the group delay functions varies strongly with the window length. The longer windows produce much smoother group delay functions with very few, but clearly defined negative zeros crossings (NZC). Note that NZC instants identify the instants of acoustic excitation. For N=101, only the first three impulses are identified. The corresponding NZC onstants are clearly defined with the local gradient that is close to the ideal* value of -1. The slight deviation from the ideal value is due to the presence of noise. For the slightly smaller window, N=61, the number of identified impulses is higher; first five as well as the last impulse are correctly identified. With further reduction in the window length, more impulses are detected, but only the closely spaced impulses are distinctly identified. For N=13 and N=21, the group delay function is much more jittery and it contains a number of spurious zero crossings. In such cases, the instances of zero crossing are less defined with the local gradient deviating further from the ideal, -1, value.

It is evident that the ability of the group delay measure to identify the instances of acoustic excitations depends on the length of the group delay window and the distances between the successive excitations. With this simple experiment, we can establish that the optimal window length, as a fraction of pitch period length, should be in the range $0.5 \le N \le 2$. If the window length is below this range, there will be moments when the group delay window will contain only noise. This will result in an increased number of spurious negative zero crossings and false detection. On the other hand, if the window length is too large, the ability to differentiate the individual excitations is impaired as more than a single excitation will be present in the group delay window at all times. However, it is interesting to observe that the last impulse in our idealized LPC residue is successfully identified even with the relatively large N=61 window. In the following section, we will investigate, in the context of a multiple-impulses problem, the effect of the excitation energy on the ability of the group delay measures to accurately detect the acoustic excitation instants.

* Group delay measures exhibit a form of shift invariance: If $w(n) \equiv 1$ and $u(r) = u(N+r) = 0$, then $d_i(r+1) = d_i(r) - 1$, where $i \in \{AV, EW\}$. As such a gradient of -1 is expected for the ideal impulses.

Group delay response to multiple impulses

The case of multiple impulses occurs when the group delay window is longer than the pitch period, or as it is often the case with the LPC residual, when the acoustic signal includes additional features, such as the false vocal cord excitations, vocal tract artifacts.

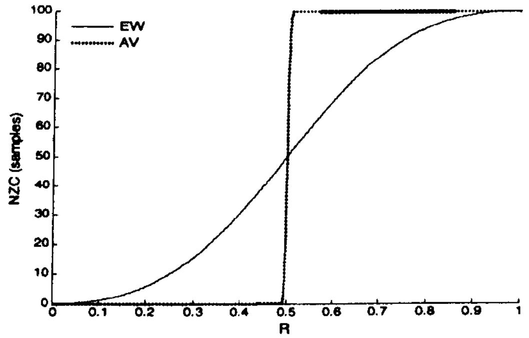


Figure 3.2: NZC instance as a function of parameter R

In order to evaluate the response of group delay measures to multiple impulses, let us consider a general two impulse case:

$$x(n) = (1-R)\,\delta(n) + R\delta(n-n_0) \tag{3.5}$$

,where n_0 corresponds to the time lag between a pair of impulses and R controls their relative intensity levels. R values are constrained to $0 \le R \le 1$. With the test signal clearly defined, analytical solution for the *energy weighted* and the *average group delay* measures is reached:

$$\begin{aligned} NZC_{AV} &= \frac{n_0}{1-b^{N/\gcd(n_0,N)}} \\ NZC_{EW} &= \frac{n_0}{1+b^2} \end{aligned} \tag{3.6}$$

where $b = 1 - R^{-1}$. *gcd* and NZC denote the *greatest common divisor* and negative zero crossing instant, respectively. Figure 3.2 shows the negative zero crossing function evaluated for a specific case, where N=151 and n_0 =100.

The results show that both group delay measures exhibit a bias towards the more intense impulse. For the *energy weighted group delay* measure, the negative crossing function exhibits a smooth transition from one impulse to another. On the other hand, *NZC* function of the *average group delay* is abrupt and can be described as a switch function with a sharp change occurring at, R=0.5. As such, the *average group delay* measure identifies the highest peak in the window. The *energy weighted group delay* is much more sensitive to the presence of extra impulses, whereby the identification accuracy is clearly compromised when the intensity of additional impulses is comparable to the intensity of the analyzed impulse.

3.2.3 GCI estimation with the wavelet de-noised LPC residue

The group delay based GCI estimation performance is evidently dependant on the quality of the acoustic signals. In the current implementation of the group delay measures, the vocal tract artifacts, aspiration noise, and other disturbances are removed from the LPC residue with the *2nd* order Butterworth low-pass filter. In Chapter 4, we have developed an optimal wavelet thresholding strategy for the glottal volume velocity derivative signals. The denoising method is based on the translation invariant hard thresholding, 6-coefficient Coiflet filter and a decomposition level-7. Butterworth filtering is effective in removing the high frequency noise, but unfortunately it induces distortions in the underlying LPC residue signal. Although, it is desirable to remove the aperiodic features and noise from the LPC residue, we acknowledge that the group-delay-based GCI estimation relays on the prominence of the rapidly varying regions in the underlying signal corresponding to the glottal pulse peaks. Wavelet thresholding is able to preserve the slow, as well as the rapid variations in the underlying signal by exploiting the compactness property of wavelets i.e localizations in time and frequency. As such, wavelet thresholding is a much more sophisticated solution to this particular denoising problem. In the next section, we will evaluate if the proposed method can enhance the performance of the group delay measures as a replacement for the Butterworth filter. Furthermore, we will attempt to optimize the GCI estimation performance with respect to the group delay window length.

3.2.4 Performance evaluation on natural speech

In this section, group delay measures are evaluated on a database containing a read speech sentence from five male and five female speakers. The corresponding electroglottographic (*EGG*) files are used as a reference source for the instants of glottal closure. Although, EGG-obtained glottal closure instants are themselves not entirely reliable, they constitute the best available GCI reference source. The performance of each GCI estimation method is evaluated in terms of *detection rate*, *identification rate* and *identification accuracy*. *Identification rate* is defined as a fraction of larynx cycles that contain exactly one negative zeros crossing instant in a group delay function. The *detection rate* measures the fraction of larynx cycles that contain any number of *NZCs*. If the correct *NZCs* could be distinguished from the redundant *NZCs*, then the *identification rate* and *detection rate* would equate. As such, *detection rate* indicates the capacity of a group delay measure to locate the acoustic excitations. Standard deviation of the identification errors (the discrepancy between the estimated and the actual acoustic excitations) is used as a measure of *identification accuracy*. *Identification accuracy* is evaluated for the glottal cycle that contain only one negative zero crossing.

Note that in the presentation of experimental results, the group delay measures will be denoted as follows.

- *AV* - average group delay
- *EW* - energy weighted group delay
- *Den. AV* - average group delay with the wavelet-denoised LPC residue
- *Den. EW* - energy weighted group delay with the wavelet-denoised LPC residue

Experiment 3.A: Fixed group delay window

In this experiment, the optimal window length value is evaluated in terms of *window length coefficients* (*WLC*). The *window length coefficient* is defined as a factor by which the average pitch period value is to be multiplied to obtain the length of a group delay window.

Using the autocorrelation-based pitch estimation, the average glottal cycle length is estimated as 5.0 ms and 8.3 ms, for female and male speakers, respectively. The window length coefficient is varied in steps of 0.1 in the range defined as $0.5 \le WLC \le 2$. The boundaries for *WLC* range are based on the qualitative evaluation of the grouped measures in Section 3.2.2. Figure 3.3 shows the GCI estimation performance in terms of *detection rate* as a function of window length coefficient. The *identification rate* and *identification accuracy* values are shown in Figure 3.4 and Figure 3.5, respectively. The results suggest that the capacity of a group delay measure to detect acoustic excitations improves with the decreasing window length coefficient. At *WLC* = 0.5, both *energy weighted group delay* measures (*EW* and *Den. EW*) achieve a peak *detection rate* of 97.9 %. The *average group delay* measures (*AV* and *Den. AV*) achieve reasonably good *detection rates* for higher values of WLC, especially around the point WLC =1.

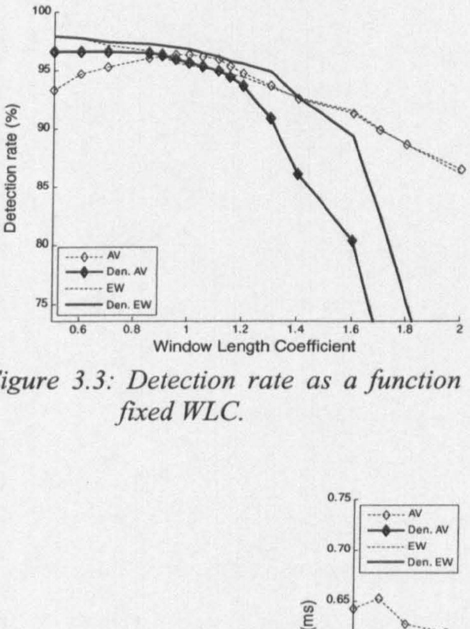


Figure 3.3: Detection rate as a function of fixed WLC.

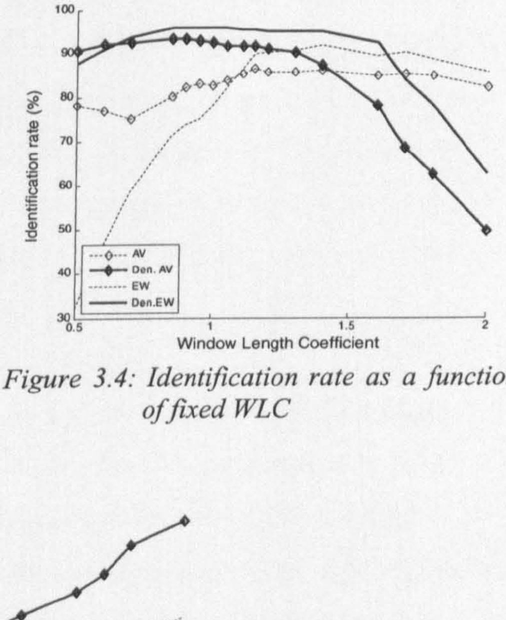


Figure 3.4: Identification rate as a function of fixed WLC

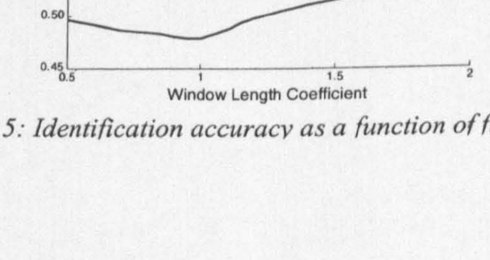


Figure 3.5: Identification accuracy as a function of fixed WLC

On the other hand, the *identification rate* functions have the same general form, irrespective of the GCI estimation method. The optimal performance is achieved for some moderate *WLC* value. As the *WLC* is increased, or lowered away from the optimal WLC value, the *identification rate* gradually deteriorates. The *average group delay (AV)* attains the peak *identification rate* of 86.6%, for *WLC*=1.15. The performance of the *energy weighted group delay* (*EW*) is considerably better. The peak *identification rate* of 92.02% is reached for *WLC*= 1.4.

The proposed change in the denoising method has had a positive effect on the performance of both group delay measures. The *identification rate* and *identification accuracy* are noticeably improved, especially for the lower window length coefficient values. These observations are in accord with out expectations, as they reflect a superior denoising performance of the wavelet thresholding method. Since, the presence of noise in the LPC residual is the principal cause of spurious NZCs in a group delay function, the frequency of false NZC instants increases when the group delay window is reduced below the glottal cycle length. Therefore, the effect of improved denoising is mostly evident for lower *WLC* values. The modified group delay measures, *Den. AV* and *Den. EW*, attain the peak *identification rates* of 93.4 % and 95.8 %, at *WLC*=0.9 and *WLC*=1, respectively. The *identification accuracy* curves closely follow the *identification rate* curves. As such the optimal WLC values are generally the same for both performance measures.

What we find interesting is that the *identification accuracy* of the *energy weighted group delay* has significant improved with the introduction of the wavelet-denoising method. In fact, the *energy weighted group delay* measure benefits considerably more from the change in the denoising method, than the *average group delay*. This observation can be explained by the multiple-impulse analysis presented in Section 3.2.2, where we have shown that in comparison to the *average group delay*, the *energy weighted group delay* measure is more sensitive to the energy levels of additional impulses within the group delay window.

Experiment 3.B: pitch synchronous group delay window

In this experiment, pitch synchronous GCI estimation is evaluated for the same range of window length coefficient values as in the pervious experiment. However, in this case, the group delay window length is defined as a product of the glottal cycle length and the window length coefficient. The duration of consecutive glottal cycles is estimated via autocorrelation-based pitch estimation. The pitch trajectory is filtered with a 5th order median filter in order to remove the effects of estimation inaccuracies and to allow gradual evolution of the group delay window size. The GCI estimation is evaluated in the same manner as in the pervious experiment. The *detection rate, identification rate* and *identification accuracy* results are reported in the Figure 3.6, Figure 3.7, and Figure 3.8, respectively.

The general performance trends with respect to window length coefficient values have remained the same as for the fixed-window group delay analysis. The *energy weighted group delay* measures outperform their *average group delay* based counterparts. Also, the modified group delay measures, *Den. EW* and *Den. AV,* outperform the standard *energy weighted* and *average group delay* measures, *EW* and *AV.* Ultimately, *Den. EW* offers the best glottal closure instant estimation performance. At *WLC*=1, it attains the peak *identification rate* of 98.59 % and the optimal *identification accuracy* of 0.423 msec.

Compared to the *energy weighted group delay* (*EW*) measure with the fixed group delay window, the modified pitch synchronous *energy weighted group delay* measure (*Den. EW*) improves the *identification rate* and *accuracy* by 6.57 % and 0.158 ms, respectively. This represents a considerable improvement in the performance, especially if we consider that 6.57% increase in the *identification rate* corresponds to 82.33 % reduction in the number of unidentified glottal excitations. As such, we will use this method to obtain the estimates of glottal closure instants throughout the thesis.

We will end this section with a few general comments about the proposed GCI estimation method. Even for the optimal length of the pitch synchronous group delay window, the *identification rate* is below the *detection rate* for all GCI estimation methods including those that implement the wavelet thresholding technique. This implies that the spurious negative zeros crossings are still present in the group delay functions. On closer inspection, we have ascertained that the vast majority of these false NZCs correspond to the peaks at glottal opening instants, and only few are due to noise. We would also like to note that in comparison to the glottal flow derivative waveform, LPC residue exhibits significantly different temporal and spectral characteristics. Thus, the proposed wavelet thresholding method might not be the optimal denoising solution for this type of signal. We would expect further, but not considerable, improvements in the GCI estimation performance if the wavelet thresholding is optimized for the LPC residual, specifically.

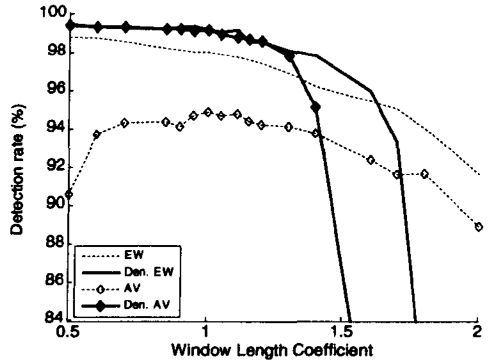


Figure 3.6: Detection rate as a function of pitch synchronous WLC

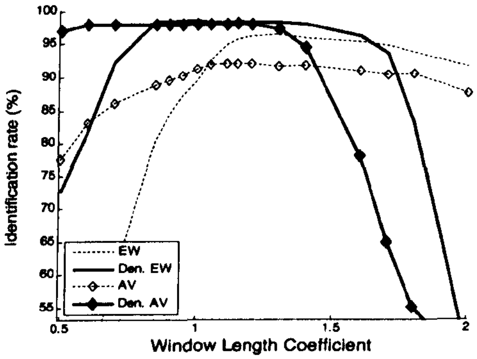


Figure 3.7: Identification rate as a function of pitch synchronous WLC

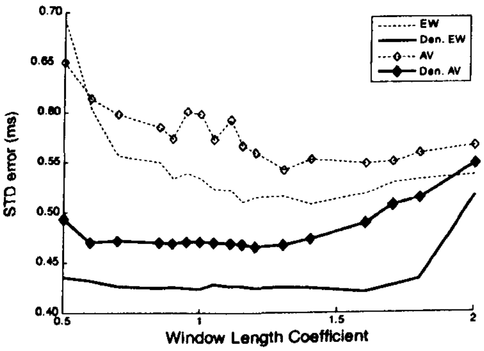


Figure 3.8: Identification accuracy as a function of pitch synchronous WLC

3.3 Source-filter deconvolution and voice source properties

Voice source estimation via inverse filtering implies strong assumptions about the glottal volume waveform and the transfer function of the vocal tract. As such, the inverse filtering results are regarded as the glottal flow estimates and not the actual glottal waveforms. It might sound trivial, but the distinction is very important when discussing the quality of voice source reconstruction and the performance of voice source parameterization. Essentially, most of the inadequacies of the source–filter model of speech production, as well as the inaccuracies in the inverse filtering implementation[†] are manifested in the voice source estimate waveforms. In voice quality profiling, whereby one seeks to obtain a parametric representation for the perceived voice textures, the vocal tract artifacts and the artifacts of nonlinear source-filter coupling are seen as forms of voice source degradations that conceal the actual voice source signal and the true voice source parameters. On the other hand, in speech synthesis, voice source estimate represents the actual signal that needs to be adequately modeled in order to achieve faithful speech reconstruction.

Another important voice source processing issue is that the manner in which the vocal fold vibrations and the corresponding glottal flow waveforms are realized varies from one speaker to another. Some speakers have widely abducted phonations, where the vocal folds may never fully close, while for other speakers, phonation may be adducted with complete and rapid glottal closures. The closure of vocal folds can occur simultaneously along the length of vocal folds, or it can occur in a zipper like manner. The extent of aperiodicity and aspiration noise can also vary between speakers. Švec *et al.* have demonstrated, via videokymography, that the deviations from the idealized vocal fold behavior, readily occur among the healthy people without any voice disorders [133]. For speakers with the distinctly adducted phonation, complex vibratory patterns can emerge, and often "ripples" are seen in the vibrations of the vocal folds. Healthy speakers with *creaky* phonation can have irregular vocal fold vibrations with sub-harmonic patterns, such as double opening. Normal larynges

[†] Inaccuracies in the inverse filtering implementation are generally related to the estimation of glottal closure instants and the closed-phase intervals in the consecutive glottal cycles.

are found to be rarely symmetric, and often there is a degree of phase delay between the vocal folds. This phenomenon can affect the voice texture, and under the extreme situations, such as high pitch and intensity, voice can sound completely hoarse. Zannger *et al.* have investigated the acoustic characteristics of the distorted tones that are commonly heard in rock music [149]. They have found that the "distorted" voice texture is associated with the vibrations of the supraglottal mucosa (including the ventricular folds, aryepiglottic folds and the anterior part of the mucosa that covers the arytenoid structure), and very complex and structurally rich glottal flow derivative realizations.

In this section, an overview of the closed-phase pitch-synchronous inverse filtering method for obtaining the glottal flow derivative estimates is presented. Subsequently, we will describe a formant modulation analysis technique. Both techniques are employed on a range of voice qualities, including two examples of pathological voices, to enable a comparative study of the temporal structure of the glottal flow derivative estimates in relation to an idealized view of voice source realizations as defined by Liljencrants-Fant model.

3.3.1 Closed-phase inverse filtering

According to the source–filter theory of speech production, illustrated in Figure 2.2, the transfer function of the voiced speech can be expressed as:

$$S(z) = A\,G(z)\,V(z)\,R(Z) = \frac{A\,G(z)\,R(z)}{H(z)} \tag{3.7}$$

where $G(z)$ denotes the z-transform of the glottal flow over a pitch period; A is the gain factor; $V(z) = 1/H(z)$ describes the minimum-phase all-pole vocal tract transfer function; $R(z)$ corresponds to the radiation load. The combined effect of glottal flow, radiation load and gain can be expressed as $b(n) = A\,g(n) * r(n)$. Since radiation load can be represented by a differencing filter, $r(n) = \delta(n) - \delta(n-1)$ [49], [95], the sequence $b(n)$ describes a scaled glottal flow derivative over one pitch period. Thus, the *voiced* speech signal, $s(n)$ can be modeled as:

$$s(n) = \sum_{k=1}^{p} h(k)s(n-k) + b(n) * \sum_{m=-\infty}^{\infty} \delta(n-mP) \tag{3.8}$$

where $P(z)$ denotes a periodic impulse train with a period P, where $p[n] = \sum_{k=-\infty}^{\infty} \delta[n-kP]$. In voice source analysis, it is generally assumed that the sequence $b(n)$ is shorter than the length of the glottal cycle, such that there exists a region C in which the difference equation (3.8) is not driven by $b(n)$. This interval corresponds to the closed-phase region of the glottal pulse cycle, minus one sample to account for the affect of the lip radiation term. During the closed-phase interval, the speech signal is related to the vocal tract coefficients as:

$$s(n) = \sum_{k=1}^{p} h(k)s(n-k), \quad n \in C \tag{3.9}$$

As such, an estimate of the glottal flow derivative can be obtained by inverse filtering the speech waveform with the all pole vocal tract model that is derived over the closed-phase interval of the glottal cycle. This is the basic principle behind the *closed-phase pitch synchronous analysis* [147], [27], [148].

The most challenging aspect of the closed-phase pitch synchronous analysis is obtaining accurate estimates of the closed-phase intervals. In regards to closed-phase estimation, a variety of approaches have been proposed, but the following methods have emerged as the most popular. Wong *et al.* [147], and Cummings and Clement [27] use a one-sample-shift sliding covariance analysis of the speech waveform and a function of the linear predictor error to obtain the closed-phase estimates. On the other hand, Plumpe and Quatieri use the sliding covariance analysis and the vocal tract formant modulation analysis to estimate a stationary formant region which is associated with the closed-phase glottal cycle intervals [114]. The closed-phase intervals can also be estimated via the analysis of electroglottographic signals [20]. In this thesis, we will adopt the closed-phase pitch synchronous inverse filtering of recorded speech as the means for obtaining the glottal flow derivative estimates. Furthermore, we will use a one-sample-shift, sliding covariance analysis and a function of the linear predictor error to obtain the closed-phase interval estimates.

3.3.2 Formant modulation analysis

Formant modulation analysis is a term that describes the study of formant frequency movement within a glottal cycle. Since formant modulation (movement) is a result of time-varying non-linear source/vocal tract coupling, it is expected to be more prominent during the glottal *open-phase* than during the *closed-phase,* when the vocal folds are closed [4]. Correspondingly, the closed-phase of a glottal cycle can be estimated as a region during which formants are relatively "stationary". In addition, the extent of formant modulation during the glottal open phase can be used to indicate the level of source/filter coupling during speech production. Here, we present a brief overview of formant modulation analysis. A more detailed discussion is given in [114].

Formant Modulation

In comparison to other formants, the first vocal tract formant exhibits the strongest dynamics after the onset of the open phase, and a higher degree of stationarity during the closed-phase interval. Thus, the formant modulation analysis is usually performed on the first formant.

The trajectory of the 1st formant is estimated over the glottal cycle duration using a one-sample-shift sliding covariance based linear prediction analysis. The analysis is initiated at one sample after an identified GCI mark, and is followed until the end of the last window reaches the next GCI mark. Hence, there are N-N_w number of windows over each glottal cycle, where N and N_w denote the pitch period and the analysis window lengths, respectively. Vocal tract coefficients are estimated for each analysis window using an all pole vocal tract model of order, p=14. 1st formant trajectory is obtained by performing a Viterbi search on a space constrained to the four lowest poles with the bandwidths less than 500 Hz;

The size of the analysis window is a crucial factor in the formant modulation analysis. The lower limit of the analysis window size is dictated by the prediction order. In order to avoid failure of Cholesky decomposition, N_w is required to be at least three samples longer than the prediction order. On the other hand, the upper constraint is governed by the amount of available data, i.e. the length of the glottal cycle. A meaningful formant modulation analysis requires the analysis window length to be a fraction of a pitch period length. Thus, the length of the analysis window is set to $N_w = N/4$, as long as the linear prediction order constraint is satisfied. Increasing the length of the analysis window beyond twice the prediction order is detrimental to the time resolution, while not having much effect on the accuracy of the formant modulation analysis.

Initial stationary region

Having obtained the formant trajectory, the next step is to identify the stationary formant region and the onset of formant modulation. In order to mark the onset of the open phase, an adaptive threshold on the degree of formant modulation is required. A fixed formant modulation threshold would not be able to provide accurate and consistent performance across the range of voice quality types and speakers.

Adaptive threshold is based on a statistical analysis of the formant values. The first step in this statistical approach is to identify a region of the formant trajectory that exhibits the highest degree of local stationarity. This glottal cycle interval is referred to as the *initial stationary formant region* (ISFR) and it can be estimated via the following algorithm:

$$ISFR = \arg\min_{n} \sum_{i=n}^{n+4} \left| F(i) - F(i-1) \right| \ , \ 1 \le n < N - N_w - 5 \tag{3.10}$$

where $F(i)$ denotes the 1st formant's value at the i^{th} sample after the instant of glottal closure. A conservative amount of data (five formant values) is used in an attempt to avoid taking formant values that are possibly outside the stationary region. Subsequently, a statistical model of formant modulation is developed for the *initial stationary* region. Gaussian distribution is used for this purpose.

Full stationary region

The next step is to expend the *initial stationary* region with the neighboring points that are statistically similar to the *initial stationary* region. The expansion is done by a one-sample-shift using the following principle: if the next formant value is less than two standard deviations away from the mean value of the statistical model, it is associated with the stationary region. As the *initial stationary* region is expended to the right, the statistical model of the stationary formant region is adapted to include the "new points". Once the formant deviation from the statistical mean exceeds the threshold of two standard deviations, the formant value is considered to be outside the stationary formant region and the expansion stops. This point marks the onset of the open phase of glottal cycle and the start of formant modulation. Subsequently, the stationary formant region is expanded to the left to identify the onset of the stationary region. Again, a threshold of two standard deviations is used. However, in the expansion to the left, the statistical model of formant modulation is not adapted as the model is already well established. Furthermore, in some cases, the movement of the 1^{st} formant prior to the initial stationary region can be very gradual and further update could lead to inaccurate threshold estimates. It is important to note that for high pitched voices, where the glottal cycle is below 7 ms, the size of the analysis window will be over-constrained. In such cases, it is assumed that both the source and the vocal tract are stationary over multiple glottal cycles and the covariance analysis window is split into two parts across two successive pitch periods.

3.3.3 A study of glottal flow derivative estimates

The closed-phase pitch synchronous inverse filtering and the vocal tract formant modulation analysis are performed on a segment of sustained vowel /a/ for 5 male speakers with different types of phonation and voice quality. The dataset includes *modal* voice, *creaky* voice, *breathy* voice, and two examples of voice disorder, *laryngeal cancer* and *vocal fold paralysis.* The dataset is sampled at 10 kHz. Vocal tract poles are estimated over the glottal closed-phase regions, as determined by the formant modulation analysis, using the covariance method of linear prediction with a 14th order predictor. The results for *creaky* voice, *breathy* voice, *vocal fold paralysis*, *laryngeal cancer*, and *modal* voices are shown in Figure 3.9 - Figure 3.13, respectively. The top panels display the speech waveforms. The 1st formant trajectories and the formant stationary regions are shown in *b)* panels. The bottom panels display the estimates of the glottal flow derivative waveforms. Note that the formant trajectory graphs correspond to 75% of glottal cycle duration, as prescribed by formant modulation analysis procedure. The time domain labeling of both panels is referenced to the identified glottal closure instant.

In each of five examples, the estimated formant trajectories exhibit clearly defined formant stationary regions. The most extensive formant modulation is observed in vocal fold paralysis examples and breathy voice. On the other end of the scale is the modal voice with the weakest formant modulation. An interesting observation is that the stationary formant regions do not always coincide with the closed-phase intervals according to the Liljencrants-Fant's representation of glottal flow derivative waveforms. In the instances of laryngeal cancer and breathy voice, the stationary formant regions extend well beyond the nominal closed-phases, whereas for modal and creaky voices, the formant modulation onsets occur prior to the nominal open phases. In the breathy voice example, inspection of the glottal flow derivative estimate suggests that vocal folds do not fully close. However, the results of formant modulation analysis reveal a clearly distinct region in which formants are stationary indicating a lack of source-filter coupling and a complete vocal fold closure. In both, modal and creaky voices, the formant modulation onset occurs before the onset of the nominal open

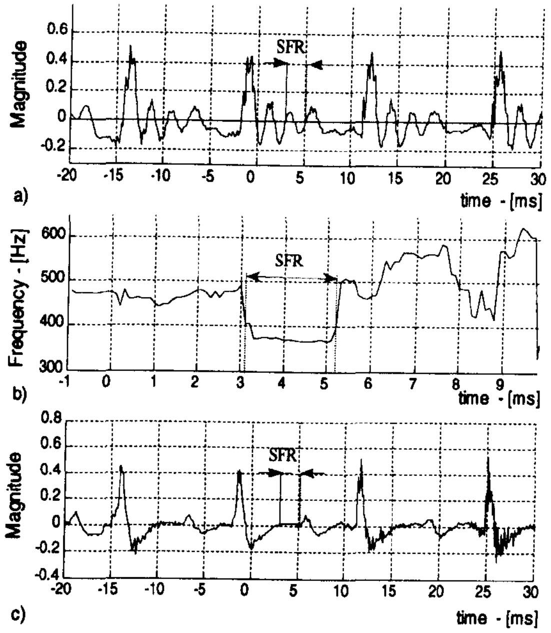


Figure 3.9: Creaky voice;
a) speech signal;
b) formant trajectory;
c) glottal flow derivative waveform.

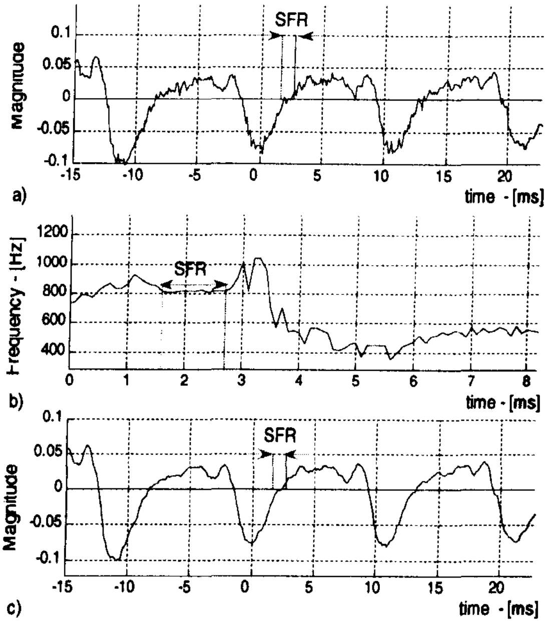


Figure 3.10: Breathy voice;
a) speech signal;
b) formant trajectory;
c) glottal flow derivative waveform.

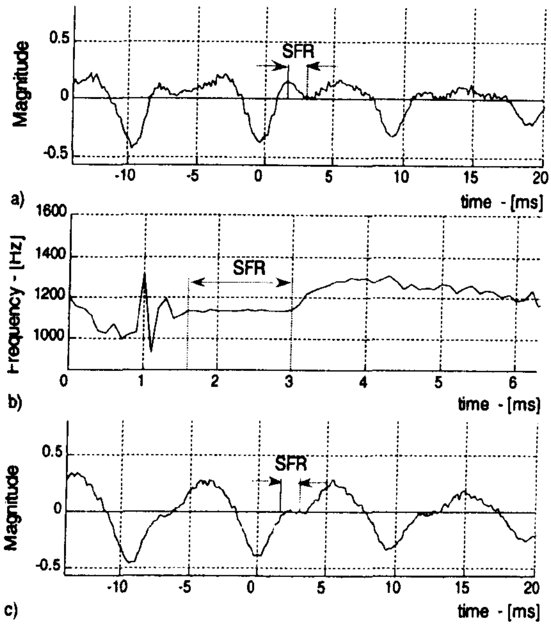


Figure 3.11: Vocal fold paralysis;
a) speech signal;
b) formant trajectory;
c) glottal flow derivative waveform.

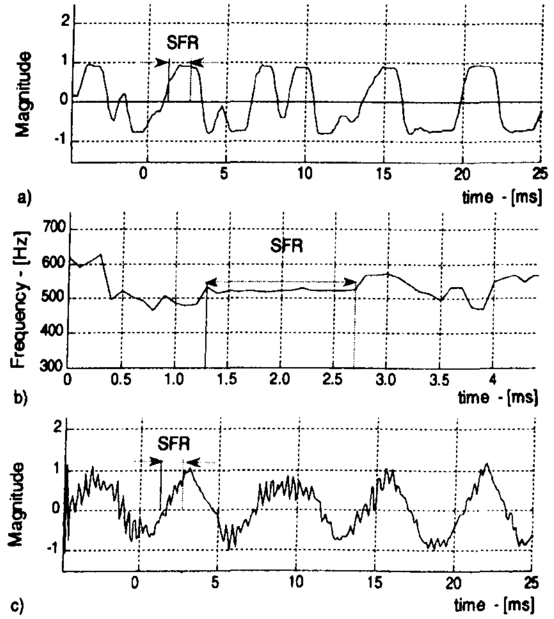


Figure 3.12: Laryngeal cancer
a) speech signal;
b) formant trajectory;
c) glottal flow derivative waveform.

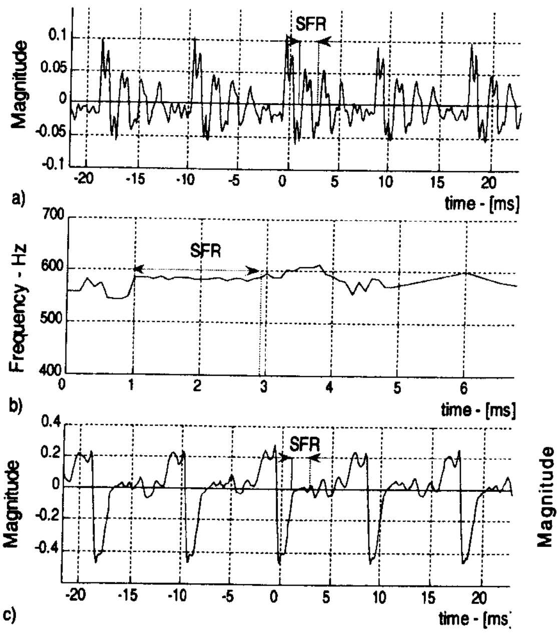


Figure 3.13: Modal voice;
a) speech signal;
b) formant trajectory;
c) glottal flow derivative

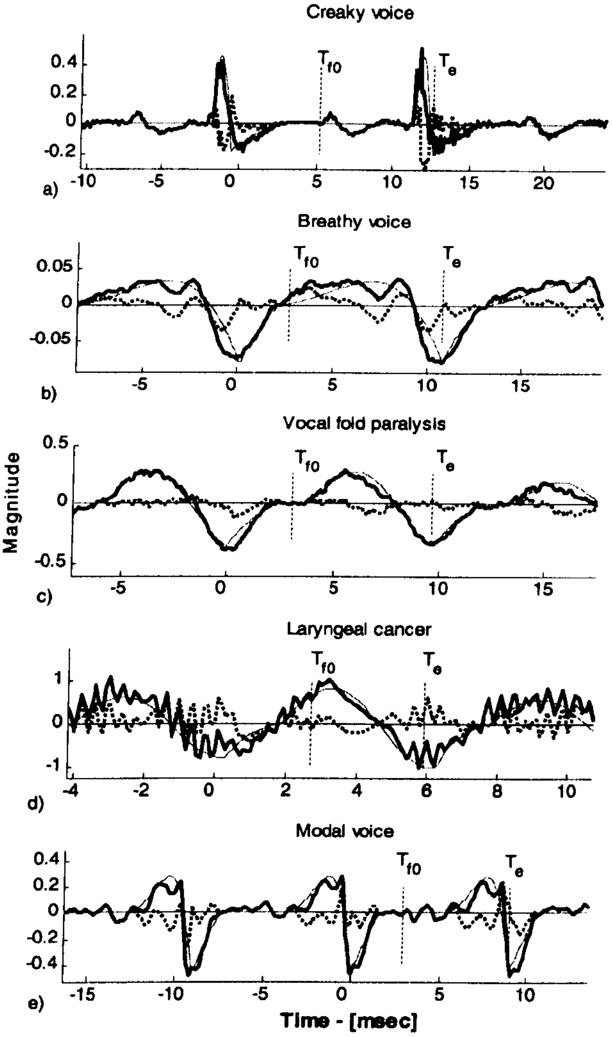


Figure 3.14: Glottal flow derivative estimates (solid thick line), synthesized LF waveforms (solid thin line), and the corresponding LF modeling residue (dotted line).

phase and coincides with the onset of formant ripple. The term "ripple" is used to describe a sinusoidal-like perturbation that overlays the glottal derivative waveform. It arises from the time-varying non-linear coupling of the glottal flow with the vocal tract, primarily with the first formant resonances [21]. Thus, this phenomenon is also referred to as the first formant ripple. Thus, we are lead to infer that the vocal folds must have been partly open during the nominal closed-phases. In other voice example, the ripple is considerably suppressed or non existent at all. The speakers also exhibit varying amounts of turbulence in the voice source realizations. In laryngeal cancer and creaky voices, a high degree of turbulence is present in the voice source estimates suggesting a narrow and parallel vocal fold opening, rather than a triangular opening. The laryngeal cancer also exhibits a specific phenomenon that is not found in any other speaker; the voice source signal is highly irregular and two distinct types of glottal flow derivative realizations can be observed. Ultimately, the reason for the varied displays of the glottal flow derivative waveforms relates to the fact that the laryngeal settings, geometry, and physiology are different for each individual [4], [44].

We have employed a signal to noise ratio measure to establish the extent by which the Liljencrants-Fant's model can be used to represent the voice source estimates. Firstly, the glottal flow derivative estimates are parameterized using Alku, and Vilkman's *direct estimation* method [2]. Manual corrections were made when deemed necessary. The results of parameterization are displayed in Table 3.1. Subsequently, the Liljencrants-Fant's waveforms are subtracted from the glottal flow derivative estimates to obtain the modeling residual signals, i.e. $v_r(n) = v_g(n) - v_{\mathrm{LF}}(n)$. The modeling SNR values are obtained as

$$SNR = 10\log_{10}(\sum_{i=0}^{N-1} v_g^2(n) / \sum_{i=0}^{N-1} \left(v_g(n) - v_{LF}(n)\right)^2 \qquad (3.11)$$

,where N refers to the glottal cycle length. For each speaker, the modeling SNR value is evaluated and presented in Table 3.2.

Table 3.1: Liljencrants-Fant's describing the synthetic waveforms in Figure 3.14, expressed as a percentage of glottal cycle duration

Stimuli	T_p [%]	T_e [%]	T_c [%]	T_a [%]	F_0 [Hz]
Creaky voice	8.17	9.15	34.13	13.74	76.34
Breathy voice	64.22	78.90	98.00	11.01	91.74
Vocal fold paralysis	43.00	61.00	86.00	15.00	103.09
Laryngeal cancer	48.00	75.00	98.33	10.20	169.50
Modal voice	36.26	40.66	59.34	9.80	111.11

Table 3.2: Modeling SNR for five speakers

Creaky	Breathy	V.f. paralysis	Cancer	Modal
4.10 dB	8.92 dB	14.49 dB	6.88 dB	9.86 dB

In Figure 3.14, we have displayed the estimated glottal flow derivative waveform, the synthesized Liljencrants-Fant's waveform, and the corresponding LF modeling residue, for each of 5 speakers. Note that T_{f0} denotes the glottal opening instant obtained via formant modulation analysis, while T_e marks the glottal closure instant. Since there is more than 10 dB difference between the best (*vocal fold paralysis*) and the worst (creaky voice) modeled voice source signal (see Table 3.2), we are inclined to suggest that the ability of the Liljencrants-Fant's model to represent the voice source signal may be speaker dependant. In order to substantiate this proposition, a study of LF residue waveforms needs to be conducted, as in [114]. In [114], the authors have focused on the *modal* voices, only. Their findings indicate that the first formant ripple and aspiration noise are the predominant features of the LF residual waveforms. Given that our study includes a wider range of voice quality types, we are able to conduct a more conclusive analysis of the residue waveforms.

Interestingly, in relation to *modal* voice, our results are in accord with those presented in [114]. However, in other examples, we have also identified the inadequacy of LF model to represent the complex temporal features in the voice source signal as jet another significant contributor to modeling error. In *breathy* voice, the ripple frequency is not anywhere near the formant frequencies as it is the case with *modal* voice. The graph in Figure 3.10 *b*) shows that there is very little formant modulation during glottal abduction. Thus, we believe that the observed "ripple" constitutes an integral part of the speaker's voice source signal. *Creaky* voice is an interesting case as well. It contains comparable amounts of first formant ripple, modeling error[‡] and aspiration noise. The first formant ripple dominates over more than the first half of the abduction phase, while the modeling error and aspiration noise occupy the regions just prior and after the glottal closure instant, respectively. All three residual elements are clearly visible and seem to exist in temporal isolation. In the *laryngeal cancer*

[‡] For the purpose of simplicity, the term, *modeling error* is from here on used to specifically denote those features of the residual signal that can not be attributed to either aspiration noise or the formant ripple.

instance, the modeling residue waveform exhibits a high degree of irregularity. The formant ripple is a dominant residue feature only for the middle of the three glottal pulses. In other pulses, high frequency aspiration noise and modeling error are the principal elements of the residual structure. The last remaining subject of our analysis, *vocal fold paralysis,* displays by far the most idealistic voice source waveform. In addition, its modeling residue does not contain any significant amounts of formant modulation artifacts nor turbulent components related aspiration noise. Thus, the modeling SNR is notably higher than in other examples. Overall, these results show that the main residual features, namely, formant ripple, aspiration noise and modeling error, are a direct consequence of an over simplistic view of vocal fold realization that is adopted by the Liljencrants-Fant's model. The relative energy distribution of the individual residual elements exhibits drastic variation across speakers and phonation types. This fact alone constitutes a notable evidence that the fine glottal flow derivative structure might be an important correlates of speaker individuality and possibly voice quality.

3.4 Conclusion

Closed-phase pitch-synchronous inverse filtering and a formant modulation analysis technique are employed on a range of voice qualities types, including two examples of laryngeal pathology, to enable a qualitative evaluation of the temporal structure of glottal excitation estimates. The results of our study suggest that due to the inherent complexity of glottal flow realizations, the inadequacies of the source–filter model of speech production and the inaccuracies in the implementation of inverse filtering, more often than not, voice source estimates do not completely comply with the idealized waveforms of Liljencrants-Fant's glottal flow derivative model. In the best of circumstances, Liljencrants-Fant's model provides enough degrees of freedom to adequately represent only the general shape or the "coarse structure" of the glottal flow derivative waveforms. The fact that Liljencrants-Fant's model can not represent complex voice source realizations nor the formant modulation ripples is a serious deficiency of this model. In the LF representation, the fine glottal flow derivative structure is discarded and correspondingly, some of the information related to the voice individuality and voice quality is inevitably lost. Furthermore, formant modulation analysis has shown that Liljencrants-Fant's parameters do not always accurately identify the significant events in the vocal fold dynamics, and thus, the process of LF-based voice source parameterization carries an inherent degree of fallibility. Presumably, these limitations are manifested in the qualities of LF-based speech synthesis and related voice quality conversion methods. Thus, we deem that a more sophisticated model is required to satisfy the requirements of the state of the art speech processing applications. In this chapter, we have also considered a group delay approach to GCI estimation. Specifically, *average group delay* and *energy weighted group delay* measures are discussed in detail. We have proposed a GCI estimation method based on a group delay algorithm and the translation-invariant hard-thresholding of LPC residue. Thresholding is performed with the 6-coefficient Coiflet filter and a primary resolution level-7. The proposed method is based on a study, presented in Chapter 4, where we have aimed to develop an optimal wavelet thresholding strategy for the glottal volume velocity derivative signals. The performances of the two group delay measures and the proposed method are evaluated for a range of fixed and pitch-synchronous group delay window lengths. We have found that in comparison to the *energy weighted group delay* measure with a fixed group delay window, the pitch synchronous *energy weighted group delay* measure with the wavelet–denoised LPC residue improves the *identification rate* and *accuracy* by 6.57 % and 0.158 ms, respectively. This represents a considerable improvement in the performance, especially if we consider that 6.57 % increase in the *identification rate* corresponds to 82.33 % reduction in the number of unidentified glottal excitations. In large, these results reflect a superior denoising performance of the wavelet thresholding method. In the standard implementation of the group delay measures, the vocal tract artifacts, aspiration noise, and other disturbances are removed from the LPC residue with the *2nd* order Butterworth low-pass filter. Unlike Butterworth filtering, wavelet thresholding is able to preserve the slow, as well as the rapid variations in the underlying signal by exploiting the compactness property of wavelets i.e localizations in time and frequency. We would expect further, but not considerable, improvements in GCI estimation performance if the wavelet thresholding is optimized for the LPC residual, specifically.

Chapter 4

Voice Source Denoising

ABSTRACT

Estimates of voice source signal obtained via closed-phase pitch-synchronous inverse filtering of recorded speech exhibit complex temporal and spectral features and to various degrees contain elements of aspiration and processing noise. In this chapter we attempt to develop an optimal wavelet-based de-noising strategy for voice source signals. Wavelet thresholding techniques are preferred over the traditional linear denoising methods, primarily because they exhibit near optimal properties in the minimax sense and offer a better rate of convergence. Our principal aim is to preserve the shape of a non-stationary signal that is observed in additive noise for further glottal excitation analysis, e.g. voice source parameterization. We acknowledge the fact that even a small degree of over-smoothing can considerably compromise the authenticity of the parametric voice quality description. Thus, denoising distortion measures, in conjunction with SNR enhancement, are employed as the performance evaluation criterion. We compare an assortment of thresholding estimators, including the classical term-by-term thresholding methods, block thresholding methods and a Bayesian method, in an extensive simulation study on six commonly cited voice quality types and a variety of priori noise levels, wavelet basis functions, and decomposition levels. The results show that the relationship between the thresholding parameters and the thresholding performance is highly non-linear. Short wavelet filters tend to have inadequate approximation properties, while more regular wavelets, corresponding to higher filter orders, have better decorrelating properties at the expense of temporal compactness. The choice of decomposition level is also found to have a strong effect on the quality of the reconstructed signal and in particular around the instants of glottal closure. Ultimately, the optimal denoising strategy is associated with translation invariant hard thresholding, decomposition level-7 and Coifl wavelet basis function. The results obtained on natural voice source data suggest that the superior performance of this denoising strategy can be attributed to its effectiveness in suppressing the pseudo-Gibbs artifacts.

4.1 Introduction

Deconvolution of the aperiodic features from the periodic components in the voice source estimates is an important aspect of voice source analysis. The problem, is to recover the underlying signal from a voice source estimate without inducing a significant level of distortions in the recovered signal. We want to preserve both, the slow and rapid variations in the glottal flow derivative waveforms in order to allow accurate parameterization of the voice source signal. In particular, we are concerned with preserving the original glottal pulse shape in the region of the glottal closure instants as we are aware that even a small degree of over-smoothing can considerably compromise the authenticity of the parametric voice quality description. The glottal flow derivative estimates obtained via closed-phase pitch synchronous inverse filtering tend to exhibit a range of features in both temporal and spectral domains. Thus, the denoising methods based on a time-frequency representation of signals should, in theory, provide better denoising performances than the standard linear methods. Unlike other standard orthonormal bases, wavelets are localized in time and frequency. Signals exhibiting rapid local changes can be well represented with just a few wavelet coefficients. Heisenberg's principle states that modeling of time-frequency phenomena can not be accurate in time domain and frequency domain simultaneously. However, by their inherent nature, wavelets provide an automatic tradeoff of the time-frequency accuracy and are able to manage the constraints related to Heisenberg's principle in a data dependant manner.

Weaver *et al.* [146] are generally regarded as the first researchers to use wavelet transforms for the purpose of suppressing noise. They have proposed a novel method, essentially, a hard thresholding scheme, for denoising magnetic resonance images. Their results highlighted preservation of edge sharpness as the principal advantage of the wavelet-based denoising. Subsequently, Donoho and Johnstone [37] mathematically derived several important properties of wavelet thresholding. They showed that wavelet thresholding exhibits near optimal properties in the minimax sense and offers a better rate of convergence than the conventional linear denoising methods. Since then, much of research effort has been dedicated to methods for threshold estimation, while the importance of other thresholding factors or specifically, the choices of a wavelet basis function and a primary resolution level are commonly neglected. Hall and Patil [61] initiated the research work in primary resolution (decomposition) level optimization. Their results unequivocally show that the choice of primary resolution can have a significant influence on the thresholding performance. Subsequently, Härdle, *et al.* [63] suggested that the primary resolution level should be asymptotically prescribed by the following equation: $j_0(n) = \log_2(\log(n)) + 1$. Nevertheless, there is a general consensus among researchers, e.g Hu & Loizou [74], Fadili & Bullmore [41], that the optimal resolution level is best determined via simulation experiments, where a full range of resolution levels is systematically evaluated. As far as the choice of wavelet basis function is concerned, researchers commonly disregard other alternatives and automatically opt for the Daubechies' family of wavelets with some moderate number of vanishing moments, e.g. Zhang & Luo [150], Lu, [96]. Although, this type of wavelet basis exhibits good approximation properties for a wide range of signals, ultimately, the optimal choice and order of wavelet basis depends on the dominant features of the underlying signal. In the context of voice source processing, there has been some effort to apply wavelet thresholding techniques on the glottal excitation signals; most notably by Hui-Ling Lu [96]. However, this PhD thesis presents a very limited study in a sense that only a single voice quality type and a single priory SNR value are considered. In addition, the author does not attempt to optimize any of the thresholding parameters for the voice source signal, specifically.

The chapter is organized as follows. In Section 4.2, an overview of wavelet thresholding is presented. In our study, we have included a range of commonly used thresholding methods such as Universal thresholding, SureShrink thresholding, Hybrid-Sure thresholding, Translation-Invariant thresholding, Hypothesis-Testing-based thresholding, Block thresholding, and Bayesian Adaptive Multi-resolution Smoother. A brief description of these methods is also provided. In Section 4.3, we have introduced a set of experiments designed to attain an optimal denoising strategy for voice source signals. Note that the optimal denoising strategy for voice source signals is first developed on simulated signals and is eventually evaluated on the natural acoustic data. The results of these experiments are reported and discussed in Section 4.4. Section 4.5 concludes the chapter.

4.2 Wavelet estimators in nonparametric regression

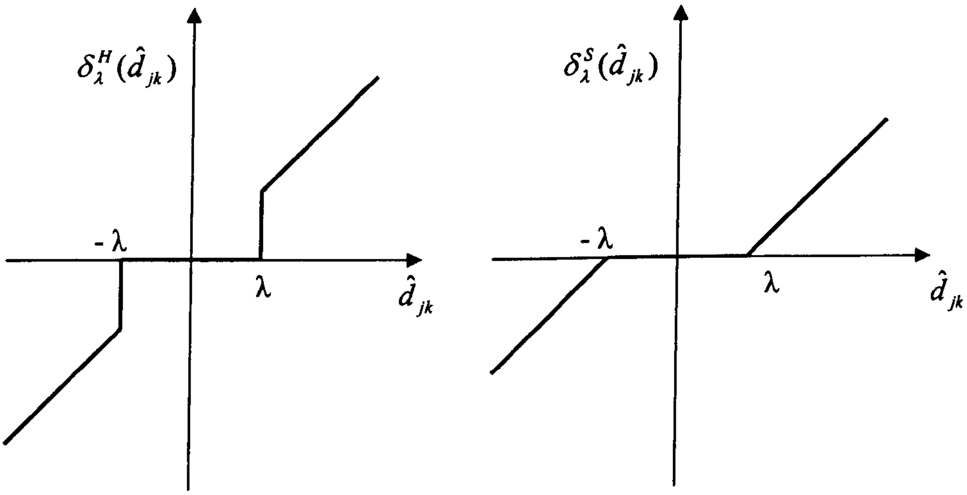
Let us consider the standard wavelet shrinkage or nonparametric regression model:

$$y_i = g_i + w_i \qquad i = 0, \ldots, N\text{-}1 \tag{4.1}$$

where the signal, g_i is corrupted with additive white Gaussian noise $w_i \sim N(0,\sigma^2)$ with zero mean and variance σ^2. The aim is to recover the underlying signal g_i from the observed noisy data y_i without assuming any parametric structure for g. The signal, g_i is only assumed to have a prescribed level of regularity. Wavelet thresholding methods exploit the fact that the energy of a signal with a certain amount of regularity is concentrated in a few coefficients in the wavelet domain, whilst the noise energy is expected to be uniformly distributed among the wavelet coefficients. Thus, the thresholding methods aim to discard the smaller coefficients associated with noise and retain the larger coefficients that represent the underlying signal g. In general, implementation of wavelet denoising is based on a three-steps procedure involving wavelet decomposition, non-linear thresholding and wavelet reconstruction. Note that for understanding the material presented in this chapter a familiarity with wavelet theory and its signal processing applications is required. The author strongly recommends the following book [145].

4.2.1 Thresholding functions

Thresholding functions are critical to the performance of wavelet based denoising methods as they dictate the level and the manner of wavelet coefficient attenuation. The two commonly used thresholding functions, *Hard* and *Soft* thresholding, are illustrated in Figure 4.1. *Hard* thresholding is based on a discontinuous function that discards the wavelet coefficients with the absolute value below a certain *threshold*, and retains all the other coefficients. It is often referred to as "keep" or "kill" rule. *Soft* thresholding also discard the coefficients with absolute values below the threshold. However, the remaining coefficients are shrunk towards zero, rather than left completely unchanged. As such, the soft thresholding rule is referred to as a "shrink" or "kill" rule. The main distinction between these two types of thresholding functions is that *soft* thresholding does not introduce discontinuities in the signal around the threshold value. The mathematical description for the *hard* and *soft* thresholding functions is presented in (4.2) and (4.3), respectively.



a) Hard thresholding *b) Soft thresholding*

Figure 4.1: Hard thresholding and soft thresholding

$$\textit{Hard thresholding} \qquad \delta_\lambda^H(\hat{d}_{jk}) = \begin{cases} \hat{d}_{jk} & \text{for } \left|\hat{d}_{jk}\right| > \lambda \\ 0 & \text{for } \left|\hat{d}_{jk}\right| \le \lambda \end{cases} \tag{4.2}$$

$$\textit{Soft thresholding} \qquad \delta_\lambda^S(\hat{d}_{jk}) = \begin{cases} \hat{d}_{jk} - \lambda & \text{for } \hat{d}_{jk} > \lambda \\ \hat{d}_{jk} + \lambda & \text{for } \hat{d}_{jk} < -\lambda \\ 0 & \text{for } \left|\hat{d}_{jk}\right| \le \lambda \end{cases} \tag{4.3}$$

, where $\hat{d}_{jk}$, $j = j_0, \ldots, J\text{-}1$; $k = 0, \ldots, 2^j\text{-}1$ corresponds to empirical wavelet coefficients.

Many other thresholding rules have been developed, such as *nonnegative garrote* thresholding [56], *firm* thresholding [55], *SCAD* thresholding [6] etc... However, in practical applications, the *soft* and *hard* thresholding are by far the most commonly employed thresholding rules, and for those reasons only they will be considered in this thesis.

A choice of threshold levels is crucial for the performance of wavelet based denoising systems. The threshold estimators aim to select a threshold value that has a high probability of being just above the maximum level of the noise coefficients. What follows is a brief overview of the thresholding methods that are considered in our attempt to develop the optimal denoising strategy for voice source signals.

4.2.2 Thresholding methods

Universal threshold

Donoho and Johnstone proposed a simple but very effective wavelet shrinkage method (*VisuShrink*) based on the *universal* threshold for each scale [35]. The *universal* threshold is defined as:

$$\lambda^U \overset{\Delta}{=} \hat{\sigma}\sqrt{2\log N} \tag{4.4}$$

,where σ^2 and N denote the noise variance and the signal length, respectively.

Since the *universal* threshold estimator relies only on noise variance value and not on the input signal, it can be efficiently implemented. However, it tends to overestimate the threshold level and in many occasions it over-smoothes the noisy signal [37].

SureShrink threshold

Donoho and Johnstone proposed a method whereby empirical wavelet coefficients are thresholded at each resolution level j based on the threshold value λ_j^S [37]. The proposed method relies on the Stein's unbiased risk criterion to obtain an unbiased estimate of l^2-risk.

Let us consider the principal denosing problem in (4.1), and suppose that $\{X_1,\ldots,X_S\}$ are independent s-dimensional random variables $N(\mu_i,1), i=1,\ldots,s$. The problem is to estimate the mean vector $\mu=(\mu_1,\ldots,\mu_S)'$ with minimum l^2-risk. Stain showed that for a nearly arbitrary nonlinear biased estimator its l^2-loss can be estimated in the unbiased fashion. For any estimator μ that can be expressed as $\hat{\mu}(X)=X+g(X)$, where function $g=(g_i)_{i=1}^S : R^S \to R^S$ is weakly differentiable, Stein states that $\|\hat{\mu}(X)-\mu\|^2$ can be formulated as:

$$E_\mu\|\hat{\mu}(X)-\mu\|^2 = s + E_\mu\{\|g(X)\|^2 + 2\nabla\cdot g(X)\} \tag{4.5}$$

where $2\nabla\cdot g(X)=\sum_i \frac{\partial g_i}{\partial g_x}$.

Applying the soft thresholding rule on (4.5) we can obtain the unbiased estimate of l^2-risk.

$$SURE(\lambda;X) = d - 2\cdot\#\{i : |X_i| \le \lambda\} + \sum_{i=1}^{d}(|X_i| \wedge \lambda)^2 \tag{4.6}$$

where $E_\mu SURE(\lambda;x) = E_\mu\|\hat{\mu}^\lambda(X)-\mu\|^2$; # denotes the cardinality of a matrix.

The SURE threshold estimator is set to minimize the l^2-risk estimate as:

$$\lambda^s = \arg\min_{0\le\lambda\le\lambda^*} SURE(\lambda;X) \tag{4.7}$$

, or equivalently:

$$\lambda^s = \arg\min_{0 \le \lambda \le \lambda^U} SURE(\lambda; \frac{\hat{d}_{jk}}{\hat{\sigma}}),\ j = j_0, \ldots, J\text{-}1;\ k = 0, \ldots, 2^j\text{-}1 \tag{4.8}$$

where $\lambda^* = \sqrt{2 \log s}$, and $\lambda^U \overset{\Delta}{=} \hat{\sigma}\sqrt{2 \log N}$ with $N = 2^j$. Donoho and Johnstone [37] proposed a robust estimate of noise level:

$$\sigma_j = \frac{MAD_{j,x}}{0.6745} \tag{4.9}$$

where $\mathrm{MAD}_{j,x}$ denotes the median absolute deviation of all wavelet coefficients at resolution level *j*. The normalization factor is derived from the fact that the expected median magnitude of a zero mean Gaussian white noise sequence of variance σ^2, is given by 0.6745σ. The estimator is very robust and accurate, and as such, it is used by most thresholding methods. Johnstone and Silverman have shown that the *SURE* threshold estimator can also be applied in the presence of correlated noise [80].

Hybrid Sure threshold

Donoho has shown that in the instances of severe sparsity of wavelet coefficients, such as when the noise dominates the input signal $X = \{x\}_{i=1}^N$, the *universal* threshold outperforms the SURE threshold [36]. For that reason Donoho and Johnstone [37] developed a *heuristic SURE* estimator where a threshold value is selected as either *SURE* or *Universal* based on a comparison of s_d^2 and λ_d values. The *heuristic SURE* estimator is given as:

$$\lambda = \begin{cases} \sigma\sqrt{2 \log N}, & s_d^2 \le \lambda \\ \lambda^S, & s_d^2 > \lambda \end{cases} \tag{4.10}$$

, where $s_d^2 = \frac{1}{N}\sum_{i=1}^N x_i^2 - \sigma^2$ and $\lambda_d = \frac{\sigma}{\sqrt{N}}(\log_2 N)^{\frac{3}{2}}$. λ^S is obtained according to (4.8).

Translation-Invariant threshold

Translation invariant thresholding was first introduced by Coifman and Donoho in [25]. The method is developed with the purpose to reduce the thresholding artifacts associated with the pseudo-Gibbs phenomena around the discontinuity points. This achieved by averaging out the translation dependence. Essentially, the process involves shifting of data, denoising of the shifted data, and back-shifting of the reconstructed data. The procedure is repeated for a full range of circulant shifts and the average of the reconstructed signal is taken to represent the denoised signal.

The translation invariant wavelet thresholding estimator for data $y = (y_1, y_2, y_3 \ldots ,y_n)$ is defined as :

$$\hat{g}^{TI} = \frac{1}{n}\sum_{k=1}^{n}(WS_k)'\,\delta_\lambda\,(WS_k y) \qquad (4.11)$$

where S_k denotes the shift matrix:

$$S_k = \begin{pmatrix} O_{k\times(n-k)} & I_{k\times k} \\ I_{(n-k)\times(n-k)} & O_{(n-k)\times k} \end{pmatrix} \qquad (4.12)$$

The term δ_λ corresponds to a thresholding rule type (e.g. *soft* or *hard*). *W* represents the *n*th order orthogonal DWT matrix; *I* and *O* denote the identity matrix and *zero* matrix, respectively. The subscripts denote matrix dimensions.

In this thesis, we will consider *translation invariant* thresholding with both *hard* and *soft* thresholding rules. Translation invariant methods will be based on the thresholding level that is recommended by Coifman & Donoho $\lambda = \hat{\sigma}\sqrt{2\log_e((n\log_2(n))}$. They have suggested that the lower thresholding levels can have counterproductive effects, whereby the extent of the thresholding artifacts is higher than for the non-invariant thresholding methods.

Thresholding as a Recursive Hypothesis Testing Problem

Ogden & Parzen proposed a hypothesis testing procedure with the level dependant thresholds [113]. In this thresholding method, wavelet coefficients are retained only when there is strong evidence that they are required for reconstruction. Let us consider $\{X_1, X_2, X_3 ...., X_s\}$ as independent random variables N(μ_i, 1), that represent the observed wavelet coefficients at any level $j = j_0, \ldots, J\text{-}1$ with $s = 2^j$. In addition, let I_s denote a non-empty subset of indices $\{1,\ldots,s\}$. The multiple hypothesis testing problem can be expressed as follows:

$$H_0 : \mu_i = 0,\ i \in I_s \ \text{ vs } H_1 : \mu_i \neq 0,\ i \in I_s \text{ and } \mu_i = 0 \text{ for all } i \notin I_s. \tag{4.13}$$

The hypothesis is tested with the standard likelihood ratio test (LRT). If I_s cardinality is known, say equal to k, then the standard likelihood ratio test statistic is the sum of squares of the k largest X_i's. In practice, the cardinality of the set I_s, is rarely known and the authors have suggested a recursive testing procedure for I_s. The critical threshold at level α for this distribution is derived as:

$$\lambda_s^\alpha = \left(\Phi^{-1}\left[\frac{(1-\alpha)^{1/s}+1}{2}\right]\right)^2 \tag{4.14}$$

where Φ denotes the cumulative distribution function of a standard normal random variable. The threshold λ_j at each level $j = j_0, ..., J\text{-}1$ is obtained recursively according to the following steps

1. Obtain λ_s^α according to (4.14) for each level and orientation and compare it to the largest X_i^2.
2. For $X_i^2 \geq \lambda_s^\alpha$ discard the X_i's with the largest absolute value, set s to $s\text{-}1$, and return to Step 1.
3. If $X_i^2 < \lambda_s^\alpha$, then the residual wavelet coefficients are deemed not to contain a strong signal. For the current level j the threshold λ_j is set equal to the largest remaining X_i in absolute value
4. Apply the inverse DWT to obtain an estimate of the function g.

This procedure has level dependent thresholds λ_j with inherent *soft* thresholding rule. At each level j, 'large' wavelet coefficients are discarded from the dataset until the residue is not distinguishable from pure noise. Having threshold λ_j equal to the maximum absolute value of the residual wavelet coefficients, it is ensured that the residual coefficients are shrunk to zero. The significant wavelet coefficients are also shrunk towards zero by the same amount. The parameter α determines the thresholds λ_j, and controls the amount of smoothness. By increasing the value of α, the likelihood of a wavelet coefficient being included in the reconstruction increases, and thus, the reconstructed signals tend to exhibit higher levels of smoothness. In our study, the recommended value $\alpha = 0.05$ will be used.

Block thresholding

In term-by-term thresholding, each wavelet coefficient is compared with a set threshold. The coefficients with the absolute value above a threshold value are retained, while others are discarded. This approach tends to remove too many terms from the empirical wavelet expansion which leads to estimation bias and a suboptimal l^2-risk convergence rate. Block thresholding methods attempt to improve the estimation accuracy by taking into account information about neighboring empirical wavelet coefficients. The empirical wavelet coefficients are thresholded in blocks rather than individually. Hence, the amount of information available for estimating the "average" empirical wavelet coefficients and for making decisions about retaining or discarding them is an order of magnitude larger than in the case of term-by-term thresholding. In our study, we will consider both the non-overlapping and overlapping block thresholding estimators.

Non-Overlapping Block threshold

Cai proposed a non-overlapping block thresholding estimator based on the ideal adaptation approach and inequality oracle [14]. At each resolution level $j = j_0, \ldots, J\text{-}1$, the observed wavelet coefficients $\hat{d}_{jk}$ are assembled into blocks of length L. When L does not divide 2^j exactly, the first few empirical wavelet coefficients can be used in deduction of the final block (augmented case), or the remaining observed wavelet coefficients can be discarded (truncated case). Within each block, wavelet coefficients are estimated simultaneously using the James-Stein thresholding rule:

$$\hat{d}_{jk}^{jb} = \max\left(0, \frac{S_{jb}^2 - \lambda L \sigma^2}{S_{jb}^2}\right)\hat{d}_{jk} \tag{4.15}$$

where S_{jb}^2 corresponds to the sum of squared empirical wavelet coefficients in the block *(jb)*; The term *(jb)* indicates the set of indices of the wavelet coefficients in the b^{th} block at level j such that $(jb) = \{(j,k) : (b-1)L + 1 \le k \le bL\}$. The function g is reconstructed by applying IDWT on a vector of thresholded wavelet coefficients $\hat{d}_{jk}^{jb}$ for k=0...2^{j0}-1 and j=j_0...J-1.

Cai has shown that block size L = *log*(n) realizes an estimator that is both locally and globally adaptive [14]. He has also recommended a threshold value for function estimation problems, λ = 4.50524. The resulting non-overlapping block thresholding estimator is commonly referred to *BlockJS.*

Overlapping Block threshold

Overlapping block thresholding estimator is essentially a variant of the non-overlapping block thresholding estimator. In the overlapping estimator, blocks are extended by $L_e = \max(1, L_0/2)$ in each direction, to include the wavelet coefficients of the neighboring blocks. L_0 refers to the length that a block would have without the overlap. Within each block, wavelet coefficients are estimated simultaneously using the James-Stein thresholding rule in (4.15). As in the case of non-overlapping block thresholding, the function g is reconstructed by applying the inverse DWT on a vector of thresholded wavelet coefficients $\hat{d}_{jk}^{jb}$ for k=0...2^{j0}-1 and j=j_0...J-1. Cai & Silverman [15] suggested using L_0 = *log*(n/2) and λ = 4.50524. The resulting thresholding procedure is referred to as *NeighBlock.*

Bayesian Adaptive Multiresolution Smoother (BAMS)

Vidakovic & Ruggeri have developed the Bayesian Adaptive Multiresolution Smoother [142]. In comparison to other Bayesian-based wavelet thresholding methods, BAMS uses simple and optimized shrinkage rules, and thus, it is computationally inexpensive. For more information on the Bayesian Adaptive Multiresolution Smoother (BAMS) refer to [142].

4.3 Developing the optimal denoising strategy for glottal flow derivative signals

In the previous section, we have described a number of denoising procedures. These procedures are listed in Table 4.1. The first two columns denote the indexes and the acronyms, while the final two columns provide the description of the thresholding methods.

Table 4.1:
A list of wavelet thresholding procedures considered in the voice source denoising study.

1	*VISU-H*	VisuShrink	Hard
2	*VISU-S*	VisuShrink	Soft
3	*SURE*	SureShrink	Soft
4	*HYB-SURE*	SureShrink	Hybrid
5	*TI-H*	Translation Invariant	Hard
6	*TI-S*	Translation Invariant	Soft
7	*THRDA1*	Hypothesis Testing	Soft
8	*BLOCKJS-A*	Block Thresholding	Augment
9	*BLOCKJS-T*	Block Thresholding	Truncate
10	*BLOCK-NEIGH*	Overlapping Block Thresholding	
11	*BAMS*	Bayesian Adaptive Multi-resolution	

Dataset

The optimal denoising strategy for voice source signals is first developed on simulated signals and is eventually evaluated on the natural acoustic data. In this section, we will describe the synthetic dataset. Synthetic dataset consists of 6 test signals. The test signals are essentially a synthesized stream of 100 glottal flow derivative pulses (Liljencrants-Fant's model) with added aspiration noise (aspiration noise model is presented in figure 2.18). Each test signal corresponds to one of the 6 voice quality types that are most commonly cited in literature: *modal, vocal fry, falsetto, breathy, tense,* and *lax* voice. As such, it is our hope that the selected data would be adequately representative of the full spectrum of naturally occupying voice source realizations.

Table 4.2:
Fitch frequency and *R*-parameters (expressed as a percentage of the glottal cycle duration) describing the synthetic dataset. The values for *modal, vocal fry, falsetto,* and *breathy* voice are obtained from Childers and Lee [20] while the values for *tense* and *lax* voices are obtained from Van Dinther [33].

Index	Voice Quality	R_a [10^{-2}]	R_k [10^{-2}]	R_0 [10^{-2}]	F_0 [Hz]
1	*Modal*	2.1	30.6	64.0	106
2	*Vocal Fry*	0.5	25.0	25.0	45
3	*Falsetto*	13.3	35.1	77.0	344
4	*Breathy*	10	44.8	84.0	200
5	*Tense*	1.1	25.0	41.0	110
6	*Lax*	2.0	51.0	82.0	110

Table 4.2 shows the parametric description of the considered voice quality types. The parameters for *modal, vocal fry, falsetto,* and *breathy* voice are obtained from Childers and Lee [20], while the parameters descriptions the *tense* and *lax* voices are taken from Van Dither [33]. Glottal flow derivative waveforms are synthesized using the Liljencrants-Fant's model, see Chapter 2, Subsection 2.2.4.2. The turbulence noise is synthesized using the model described in Figure 2.18, using the Hamming window with N = 0.5 duty cycle. The Hamming window is centered around the glottal closures instant, and thus $L = 0$; the noise floor is set to 40.0 % of the maximum noise envelope amplitude, $A_{NF} / A_M = 0.4$. The actual

values for the two amplitude parameters are obtained from the priori SNR value. Examples of synthetic dataset are provided in Appendix B, Figures B.1 – B.6. We have to stress that the synthetic dataset represents only a crude approximation to the actual voice source estimates as the lack the fine glottal flow derivative structure as well as a range of degradations due to the imperfect source-filter deconvolution.

Evaluation criteria

The performance of thresholding methods is evaluated using the following performance criteria: *signal to noise ratio* and *maximum deviation*. Signal to noise ratio (SNR) is defined as the energy ratio between the "clean" test function (a stream of glottal derivative pulses prior to addition of turbulence noise) and the error in the reconstructed (denoised) signal. SNR values are obtained as:

$$SNR \equiv 10\log_{10}\sum_{n=1}^{N}\left(\frac{f^2(n)}{(f(n)-\hat{f}(n))^2}\right) \tag{4.16}$$

where $f(n)$ and $\hat{f}(n)$ correspond to the "clean" test function and the reconstructed signal, respectively. N denotes the number of samples in each of the two functions.

Maximum deviation (*MXDV*) is a measure of distortion in the recovered signal. *Maximum deviation* is obtained for each glottal cycle and a statistical model of MXDV distribution over a stream of glottal pulses is used as a performance measure. Four statistical parameters of *maximum deviation* (MXDV) are considered: the average value, standard deviation, and its upper and lower limit. MXDV for the i^{th} glottal flow derivative pulse is estimated as

$$MXDV_i \equiv \max_{(i-1)\times T \le n \le T\times i-1}\left|f(n)-\hat{f}(n)\right| \tag{4.17}$$

where T denotes the length of glottal cycle. $MXDV_i$ indicates the maximum deviation value for the i^{th} glottal pulse in the glottal pulse sequence.

Experiments

In the first experiment, the performance of each thresholding method is optimized with the respect to the wavelet basis family, decomposition level, and wavelet filter length. For that purpose, a set of orthogonal wavelet families (*Daubechies'*, *Symmlets* and *Coiflets)*, a wide range of filter lengths (4-20 for *Daubechies'*, 4-16 for *Symmlets*, 6-30 for *Coiflets*), and a range (2-9) of decomposition levels are considered. During the optimization process, average *SNR* value across the test functions is used as a primary performance criterion. However, when two or more denoising strategies are found to have similar SNR performances, then their MXDV results are used to indicate the best denoising option. In the second experiment, the performance of the optimized thresholding methods is evaluated on the individual voice quality types. Our final judgment on the optimal voice source denoising strategy is made upon consideration of the performance of each thresholding strategy across a range of priory SNR levels. Finally, the optimal voice source denoising strategy is evaluated on natural acoustic data. In order to simplify the presentation of results, the labeling of some graphs is based on the numerical indexes, rather than on the actual names of the denoising procedures and test functions.

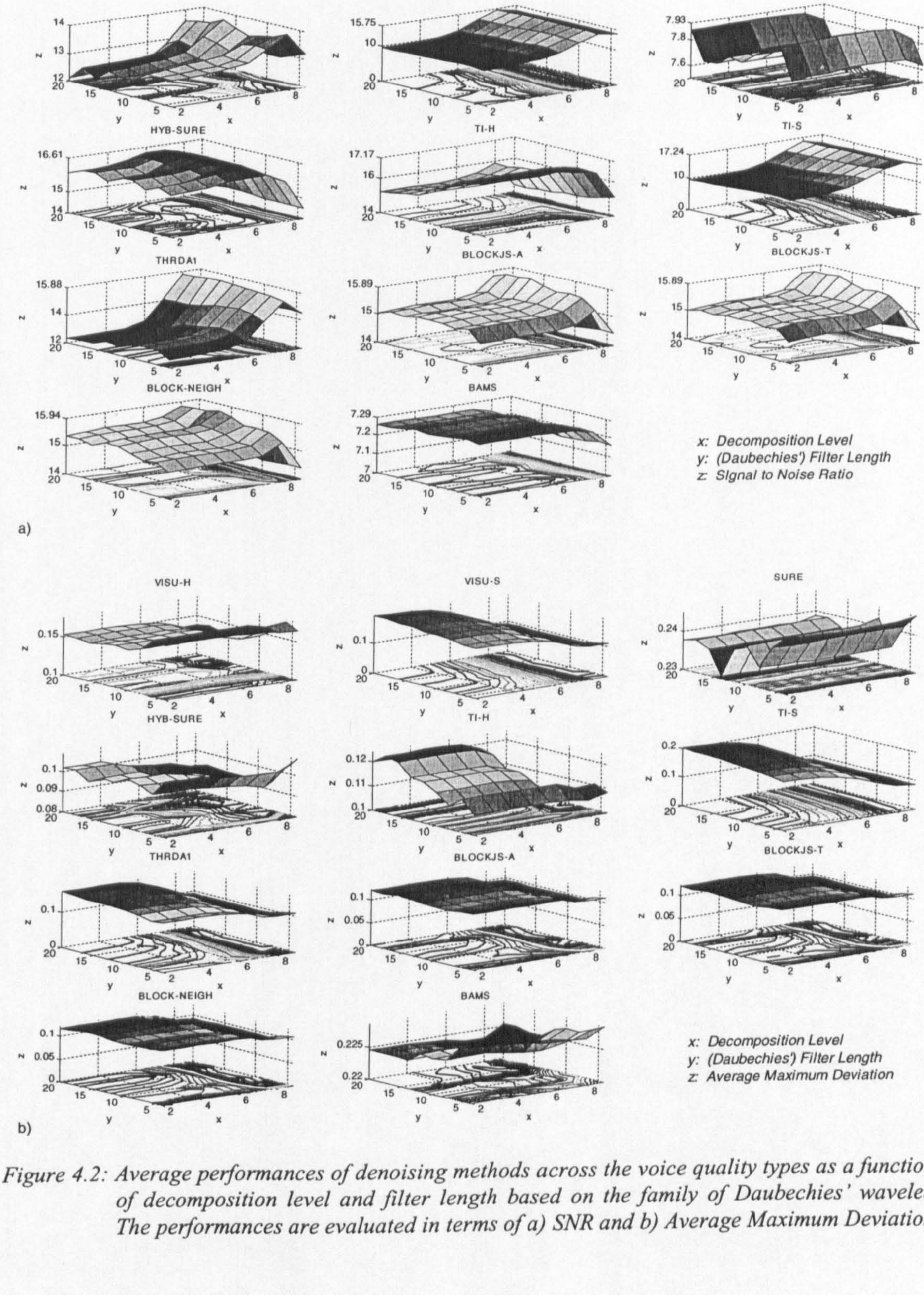


Figure 4.2: Average performances of denoising methods across the voice quality types as a function of decomposition level and filter length based on the family of Daubechies' wavelet. The performances are evaluated in terms of a) SNR and b) Average Maximum Deviation

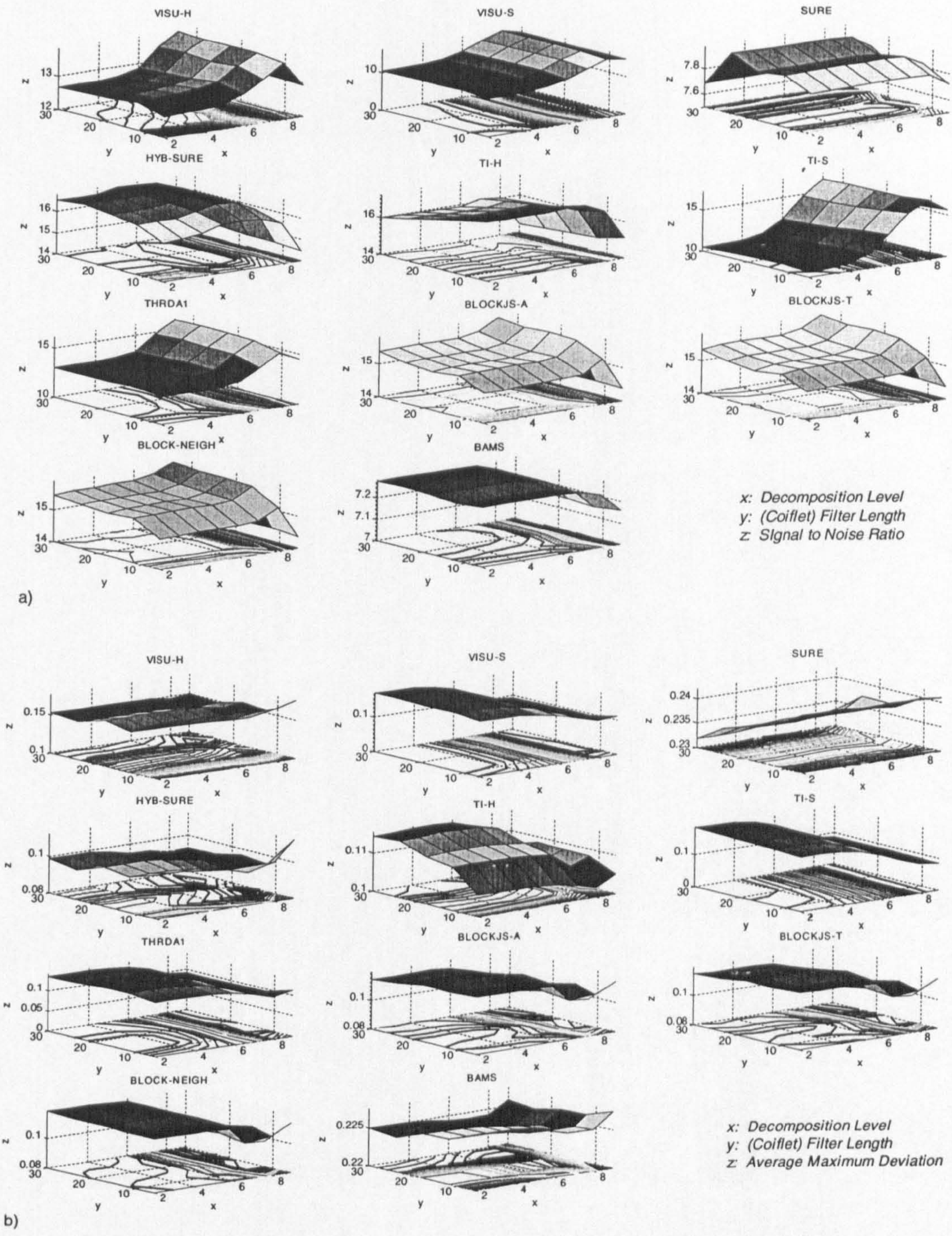


Figure 4.3: Average performances of denoising methods across the voice quality types as a function of decomposition level and filter length based on the family of Coiflet wavelets. The performances are evaluated in terms of a) SNR and b) Average Maximum Deviation

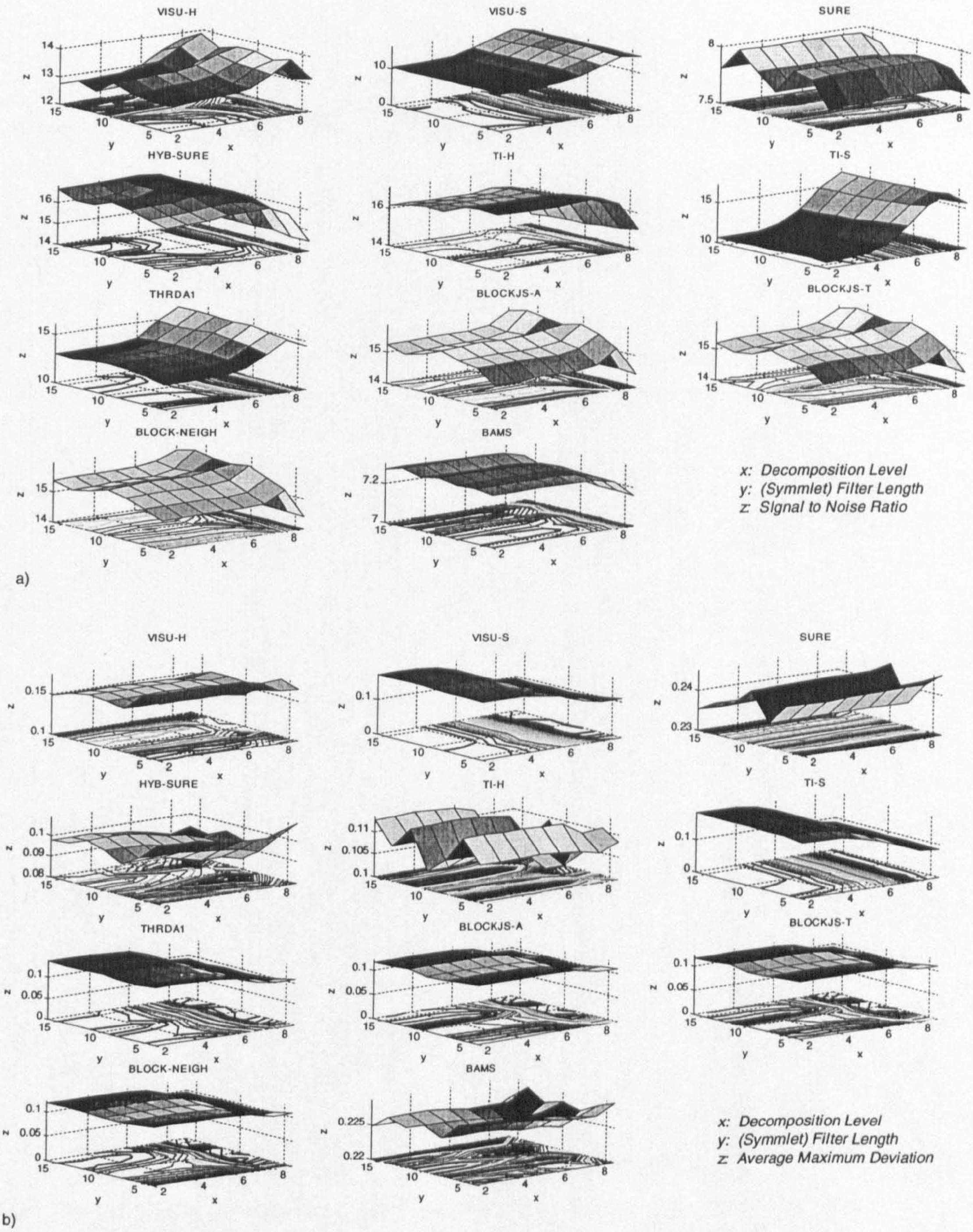


Figure 4.4 Average performances of denoising methods across the voice quality types as a function of decomposition level and filter length based on the family of Symmlet wavelets. The performances are evaluated in terms of a) SNR and b) Average Maximum Deviation

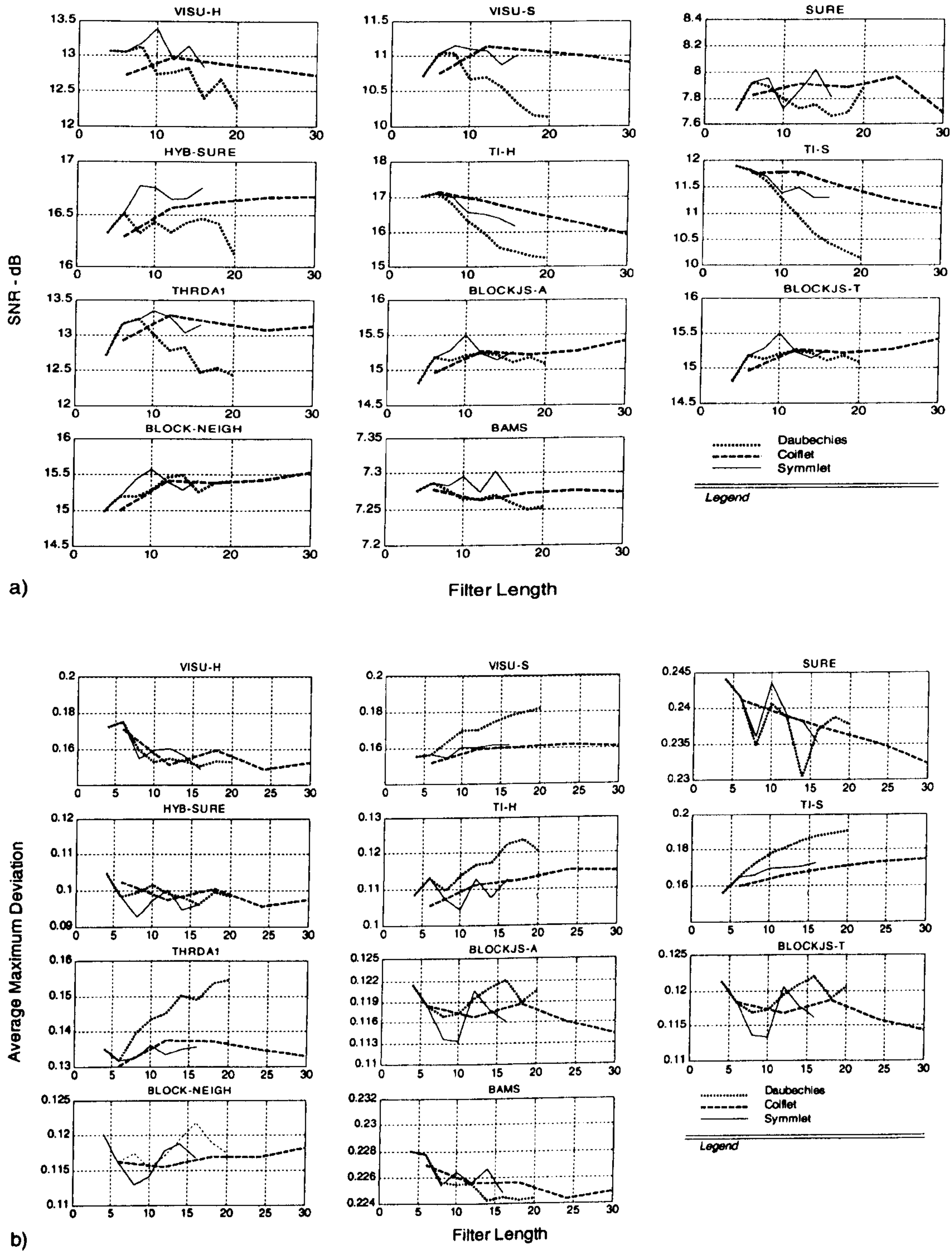


Figure 4.5: Average performances of denoising methods across the voice quality types as a function of filter length. The decomposition level is fixed at 5.

4.4 Results and discussion

Optimizing the performance of each thresholding method

The performance of each thresholding method is evaluated as a function of filter length and decomposition level. The test signals are synthesized for SNR = 6dB. The chosen SNR value is considerably lower than what would be the realistic aspiration noise level in the voice source estimates. However, low SNR value puts a higher strain on denoising methods and enables easier performance evaluation. The results for Daubechies', Coiflet, and Symmlet wavelets are shown in Figure 4.2, Figure 4.3, and Figure 4.4, respectively. The panel *a)* corresponds to average SNR results, while panel *b)* displays the average *maximum deviation* results across the voice quality types. The graphs show how the performance of each thresholding method varies as a function of two parameters wavelet filter length and decomposition level.

Visual inspection of these performance surfaces and their respective contour lines suggest that the relationship between the thresholding parameters and the thresholding performance is highly non-linear. Clearly, completely different combinations of decomposition levels and wavelet filter lengths can lead to similar thresholding performances. It is also evident that this relationship differs from one thresholding method to another. Some thresholding methods are more sensitive to wavelet filter length, whereas other methods are more sensitive to the choice of decomposition level. For instance, VISU-S, TI-S and THRDA1 are much more sensitive to the choice of decomposition level than they are to filter length while in contrast, the performances of TI-H and SURE are almost exclusively dependant on the choice and order of wavelet basis.

In order to further investigate how the choice and order of the wavelet basis affects the thresholding performance we have reported the performance of each thresholding method at decomposition level $J_c = 5$ as a function of wavelet filter length, see Figure 4.5. This figure allows a comparative study between Daubechies, Coiflet and Symmlet wavelets. Again, SNR and average MXDV results are shown in panels *a)* and *b),* respectively. Although the results suggest that each thresholding method is uniquely affected by the choice of wavelet filter length, some generalizations can still be made. First, we will note that the length of the filter determines the number of vanishing moments and the regularity of the wavelet. Vanishing moments correspond to the degrees of the polynomials representing a linear combination of the smoothing function and its translation. As such, the number of vanishing moments determines the rate of wavelet convergence. In Figure 4.5, we can clearly see that wavelets with short filter lengths tend to perform poorly. They do not exhibit a sufficient degree of regularity to provide an adequate representation of the voice source signal. On the other hand, more regular wavelets, corresponding to higher filter orders, have better decorrelating properties at the expense of temporal compactness. Clearly, as the length of the wavelet filters is increased beyond the optimal value the performance of the denoising methods deteriorates. When the wavelets become too regular, the rapidly varying components of the underling signal are being over-smoothed. This phenomenon is easily detected by the sharp increase in MXDV values. In the vast majority of the considered thresholding methods, a reasonable compromise between temporal compactness and regularity is found to exist for some moderate filter length.

Our observations of the reconstructed voice source signals have suggested that the choice of decomposition level has a significant effect on the thresholding performance. Generally, at near optimal decomposition level, the reconstructed signal contains temporally localized distortions in form of aspiration noise residue and pseudo-Gibbs artifacts. The pseudo-Gibbs artifacts arise due to the poor alignment between the discontinuities in the signal and wavelet features. We have to note that in contrast to the classical Gibbs phenomena corresponding to the Fourier based analysis, the pseudo-Gibbs phenomena are considerably better behaved. Pseudo-Gibbs artifacts exhibit a high level of temporal localization and low amplitudes of oscillation. At inadequate decomposition levels, and especially when the decomposition is too-deep, the reconstructed signal contains significantly higher levels of distortion. Low decomposition levels cause "over-smoothing" of the underlying signal around the glottal closure instant. The average maximum deviation results in Figure 4.2, Figure 4.3, and Figure 4.4, are in accord with these observations.

Table 4.3:
The worst and the optimal combination of thresholding parameters for each denoising method. The denoising strategies are arranged in a descending order according to the maximum SNR level

Method	Max SNR				Min SNR				Optimization Capacity [dB]
	Wavelet	Filter length	Decomp. level	SNR [dB]	Wavelet	Filter length	Decomp. level	SNR [dB]	
BAMS	Sym 7	14	6	7.31	Db 9	18	9	7.14	0.17
SURE	Sym 7	14	7	8.02	Db 8	16	9	7.59	0.43
VISU-H	Sym 7	14	8	14.15	Db 10	20	2	12.11	2.04
BLOCKJS-A	Db 7	14	8	15.89	Db 4	8	9	14.34	1.54
BLOCKJS-T	Db 7	14	8	15.89	Db 4	8	9	14.34	1.54
VISU-S	Coif 3	18	8	15.93	Db 10	20	2	9.62	6.31
BLOCK-NEIGH	Sym 7	14	8	15.95	Db 4	8	9	14.35	1.60
THRDA1	Coif 3	18	8	16.05	Db 10	20	2	12.34	3.70
HYB-SURE	Sym 4	8	6	16.79	Coif 1	6	9	14.47	2.32
TI-H	Coif 1	6	7	17.23	Db 9	18	9	14.42	2.81
TI-S	Db 1	2	8	17.24	Db 10	20	2	9.61	7.63

In order to demonstrate the importance of selecting adequate decomposition levels and wavelet basis functions, we have evaluated *optimization capacity* of each thresholding method. The *optimization capacity* is defined as the SNR difference between the best and the worst SNR performance for the considered range of thresholding parameters. The results are presented in Table 4.3. TI-S, with optimization capacity of 7.63 dB, is found to have the highest level of sensitivity to the choice of thresholding parameters. VISU-S exhibits the second highest *optimization capacity* of 6.31 dB, while the majority of other methods have the *optimization capacity* just above 1.5 dB. The results also show that the two optimized translation invariant procedures by far outperform the other thresholding strategies. Our informal evaluation of the reconstructed voice source signals indicate that the superior performance of these methods can be attributed to their effectiveness in suppressing pseudo-Gibbs artifacts. The results in Table 4.3 also show that the optimal choice and order of wavelet basis function varies significantly between the thresholding methods. In fact, all three wavelet families, i.e. Daubechies, Coiflet and Symmlet wavelets, constitute an optimal choice for some thresholding method. Since, more than a third of the considered thresholding methods achieve their optimal performance for Symmlet 7 wavelet, we can infer that this wavelet basis has similar levels of smoothness as the voice source waveforms. On the opposite extreme is the Daubechies wavelet with the filter length 20. This wavelet basis function is often found to produce poor denoising performances. With respect to the decomposition level, the thresholding methods perform very poorly for the extreme values of the considered decomposition level range. On the other hand, moderately high decomposition levels are found to provide the best denoising options. Indeed, all of the considered thresholding methods attain their optimal performances for the decomposition levels ranging from 6-8.

Performance evaluation: optimized thresholding methods across voice quality types

The performances of each optimized thresholding method across the voice quality types are reported in Figure 4.6. We will remind that the x-axis indices $\{1, 2, 3, 4, 5, 6\}$ correspond to the following voice quality types: *modal*, *vocal fry*, *falsetto*, *breathy*, *tense* and *lax* voice, respectively. Panel *a)* corresponds to SNR results, while panel *b)* displays the *maximum deviation* results. Four statistical measures of *maximum deviation* (MXDV) distribution are displayed: the average value, standard deviation, and its upper and lower limit. The rectangular box is centered at the mean of *maximum deviation* and its length indicates a distance of two standard deviations. The vertical line corresponds to the distribution span.

The SNR results show that the thresholding performances are relatively consistent across the voice quality types. Only *tense voice* slightly stands out with somewhat lower SNR values compared to the other voice quality types. We have found *maximum deviation* results much more interesting. Figure 4.6b) suggests that the *falsetto, breathy* and *lax* voices exhibit larger distortion levels than vocal *fry* and *tense* voices. Based on our experience with wavelet thresholding, we have come to expect higher distortion levels from rapidly varying signals with intense spikes and more discontinuities. In case of voice source signals, the adducted phonations, such as *vocal fry* and *tense* voices, are associated with long closed-phases, sharp and intense pulses around the glottal closure regions. On the other hand, the abducted phonations, such as *falsetto, breathy* and *lax voices*, vary more gradually and the glottal closure pulses are less intense and less rapid. As such, we would expect the reconstructed waveforms corresponding to abducted phonations to exhibit considerably lower distortion levels. The fact that the opposite is true can be explained by considering the temporal structure of glottal flow derivative pulses and turbulence noise signals. The *abducted* phonations have short closed-phases, and thus, they have a uniformly distributed glottal flow derivative energy. Conversely, the *adducted* phonations, with long closed phases, have a very asymmetrical energy distribution over the glottal cycle length; much like the aspiration noise, most of the glottal flow derivative energy is concentrated in a close proximity of the glottal closure instants. Thus, the effective SNR level in the vicinity of glottal closure instants is lower in *adducted* phonations and consequently, the reconstructed voice source signals exhibit higher levels of aspiration noise residue and higher distortion levels. Note that the noise residue corresponds to the high amplitude components of aspiration noise that were not suppressed as their amplitude exceeded the estimated threshold value.

Optimized thresholding methods across a range of priori SNR level

In this section, we have evaluated the performance of the optimized thresholding strategies for a priori SNR range of 0 dB - 21 dB, in increments of 3 dB. The results are reported in Figure 4.7. Let us remind that we have already established that TI-S provides the best denoising option for prior SNR = 6 dB. The results in Figure 4.7 show that the translation invariant soft thresholding does not retain a superior performance across the range of SNR levels. This phenomenon can be explained by a very small thresholding parameter space for which TI-S thresholding achieves a near optimal performance. Figure 4.4a) clearly illustrates that TI-S is very sensitive to decomposition levels and performs considerably better at level 8 than for any other level. On the other hand, TI-H thresholding is almost insensitive to decomposition levels and performs at near optimal performance for a wide range of thresholding parameters. It provides almost a constant SNR enhancement of around 10.5 dB across the considered noise range. Furthermore, in relation to most other methods, its performance improves with lower and more realistic noise levels (SNR>6 dB). As such, the translation invariant hard thresholding, based on a wavelet basis function Coif1 and decomposition level 7, constitutes the optimal thresholding strategy for voice source signals.

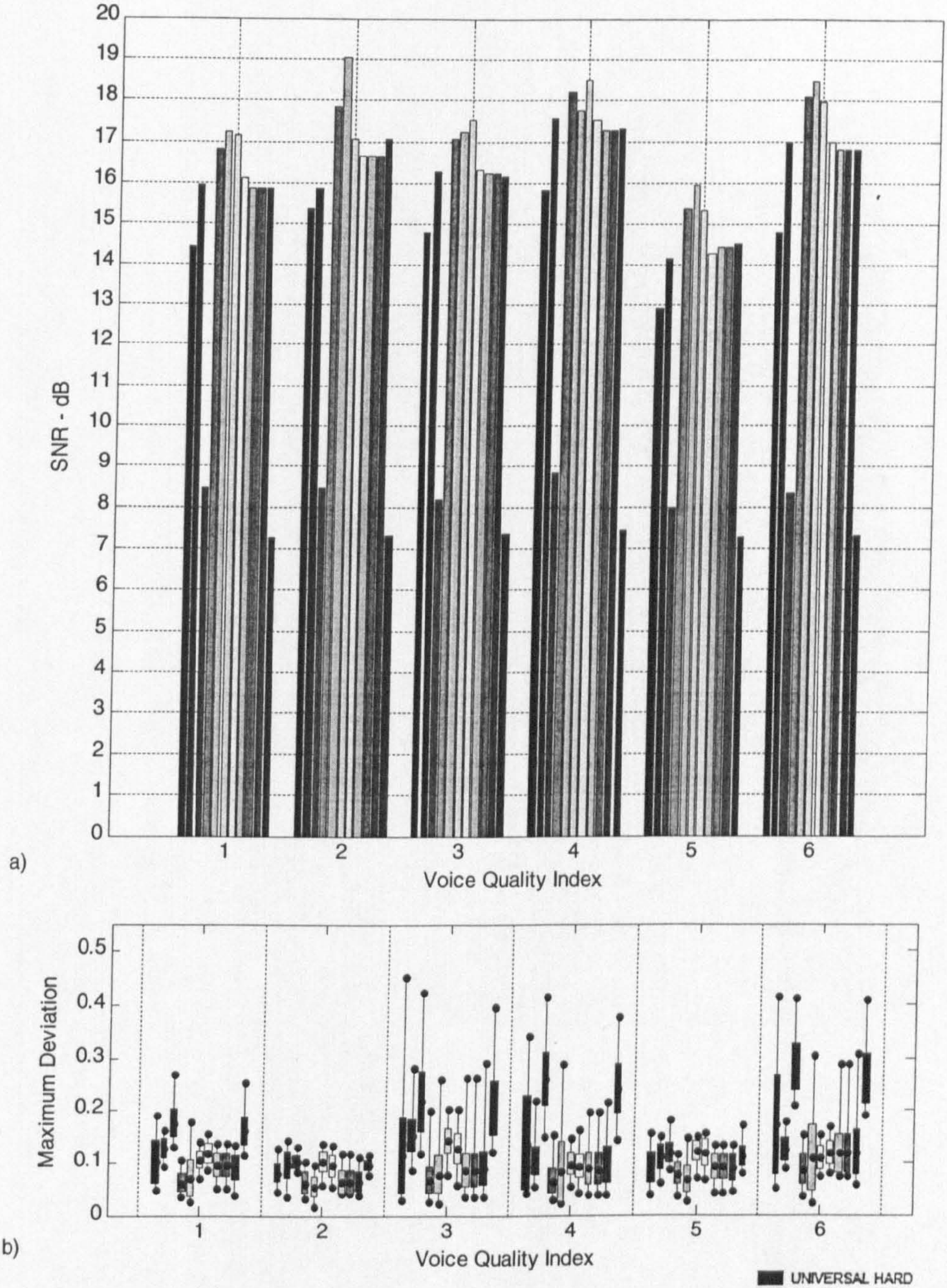


Figure 4.6: Thresholding performances as a function of voice quality. Each denoising method is optimized to yield the highest average SNR across the test functions. Input SNR = 6 dB. The performances are evaluated in terms of a) SNR b) Maximum Deviation

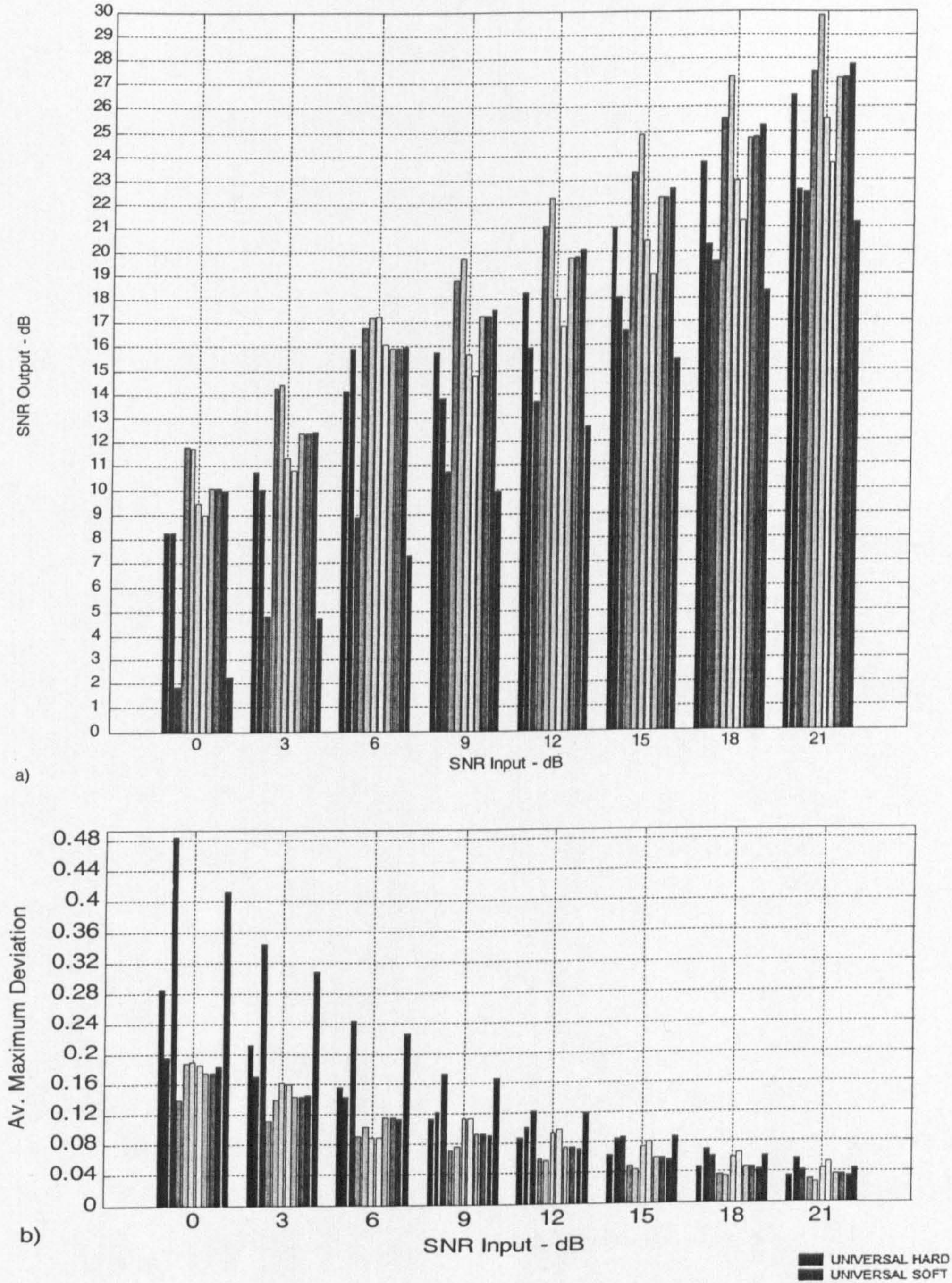


Figure 4.7: Average performances of optimized denoising methods across the voice quality types. The performances are evaluated in terms of a) SNR and b) Average Maximum Deviation

legend

Performance evaluation on natural acoustic data: optimized TI-H thresholding

Since the development of an optimal denoising strategy is based on a database of 6 different voice quality types, we would expect the optimized TI-H thresholding to adequately cope with a range of voice quality realizations. However, the simulated voice source signals do not contain "fine" structural elements* that are found in the real acoustic data. In this section, the performance of translation invariant hard thresholding, based on the 6-coefficient Coiflet filter and a primary resolution level-7, is evaluated on natural acoustic data. The test data corresponds to a read speech sentence: *"She had your dark suit in greasy wash water all year"*, sampled at $F_s= 10$ *kHz*. The waveform belongs to a male speaker with *modal* phonation - *WSJCAM0* database. An estimate of voice source signal is obtained via closed-phase, pitch-synchronous inverse filtering of the speech waveform. In the process of blind deconvolution, the vocal tract frequency response is modeled with 14 coefficients obtained through a covariance based linear prediction analysis. The glottal closure instant estimation is based on the *energy weighted* group delay algorithm and the wavelet-denoised LPC residue. The GCI estimation algorithm is provided in Chapter 3. Subsequently, translation-invariant hard-thresholding is applied on the voice source estimate to remove the turbulent components from the underlying signal. The results corresponding to a vowel /a/ in the word "*all*" are reported in Figure 4.8. Panel *a)* shows the voice source estimate waveform. The corresponding denoised waveform is displayed in panel *b)*. The extracted noise is shown in panel *c)*.

This example is a particularly aspirated voice source signal which was chosen deliberately in order to allow a better qualitative evaluation of TI-H performance. Apart from the "coarse" glottal flow derivative structure [114] and aspiration noise, the voice source estimate contains elements of the first formant ripple. The ripple describes a sinusoidal-like perturbation in the glottal derivative waveform due to the time-varying non-linear coupling of the glottal flow and the vocal tract, [21]. Overall, the temporal structure of the voice source estimate complies with Ananthapadmanabha and Fant's [4] predictions and thus, it constitutes a valid dataset for our study. The results of voice source thresholding are rather pleasing.

* A discussion on the temporal glottal excitation structure is presented in Chapter 3, Subsection 3.3.3;

Aspiration noise is almost completely absent in the reconstructed signal. On the other hand, the rapidly varying components in the glottal flow derivative waveforms are clearly preserved. We can visually confirm that the glottal excitation peaks corresponding to the glottal closure instants are denoised without any observably traces of over-smoothing. Furthermore, the elements of formant modulation ripple are also suppressed as evidenced by the prominent and well defined closed-phase regions in the reconstructed signal. Nevertheless, a certain amount of distortion, in the form of isolated spikes can be noticed around the closed-phase intervals of glottal cycle, and in particular around the closed-phase onset. These distortions are an amalgam of the residual pseudo-Gibbs artifacts and the high amplitude components of aspiration noise, which were not suppressed as their amplitude exceeded the estimated threshold value. We deem that the presence of these distortions would not be detrimental to process of voice source parameterization as they are temporally localized and small in amplitude. In Appendix B, Figures B.1- B.6, we have presented the thresholding results corresponding to the synthetic voice source signals. Clearly, the thresholding performance does not differ between the synthetic and natural acoustic data. The figures in the Appendix B are presented to provide an additional support for the arguments made earlier in this section.

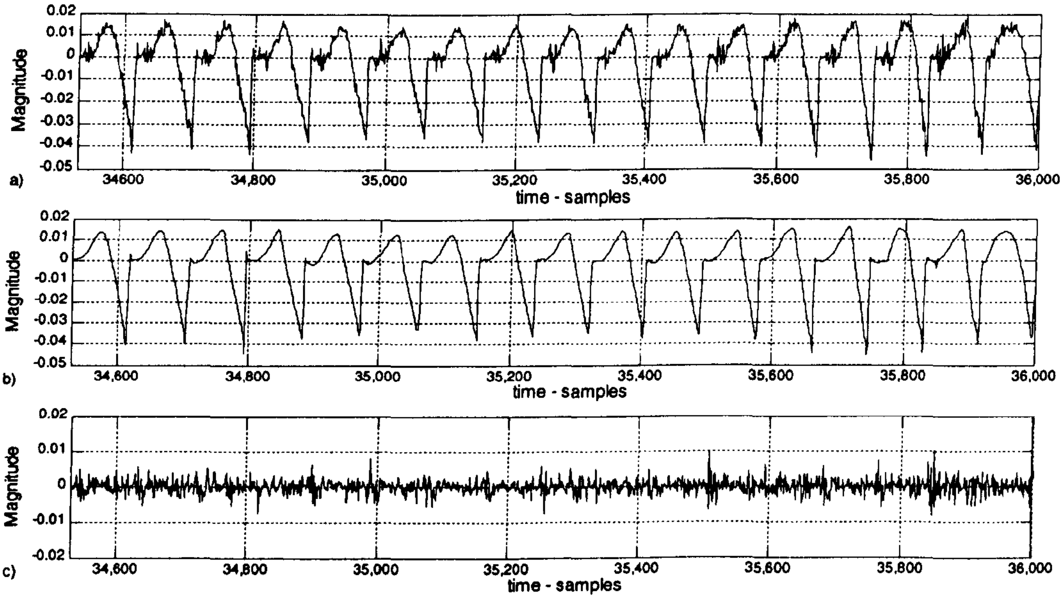


Figure 4.8: a) Voice source estimate; b) Denoised voice source estimate; c) Extracted noise; Waveforms correspond to the utterance /a/ from "She had your dark suit in greasy wash water all year". Male speaker; F_s= 10 kHz;

4.5 Conclusion

In this chapter we have aimed to develop the optimal thresholding strategy for glottal flow derivative signal. The following methods have been considered: Universal thresholding, SureShrink thresholding, Hybrid-Sure thresholding, Translation-Invariant thresholding, Hypothesis-Testing-based thresholding, Block thresholding, and Bayesian Adaptive Multi-resolution Smoother. We have systematically investigated the thresholding performance as a function of two thresholding parameters - the choice of wavelet basis function and the coarsest level of the wavelet decomposition. The main problem that we have encountered is the fact that the relationship between the thresholding parameters and the thresholding performance is highly non-linear. Furthermore, this relationship differs from one thresholding method to another. However, some rather crude trends were made apparent. Short wavelet filters, i.e. wavelets with a small number of vanishing moments, tend to have inadequate approximation properties, and as such, the quality of the reconstructed signal is often poor. On the other hand, more regular wavelets, corresponding to higher filter orders, have better decorrelating properties at the expense of temporal compactness. In the vast majority of the considered thresholding methods, a reasonable compromise between these effects is found to exist for some moderate filter length. A choice of decomposition level is also found to have a strong effect on the quality of the reconstructed signal. In most cases, if the decomposition level is too deep, the reconstructed signal is found to contain distortions around the instant of glottal closure (over-smoothing of glottal peak).

Ultimately, the optimal denoising strategy was associated with the translation invariant hard thresholding based on the decomposition level-7 and Coifl wavelet basis function. It is important to note that the optimal decomposition level is dependent on the sampling rate. In our case, the analysis is performed at 10 KHz. For higher sampling rates, the decomposition level should be increased and vice-versa. The translation invariant thresholding performs better than other considered thresholding methods as it is able to minimize the thresholding artifacts associated with misalignments between the sharp changes in the signal and the features of the wavelet. When the TI-H thresholding is applied on a voice source estimate obtained from the natural speech via inverse filtering, the results were very pleasing. Turbulent components in the voice source estimate were almost completely removed. On the other hand, the rapidly varying components in the glottal flow derivative waveforms were clearly preserved. We could visually confirm that the glottal excitation peaks corresponding to the glottal closure instants were denoised without any observably traces of over-smoothing. However, a certain amount of distortion, in form of isolated spikes, was noticed around the closed-phase intervals of glottal cycle, and in particular around the closed-phase onset. In large, these distortions correspond to the high amplitude components of the aspiration noise that were not suppressed as their amplitude exceeded the estimated threshold value. However, the distortions around the closed-phase onset might correspond to the pseudo-Gibbs phenomena induced by the thresholding process. However, we deem that these distortions will not be detrimental to voice source parameterization as they are temporally localized and small in amplitude.

Chapter 5

Voice Source Parameterization with application to Voice Quality Profiling

ABSTRACT

In this chapter, we propose Characteristic Glottal Pulse Waveform Parameterization and Modeling (CGPWPM), which is a novel fully automatic source-filter based framework for voice source analysis, parameterization and reconstruction. The proposed method is not constrained to the idealized glottal waveform approximations (e.g. Liljencrants-Fant's model), but instead relies on the estimates of Characteristic Glottal Pulse Waveforms to parametrically describe voice source dynamics and perform adaptive voice source reconstruction. CGPWPM enables representation of both, the "coarse" and the "fine" voice source structures and correspondingly, it facilitates the analysis of those voice source features that can not be accurately or efficiently represented by a deterministic glottal model. Furthermore, unlike any other method that we are aware of, it enables accurate estimation of statistical properties of turbulent components related to aspiration noise. In the design of CGPWPM, specific measures are taken to prevent pathological voice source parameterization and to mitigate the effects of inaccurate glottal closure instant estimation. In this chapter we present the voice source parameterization aspect of the proposed method and its application to voice quality profiling, while the voice source reconstruction aspect and its applications are discussed in the following chapter. Voice source parameterization is evaluated on both synthetic and natural acoustic signals and the results demonstrate the accuracy and robustness of the proposed method. The voice source parameters obtained on natural acoustic data attain realistic values and are generally free from outliers. We have used our voice quality profiling results to derive a surprising simple relationship between the glottal shape parameter R_d and voice quality. The relationship depicts a harsh and falsetto voices as the two extremes of voice quality spectrum. Creaky and Breathy voices exhibit somewhat more moderate deviation from modal voice while lax and tense voices attained a glottal pulse shape values that were the closest to modal voce.

5.1 Introduction

A robust modeling and parameterization of voice source signal estimates has applications in all areas of speech processing, but in particular for speech synthesis, speaker verification, speaker identification, voice morphing, voice quality analysis and clinical research. In the early days of speech synthesis, voice source signals were modeled as a series of impulses spaced at the fundamental frequency of vocal fold vibrations. It was soon realized that more accurate modeling of glottal waveforms is required to synthesize a natural sounding human voice. Although much research work has been done since, the modeling and the analysis of glottal pulse waveforms remains as one of the most important, difficult and relatively under-explored aspects of speech processing. In this introductory section, we will briefly describe some of the most important issues related to voice source parameterization and modeling.

Voice source parameterization can be approached via analysis of temporal [128], [52], [123] or spectral features [32], [54], [3], [73] in the estimated voice source signal. There have also been some attempts to develop a hybrid approach [82], [17], [18], [45]. However, it is undeniable that the temporal-analysis-based voice source parameterization methods have emerged as more successful. These methods can be classified into two broad groups, *direct estimation* and *fit estimation* methods. *Direct estimation* methods [1], [2], [88], [57] attempt to estimate the characteristic voice source features via simple programming procedures, such as: minima, maxima and zero crossings, whereas *fit estimation* methods [5], [77], [116], [105], [81] employ fitting of a mathematically defined glottal waveform to the observed data. In each case, the choice of a glottal pulse model is the single most important performance factor. When considering this issue, it is also important to appreciate that humans produce a very extensive range of vocal fold realizations [133], [149], and that any approach to voice source estimation inherently carries a variety degradations with varying spectral and temporal properties [119], [4], [21]. The most severe forms of voice source signal degradations are induced by the imperfect deconvolution of the voice source signal and the vocal tract filter from speech. The artifacts of source-vocal tract interaction, such as the formant-like ripple, and the induced glottal pulse skewness to the right due to the inertive loading by the subglottal and supraglottal acoustic systems can make a voice source signal particularly difficult to parameterize. Current glottal pulse models, e.g. [22], [105], [141], [65], [86], [53], including the popular Liljencrants-Fant's (LF) model [43] have adopted overly simplistic interpretations of voice source dynamics, and yet, they do not offer the desired levels of modeling accuracy and parameterization robustness. In Chapter 3, we have demonstrated the extent of diversity in the temporal structure of the glottal flow derivative estimates and have shown that LF model can adequately describe only the general shape of the glottal flow derivative waveforms, leaving the fine structural elements [114] of glottal pulse realizations unrepresented. In instances when the glottal model waveforms substantially differ from the actual voice source signal, the concept of parameterization accuracy looses its meaning altogether.

There is a need for a more sophisticated voice source model that can represent a wider scope of vocal fold realizations, preserve the "fine" structural elements in glottal pulse waveforms and ultimately, enable a high quality voice source analysis and source-filter-based speech synthesis. With these motivations, we have developed Characteristic Glottal Pulse Waveform Parameterization and Modeling (CGPWPM) as a more robust and more accurate alternative to Liljencrants-Fant's model. In fact, CGPWPM offers an entirely novel framework for voice source analysis, parameterization and reconstruction. The proposed method is not constrained to the idealized glottal waveform approximations, but instead relies on the estimates of Characteristic Glottal Pulse Waveform to parametrically describe glottal flow dynamics and perform adaptive voice source reconstruction. The novelty of this approach requires an extensive elaboration on a range of issues. As such, the problems of voice source parameterization and voice source reconstruction are treated separately. In this chapter we will focus on voice source parameterization, whereas in Chapter 6, we will describe how the principles behind the Characteristic Glottal Pulse Waveform Parameterization can be extended to voice source reconstruction, speech synthesis and voice quality conversion.

Section 5.2 describes the CGPWPM approach to voice source parameterization. In Section 5.3, the CGPWPM performance is evaluated on the synthetic and natural speech datasets. In Section 5.4, we aim to develop a parametric voice quality description for a range of voice quality types. Section 5.5 concludes the chapter.

5.2 Characteristic Glottal Pulse Waveform Parameterization

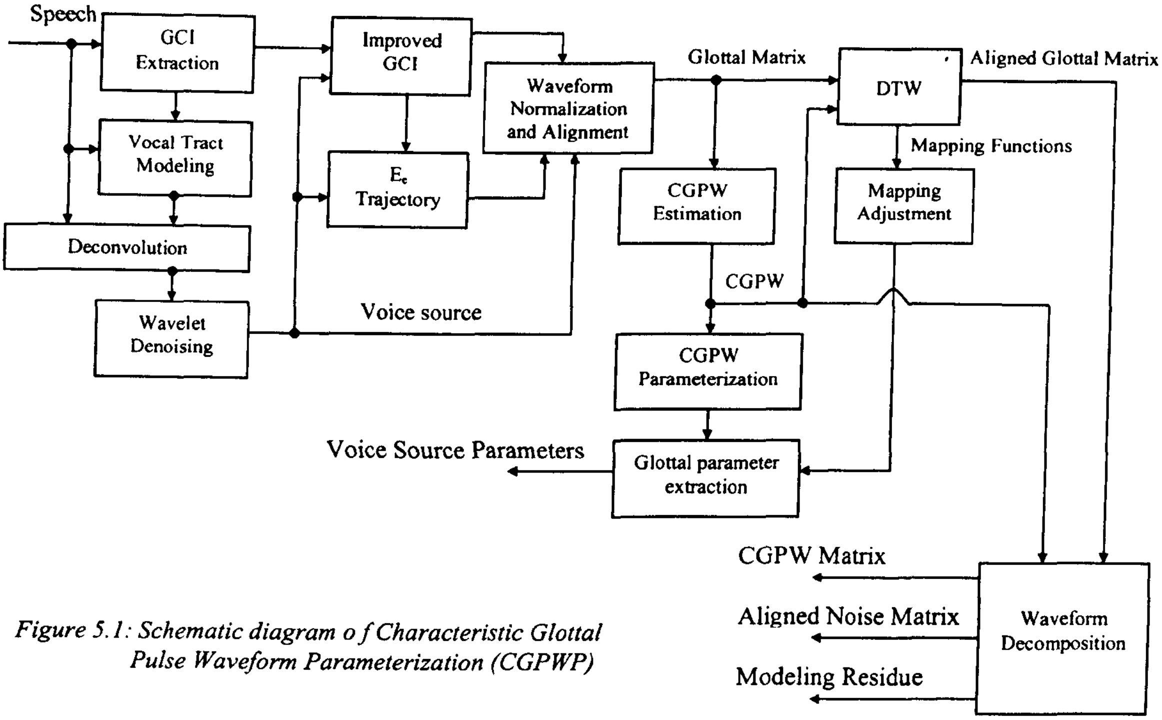


Figure 5.1: Schematic diagram o f Characteristic Glottal Pulse Waveform Parameterization (CGPWP)

The proposed voice source parameterization procedure is outlined in Figure 5.1. The figure shows the voice source deconvolution components that are described in Chapter 3. We will remind that the voice source estimates are obtained via closed-phase pitch-synchronous inverse filtering of speech signals. In the process of blind deconvolution, the vocal tract frequency response is modeled with 14 coefficients obtained through a covariance based linear prediction analysis. The glottal closure instant estimation is based on the *energy weighted* group delay algorithm and wavelet-denoising of the LPC residue. The improved estimates of glottal closure instants and the glottal excitation strength (E_e) contour are obtained using the standard peak-picking procedure. The glottal flow derivative estimates are denoised via translation invariant hard thresholding based on the 6-coefficient Coiflet filter and a decomposition level-7. In Chapter 4, the above mentioned method is found to be the optimal wavelet thresholding strategy for the glottal flow derivative signals. Before we describe the remaining components of the Characteristic Glottal Pulse Waveform Parameterization (CGPWP), we will outline the main principles behind the parameterization process.

The estimates of glottal flow derivative waveforms are normalized in both, time and amplitude domains, and subsequently, sequentially arranged to form a matrix of glottal pulses that we refer to as *glottal matrix*. In the next stage, Characteristic Glottal Pulse Waveform (CGPW) is estimated from the *glottal matrix* via a modified Euclidian distance measure. CGPW denotes a single glottal flow derivative pulse that "best" describes the entire database of *glottal matrix* pulses and correspondingly represents a typical oscillatory cycle in the voice source estimate. In that respect, each *glottal matrix* pulse is treated as a potential candidate for Characteristic Glottal Pulse Waveform. Under the proposed framework, CGPW is used as a glottal flow derivative model, rather than an opportune combination of mathematical functions. Thus, the voice source model is of adaptive nature and its form and structure is directly dependant on the observed voice source signal. Accordingly, voice source parameterization objective is redefined as obtaining a parametric description for the non-linear temporal evolution of CGPW through a sequence of *glottal matrix* pulses. The first step in the parameterization process is to obtain a parametric description for the Characteristic Glottal Pulse Waveform, itself. For this purpose we employ a simple, but a robust *direct estimation* method [2] to produce a set of modified LF parameters, which we denote as *glottal* m*atrix* (GM) parameters. In the same way as LF parameters, the GM parameters signify the important events in the temporal structure of the glottal flow derivative model and thus, they can be readily used to derive the established LF-based voice quality quantifiers (e.g. speed quotient, open quotient…). In the second stage, the temporal relations between CGPW and other *glottal matrix* pulses are established via Dynamic Time Warping (DTW) algorithm. The output of DTW block or specifically, a surface of optimal non-linear alignment functions is utilized to extend the GM parameters from CGPW to other *glottal matrix* pulses. As such, the parameterization results for an entire glottal matrix are referenced to the parametric description of a single glottal pulse waveform. This property enables CGPWPM to be very effectively employed in a semi-automatic manner, e.g. the parametric description of CGPW can be manually altered, and GM parameters are automatically extended to the entire voice source signal. The waveform decomposition block is used to quantify the quality of voice source parameterization. More importantly, this block provides the means to estimate the statistical properties of non-stationary turbulent components related to aspiration noise. As we will demonstrate in the following chapter, aspiration noise can also be successfully synthesized and integrated with the glottal flow derivative to realize a faithful voice source reconstruction.

Waveform Normalization and Alignment, and glottal matrix estimation

Waveform normalization and alignment provides a platform for voice source parameterization independent of glottal excitation strength and pitch frequencies. The amplitude normalization is performed in the following manner. The glottal excitation strength estimates (E_e) are interpolated, via monotone piecewise cubic interpolation [51] over the entire duration of voiced source signal to obtain a E_e envelope. E_e envelope is subsequently used to scale the voice source magnitude so that the instants of main voice source excitations, i.e. GCIs, have the identical amplitudes, $\widetilde{E}_e = -1$.

The temporal normalization is achieved by re-sampling the individual glottal pulse to a normalized pitch period length, T_N. The re-sampling factors are obtained as ratios of the pitch period estimates and T_N. . The pitch trajectory is estimated directly from the GCI estimates. In the process of constructing a *glottal matrix,* each glottal pulse is extracted from the voice source signal with a rectangular window centered in-between the successive glottal closure instants. The boundaries of the window are extended beyond the glottal closure instants, such that the *glottal matrix* frame length is 20 % longer than the length of the normalized pitch period length. This is done in order to enable CGPWP to cope with potential GCI errors. This issue is further explored in the later sections of this chapter. When the observed glottal flow derivative pulses are normalized and sequentially aligned to form a *glottal matrix,* $\mathbf{G}$, then, the i^{th} normalized glottal pulse, $\hat{\mathbf{e}}_i$ is defined as:

$$\hat{\mathbf{e}}_i(n) = \mathbf{G}(i, n+\delta+1), \qquad 0 \le n < T_N \tag{5.1}$$

The δ value is related to the normalized pitch period length T_N and the *glottal matrix* frame length L as: $\delta = (L - T_N)/2$. The *glottal matrix* is an N-by-L matrix; where N and L denote the number of *glottal matrix* pulses and the glottal matrix frame length, respectively. Throughout this paper CGPWPM uses L=120, T_N =100, and δ =10. The normalized pitch period length value, T_N=100, represents a compromise between the desired temporal resolution of glottal flow derivative waveforms and the computational efficiency requirements (limited by DTW).

Examples of *glottal matrixes* are provided in *a)* panels of the following figures: Figure 5.14, Figure 5.19, Figure 5.24, Figure 5.29, Figure 5.34, and Figure 5.39. The graphs correspond to the voiced segment of a read speech sentence, "Don't ask me to carry an oily rag like that", from each of 3 male and 3 female speakers from the *WSJCAM*0 database. Note that the x-axis of these graphs (labeled as *glottal cycle index)* indicates the sequential order of the individual glottal pulses as they appear in the voice source signal.

Characteristic Glottal Pulse Waveform estimation

Characteristic Glottal Pulse Waveform (CGPW) estimation aims to select the most typical glottal flow derivative realization from a database of normalized glottal pulse waveforms. The Characteristic Glottal Pulse Waveform is estimated as a glottal pulse waveform that attains the minimal cumulative Euclidian distance across the candidate waveforms. In that respect, every single glottal pulse in the *glottal matrix* is treated as a Characteristic Glottal Pulse Waveform candidate. The estimation is conducted using the following algorithm:

$$CGPW = \hat{e}_i = \min_{\arg i} \sum_{k=1}^{N} (\hat{e}_i - \hat{e}_k)^T (\hat{e}_i - \hat{e}_k) \ , \ 1 \le i \le N \qquad (5.2)$$

where $\hat{e}_i$ is a column vector representing the i^{th} glottal pulse waveform in the *glottal matrix*. Since the glottal pulse waveforms are normalized in both, time and amplitude domains, the selection is based, exclusively, on the waveform shape. Thus, provided that the normalized glottal pulse waveforms are correctly aligned in a *glottal matrix*, CGPW estimation will select a glottal pulse waveform with the temporal structure that is most frequently produced by a speaker. CGPW estimation involves the glottal matrix pulse regions in-between the two glottal closure instants to ensure that only exact glottal cycle periods are used. The remaining 20% of glottal matrix frame length is neglected to avoid introducing estimation bias. It is also important to note that apart from the general glottal pulse shape, the fine glottal flow derivative structure can also have a significant influence on the Characteristic Glottal Pulse Waveform estimation process.

Dynamic Time Warping - DTW

CGPWPM requires an automatic temporal alignment between the Characteristic Glottal Pulse Waveform and other *glottal matrix* waveforms. For this purpose we use DTW, as it provides the optimal solution for the time series alignment problems [115]. In DTW framework, CGPW and an analyzed glottal pulse are viewed as a reference signal R and a test signal T, respectively.

$$T = t_1, t_2, t_3, \ldots, t_i \ldots t_n \tag{5.3}$$

$$R = r_1, r_2, r_3, \ldots r_j \ldots r_m \tag{5.4}$$

In the first stage of DTW a, an n-by-m matrix is constructed, where the (i^{th}, j^{th}) element of the matrix contains the distance $d(t_i, r_j) = (t_i - r_j)^2$ between the points t_i and r_j. The matrix element (i, j) represents the alignment between the elements t_i and r_j. An alignment path, W

$$W = w_1, w_2, w_3, \ldots w_K \tag{5.5}$$

is a contiguous set of matrix elements that describes the mapping between T and R. The k^{th} element in the alignment path corresponds to $w_k = (i,j)_k$. The optimal alignment is found using the dynamic programming to evaluate the recurrence in (5.6), which defines the cumulative distance $D(i,j)$ as a summation of the distance in the currant cell $d(i,j)$ and the minimum cumulative distances of the adjacent cells. The alignment path is typically subjected to several constraints, namely, boundary constraint, monotonicity constraint, and continuity constraint.

$$D(i,j) \;=\; d(t_i, r_j) \;+\; \min\begin{Bmatrix} D(i-1, j-1) \\ D(i-1, j) \\ D(i, j-1) \end{Bmatrix} \tag{5.6}$$

Boundary constraint: forces the alignment path to begin and end in the in the diagonally opposite corner cells of the alignment matrix: w_1=(1,1) and w_k=(n, m).

Monotonicity constraint: requires the elements of the alignment path to be monotonically spaced in time. If w_k = (x, y) and the w_{k-1} = (x', y'), then the monotonicity constraint imposes x-x' ≥ 0 and y-y' ≥ 0.

Continuity constraint: restricts the allowable steps in the warping path to adjacent cells. This includes the diagonally adjacent cells. If w_k = (x, y) and the w_{k-1} = (x', y'), then the continuity constraint imposes x-x' ≤ 1 and y-y' ≤ 1.

In addition to the boundary, continuity and monotonicity constraints, the global constraints in form of Sakoe-Chiba Band [121] are imposed on warping path in order to prevent pathological alignments, as well as to increase the computational efficiency of DTW. Pathological warping arises when a relatively small section of one sequence maps onto a relatively large sequence of another. On the other hand, if the global constrains are too "severe" the alignment process could become meaningless. In Section 5.3.2, we have evaluated the optimal size for the DTW alignment window. An alignment function example between CGPW and a glottal pulse is shown in Figure 5.2. CGPW is obtained from the utterance "We were <u>a</u>way a year ago", sampled at 10 kHz. A displayed glottal pulse belongs to a vowel /a/. Voice source parameterization is performed for $L=120$ and $T_N=100$. The acoustic data corresponds to a female speaker and the *WSJCAM*0 database.

Mapping adjustment

Before voice source signal could be parameterized, the effect of glottal closure instant estimation errors on DTW alignment functions needs to be considered. The glottal closure instants play an import role in CGPWP parameterization process. They provide the means to estimate pitch trajectories, and in turn, to estimate the re-sampling factor by which glottal waveforms are temporally normalized. Furthermore, in the process of *glottal matrix* synthesis, they enable sequential glottal pulse alignments. As such, GCI estimation errors can have a significant impact on the voice source parameterization performance.

In this section, we will describe a measure that attempts to reduce the effect of GCI errors. Let us assume that CGPW is bounded by accurate GCIs. Since CGPW estimation algorithm, (5.2), is highly sensitive to inaccurate glottal pulse alignment and normalization, this assumption will hold true even for the moderately-performing GCI estimation algorithms. Let us also adopt an analogue representation of alignment functions, such that the alignment path between the i^{th} glottal matrix pulse and CGPW can be described as $W_i(r)$, where $0 \le r < T_N$. Thus, in the instance when the two GCIs bounding a *glottal matrix* pulse are correctly estimated, the alignment path between the glottal pulse and CGPW will contain the following two points (GCI_1, GCI_1) and (GCI_2, GCI_2). Conversely, any discrepancy between GCI_1 and $W_i(GCI_1)$ or GCI_2 and $W_i(GCI_2)$ is an indication of GCI error. In such instance, the alignment function is adjusted by the following algorithm:

$$\hat{W}_i = W_i + r - W_i(GCI_1) + \frac{(GCI_1 - r)(W_i(GCI_2) - W_i(GCI_1))}{GCI_2 - GCI_1} \qquad (5.7)$$

where $\hat{W}_i$ and W_i denote the adjusted and DTW obtained alignment functions, respectively.

Essentially, the adjusted alignment function describes the optimal alignment function that would exist between CGPW and a glottal pulse if the glottal pulse was accurately normalized and aligned. The discrepancies between GCI_1 and $W_i(GCI_1)$, and GCI_2 and $W_i(GCI_2)$ are multiplied by the re-sampling factor that is used in temporal glottal pulse normalization to obtain the actual GCI error values and the improved pitch period estimates. In Figure 5.3, we have illustrated the mapping adjustment process. In this example, the relative error in the GCI estimates has caused the pitch period to be overestimated, the *glottal matrix* pulse is represented with fewer than T_N=100 samples, the alignment function is shifted in r domain and its average slope, in the regions between GCI_1 and GCI_2, has decreased bellow 1. The figure shows that the adjusted alignment function contains points: (GCI$_1$, GCI$_1$) and (GCI$_2$, GCI$_2$).

Voice source parameterization

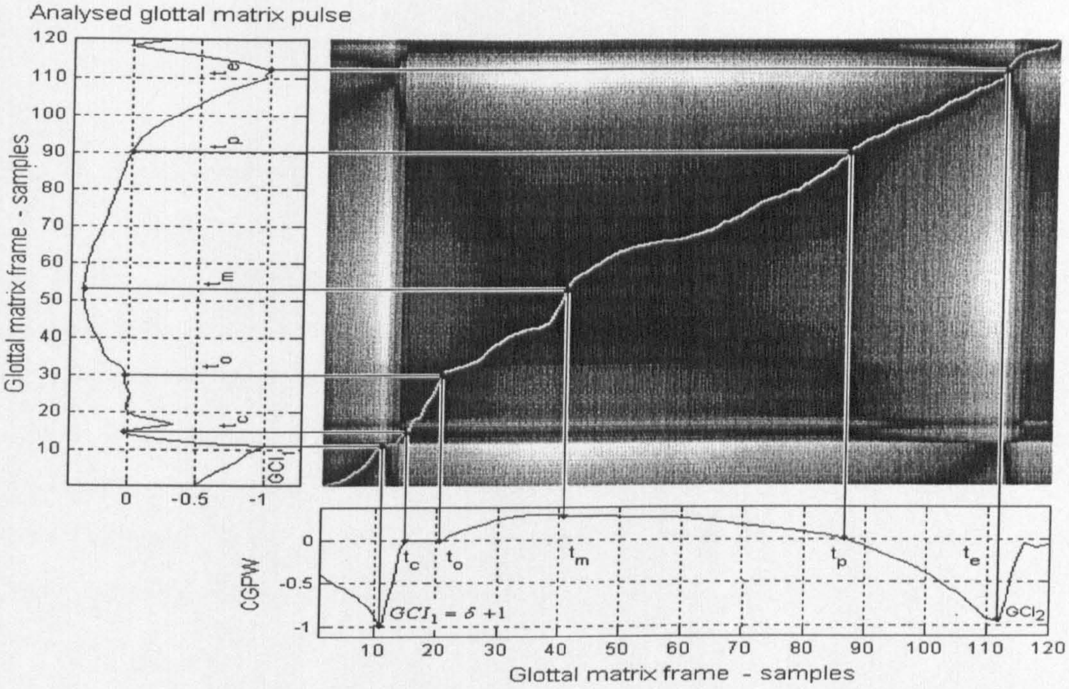


Figure 5.2: Optimal alignment function is used to map GM -parameters from CGPW to a glotta matrix pulse. Darker colors indicate a lower cost in the DTW *cost matrix. Since in thi case, GCI estimation is accurate, the adjusted alignment function is identical to th optimal* DTW *alignment path.*

In the parameterization of Characteristic Glottal Pulse Waveform we have opted for a computationally efficient and a highly robust *direct estimation* method[*] described in [2]. In addition to the timing parameters of the Liljencrants-Fant glottal flow derivative model, the maximum glottal flow derivative instant, t_m (see Figure 2.7) is include in the GM parameter set to ensure a sufficient temporal resolution of CGPW for adequate CGPWPM-based voice source reconstruction; This issue is further clarified in Chapter 6.

Since the temporal relationships between CGPW and the other glottal pulses in the *glottal matrix* are described by a surface of DTW functions, the entire *glottal matrix* can be

[*] Since CGPW is just a single glottal pulse waveform, we can afford to manually verify and alter the parameterization results. However, in our experience, manual corrections are hardly ever required.

parameterized by the following algorithm:

$$P_i = \hat{W}_i \; (P_{CGPW}) \qquad P_{\mathrm{CGPW}} = \{\delta+1,\, t_c,\, t_o,\, t_m,\, t_p,\, t_e\} \tag{5.8}$$

where $\hat{W}_i$ refers to the adjusted alignment function between CGPW and the i^{th} *glottal matrix* pulse. P_{CGPW} and P_i denote the GM-parametric descriptions of CGPW and the i^{th} glottal matrix pulse, respectively. Since the alignment path is in reality a discrete-valued function, P_i values are obtained via interpolation. For this purpose, monotone piecewise cubic interpolation is employed as this method retains the shape and monotonicity of the DTW alignment functions. Note that in the parameterization process, the adjusted alignment functions are used, rather than the DTW alignment functions, as the adjusted functions are more accurate with the respect to GCI estimates. The algorithm (5.8) enables the GM parameters to be extended from the CGPW to other *glottal matrix* pulses using the surface of adjusted alignment functions, only.

Figure 5.2 illustrates how the adjusted alignment functions are used to map the GM parameters from CGPW to other *glottal matrix* pulses. In Figure 5.4 we have shown a sequence of adjusted alignment functions and GM-trajectories for the first 80 glottal pulses in the utterance "*We were away a year ago*", articulated by a female speaker. Alignment functions depict a nonlinear evolution of CGPW through a sequence of *glottal matrix* pulses, whereas, GM-trajectories illustrate the dynamics of important features in the glottal flow derivative waveform. The GM-trajectories also coincide with the characteristic alignment function features. This property will be later exploited in voice source reconstruction. Due to the nature of the proposed method, the two glottal closure instants inside each *glottal matrix* frame are fixed at $GCI_1=\delta+1$ and $GCI_2=L-\delta+1$. Thus, the non-linear CGPW evolution is described by only four variable parameters $\{t_c, t_o, t_m, t_p\}$. Note that the return coefficient, T_a is not included in the GM parameter set as it is a measure of effective closed-phase abruptness rather then a specific event in the glottal derivative waveform. Nevertheless, we acknowledge the fact that the return coefficient is an important voice quality correlate and thus, we employ the exponential fit procedure on *glottal matrix* pulses, in the regions between $\delta+1$and t_c, to obtain T_a estimates.

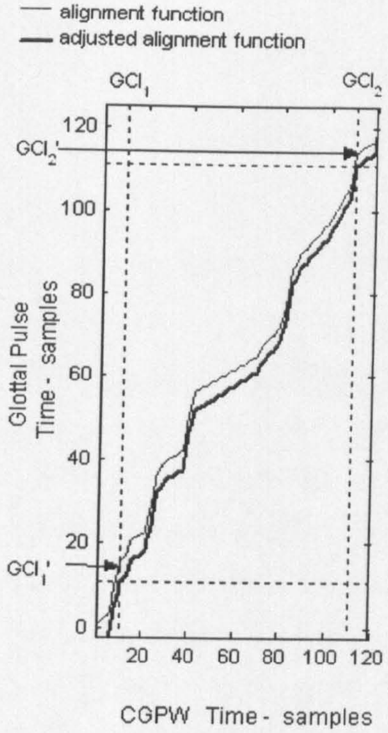


Figure 5.3: Alignment function is adjusted to remove the effect GCI errors on a glottal pulse parameterization. The data corresponds to the first frame of the alignment surface shown in Figure 5.4

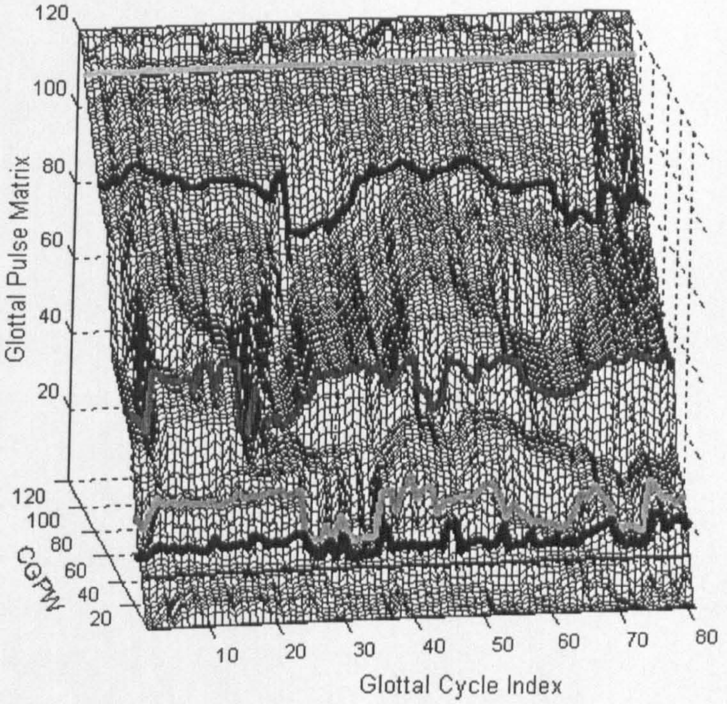


Figure 5.4: *A segment of GM trajectories (in increasing order: $\delta+1$, t_c, t_o, t_m, t_p, t_e) and the corresponding adjusted alignment functions obtained on the first 80 glottal pulses in the utterance "We were away a year ago". Female speaker - WSJCAM*0 *database*

Waveform Decomposition

Waveform decomposition provides the means to estimate the aspiration noise envelope, which is the principal statistical description of the non-stationary turbulent components related to aspiration noise [23]. The procedure requires the wavelet denoising block to be omitted from CGPWP parameterization system. The first stage in this process is to produce *aligned glottal matrix* by aligning the *glottal matrix* towards CGPW via alignment functions provided by the DTW block. Some examples of the *aligned glottal matrixes* are provided in the *b)* panels of the following figures: Figure 5.14, Figure 5.19, Figure 5.24, Figure 5.29, Figure 5.34, and Figure 5.39. Visual inspection of these graphs, confirms the assumption that CGPW can be used to adequately represent other glottal pulse waveforms via the constrained non-linear optimal time warping functions.

In the second stage, the estimated CGPW is subtracted from each *aligned matrix* frame to produce *residue matrix*, $\mathbf{R}$.

$$\mathbf{R}(i,n) = \mathbf{G}'(i,n) - \mathbf{M}_{CGPW}(i,n) \tag{5.9}$$

where $\mathbf{G}'$ denotes the *aligned glottal matrix* and $\mathbf{M}_{CGPW}$ represents an N-by-L matrix containing N repetitions of the Characteristic Glottal Pulse Waveform. The *residue matrix* is composed of aspiration noise and the slowly evolving *glottal matrix* features that can not be represented via CGPW and the monotonically increasing DTW functions, alone. In a similar manner as in Enhanced Waveform Interpolative Coding [†] [60], i.e. via simple high-pass/low-pass filtering, the *residue matrix* is then decomposed into *modeling residue* and *aspiration noise* matrices. *Modeling residue matrix* represents the true modeling error, whereas *aspiration noise matrix* represents the turbulent components related to aspiration noise and to a much lesser degree, processing noise. Aspiration noise envelope is related to a waveform of RMS values obtained across the glottal cycle index of *aspiration noise* matrix. Some examples of *aspiration noise matrixes* and *modeling residue matrixes* are provided in *d)* and *e)* panels, respectively, of the following figures: Figure 5.14, Figure 5.19, Figure 5.24, Figure 5.29, Figure 5.34, and Figure 5.39. Examples of aspiration noise envelope are provided in the following figures Figure 5.16, Figure 5.21, Figure 5.26, Figure 5.31, Figure 5.36, and Figure 5.41. Note that only the sections in-between the two GCIs are shown. As such, the presented aspiration noise envelopes correspond to exactly one normalized pitch cycle, T_N=*100* samples. The aspiration noise envelopes exhibit strong peaks in the regions of glottal closure and have relatively low energy levels in the remaining segments of the glottal pulse cycle. In case of *M3* and *F2* speakers, the aspiration noise shows high intensity levels around the glottal opening instant. This phenomenon is also in line with the naturally occurring aspiration noise behavior. In Chapter 6, described aspiration noise envelope estimation procedure is evaluated via speech synthesis experiments.

[†] The rapidly evolving component (REW) of the EWI coders can not be used to represent the aspiration noise for the following reasons. In EWI coders, the acoustic waveforms are not aligned in the manner that could account for the non-linear evolution of the glottal flow derivative waveforms. Secondly, the system is usually applied on the *LPC*-residue or the speech signal, rather than on the credible voice source estimates.

5.3 Performance evaluation

Evaluation of voice source parameterization techniques is not a trivial task and there is neither a generally accepted nor standardized way in assessing the parameterization performance. The validity of experiments conducted on natural speech is hindered by a simple fact that the correct voice source parameters are not known. Also, it is important to appreciate that the voice source parameterization is ultimately affected by the quality of voice source estimates. In practice, voice source signals often contain numerous disturbances and degradations caused by the imperfect deconvolution of a voice source signal and a vocal tract filter from speech. First formant ripple is often observed in the glottal flow derivative signal obtained via inverse filtering [21]. The glottal flow derivative waveforms may be skewed to the right due to the inertive loading by the subglottal and supraglottal acoustic systems [119]. A nonlinear increase in the voice source strength can occur when the frequency of the first vocal tract resonance is near an integral multiple of glottal cycle frequency [4]. The extent in which these degradations are present in the voice source signal, as well as their effect on the parameterization performance is not easy to quantify. Thus, the experiments on natural speech are usually constrained to sustained vowels or the short term segments of "well behaved' voice source signal [52], [77], [105], [54]. The most common form of parameterization performance evaluation is of qualitative nature, an example of which is the visual comparison between the glottal flow derivative signals and the waveforms generated by a glottal model [32], [116], [73], [43]. Usually, if not exclusively, the visual inspections are performed over a small number of glottal cycles. In addition, listening tests can be set up to asses the "naturalness" of synthesized speech, where the voice source signal is reconstructed from the estimated voice source parameters. However, neither of the two methods is effective enough in distinguishing the levels of voice source parameterization accuracy that are required in the fundamental research on speech production, clinical research and voice quality research.

Another approach in assessing the parameterization performance is based on the synthetic speech databases [28], [129]. The main advantage of this approach is that the voice source parameters are known in advance and thus, the voice source parameterization performance can be quantitively evaluated. However, the synthetic glottal pulses lack the structural complexities of the natural acoustic signals as well as a range of degradations that are normally imposed by the voice source estimation procedures. As such, the performance results that are obtained on these well behaved mathematically idealized voice source signals can not be taken as a faithful reflection of the parameterization performance on the natural acoustic data.

In order to provide a thorough assessment of the Characteristic Glottal Pulse Waveform Parameterization (CGPWP) performance, evaluation experiments are conducted on natural and synthetic speech datasets. At this stage, we would like to note that due to the nature in which CGPW parameterization is realized we are able to provide both, qualitative and quantitive evaluation, on natural datasets.

5.3.1 Performance evaluation - synthetic dataset

Strik *et al.* have conducted a set of experiments to study the effects of various types of signal distortions on the performance of two voice source parameterization techniques [130]. Our evaluation of the Characteristic Glottal Pulse Waveform Parameterization is roughly based on these experiments. However, we have introduced certain refinements in order to enable a more thorough and more objective performance assessment. In addition to the experiments designed to study the effect of non-integer glottal pulse parameter values and the effect of low pass filtering, we have also introduced an experiment that studies how the inaccuracies in the glottal closure instant estimates affect the parameterization performance.

The CGPWP performance will be compared to the voice source parameterization methods considered in [130]. These two methods are prominent examples of *fit estimation* (FE) and *direct estimation* (DE) approach to voice source parameterization. The *fit estimation* method is a three stage parameterization procedure involving the initial parameter estimation, simplex search algorithm, and Levenberg-Marquardt algorithm. The *direct estimation* approach is represented by Alku and Vilkman's parameterization technique [2]. The performances of these two methods will be used as a performance benchmark for the CGPWPM-based voice source parameterization.

For effective comparison, we will use the same performance criteria and the same experimental setup as those in [130]. The parameterization performance is evaluated with the respect to the individual voice source parameters, and the median absolute error is used as a performance measure. The test voice source signals are synthesized via Liljencrants-Fant glottal pulse model. Also, the glottal cycle duration is kept at a constant value of 10 ms and the acquisition (sampling) rate of 10 kHz is employed. Unlike in [130], we will also evaluate CGPWP performance with the respect to the instant of complete glottal closure t_c, as we deem that its value is important in estimation of certain voice quality correlates, e.g. *open quotient* and *speed quotient*. Note that Strik has justified the exclusion of this voice source parameter by citing its insignificant role in voice source synthesis[*‡]. Table 5.1 shows the parametric descriptions for 11 types of glottal pulses used in Strik's performance evaluation experiments. We think that the size of this dataset does not adequately reflect the full range of naturally occurring vocal fold realizations. It is important to note that the shape of the glottal pulse is a critical factor in any voice source parameterization related experiments and a study of signal distortion effects is not an exception. A constrained dataset such as this is likely to introduce a bias in the experimental results.

Table 5.1:
The parameter values of test glottal pulses expressed as a percentage of pitch period

Test pulse index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
T_p – [%]	40	40	60	60	60	60	40	40	52	52	52
T_e – [%]	52	52	72	72	88	88	60	60	72	72	72
T_c – [%]	100	100	100	100	100	100	100	100	100	100	100
T_a – [%]	4	16	4	16	4	8	4	16	4	10	16

[*‡] The end of closing phase of glottal cycle can be ascertained, with a certain degree of accuracy, from the t_e and t_a values, Childers and Ahn [23], Childers and Hu [22].

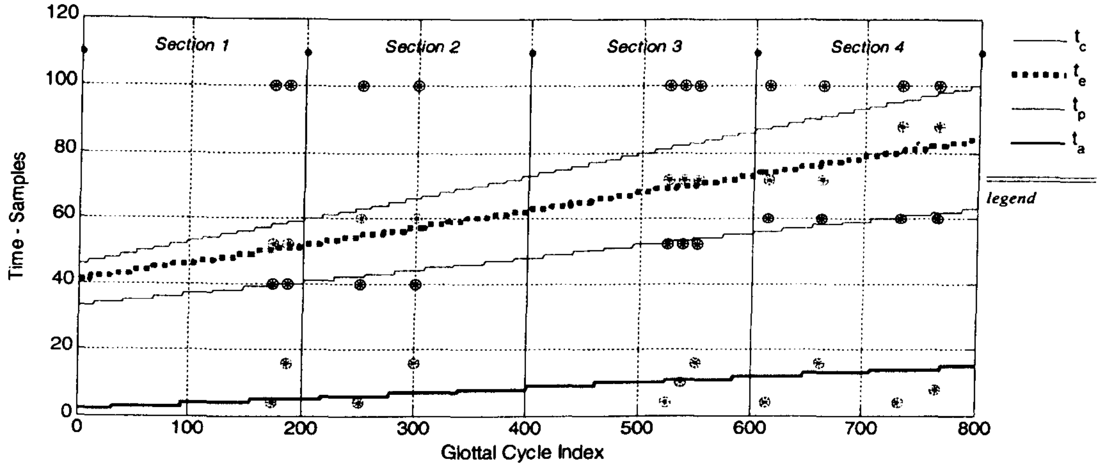


Figure 5.5: Eight hundred sets of voice source parameter values defining the glottal pulse waveforms used to synthesize four voice source signals (solid lines). The graph is partitioned into four sections each corresponding to a particular voice source signal. The parameters of eleven base pulse used in the Strik's experiments are also displayed (circles).

In our study, we have used a dataset of 800 glottal pulses. The parametric description of the dataset is presented in Figure 5.5. The x-axis of the graph corresponds to the index of individual glottal pulses while the y-axis shows the parameter values for each glottal pulse, expressed as a percentage of pitch period length. The parametric description for these pulses is obtained using the linear interpolation of the parameters denoting examples of *vocal fry* voice and *breathy* voice. In Figure 5.5, we have also displayed a dataset from Strik's experiments (as circles), and it is evident that this data exhibits a bias towards *breathy* voice.

The set of 800 parameters is divided into four equal size sections. Based on these sections, four separate voice signals are synthesized with continually varying glottal pulse waveforms. The dataset is partitioned in order to enable a reduced size of the DTW window, and correspondingly improve the computational cost. Note that the size of the required alignment window is proportional to the extent of glottal waveform variation. In natural speech, it is unlikely that a speaker will exploit a full voice quality range that is presented in Figure 5.5.

Prior to describing the evaluation experiments, we want to remind that in the process of CGPW parameterization, the parametric description of CGPW is automatically extended to all the other *glottal matrix* pulses via DTW algorithm. In order to demonstrate that the *glottal matrix* parameterization can be performed accurately, even in the instances of severe voice source signal degradation, in each of the following experiments, we will provide accurate voice source parameter values to describe the Characteristic Glottal Pulse Waveform, only.

Time Shift Experiment

Commonly, the values of the estimated voice source parameters are rounded off to a nearest sample value. However, in practice, it is very unlikely that the voice source parameters will coincide with the sample values. In this experiment, the voice source signals are shifted from 0.0 to 0.1 msec, in increments of 0.01 msec. For each shift value, the parameterization performance is evaluated and the results are reported. The experiment essentially assesses the effects of the induced quantization error on the accuracy of the voice source parameterization methods.

The process of rounding off the voice source parameter values to a nearest integer, introduces a uniformly distributed estimation error with the expected mean absolute value of $25\mu s$. This is the theoretically minimum estimation error that can be achieved by any parameterization method that employs the rounding off procedure. As such, it will be used as the performance benchmark.

Figure 5.6 shows the median absolute estimation error values for the individual voice source parameters and for a range of time shifts. As expected, with the respect to each voice source parameter, both *fit estimation* (FE) and *CGPWP* methods are able to achieve higher accuracy levels than the $25\mu s$ benchmark value. Since CGPW parameterization is based on the non-linear optimal alignment functions, it responds particularly well to the signal distortions caused by the temporal shifts of the voice source signal. The CGPWP outperforms the FE method with the respect to the instant of vocal fold abduction t_o, the instant of maximum glottal airflow t_p and the instant of vocal fold closure onset t_e. Somewhat worse performance is achieved for parameter that describes the effective duration of return phase t_a. This is due to the fact that unlike other voice source parameters, t_a estimation is not directly based on DTW functions, but it instead relies on the exponential fit procedure. Nevertheless, the performance of the fit procedure is improved by the accuracy levels in the t_e and t_c estimates and as a result, t_a values are also estimated with a reasonable level of accuracy.

In the case of CGPWP, when the voice source parameters do not coincide with integer sample values, they are obtained via the monotone piecewise cubic interpolation of the alignment functions. Since this form of interpolation preserves the shape of data and respects the monotonicity of the alignment path, CGPWP performance did not exhibit any significant dependency on the size of the induced quantization error. As such there is no observable trend between the time shift values and the estimation accuracy levels for any of the voice source parameters. Note that the performance results for the *direct estimation* (DE) method are not reported as its performance was not comparable to the other two methods.

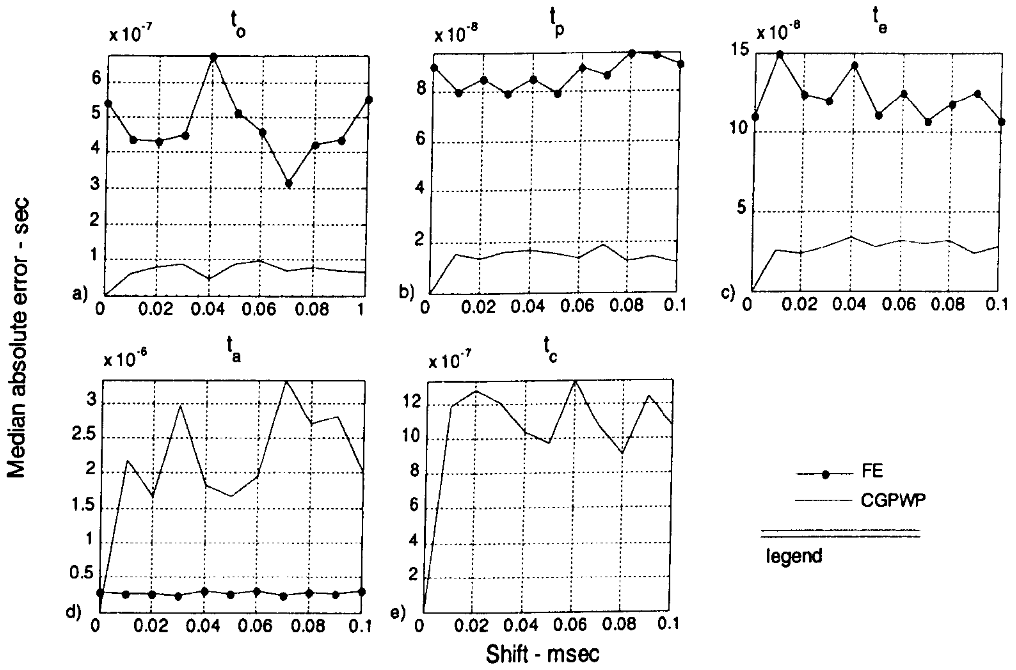


Figure 5.6: The performances of fit estimation and CGPWP with the respect to the individual voice source parameters and a range of time shift values.

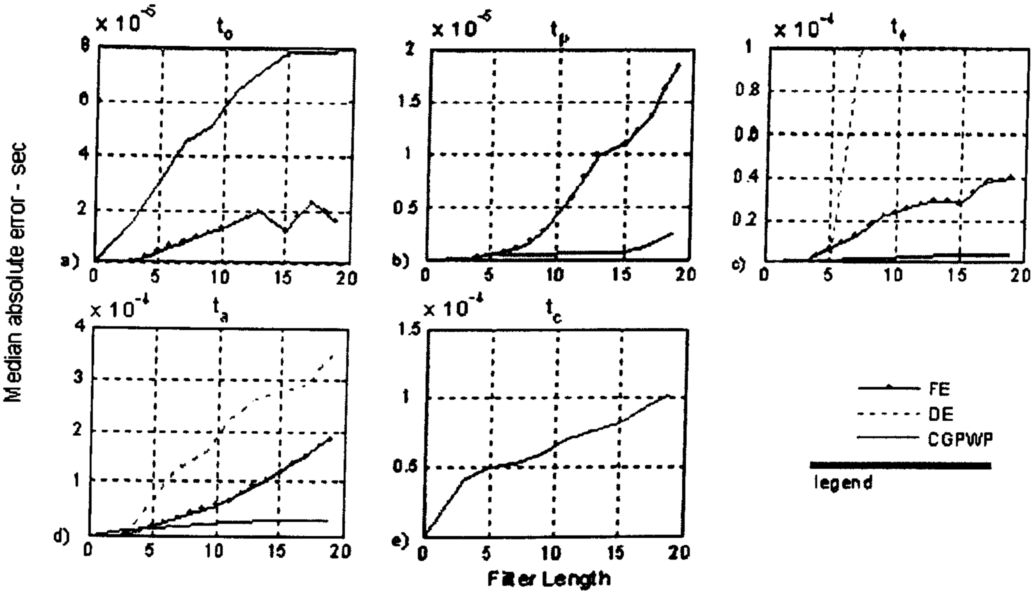


Figure 5.7: Comparative evaluation of fit estimation, direct estimation, and CGPWP on a voice signal distorted by Blackman window convolution.

Low-pass filtering Experiment

To various degrees, the estimates of the voice source signal contain the disturbances in form of aspiration noise and artifacts arising from the imperfect source-vocal tract deconvolution. In order to reduce the effect of these disturbances on the voice source parameterization, it is a common practice to apply a low pass filter on the voice source signal prior to parameterization.

The effect of low pass filtering on the parameterization performance is studied by convolving the synthesized voice source signals with the Blackman window. The extent of low pass filtering is varied by changing the length of the Blackman window from 3 to 19, in steps of 2. The voice source parameterization performance is evaluated in the same manner as in the previous experiment. Figure 5.7 shows the performance results for the *direct estimation* (DE), *fit estimation* (FE) and *CGPWP* across the range of window lengths. The results of the *direct estimation* are shown only for those parameters where its performance was comparable to the other two methods. Compared to the other two methods, CGPWP achieves a superior performance across the filter lengths and across the voice source parameters, except for t_0.

With the respect to individual parameters, the highest estimation accuracy is obtained for t_p and t_e. Glottal pulse is usually associated with the low energy values in the immediate region around the glottal opening instant t_0. Therefore, this region of the glottal pulse cycle has lower significance in determining the optimal alignment path via DTW. As a result, t_0 tends to be estimated with the lower levels of accuracy. To a lesser degree, the same is true for the onset of closed-phase, t_c. The effect of removing the high frequency components of glottal pulse waveform is such that it causes the glottal pulse to become "breathier". For very long window lengths, such as 19, the extent of glottal pulse degradation is so severe that the degraded pulses associated with the *vocal fry* are observed as *modal* glottal pulses. Nevertheless, CGPWP is able to achieve high estimation accuracy levels by exploiting the fact that the extent of degradation is similar for all glottal pulses in the *glottal matrix*, including the Characteristic Glottal Pulse Waveform.

GCI Experiment

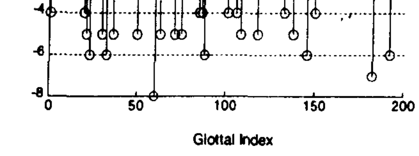


Figure 5.8: An example of the imposed GCI errors.

One of the most import issues that need to be considered is the effect of GCI estimation errors on the CGPWP performance. The glottal closure instants play an import role in CGPWP parameterization process. They are employed in temporal normalization and temporal alignment of the glottal flow derivative pulses inside the *glottal matrix.* In the design of CGPWP, we have introduced certain measures to deal with the potentially incorrect GCI estimates, namely, the mapping adjustment and the extended size of the *glottal matrix.* In the following experiment, the effectiveness of these measures is evaluated. We will only consider GCI errors of up to eight samples long, as in practice it is very unlikely that the GCI estimation would be of such poor accuracy. For the purpose of this experiment, the alignment path slope constrains are relaxed to 5:1 and 1:5 ratios. The size of the Dynamic Time Warping window is correspondingly extended. The GCI errors are introduced in form of Gaussian noise with its values rounded to

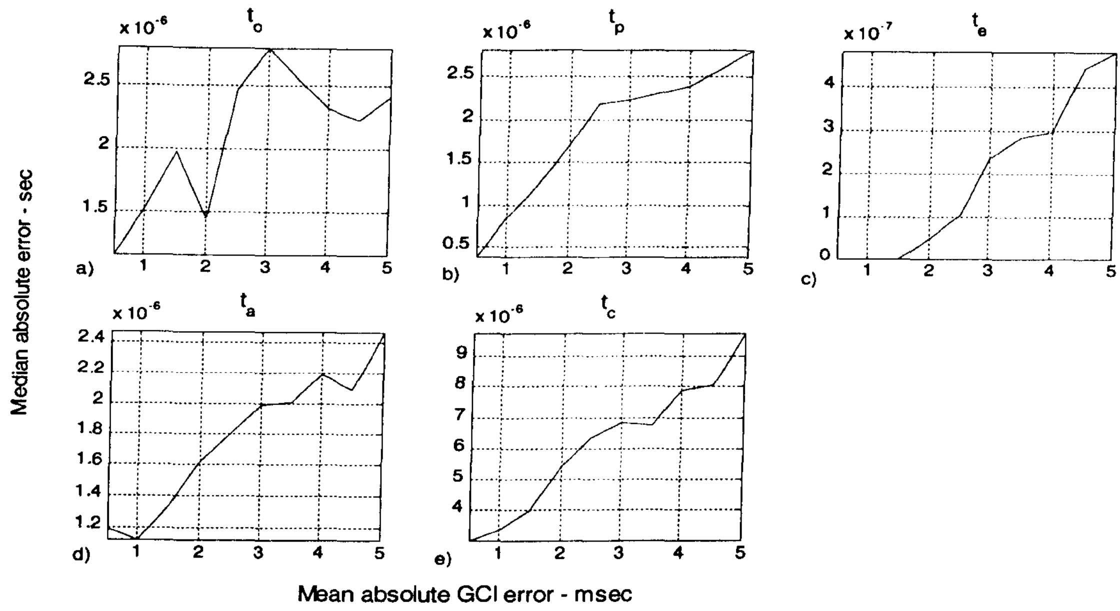


Figure 5.9: The performance CGPWP as a function of mean absolute error in the glottal closure instant estimates.

a nearest sample. Figure 5.8 shows an example of GCI error signal with the mean absolute value of 3 samples. The figure shows GCI error for each glottal pulse and the position of CGPW in the *glottal matrix*. Clearly, the estimated CGPW is bounded by the accurately estimated GCIs, as we have assumed in the CGPW estimation section. The performance results of CGPWP with the respect to various levels of GCI error are displayed in Figure 5.9. These results suggest that there is almost a linear relationship between CGPWP performance and the levels of error present in the GCI estimates. Increase in the levels of GCI error leads to a proportional decrease in the voice source parameterization accuracy.

However, the deterioration in the performance is not so much caused, directly, by the individual GCI errors, but rather by the consequential inaccuracies in the pitch period estimates. When the pitch periods are overestimated, the corresponding glottal pulses are represented by a reduced number of samples in the *glottal matrix* frames. This in turn reduces the effective resolution of the DTW alignment functions, which ultimately leads to a deteriorated parameterization performance. However, to put the results in perspective, CGPWP exhibits better performance for the highest levels of the considered GCI errors than for the lowest levels of low pass filtering. As long as the entire glottal pulse is present inside the *glottal matrix* and it is "covered" by the DTW window, we can expect CGPWP to adequately cope with even unrealistically high GCI errors. Let us remind that in the design of CGPWP, we have assumed that the estimated Characteristic Glottal Pulse Waveform will be bounded by the accurately estimated glottal closure instants. Since this assumption has held true throughout the experiments, no additional measures will be taken.

In the conclusion of this section, we want to stress that all of the considered performance evaluation experiments were inherently biased toward the *direct estimation* and *fit estimation* methods. Unlike these methods, CGPWP does not take a premise that the analyzed voice source signal behaves according to Liljencrants-Fant's model. Its performance is equally good if the voice source signal deviates from the ideal perspective of vocal fold realization. When applied on the natural voice source signal, the difference in the performances between CGPWP and the other two methods is even more substantial.

5.3.2. Performance evaluation - natural speech dataset

The modeling accuracy of CGPWPM can be evaluated via a modified segmental SNR measure:

$$SNR\_seg = \frac{10}{N}\sum_{i=1}^{N}\log_{10}\left(\sum_{n=\delta+1}^{L-\delta}\frac{\mathbf{G'}_i^2(n)}{\mathbf{E}_i^2(n)}\right) \qquad (5.10)$$

where $\mathbf{G'}$ and $\mathbf{E}$ denote *aligned glottal matrix* and a m*odeling residue matrix*, respectively. N and L represent the number of *glottal matrix* pulses and the length of each *glottal matrix* frame, respectively. The parameter δ is defined as $\delta = (L - T_N)/2$, where T_N denotes the normalized pitch period length. In the similar fashion, the extent of aspiration noise in a voice source estimate can be quantified as:

$$SNR\_aspiration = \frac{10}{N}\sum_{i=1}^{N}\log_{10}\left(\sum_{n=\delta+1}^{L-\delta}\frac{\mathbf{G}'^{2}_{i}(n)}{\mathbf{A}^{2}_{i}(n)}\right) \tag{5.11}$$

where $\mathbf{G}'$ and $\mathbf{A}$ denote *aligned glottal matrix* and *aspiration noise matrix*, respectively.

The segmental SNR values reflect the extent by which the constrained non-linear temporal warping of the Characteristic Glottal Pulse Waveform can be used to represent the other glottal waveforms in a *glottal matrix*. Since the voice source parameters are directly estimated from the alignment functions, segmental SNR values to a large extent reflect the quality of voice source parameterization performance. An important attribute of this measure is that it evaluates the overall representation of the glottal pulse shapes, and as such it accounts for the entire set of voice source parameters, simultaneously. If the SNR_*seg* values are low, it would indicate that CGPWPM is not able to adequately represent a large fraction of glottal pulses, and consequently, the parameterization results might not be credible. On the other hand, high SNR_*seg* values imply that an accurate relationship between the Characteristic Glottal Pulse Waveform and the other glottal pulses in the *glottal matrix* has been established and thus, we can be confident in the voice source parameterization results, as well. In the following experiment, we will use the segmental SNR measure to estimate the optimal size for the DTW window. Subsequently, we will use the segmental SNR, and the optimal alignment window to indirectly evaluate the quality of voice source parameterization across a range of voice quality types.

Estimation of the optimal window size

The experiment is performed on a subset of a *WSJCAM0* database containing two read speech sentences from 20 male and 20 female speakers. Subsequently, CGPWPM is applied on each acoustic stimulus to produce a set of *glottal matrix*es and *modeling residue* matrixes. The speech signals are sampled at 10 kHz and CGPWPM is performed for $L=100$ and $T_N=120$. We varied the size of the DTW window from 0.0 % (Euclidean distance) to 100 % (no global constraints) in increments of 0.833 % (1 sample). Each time, the mean segmental SNR is evaluated across the dataset. The results are reported in Figure 5.10. Note that only a section of the graph below the SNR saturation value is displayed.

The graph indicates a non-linear relationship between the SNR and the window size. An increase in SNR value with the respect to a unit increment in the window size is higher for the lower window lengths. At 21.67 % window length, SNR function reaches a saturation value of 22.11 dB. Let us remind that DTW algorithm exhibits *$O(n^2)$* time complexity. As such, the optimal alignment window length is estimated as a window length for which the SNR function is 1 % below the saturation value. In doing so, the computational cost is significantly reduced for only a slight sacrifice in the modeling accuracy. The optimal window length corresponds to 11.67 % (14 samples), for which the modeling accuracy is at *21.87 dB*. In the final comment on this experiment, we want to point out that even in the instance when the alignment window size is reduced to 0.0 %, a reasonably high SNR level is achieved. This result confirms that an estimated Characteristic Glottal Pulse Waveform is indeed an adequate representation of a typical glottal flow derivative waveform of a speaker.

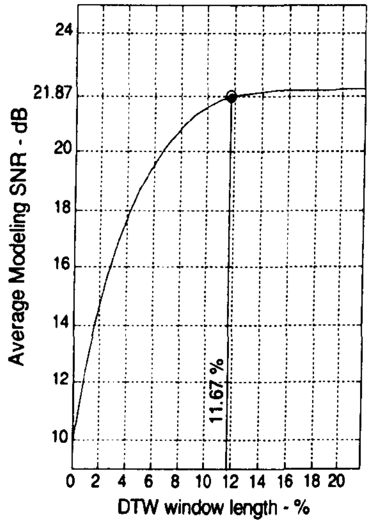


Figure 5.10: Modeling SNR as a function of DTW window length (solid line). The optimal window length and the corresponding SNR value (circle).

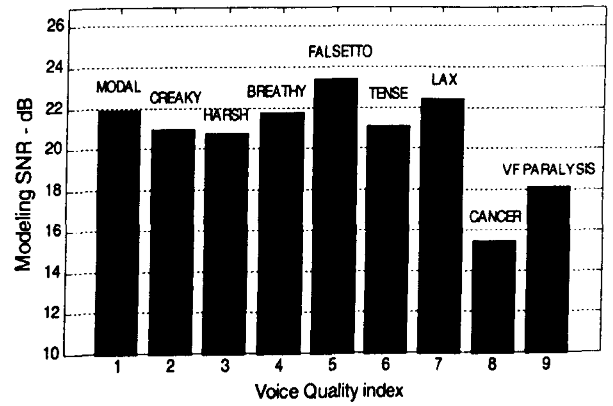


Figure 5.11: Modeling SNR for various voice quality types. The values are obtained for the optimal DTW window length.

Segmental SNR across voice quality types

In this section, a segmental SNR measure is obtained for a range of voice quality types. Each voice quality type is represented with a database of 5 male and 5 female speakers. Two read speech sentences are obtained from each speaker. The speech files are sampled at 10 kHz. CGPWPM is performed for L=100, T_N=120 and for the optimal alignment window size of 11.67% (14 samples). The local, alignment slope constraints are kept at 3:1 and 1:3 ratios. Figure 5.11 displays the segmental SNR values for *model*, *creaky*, *harsh*, *breathy*, *falsetto*, *tense*, *lax* voice and two types of voice pathologies: *cancer* and *vocal fold paralysis*.

As we have expected, the lowest modeling accuracy is obtained for the pathological voices. Difficulties in modeling pathological voices, especially those associated with the laryngeal cancer arise from the speakers' inability to maintain a reasonable level of regularity in the glottal flow derivative waveform realizations, as well as from the high aspiration noise levels. Conversely, the *SNR_seg* values across the healthy voice types are consistently high. However, *breathy*, *lax* and *falsetto* voice are slightly better modeled than the other voice types. We believe that this is related to the fact that the abducted phonations, associated with the slowly varying glottal pulse shapes, are less prone to quantization error.

In the remainder of this section we present the results of Characteristic Glottal Pulse Waveform Parameterization for three male (*M1*, *M2*, *M3*) and three female speakers (*F1*, *F2*, *F3*). The presentation for each speaker includes the following:

- Parametric voice quality description of the estimated CGPW
- GM and LF trajectories across the read speech sentence: "Don't ask me to carry an oily rag like that."
- A 0.08 second segment of a voice source estimate and the corresponding synthesized waveform, starting with the first identified glottal closure instant.
- Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aspiration Noise Matrix and Modeling Residue Matrix.
- Phase-plane plot of the Characteristic Glottal Pulse Waveform.
- RMS values obtained on *aspiration noise* and *modeling residue matrices* across glottal cycle index.

Tables 5.2-5.7 show the parametric voice quality description and the *SNR_seg* value for each of six speaker. The GM and LF parameter trajectories are presented in Figure 5.12, Figure 5.17, Figure 5.22, Figure 5.27, Figure 5.32, Figure 5.37, whereas the corresponding aspiration noise envelopes are shown in Figure 5.16, Figure 5.21, Figure 5.26, Figure 5.31, Figure 5.36, Figure 5.41. The relevant examples of voice source estimates and the corresponding synthesized waveforms are shown in, Figure 5.13, Figure 5.18, Figure 5.23, Figure 5.28, Figure 5.33, Figure 5.38. The qualitative evaluation of the synthesized voice source signals is in accord with high SNR_*seg* values. Even in the cases of speakers *M1* and *M3*, where the voice source estimates do not conform to the idealistic view of glottal flow derivative waveform, CGPWPM is robust enough to parameterize every single glottal pulse in the voice source estimate and to provide accurate voice source reconstruction. Furthermore, Figures 10-13 show that GM and LF trajectories do not have extreme outliers, and that the voice source parameters attain realistic values. This CGPWPM quality can be attributed to the global and local constraints in DTW that preclude pathological alignment and parameterization. Observation of voice source parameter trajectories indicates that the nature of their general movement across the same utterance is to some extent shared by all speakers. These results lead us to postulate that the voice source signal might have influence on the linguistic layer of speech communication. Although during our extensive work in voice source parameterization we gathered further evidence to support the above mentioned claim, until further experiments are conducted we will leave the issue with this postulation.

The estimated aspiration noise envelopes also comply with their expected form [23]. Except for the strong peaks around the glottal closure and glottal opening instants, aspiration noise envelopes realize an almost constant value over the glottal cycle duration. Note that the presented aspiration noise envelopes correspond to the sections in-between the GCIs, or exactly one normalized pitch cycle length. As far as we are aware, this is the only method that enables accurate and robust estimation of statistical properties of the turbulent components related to aspiration noise. The results clearly show that the aspiration noise envelopes and the energy distribution of turbulent components over the glottal cycle duration vary considerably across speakers. In Chapter 6 the estimates of aspiration noise envelope, rather than the idealistic approximations, e.g Hamming window, will be used in the process of aspiration noise synthesis to enable faithful voice source reconstruction.

The phase-plane plots can be used to evaluate the quality of voice source deconvolution from the vocal tract filter [40]. A successful inverse filtering would remove all the vocal tract resonance information from the glottal waveform estimates. Therefore, the phase-plane plot of the voice source estimates should produce a single closed-loop with no self-intersections. If a phase-plane plot exhibits more than one closed loop or displays self-intersections, it would be an indication that the vocal tract resonances are present in the voice source estimate. By applying the phase-plane analysis on the Characteristic Glottal Pulse Waveform, one can obtain a quick an objective assessment of the glottal flow derivative estimate quality, and thus avoid a tedious process of evaluating the quality of each glottal pulse individually. Further justification for the phase-plane analysis of the voice source signal quality is provided in Appendix C. The inspection of the phase-plane plots shown in following figures: Figure 5.15, Figure 5.20, Figure 5.25, Figure 5.30, Figure 5.35, and Figure 5.40, confirms the quality of voice source estimation as no additional loops or self intersections can be observed.

Apart from providing the means to evaluate the quality of voice source estimates, phase-plane plots also offer an alternative perspective on glottal flow dynamics. The phase-plane plots of the Characteristic Glottal Pulse Waveform display significantly different characteristics for each of the considered six speakers indicating that the temporal structure of the CGPW is specific for each speaker. This argument is of course supported by varying CGPW parametric descriptions in Tables 5.2-5.7. However, the CGPW parameters are only related to the "coarse" or the general behavior of the glottal pulse shape, whereas the nature of phase-plot realization also reflects the fine glottal flow derivative structure.

Table 5.2:
Subject M1: Parametric voice quality description and the modeling SNR value

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	Aspiration SNR [dB]	Modeling SNR [dB]
2.00	48.36	71.00	1.33	74.00	1.83	13.78	18.76

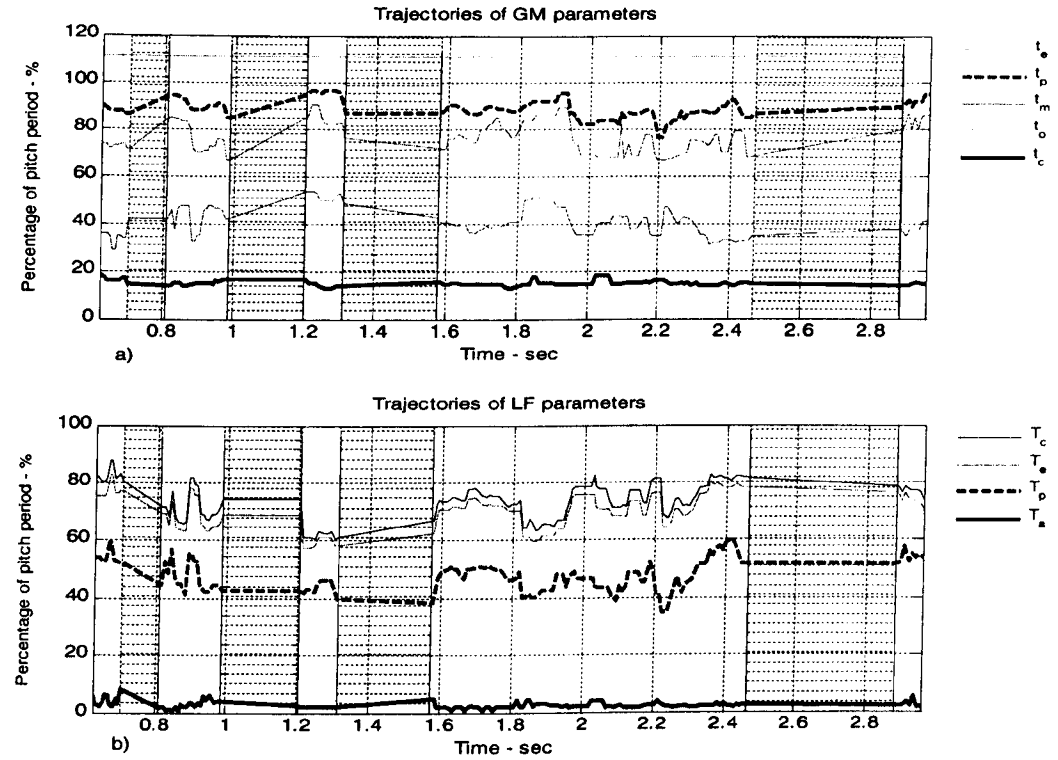


Figure 5.12: Subject M1: *The trajectory of the GM and LF parameters across a sentence.*

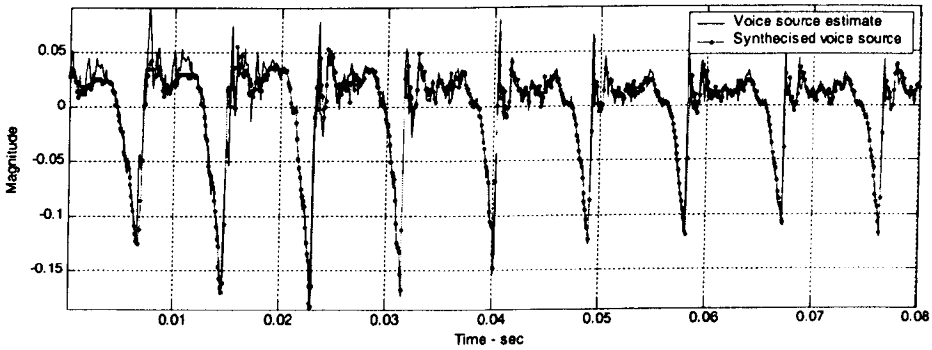


Figure 5.13: Subject M1: *A segment of voice source estimate and its synthetic version.*

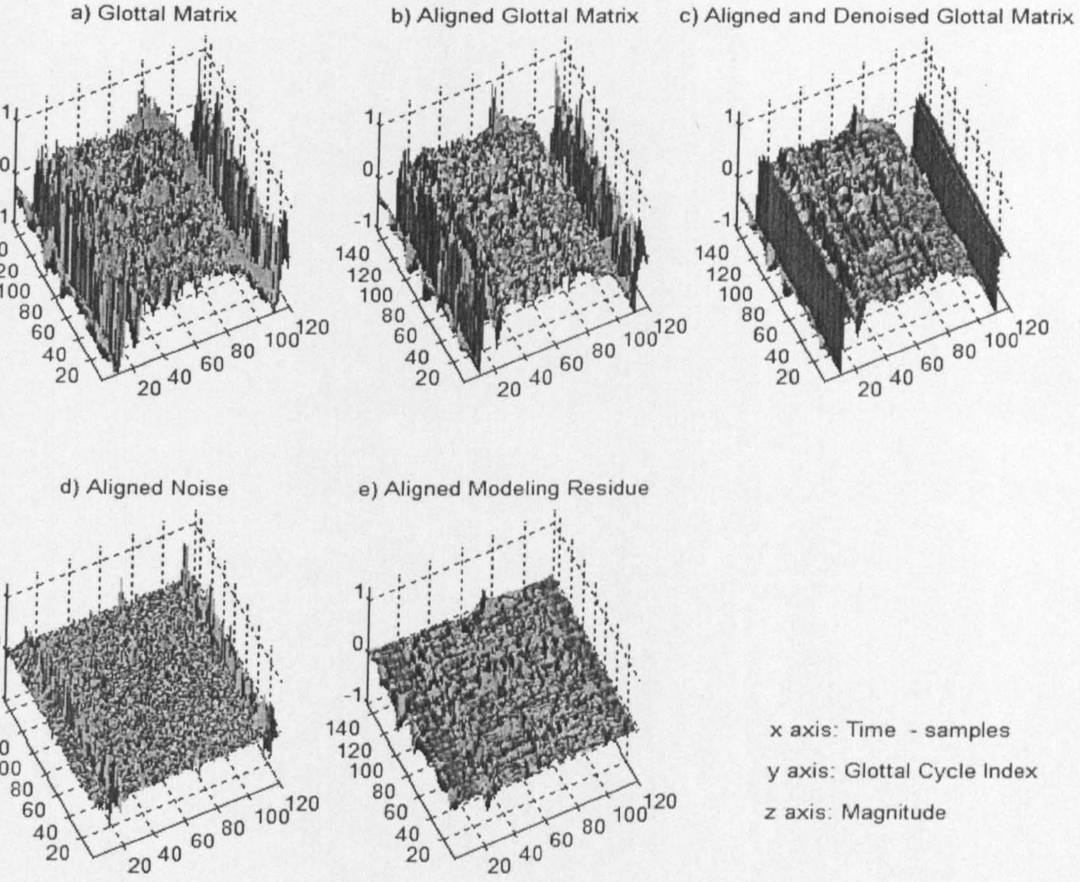
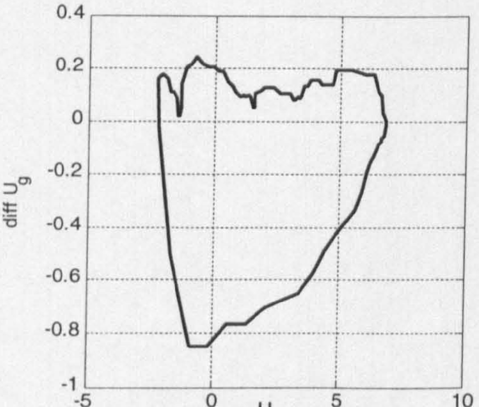


Figure 5.14: Subject M1: Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aligned Aspiration Noise estimate and Modeling Residue.



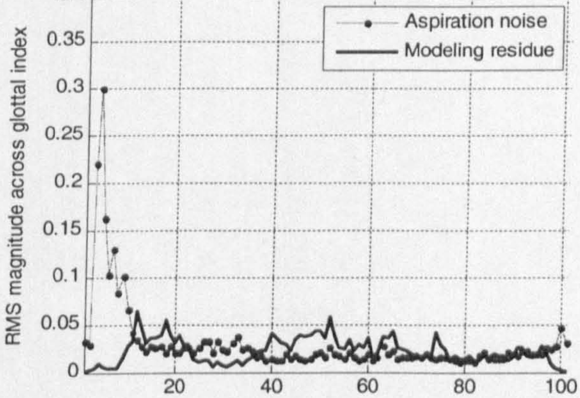


Figure 5.15: Subject M1: Phase-plane plot of the Characteristic Glottal Pulse Waveform

Figure 5.16: Subject M1: RMS values across glottal cycles of Aspiration Noise, and Modeling Residue Matrices

Table 5.3:
Subject M2: Parametric voice quality description and the modeling SNR value

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	Aspiration SNR [dB]	Modeling SNR [dB]
0.40	57.14	66.61	1.35	68.61	1.62	18.10	22.84

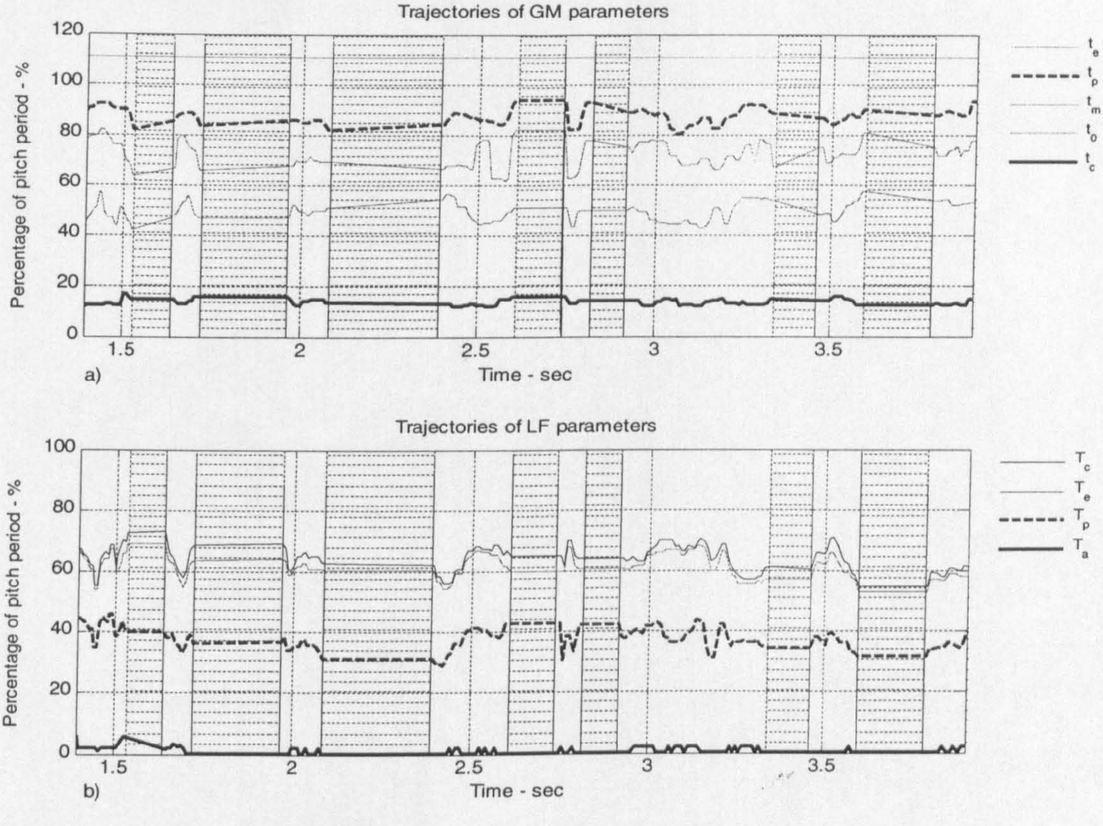


Figure 5.17: Subject M2: *The trajectory of the GM and LF parameters across a sentence.*

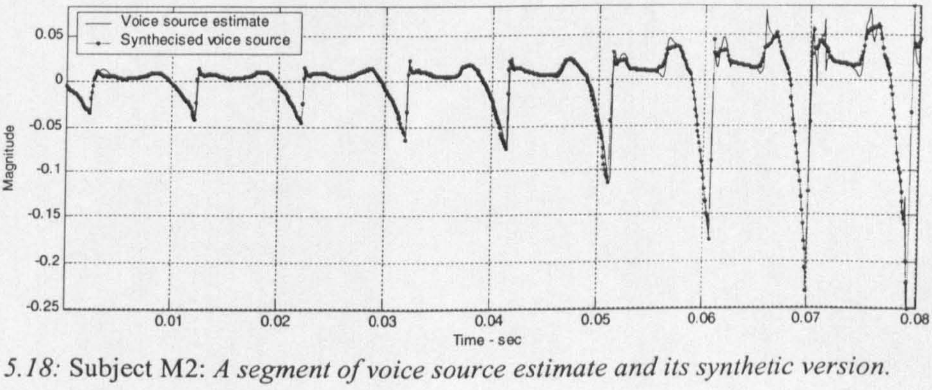


Figure 5.18: Subject M2: *A segment of voice source estimate and its synthetic version.*

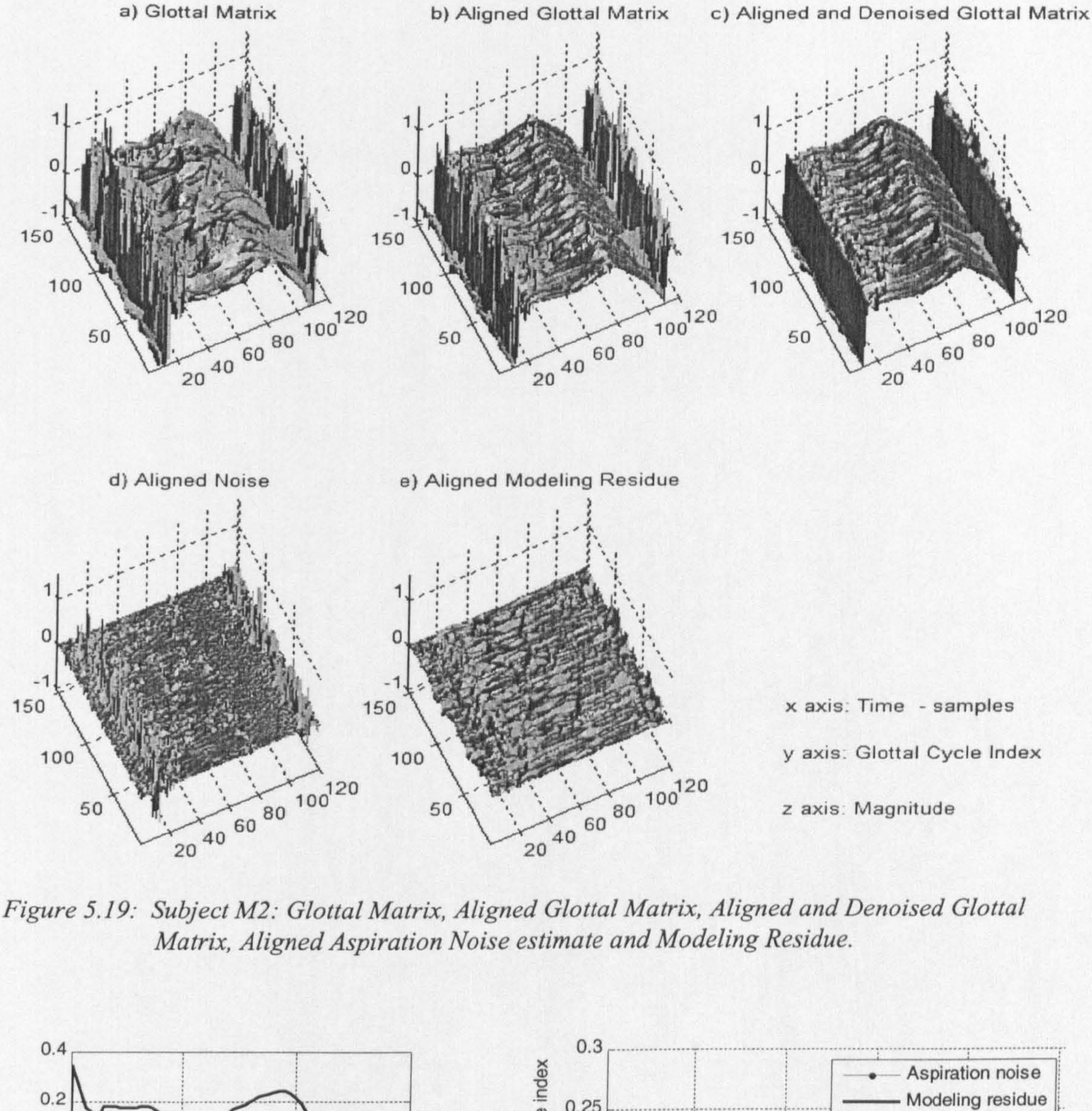


Figure 5.19: Subject M2: Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aligned Aspiration Noise estimate and Modeling Residue.

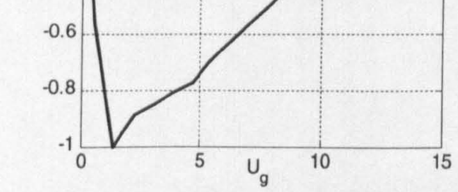


Figure 5.20: Subject M2: Phase-plane plot of the Characteristic Glottal Pulse Waveform

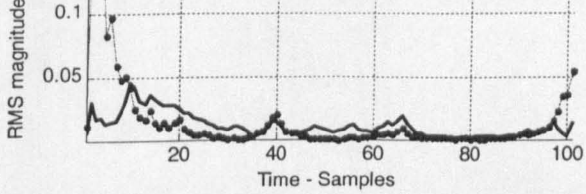


Figure 5.21: Subject M2: RMS values across glottal cycles of Aspiration Noise, and Modeling Residue Matrices

Table 5.4:
Subject M3: Parametric voice quality description and the modeling SNR value

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	Aspiration SNR [dB]	Modeling SNR [dB]
3.40	78.27	56.00	2.05	62.00	1.03	20.3	24.22

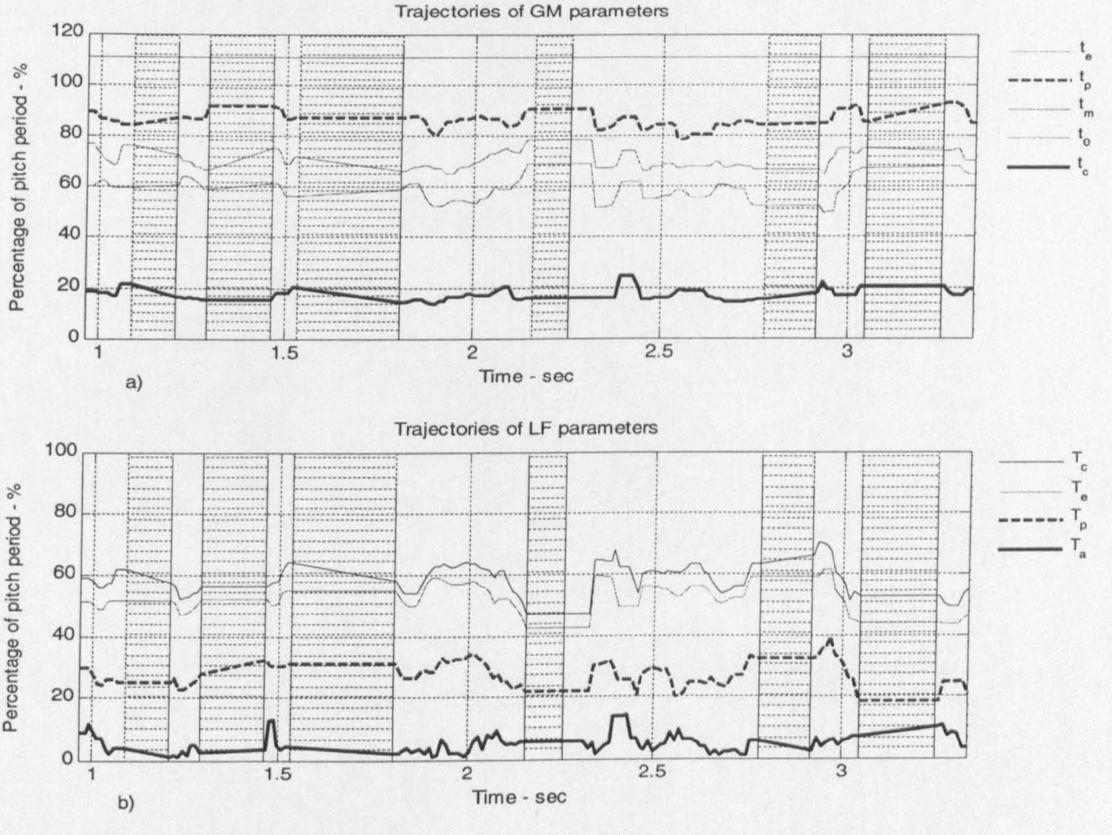


Figure 5.22: Subject M3: *The trajectory of the GM and LF parameters across a sentence.*

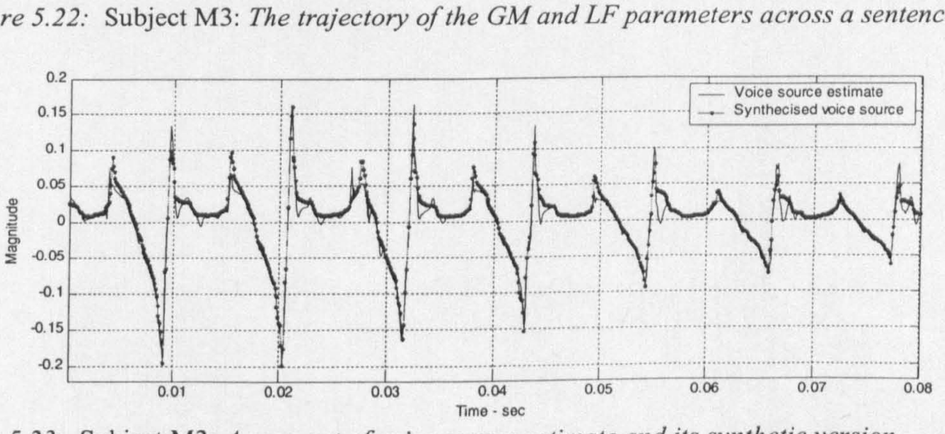


Figure 5.23: Subject M3: *A segment of voice source estimate and its synthetic version.*

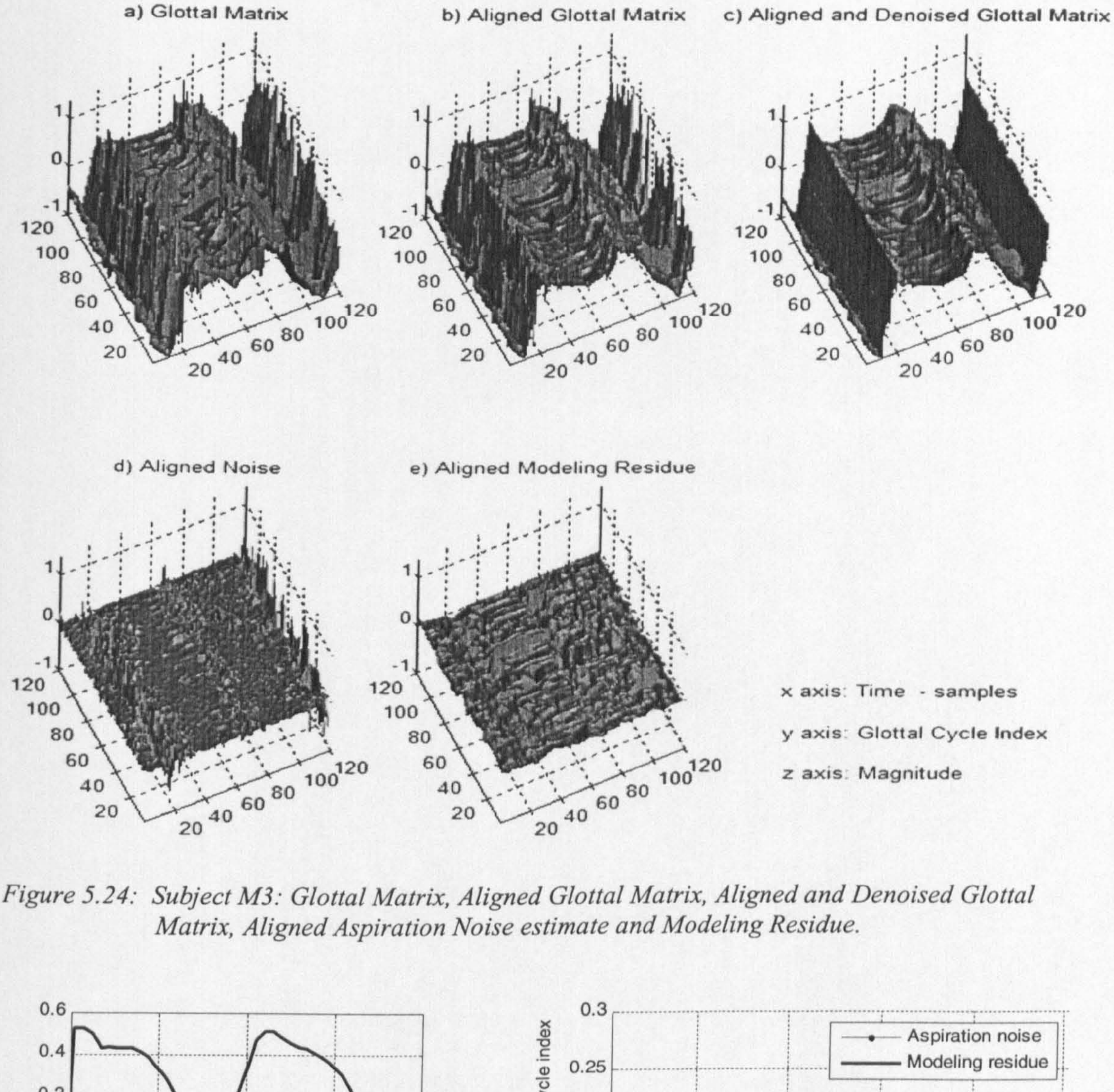


Figure 5.24: Subject M3: Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aligned Aspiration Noise estimate and Modeling Residue.

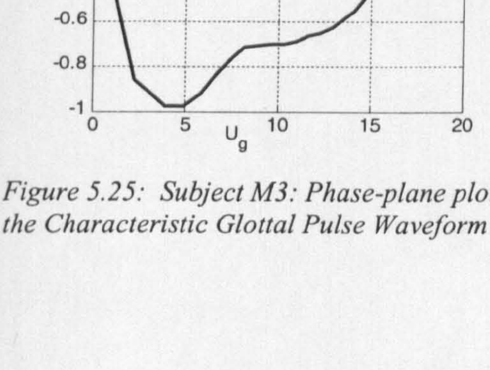


Figure 5.25: Subject M3: Phase-plane plot of the Characteristic Glottal Pulse Waveform

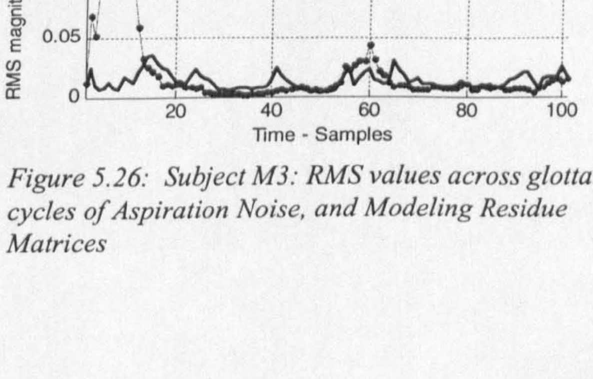


Figure 5.26: Subject M3: RMS values across glottal cycles of Aspiration Noise, and Modeling Residue Matrices

Table 5.5:
Subject F1: Parametric voice quality description and the modeling SNR value

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	Aspiration SNR [dB]	Modeling SNR [dB]
4.37	36.72	91.28	1.42	96.30	2.26	17.88	21.65

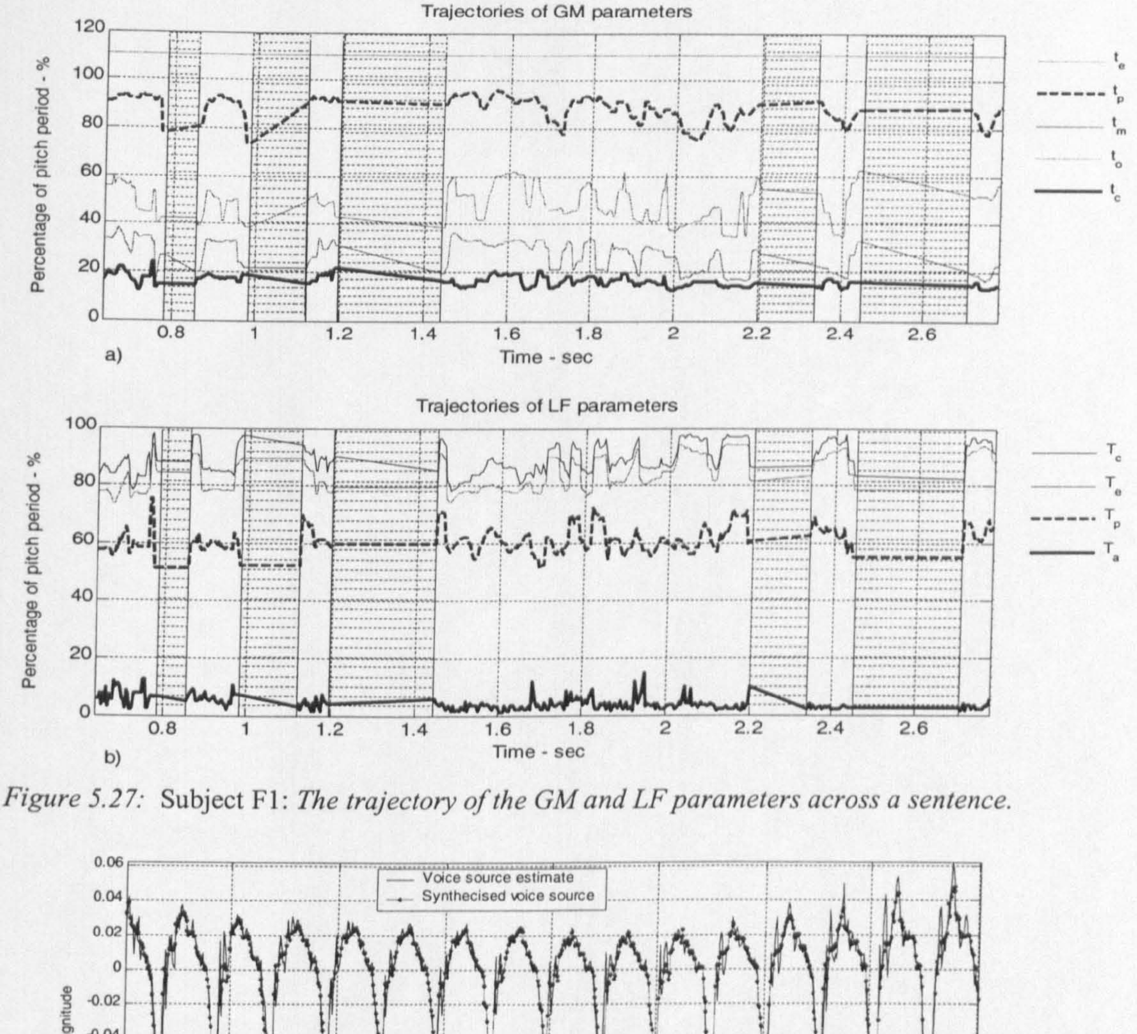


Figure 5.27: Subject F1: *The trajectory of the GM and LF parameters across a sentence.*

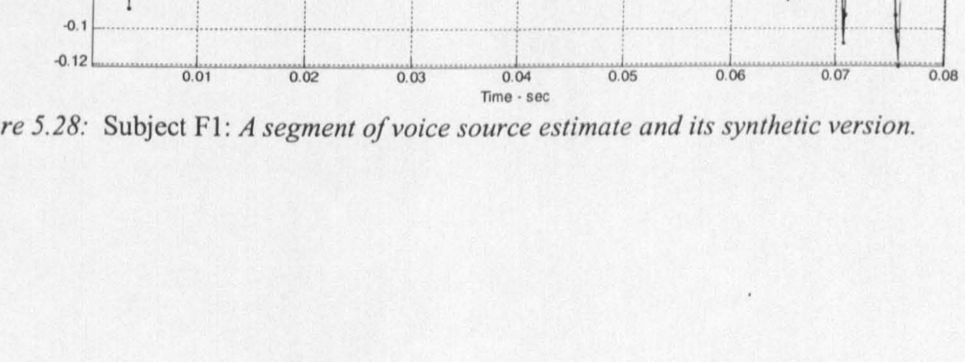


Figure 5.28: Subject F1: *A segment of voice source estimate and its synthetic version.*

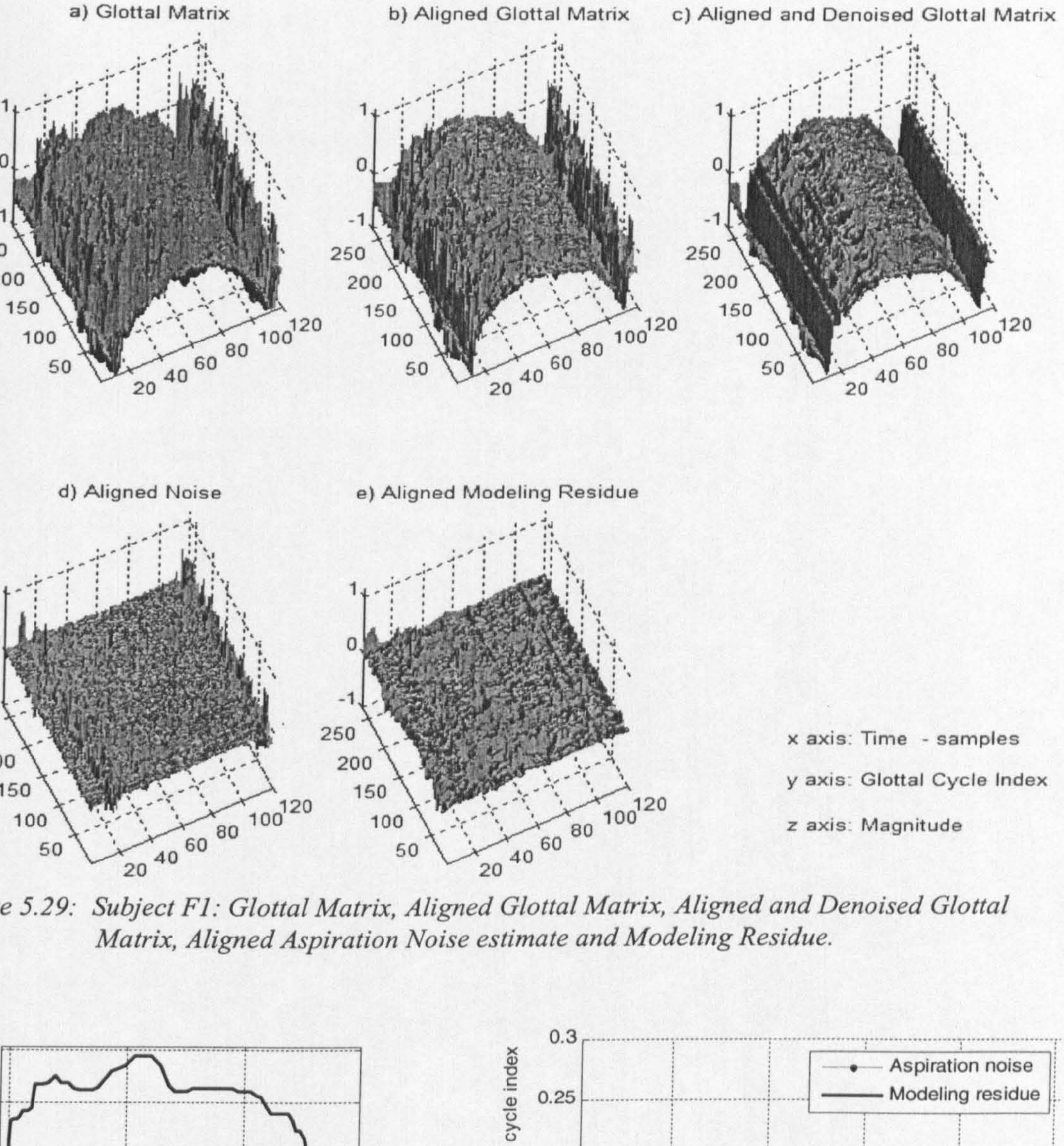


Figure 5.29: Subject F1: Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aligned Aspiration Noise estimate and Modeling Residue.

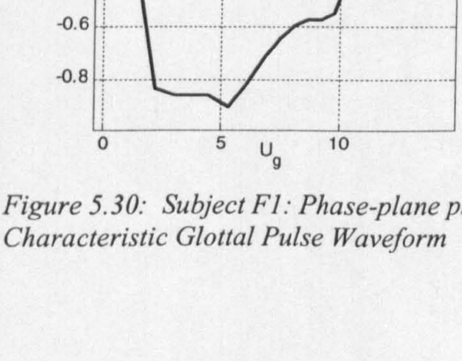


Figure 5.30: Subject F1: Phase-plane plot of the Characteristic Glottal Pulse Waveform

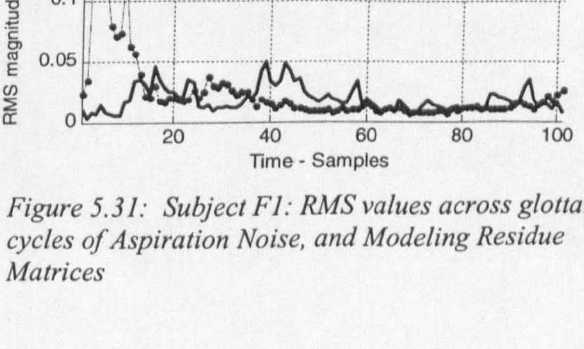


Figure 5.31: Subject F1: RMS values across glottal cycles of Aspiration Noise, and Modeling Residue Matrices

Table 5.6:
Subject F2: Parametric voice quality description and the modeling SNR value

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	Aspiration SNR [dB]	Modeling SNR [dB]
7.53	37.35	86.15	1.66	92.44	2.11	27.74	26.02

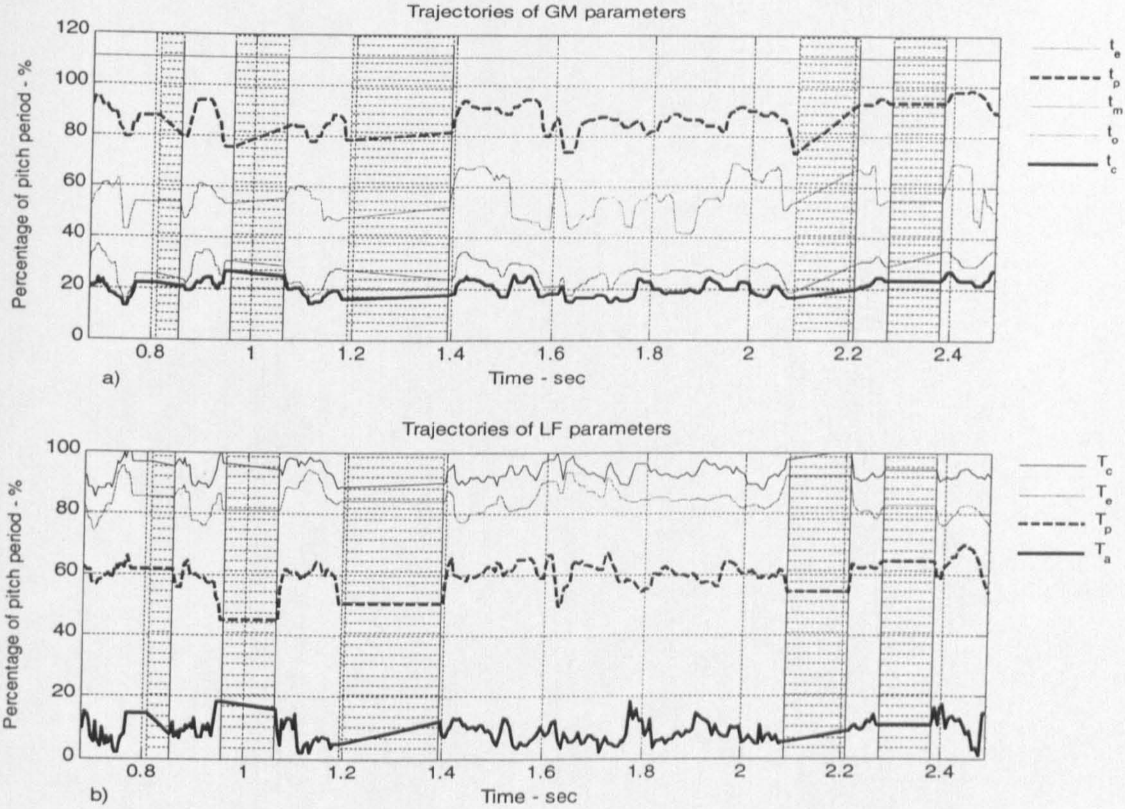


Figure 5.32: Subject F2: *The trajectory of the GM and LF parameters across a sentence*

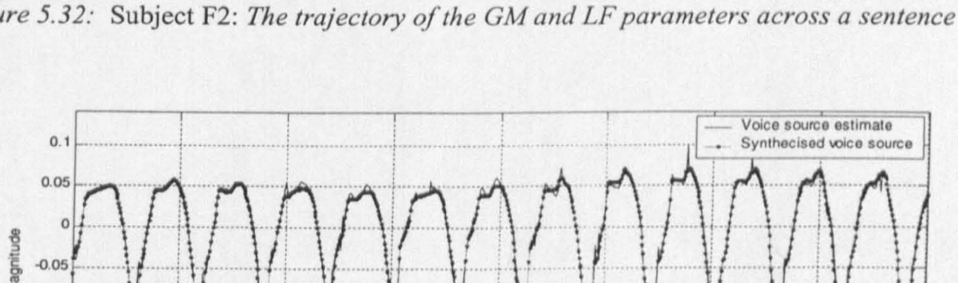


Figure 5.33: Subject F2: *A segment of voice source estimate and its synthetic version.*

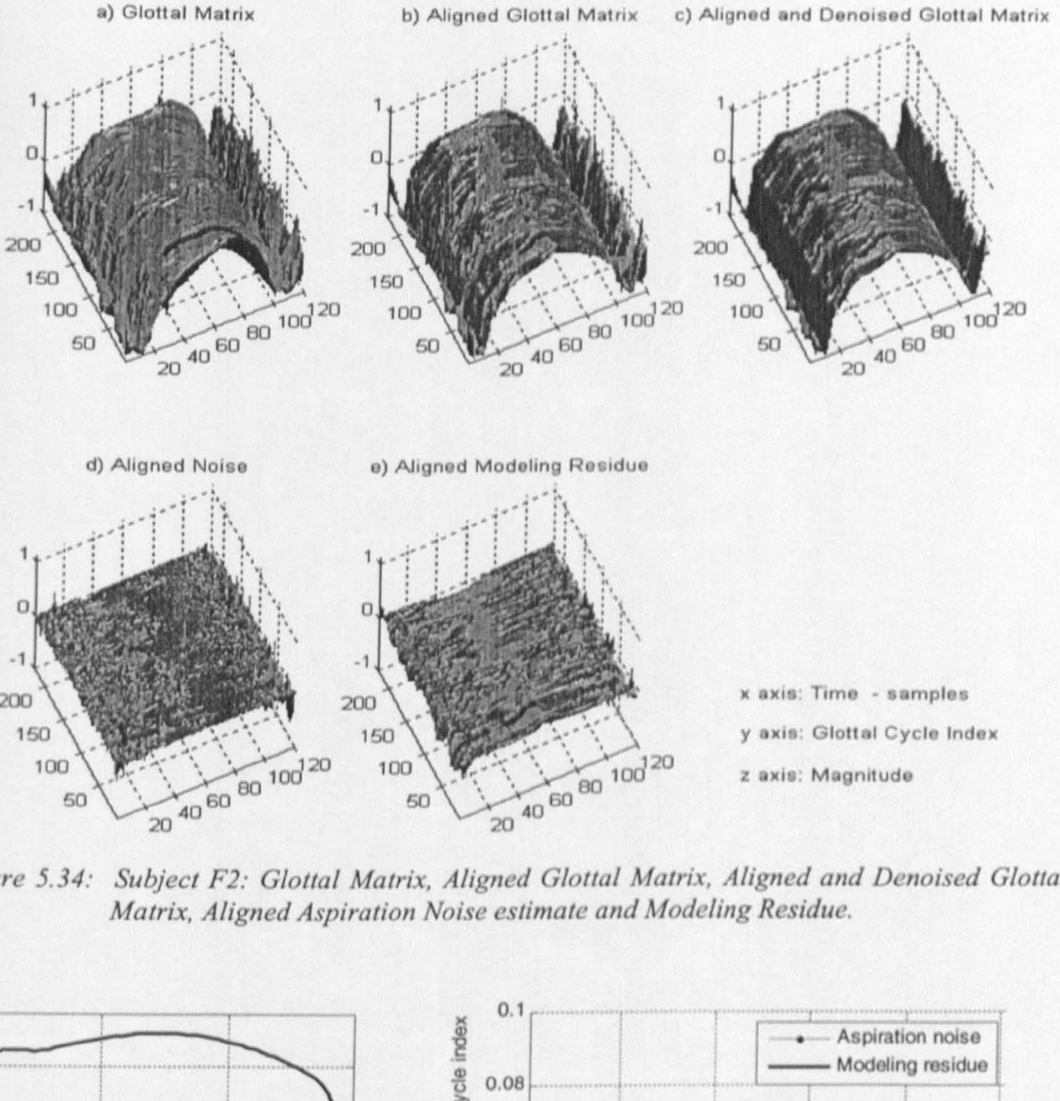
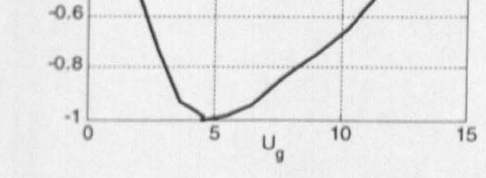


Figure 5.34: Subject F2: Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aligned Aspiration Noise estimate and Modeling Residue.



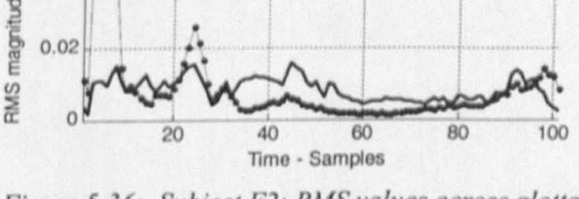


Figure 5.35: Subject F2: Phase-plane plot of the Characteristic Glottal Pulse Waveform

Figure 5.36: Subject F2: RMS values across glottal cycles of Aspiration Noise, and Modeling Residue Matrices

Table 5.7:
Subject F3: Parametric voice quality description and the modeling SNR value

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	Aspiration SNR [dB]	Modeling SNR [dB]
12.77	28.23	76.0	1.61	87.00	2.14	26.13	23.12

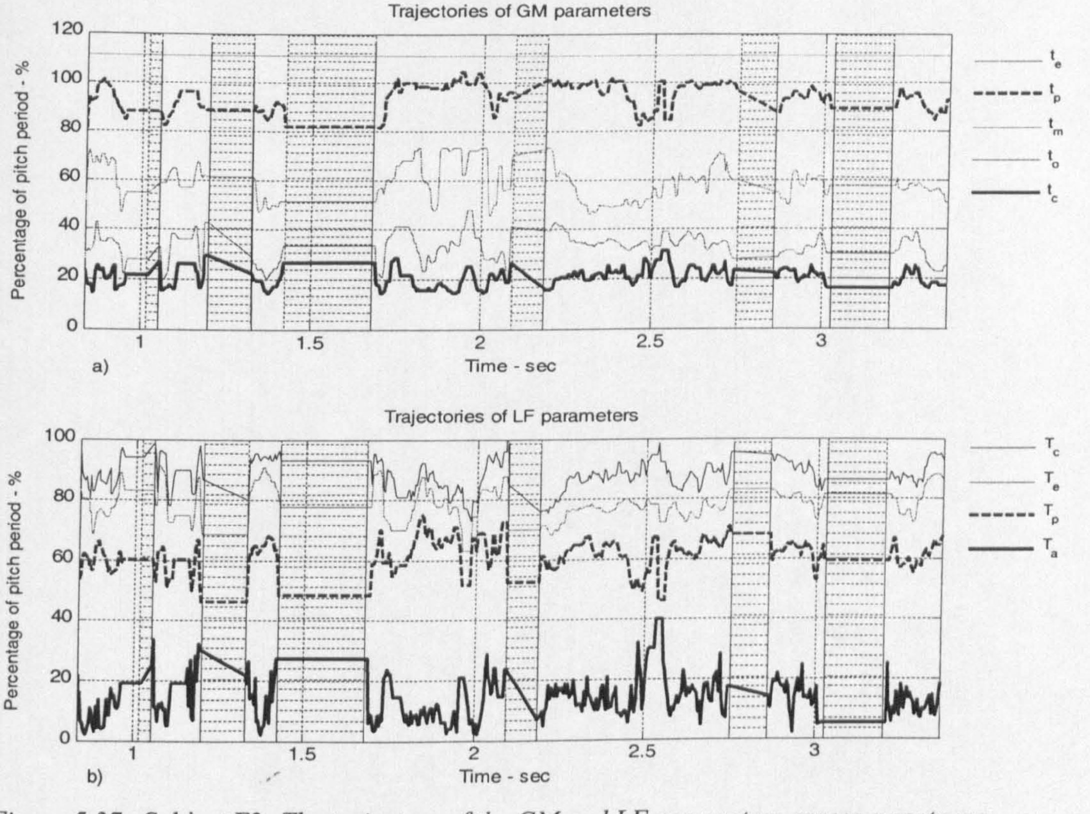


Figure 5.37: Subject F3: *The trajectory of the GM and LF parameters across a sentence.*

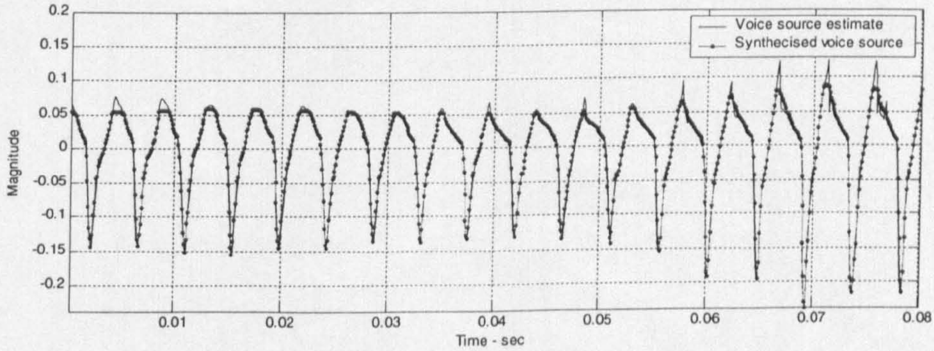


Figure 5.38: Subject F3: *A segment of voice source estimate and its synthetic version.*

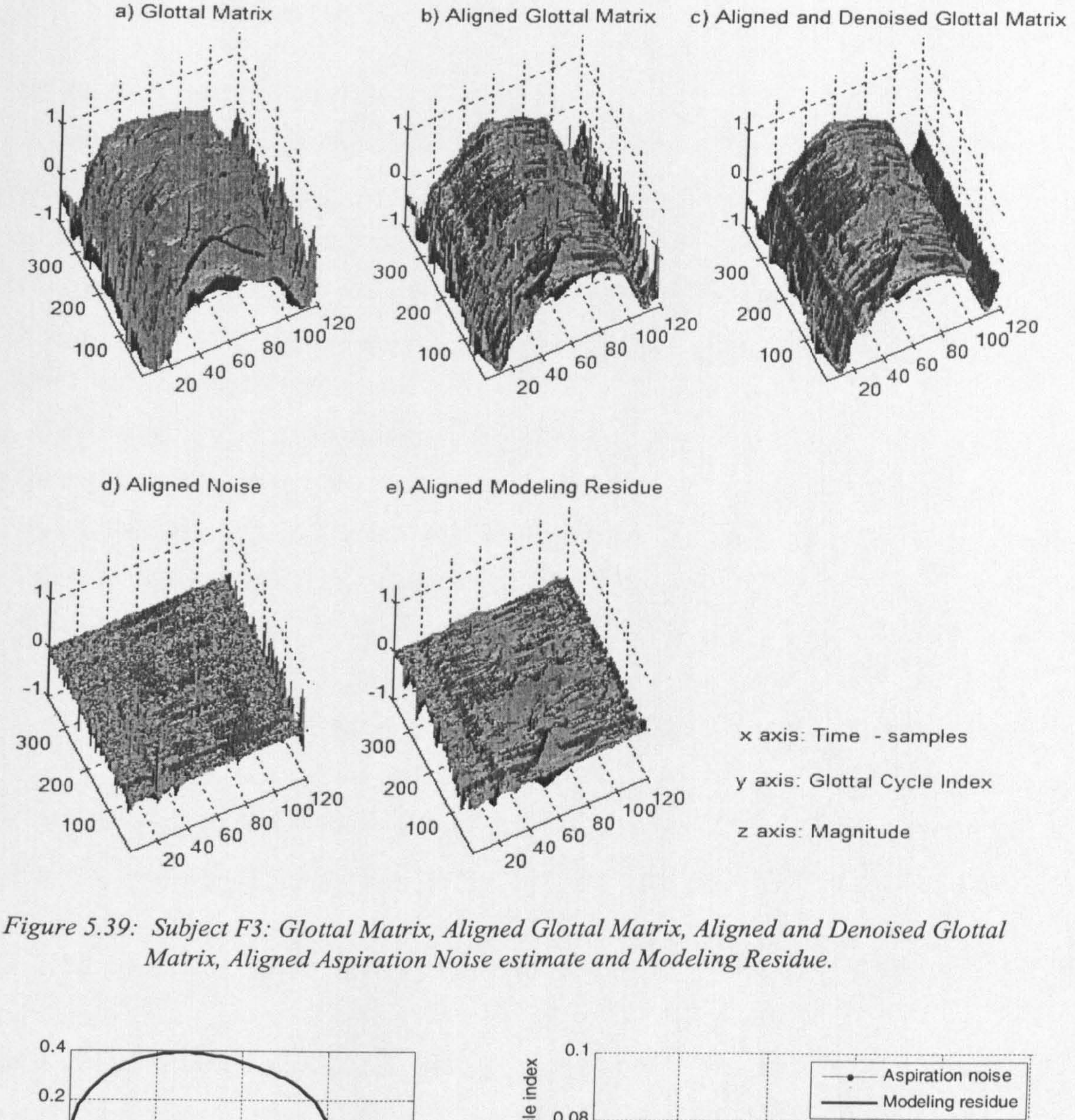
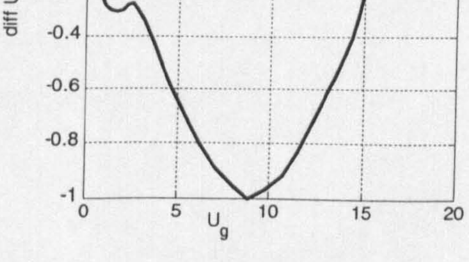


Figure 5.39: Subject F3: Glottal Matrix, Aligned Glottal Matrix, Aligned and Denoised Glottal Matrix, Aligned Aspiration Noise estimate and Modeling Residue.



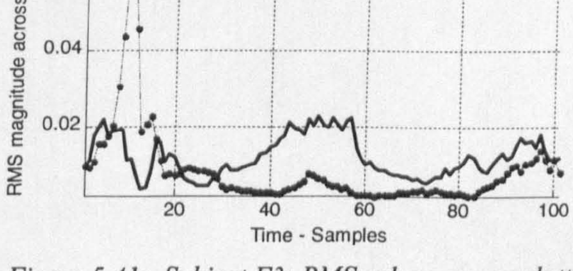


Figure 5.40: Subject F3: Phase-plane plot of the Characteristic Glottal Pulse Waveform

Figure 5.41: Subject F3: RMS values across glottal cycles of Aspiration Noise, and Modeling Residue Matrices

5.4 Voice quality profiling

As we have discussed in the Chapter 1, Subsection 1.2.2, voice quality can be broadly classified into five categories: *modal*, *creaky*, *harsh*, *breathy*, and *falsetto* phonation. Two additional phonation categories, *tense* voice and *lax* voice, are commonly cited in literature as means of describing the overall laryngeal and supralaryngeal muscular tension settings. These seven phonation types are the subject of our voice quality profiling study. The voice quality is an amalgam of many parameters and the degree and the order of importance among them differs for each type of phonation. As such, we will consider, not just the glottal shape parameters as in [33], but also signal to aspiration noise ratio, and the aperiodic acoustic features (shimmer, jitter and perturbation of glottal shape parameter, R_d). The aperiodic features are thought to be a product of non-linear behavior in the speech anatomy, whereby successive cyclic variations may alternate on each cycle of vocal fold vibrations [134], or they may appear random while in reality being the product of an underlying deterministic system [67]. Pitch, amplitude and glottal shape perturbations over the successive glottal cycles give the vowels a perception of "naturalness". On the other hand, a monotonic voice source signal is usually perceived as a machine like sound. Varying levels of jitter and shimmer can give voice specific perceptual characteristics, and thus, they are important features that need to be considered in any voice quality analysis or voice pathology classification study.

Over the years, considerable research effort has been committed to developing a parametric description for various voice quality types [83], [20], [33]. However, due to the fact that the voice quality types are not mutually exclusive, and that the perception of voice quality is a highly subjective matter, the parametric voice quality descriptions tend to differ from researcher to researcher. A common approach to voice quality profiling is to estimate an average value of a voice source parameter across pitch-scales, vowel types and speakers. However, voice source parameters are not independent, and consequently, the result obtained by averaging the individual parameters in isolation can be very misleading.

In this thesis, we have attempted to extend the principles behind the Characteristic Glottal Pulse Waveform estimation to voice quality profiling. Let us remind that CGPW estimation is a process whereby a database of normalized glottal pulse signals is analyzed, and a particular glottal flow derivative waveform is selected as being the most representative of the entire database. Since the glottal pulse waveforms are normalized in both, time and amplitude domains, the selection is based exclusively on the waveform shape. As such, the process takes into account all of the voice source parameters simultaneously. Invariably, the voice source parameters that describe the Characteristic Glottal Pulse Waveform, also describe the true average voice quality of a database. As such, if the *glottal matrix* is extended to contain the glottal pulses of a group of speakers of the same voice quality type then, the estimated Characteristic Glottal Pulse Waveform would represent the most typical glottal flow derivative realization for that particular voice quality type.

In our study, the voice quality profiling experiment is conducted in the following manner. Each voice quality type is represented with a 10 speaker database. For each speaker, a 250 long sequence of glottal flow derivative pulses is used. Voice source estimates are obtained via closed-phase pitch synchronous-inverse filtering of speech using 14th order liner prediction coefficients to model the frequency response of vocal tract. The acoustic stimuli correspond to read speech sentences sampled at 10 kHz. The entire database of 2500 glottal pulse signals is normalized and aligned to form an intra-speaker *glottal matrix*. Finally, the Characteristic Glottal Pulse Waveform is estimated from each voice quality database and subsequently parameterized via a *direct estimation* method described in [2]. Since only one glottal pulse waveform needs to be parameterized for each voice quality type, the results of parameterization are verified and, if judged necessary, corrected manually. For each phonation type, the aperiodicity features and the signal to noise ratio value are obtained by applying CGPWPM on the speech file from which the Characteristic Glottal Pulse Waveform is estimated. Since the CGPW belongs to one of the 10 speakers involved in the extended *glottal matrixes*, only the speech file for that particular speaker is parameterized.

The jitter values are estimated using the perturbation quotient of a sequence according to (5.12). Jitter is defined as the perturbation cycle-by-cycle of pitch, or as the change in the periodicity of the glottal cycle. Shimmer is defined as the perturbation of the glottal excitation amplitude and it is estimated according to (5.13). We have also introduced the perturbation of the glottal flow derivative shape as a voice quality correlate parameter. It is estimated as a perturbation quotient of R_d trajectory (5.14). The perturbation quotients for all of the considered aperiodicity measures are estimated for K=3.

$$Jitter = \frac{100}{N-K+1} \sum_{i=\frac{K-1}{2}+1}^{N-\frac{K-1}{2}} \frac{\left| F(i) - \frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} F(i+k) \right|}{\frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} F(i+k)} \tag{5.12}$$

$$Shimmer = \frac{100}{N-K+1} \sum_{i=\frac{K-1}{2}+1}^{N-\frac{K-1}{2}} \frac{\left| A(i) - \frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} A(i+k) \right|}{\frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} A(i+k)} \tag{5.13}$$

$$RDPQ = \frac{100}{N-K+1} \sum_{i=\frac{K-1}{2}+1}^{N-\frac{K-1}{2}} \frac{\left| R_d(i) - \frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} R_d(i+k) \right|}{\frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} R_d(i+k)} \tag{5.14}$$

Table 5.8:
The voice quality profile in terms of R-parameters, open quotient, speed quotient, aspiration SNR, and perturbation quotients

R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	O_q [10^{-2}]	S_q	SNR [dB]	RDPQ. [%]	Shimmer [%]	Jitter [%]	VOICE QUALITY
2.6	40.8	61.2	1.03	67.2	1.83	32.0	0.75	1.96	0.20	MODAL
1.0	32.1	37.0	0.44	41.3	2.11	20.4	1.73	2.73	0.44	CREAKY
0.8	23.3	34.9	0.29	40.0	2.42	8.9	1.92	2.99	1.94	HARSH
7.5	45.9	71.5	1.79	84.3	1.39	19.1	0.74	2.06	0.18	BREATHY
11.5	51.0	83.7	2.59	99.1	1.27	34.4	0.53	1.85	0.24	FALSETTO
1.6	28.4	45.8	0.51	52.9	2.07	21.5	1.71	2.92	1.83	TENSE
3.1	45.1	74.1	1.38	82.0	1.65	25.2	0.76	1.88	0.31	LAX

Table 5.8 displays our parametric voice quality profile for *modal*, *creaky*, *harsh*, *breathy*, *falsetto*, *tense* and *lax* phonation. Note that the definitions for R-parameters, *open* and *speed quotients* are provided in Chapter 2, Subsection 2.2.2. In Table 5.9 we have also presented the voice quality profiling results obtained from the following references: 1-18 Karlsson and Liljencrants [83], 19-27 Childers and Lee [20], *28-33 van Dinther* [33]. Effective comparison between the results our study and those of Karlsson and Liljencrants, Childers and Lee, and van Dinther is difficult to make as they have reported a range of different parametric descriptions for the identical voice quality types. This is quite understandable as the parameterization results depend on the choice of speech files involved in the profiling database. Voice quality is not constrained to a narrow region of the auditory continuum, and it is expected that a range of parameters would fit under the same voice quality category. Furthermore, phonation types are not mutually exclusive and some of them can combine to form the compound phonations.

Nevertheless, the relative trends in R-parameter values, in relationship to voice quality types, are shared by all four voice quality profiles, including ours. In comparison to modal voice, R_a values tend to be higher for *lax, falsetto*, and *breathy* voice. On the other hand *tense, vocal fry* and *creaky* voices are characterized by abrupt glottal closures, and they generally have lower R_a values. The parameter R_o attains the lowest value for *harsh* voice, whereas the highest R_o values are found in *falsetto* and *breathy* voices. Our results indicate that the glottal pulse skewness is low for *falsetto* and *breathy* voice. On the other end of the scale, *harsh* voice has the highest speed quotient value.

In regards to the aperiodic features, the following trends can be observed. *Harsh* and *creaky* voices exhibit the highest degree of aperiodicity, while *falsetto* and *breathy* voices exhibit the lowest perturbation level. The relative trend in R_d perturbation values, in relation to voice quality types, is similar to those of jitter and shimmer. Generally, results for all three measures are generally consistent. Perhaps the greatest degree of inconsistency was found in *breathy* voice, which displays the lowest jitter among the considered phonation types, second lowest R_d perturbation and the 4th highest shimmer. The results in Table 4.8 confirm that the *signal to aspiration noise ratio* is an effective parameter in distinguishing between the individual voice quality types. The lowest SNR values are obtained for *harsh* and *breathy* voice. Conversely, *falsetto* voice contains the lowest aspiration noise levels among all of the considered voice quality types.

Although, all of the described measures are important voice quality correlates, we are particularly interested in studying the relationship between the glottal shape parameter R_d and voice quality. It has been shown that R_d is one of the most effective parameters for quantifying the shape of a glottal flow derivative waveform with a single numerical value [47]. Low R_d values indicate extremely tight, adducted phonations with low *open quotient* values and abrupt glottal closures, whereas high R_d values describe breathy and abducted phonations with higher O_q and R_a values. Figure 5.42 shows a plot of R_d values, arranged in the ascending order, against the corresponding voice quality types. The data is obtained from Table 5.8.

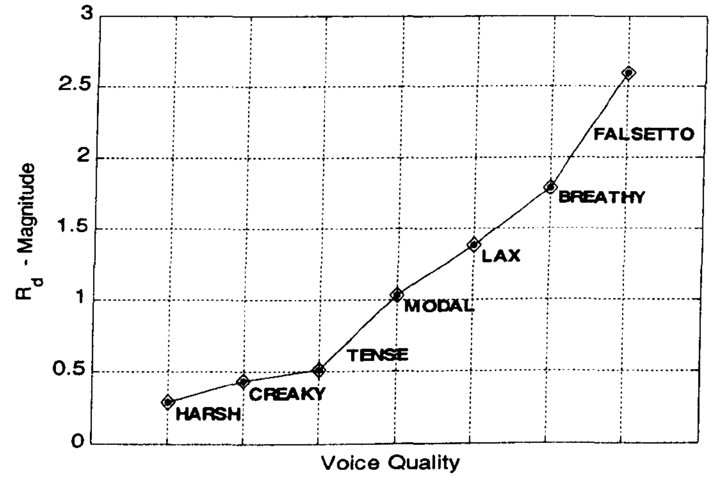


Figure 5.42: The glottal shape parameter, R_d, for various voice quality types

For the lack of better alternative, we have displayed the individual voice quality types as being equidistant from each other. Figure 5.42 describes *harsh* and *falsetto* voices as the two extremes of the voice quality spectrum, whereas *creaky* voice and *breathy* voice appear to represent a slightly more moderate deviation from the neutral voice. The remaining two voice quality types, *tense* voice and *lax* voice, exhibit the least amount of deviation from *modal* voice.

In spite the fact that the assumption about the successive voice quality types being equidistant from each other is rather crude, Figure 5.42 displays a surprising amount of regularity and depicts a clear relationship between the R_d parameter and voice quality. Since the results of our informal listening tests are in accord with the illustrated relationship, we can conclude that the glottal shape parameter R_d, is indeed an effective parameter in quantifying the temporal glottal flow derivative structure and the perceptual characteristics of the voice source signals.

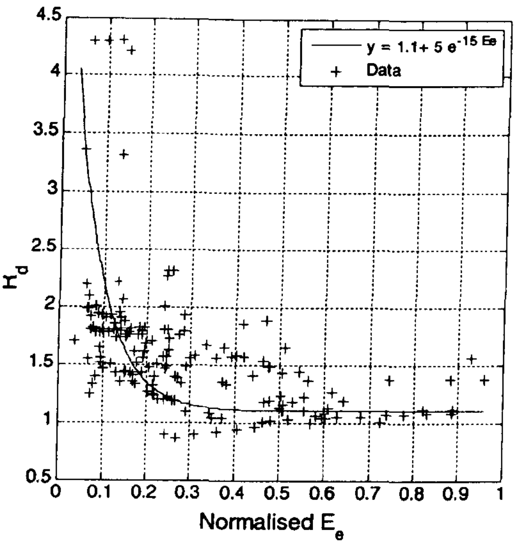


Figure 5.43: Scatter plot R_d vs. E_e and the optimal exponential fit obtained via regression analysis.

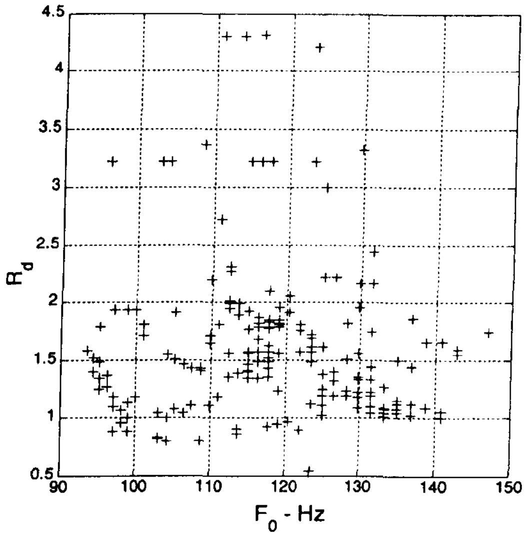


Figure 5.44: Scatter plot R_d vs F_0

Hue-Ling Lu has investigated the extent of correlations between the *LF* parameters of glottal excitation model for baritone recordings of sustained vowels over a range of voice quality types [96]. She has reported that the glottal shape parameter R_d and the normalized glottal excitation strength, $\tilde{E}_e$ are highly correlated, and has suggested that the glottal shape parameter can be approximated as an exponential function of $\tilde{E}_e$.

We have made an attempt to extend her study on natural, read speech acoustic data. Unfortunately, we have very quickly reached a conclusion that the glottal shape parameter R_d, is largely independent of glottal excitation strength. Figure 5.43 shows a scatter plot of R_d vs. $\widetilde{E}_e$ obtained for a male speaker of *modal* phonation over the *voiced* segments of the utterance *"She had your dark suit in greasy wash water all year"*. This data corresponds to an example where one of the highest correlation coefficients has been obtained in our investigation. Nevertheless, it is clear that even in this case, the glottal excitation strength could not be used to adequately predict the glottal shape parameter data. The exponential function saturates at low $\widetilde{E}_e$ values and the majority of R_d data is located at a very narrow region of the exponential function. Our study has also shown that there is even less correlation between the glottal shape and the glottal cycle duration. Figure 5.44 shows a scatter plot of glottal shape vs. frequency of vocal fold oscillations for the same speech file used in producing Figure 5.43. It is evident that the two parameters do not exhibit any observable correlation. Similar results are obtained for all the other investigated files. As such, we have deduced that the shape of the glottal flow derivative waveform can be considered, for all practical considerations, as being independent of both the frequency of vocal fold oscillations and the glottal excitation strength.

We will conclude this section of the chapter by stressing the main advantages of voice quality profiling via the Characteristic Glottal Pulse Waveform Estimation approach. The glottal shape profiling requires only one glottal pulse waveform to be parameterized for each voice quality type. On the other hand, voice quality profiling in terms of aperiodic features and *signal to aspiration noise ratio* requires a single speech file for each phonation. As such, the profiling procedure allows processing of large databases in a very short amount of time. Furthermore, the profiling accuracy increases with the database size and the results of profiling can be controlled manually.

Table 5.9:
R-parameters and frequency values and corresponding voice qualities obtained from: 1-18 Karlsson and Liljencrants [83], 19-27 Childers and Lee [20], 28-33 van Dinther [33]

	R_a [10^{-2}]	**R_k [10^{-2}]**	**R_o [10^{-2}]**	**F_0 [Hz]**	**VOICE QUALITY**
1	2.0	37.7	54.0	126	NORMAL – male
2	5.0	51.7	71.0	246	NORMAL - female
3	2.6	42.5	61.0	102	LOW F_0 - male
4	5.1	41.9	76.0	190	LOW F_0 - female
5	1.5	45.0	56.0	131	MEDIUM F_0 - male
6	4.2	48.0	71.0	250	MEDIUM F_0 - female
7	9.9	32.1	87.0	288	HIGH F_0 - male
8	3.0	48.7	65.0	360	HIGH F_0 - female
9	2.7	40.7	69.0	129	LOW LEVEL - male
10	10.5	57.1	81.0	249	LOW LEVEL - female
11	1.9	45.0	57.0	127	MEDIUM LEVEL - male
12	3.7	51.2	68.0	258	MEDIUM LEVEL - female
13	1.6	37.7	49.0	132	HIGH LEVEL - male
14	1.9	52.2	64.0	257	HIGH LEVEL - female
15	4.6	51.0	65.0	131	BREATHY - male
16	8.1	48.3	79.0	254	BREATHY - female
17	1.3	39.5	41.0	128	PRESSED - male
18	3.2	49.9	71.0	261	PRESSED - female
//					
19	2.1	30.6	64.0	106	MODAL
20	2.5	34.0	71.0	127	MODAL
21	1.5	33.3	68.0	154	MODAL
22	0.8	28.6	63.0	84	SLIGHT VOCAL FRY
23	0.5	25.0	25.0	45	VOCAL FRY
24	13.3	35.1	77.0	344	FALSETTO
25	4.3	43.6	89.0	213	FALSETTO
26	6.8	41.7	68.0	137	BREATHY
27	10.0	44.8	84.0	200	BREATHY
//					
28	0.3	30.0	52.0	110	TENSE
29	0.6	50.0	69.0	110	MODAL
30	2.0	51.0	82.0	110	LAX
31	1.1	25.0	41.0	110	TENSE
32	1.8	37.0	54.0	110	MODAL
33	3.5	43.0	65.0	110	LAX

5.5 Conclusion

Characteristic Glottal Pulse Waveform Parameterization and Modeling offers a novel framework for voice source analysis, parameterization and reconstruction of voice source signals. The proposed method is not constrained to the idealized glottal waveform approximations (e.g. Liljencrants-Fant's model), but instead, relies on the estimates of the Characteristic Glottal Pulse Waveform to obtain the voice source parameters. It uses a set of modified LF parameters and the DTW algorithm to track the nonlinear evolution of CGPW in time. The constrictions of the optimal non-linear time alignment that are employed by DTW algorithm are also extended to voice source parameterization and as such, pathological parameterization is precluded. The design of the method is motivated by the fact that the parameterization performance is linked to the extent of similarity between the glottal model and the analyzed voice source signal. CGPWPM provides the means to model both, the course and the fine structural elements of the glottal flow derivative realizations. Thus, the proposed method enables a study of voice source signal features and their temporal behavior that could not be efficiently or accurately represented by a poorly deterministic glottal flow derivative model. Another useful characteristic of this method is that the parameterization of consecutive glottal pulses across the voiced source signal is referenced to a parametric description of a single glottal flow derivative realization. As such, CGPWPM method can be used very effectively in a semi-automatic manner, as well as in the fully automatic mode. The results of the synthetic dataset based experiments have shown that the performance of the Characteristic Glottal Pulse Parameterization is virtually insensitive to the inaccuracies in the glottal closure instant estimates. On natural speech, CGPWP exhibits a robust performance even for the significant presence of disturbances in the voice source signal estimates. Overall, CGPWP exhibits a superior performance over the standard *fit estimation* and *direct estimation* methods. The results of the voice quality profiling experiment were in a general agreement with those obtained by Karlsson and Liljencrants, Childers and Lee, and van Dinther. We have used the voice quality profiling results to derive a surprisingly simple relationship between the glottal shape parameter R_d, and voice quality.

Chapter 6

Voice Source Reconstruction with applications to Speech Synthesis and Voice Quality Conversion

ABSTRACT

In this chapter, the voice source reconstruction aspect of the Characteristic Glottal Pulse Waveform Parameterization and Modeling system is presented. The novelty of this voice source reconstruction method is that it provides effective means of modeling complex glottal flow derivative realizations. Unlike the Liljencrants-Fant's model, it is able to represent both, the course and the fine structure of the glottal flow derivative waveforms. The results of our speaker identification experiments have established that the fine structural elements of the glottal flow derivative waveforms contain a notable level of speaker-dependant information. CGPWPM is applied under the source-filter model of speech production to develop the speech synthesis and voice conversion methods. The quality of CGPWPM-based speech synthesis is formally evaluated via the Mean Opinion Scores and Degradation Mean Opinion Scores. The DMOS results, obtained for CGPWPM-based synthesis of pathological voices are very high, suggesting that CGPWPM is adaptable enough to cater for very complex and structurally rich forms of glottal flow derivative realizations. With the respect to healthy, *modal* phonations, the MOS values reveal that CGPWPM-based speech synthesis rates highly on the scale of absolute perceptual acceptability. The corresponding DMOS values confirm that the speech is faithfully reconstructed on consistent basis. Triadic listening tests are used to evaluate the performance of a CGPWPM-based voice quality conversion method. The results have shown that the proposed method is able to successfully modify the glottal shape parameters, aspiration noise characteristics, and the aperiodic features of a voice source signal and consequentially, to achieve the desired perceptual effects.

6.1 Introduction

Speech synthesis can be defined as a process of generating an acoustic replica of a speech signal or of a typed text. In 1791 a Wolfgang Von Kempelen invented a mechanical machine that could produce entire phrases in French and Italian. This was one of the pioneering speech synthesis attempts. With the development of the electronic instruments, such as the oscilloscope, the researchers began to gain more insight into speech acoustics. This knowledge enabled Dudley to produce a first electronic speech synthesizer, a 10-channel vocoder that was superior to the available mechanical synthesizers at the time [39]. With the development of digital computers in the 1960, a computer software approach to speech synthesis was made possible. Since then, researchers have used computers to develop and evaluate a range of speech synthesis designs. However, it would be a fair assessment to say that much of the research effort has been focused on the spectral properties of the vocal tract filter, and less attention has been paid to the voice source signals. In the last decade, it has become evident that the development of an accurate and robust glottal excitation model is the key for obtaining a state of the art high quality speech synthesizer. A sophisticated glottal model and the high quality voice source reconstruction can be used under the source-filter model of speech production to enable an intuitive control of acoustic parameters and ultimately, to produce a high quality voice conversion and morphing systems. The following references provide extensive reviews of speech synthesis techniques and their corresponding applications: [7], [16], [122].

In Section 6.2, we will describe the voice source reconstruction aspect of the Characteristic Glottal Pulse Waveform Parameterization and Modeling (CGPWPM) system. A comparative assessment between the Characteristic Glottal Pulse Waveform model and the Liljencrants-Fant's glottal flow derivative waveform model is presented in Section 6.3. Subjective A/B listening test are used to establish which of the two models provides a perceptually more acceptable synthetic speech. These two models are also evaluated in the context of speaker identification. The aim of the speaker identification experiment is to determine whether the fine structural elements of the glottal flow derivative waveforms (which can be modeled by

CGPWPM system) contain speaker-dependant information, as we have initially claimed in Chapter 3. In Section 6.4, the quality of CGPWPM-based speech synthesis is formally evaluated via Mean Opinion Scores and Degradation Mean Opinion Scores. Voice quality conversion experiment is presented in Section 6.5. The experiment demonstrates that the CGPWPM can be used to modify the voice source parameters and to consequentially achieve the desired perceptual effects. Section 6.6 concludes the chapter.

6.2 Voice source reconstruction and speech synthesis

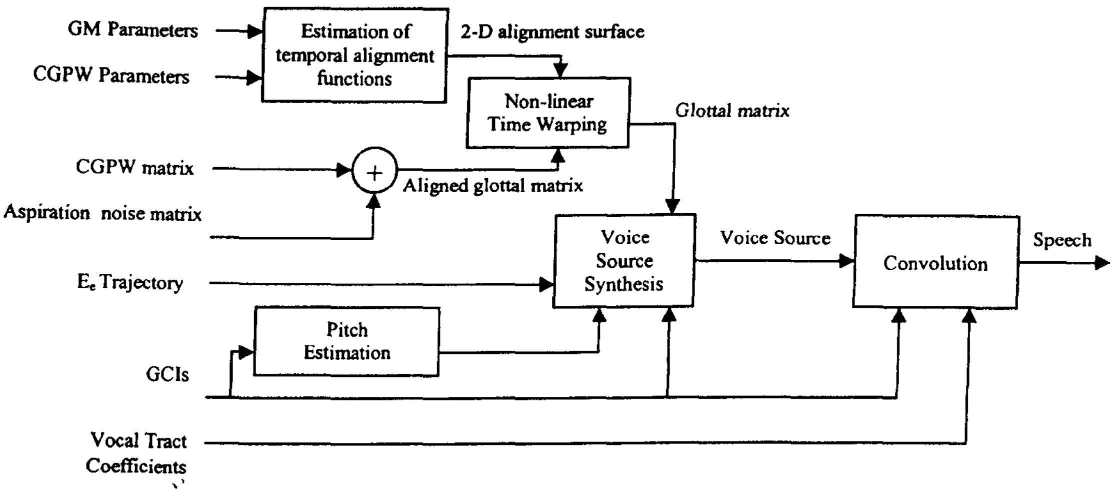


Figure 6.1: Schematic diagram of CGPWPM-based Speech Synthesis

Figure 6.1 shows a schematic diagram of CGPWPM-based speech synthesis system. The speech synthesis involves two stages, voice source reconstruction and voice source convolution with the vocal tract filter. Voice source reconstruction requires estimates of GM parameter trajectories, Characteristic Glottal Pulse Waveform, and aspiration noise envelope.

The GM parameter trajectories and the parameters describing the Characteristic Glottal Pulse Waveform are used to synthesize a 2-dimensinal surface of alignment functions. The estimated alignment functions represent the temporal relationships between the Characteristic Glottal Pulse Waveform and the consecutive glottal pulses in the *glottal matrix*. Let us remind that in Chapter 5, we have defined *glottal matrix* as a set of normalized (both in time and frequency) glottal flow derivative waveforms, arranged in a sequential order as they appear in the voice source signal. We have also shown that when a *glottal matrix* is aligned using the surface of alignment functions towards a Characteristic Glottal Pulse Waveform, it can be decomposed into the two parts, a repetitive sequence of the Characteristic Glottal Pulse Waveforms, and the additive *aspiration noise matrix*. In voice source synthesis, this process is reversed. *Aspiration noise matrix* is synthesized using the estimated aspiration noise envelope, and added to a repetitive sequence of the Characteristic Glottal Pulse Waveform to produce *aligned glottal matrix*. *Aligned glottal matrix* is subsequently non-linearly warped with a surface of synthesized alignment functions to obtain an approximation to an original *glottal matrix*. From this stage on, the process of voice source reconstruction and speech synthesis is rather straight forward. The normalized glottal pulses are extracted from the *glottal matrix*, temporally scaled according to pitch estimates, and arranged in a sequence according to GCI estimates. Subsequently, the glottal pulse sequence is scaled in amplitude with the synthesized excitation strength envelope to produce the reconstructed voice source signal. The excitation strength envelope is generated from GCI and E_e estimates via the monotone piecewise cubic interpolation. Finally, the speech signal is obtained by convolving the voice source signal with the vocal tract filter. It is important to note that CGPWPM is used for *voiced* speech, only. In this thesis, the unvoiced segments of speech are added from the original speech to complete the acoustic signal and thus, enable perceptually-based performance evaluation experiments.

What follows is a description of two essential CGPWPM components, namely, **estimation of temporal alignment functions** and **aligned glottal matrix synthesis**. Since the other elements of CGPWPM-based synthesis are standard signal processing procedures, we deem that no further elaboration is required. Note that some examples of voice source estimates and the corresponding synthesized voice source signals are presented in Chapter 5, in the following figures, Figure 5.13, Figure 5.18, Figure 5.23, Figure 5.28, Figure 5.33, and Figure 5.38.

Synthesis: 2-D alignment surface

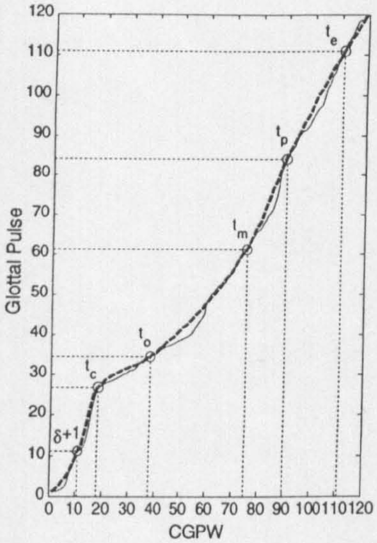


Figure 6.2: An example of DTW alignment function (solid line) and the corresponding synthesized version (dashed thick line). Data corresponds to the first frame of Figure 6.3.

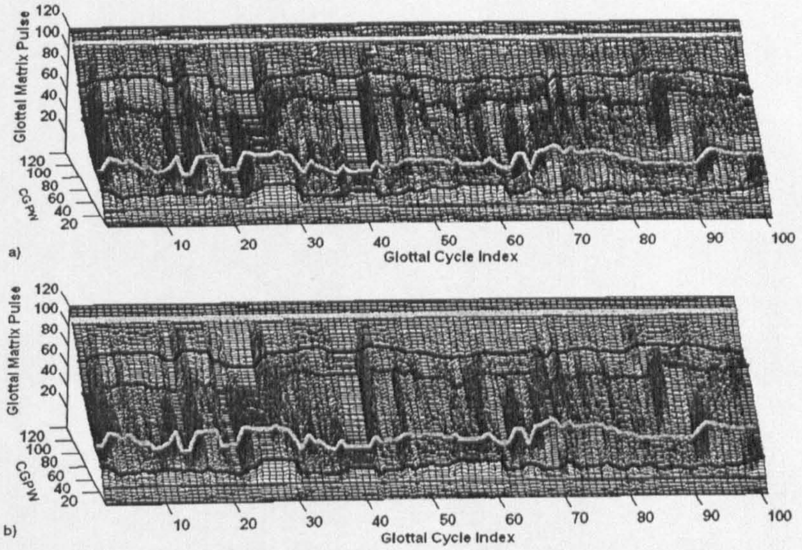


Figure 6.3: a) Adjusted DTW alignment functions and the corresponding b) synthesized alignment functions. The data corresponds to the first 100 glottal pulses in the utterance "We were away a year ago", articulated by to a male speaker (WSJCAM0 database). F_s=10 kHz. The estimated GM trajectories are superimposed over both graphs. CGPWPM is based on L=120, T_N=100, δ =10.

In the process of voice source parameterization, DTW alignment functions are used to establish the temporal relationships between the Characteristic Glottal Pulse Waveform and the other *glottal matrix* pulses. Subsequently, the 2-D surface of alignment functions is used to map the voice source parameters from the CGPW to the other glottal pulses in the *glottal matrix*, and thus obtain a parametric representation for the non-linear temporal evolution of the CGPW through the *glottal matrix*. In the reverse process, the alignment functions are estimated from the GM parameter trajectories and the CGPW parameters using the monotone piecewise cubic interpolation. The following parameters define the interpolation process {1, δ+1, t_c, t_o, t_m, t_p, t_e, L}. Note that the parameters. {1, L} are added to the GM parameter set to ensure that the *boundary constraint* that was originally imposed on Dynamic Time Warping algorithm is satisfied. Figure 6.2 and Figure 6.3 illustrate the process of alignment function synthesis. Figure 6.3 shows a segment of alignment functions surface obtained via a DTW[*] algorithm, together with its synthetic counterpart. The data corresponds to the first 100 glottal pulses of the utterance "We were away a year ago". The speech signal belongs to a male speaker (*WSJCAM0* database), sampled at 10 kHz. CGPWPM is performed for $L=120$, $T_N=100$ and $\delta=10$.

Visual inspection of the two figures, suggests that GM parameters are very effective in describing the temporal evolution of Characteristic Glottal Pulse Waveform through a *glottal matrix*. Only, minor differences between the synthetic and DTW-obtained alignment functions can be observed and it is unequivocal that the general structure of the DTW alignment surface is accurately represented by its synthetic counterpart. Figure 6.3 also demonstrates why it is necessary to include the maximum positive glottal flow derivative instant t_m, in the set of *glottal matrix* (GM) parameters. Without the parameter t_m, the alignment functions could not be adequately synthesized as there would not be a sufficient temporal resolution of the glottal flow derivative structure.

Note that the voice source reconstruction procedure does not require a complete *glottal matrix* frame. In fact only the segments in-between $GCI_1 = \delta + 1$ and $GCI_2 = L - \delta + 1$ are necessary to completely represent glottal pulse cycles, starting and ending with the glottal closure instants. In Figure 6.2 and Figure 6.3, we have also shown the "un-required" elements of the alignment functions in order to make the presentation of results more compatible with those in Chapter 5. In the actual CGPWPM-based speech synthesis, alignment functions are synthesized only for the interval bounded by the two glottal closure instants. The same applies for the *aligned glottal matrix*.

[*] In this chapter, we refer to the adjusted DTW alignment functions as simply DTW alignment functions. The prefix "adjusted" is omitted in order to avoid unnecessary confusion. DTW functions were adjusted in order to remove the effect of GCI estimation errors on the voice source parameterization performance.

Synthesis: aligned glottal matrix

The *aligned glottal matrix* is synthesized as follows:

$$\mathbf{G}'(i,n) = \mathbf{A}(i,n) - \mathbf{M}_{CGPW}(i,n) \tag{6.1}$$

where $\mathbf{G}'$ and $\mathbf{A}$ denote *aligned glottal matrix* and *aspiration noise matrix*, respectively. $\mathbf{M}_{CGPW}$ is an N-by-L matrix containing N repetitions of CGPW. N and L signify the number of *glottal matrix* pulses and the *glottal matrix* frame length, respectively. *Aligned glottal matrix* synthesis is essentially a reverse process to that used in the waveform decomposition block, see Chapter 5. *Aspiration noise matrix* is generated in the following manner. An N-by-L matrix of random Gaussian noise with unit variance and zero mean is formed. Subsequently, each noise matrix frame is modulated by the estimated aspiration noise envelope to produce *aspiration noise matrix.* This aspiration noise model is conceptually identical to the one presented in Chapter 2, Figure 2.18. However, in CGPWPM, the aspiration noise is even further integrated with the glottal pulse waveforms. In fact, the energy distribution of the turbulent components over a glottal cycle is being treated in the same way as the glottal pulse waveform itself. They are both non-linearly warped on a cycle-by-cycle basis according to the estimated voice source parameter trajectories.

6.3 Comparative evaluation: LF vs. CGPWPM

In this section, we present a comparative assessment between the Characteristic Glottal Pulse Waveform model and the Liljencrants-Fant's glottal flow derivative waveform model. In Subsection 6.3.1, subjective A/B listening tests are used to establish which of the two models produces perceptually more acceptable synthetic speech. In Subsection 6.3.2, these two models are evaluated in the context of speaker identification. The speaker identification experiment is used to establish whether the fine structural elements of the glottal flow derivative waveforms (which can be modeled by CGPWPM system) carry speaker-dependant information.

6.3.1 Subjective A/B listening tests

We have conducted a subjective *A/B* test to establish which of the two glottal flow derivative pulse models, LF or CGPWPM, produces a more natural synthetic speech. In this instance, the term "naturalness" is simply defined as human sounding. The comparative evaluation is performed on a subset of the *WSJCAM0* database. The test data includes 20 read speech sentences sampled at 10 *kHz*, 10 of which are of male speakers, and 10 of female speakers. The test database is subsequent expanded with the corresponding 20 LF-based synthetic speech files and 20 CGPWPM-based synthetic speech files. In order to enable a fair comparison between the two models, the LF synthesis of voice source signals is based on the parameters obtained via Characteristic Glottal Pulse Waveform Parameterization. LF synthesis is based on the following set of equations (2.12) and (2.13). Note that the Characteristic Glottal Pulse Waveform Parameterization is performed for the following parameter values; L=120, T_N=100 and δ =10. 40 listeners have participated in the test.

For each speaker, the two synthetic speech files are compared with each other, and with the original speech file, excluding the same sentence comparison. This results in 3 possible pairs per speaker. Each pair is presented twice in forward and twice in reverse order. As such, the listening presentation contains a total of $3\times4\times20=240$ pairs. The order of pairs is randomized and the listening test is presented as follows. A 500 *Hz* tone is used to alert the listeners that the speech material is to follow. Each pair of sentences (*A, B*) is presented twice consecutively as {(*A, B*), (*A, B*)} with a pause of 2 seconds in-between repetitions. Also, a 1-second pause is inserted in-between each *A* and *B*. Prior to the presentation of a next pair, listeners are provided with a 4-second interval to form and report their preference scores. Undecided or equal scores were not available options. Halfway through the presentation, the listeners were provided with a 10 minute break. In order to ensure the validity of the collected data, the performance of each listener is graded. Based on the scores obtained for the forward and the reverse presentation of each pair, one can evaluate the ability of a listener to make the required discriminations in a consistent manner. Four listeners were found to have a consistency level in their preference choices below the threshold level of

75%. Since these four listeners were not able to discern adequetly the levels of naturalness in the presented speech tokenss, their results are not included in this study. The average consistency level among the remaining speakers is 85.8 %. Their results are summarized in Table 6.1.

Table 6.1:
Preference scores indicating how many times one type of acoustic data is preferred over another, in term of perceived levels of naturalness.

	Average Preference scores [%]	
CGPWPM vs. LF	66.7	33.3
CGPWPM vs. Natural	47.2	52.8
LF vs. Natural	25.0	75.0

The average preference scores demonstrate that the subjective quality of the CGPWPM-based speech synthesis exceeds that of Liljencrants-Fant's model, and it is just below the natural speech itself. The fact that CGPWPM-based synthetic speech is at times preferred to natural speech is primarily attributed to the higher levels of regularity in the consecutive glottal flow derivative waveforms that is imposed by CGPWPM synthesis. This issue is further clarified in Section 6.4. In addition, we have found that 83.3 % of the inconsistencies in the listeners' choices involve the sentence pairs consisting of CGPWPM-based synthetic speech and the natural speech. This is a clear indication of the extent of difficulty that the listeners experienced in discriminating the natural and the CGPWPM-synthesized speech.

Note that the results of the subjective A/B listening test do not significantly deviate between genders, and as such, we did not produce separate reports for male and female speakers.

6.3.2 Speaker identification

In this section of the chapter, we want to ascertain whether the extra features in the CGPWPM-synthesized voice source signal contain any significant levels of speaker-dependant information. In relation to this objective, we will conduct an experiment using an established speaker identification system, rather than attempt to develop an optimal solution specifically for the voice source signals.

Over the years, researchers have made some attempts to establish the levels of speaker-dependant information in the voice source signals, e.g. [64], [135]. However, these experiments were based on the LPC-residuals, rather than on the voice source estimates obtained via the pitch-synchronous deconvolution of voice source and vocal tract from speech. On the other hand, in [114], the authors have developed a parametric representation of the fine glottal flow derivative structure, which was together with the coarse glottal pulse features (LF-model), evaluated in the context of speaker identification experiments. However, they have made no attempt to include some of the fine structural elements of the glottal pulse waveforms in the process of voice source reconstruction.

Preliminary evidence

Before we describe the speaker identification experiment, we will present some preliminary evidence that suggests that speaker-dependant information is indeed carried by the voice source signals. The evidence is based on the distributions of voice source parameters and on the *phase-plane* analysis of the Characteristic Glottal Pulse Waveform estimates.

In Figure 6.4, we have presented a set of voice source parameter distributions for two female speakers. Parameters include *open quotient*, *speed quotient*, *R-parameters* of the glottal pulse model and the glottal pulse shape parameter R_d. The results were obtained on 2 seconds of voiced speech data, from each speaker. Characteristic Glottal Pulse Waveform Parameterization is performed for the following parameter values: L=120, T_N=100 and δ =10. Visual inspection of the histograms in Figure 6.4 clearly indicates that two speakers produce considerably different sets of voice source parameter distributions. The distributions related to Subject 1 are noticeably smoother than those of Subject 2. Withstanding the *speed quotient*, the parameters of Subject 1 also exhibit more symmetrical distributions. The glottal pulse shape parameter related to Subject 1 decays gradually on either side of median distribution value, whereas for Subject 2, R_d drops sharply on the left side and falls somewhat more gradually on the right side. Another obvious difference is that Subject 2 is able to maintain a constant R_a value for a range of vocal fold realizations, while Subject 1's R_a exhibits quite strong deviations from its median value. Both speakers seem to have reasonably well defined and most importantly very characteristic voice source parameter distributions.

In Figure 6.5, we have shown the results of phase-plane analysis for six female speakers. For each speaker, two Characteristic Glottal Pulse Waveforms (CGPWs) are obtained from two linguistically different read speech sentences. Subsequently, phase-plane plots of the two CGPWs are superimposed, to enable a visual comparison. As we have hoped, the phase-plane plots do not show significant variation with the respect to the different linguistic contents, but instead vary extensively across the speakers. It is important to note that the general shape of the phase-plane plots is strongly influenced by the fine temporal structure of the Characteristic Glottal Pulse Waveforms. The results in figure 6.5 would indicate that the fine glottal flow derivative structure contains speaker dependant information. As such, these results constitute compelling evidence that CGPWPM-based voice source synthesis might carry a significantly higher content of speaker-dependant information than the Liljencrants-Fant's model, which is only able to represent the coarse glottal flow derivative structure and leaves the fine glottal flow derivative structure unrepresented.

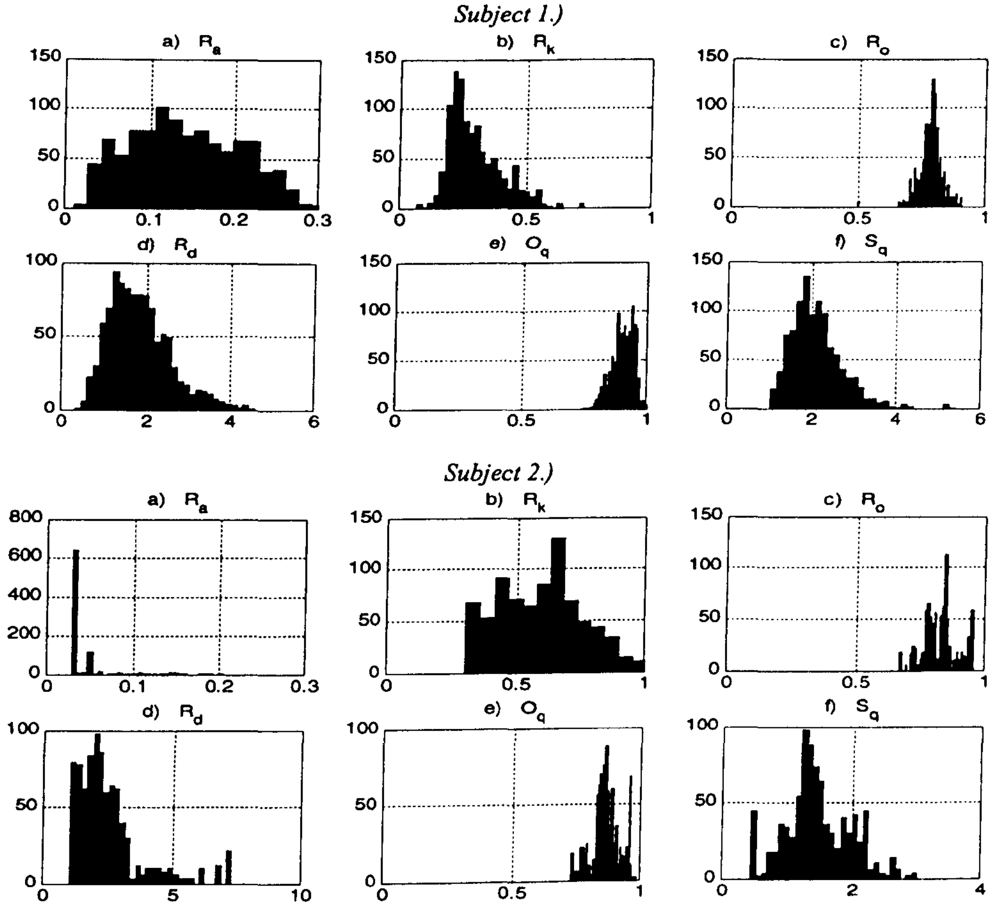


Figure 6.4: Comparison of voice quality parameter histograms for two female speakers. Parameters include R-parameters of the glottal pulse model, glottal pulse shape parameter R_d, open quotient and speed quotient.

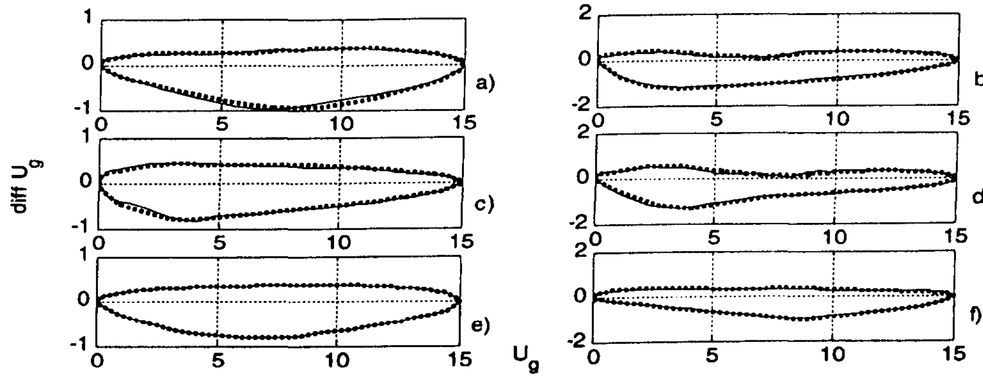


Figure 6.5: Two overlaying phase-plane plots of Characteristic Glottal Pulse Waveforms (CGPWs), obtained from linguistically different read speech sentence, for six female speakers.

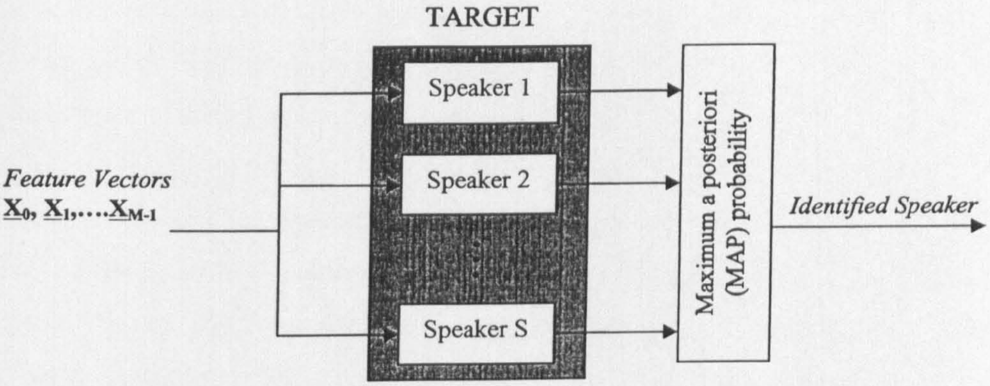

TARGET
Feature Vectors
$\underline{X}_0, \underline{X}_1, \ldots \underline{X}_{M-1}$
Speaker 1
Speaker 2
Speaker S
Maximum a posteriori (MAP) probability
Identified Speaker

Figure 6.6: Schematic diagram of the GMM based speaker identification system

Speaker identification experiment

The speaker identification experiment is performed with a Gaussian Mixture Model (GMM)[†] illustrated in Figure 6.6. 32 Gaussian mixtures are used. Each Gaussian mixture component is assumed to be characterized by a diagonal covariance matrix. Maximum Likelihood (ML) parameters are estimated using the Expectation-Maximization (EM) algorithm [31] with 10 iterations.

The speaker identification is evaluated on four types of acoustic signals:

- Speech
- Glottal flow derivative estimate
- LF-Synthesized glottal flow derivative
- CGPWMP - Synthesized glottal flow derivative

The speaker identification experiments are performed on a subset of *WSJCAM0* database involving 80 male and 80 female speakers. A Gaussian Mixture Model is associated to each speaker. Segments of 15 and 5 seconds long data are used for training and testing of the

[†] GMM is conceptually similar to the widely used Hidden Markov Model (HMM). The main difference between the two systems is that GMM ignores the temporal information of the acoustic observation sequence.

system, respectively. A 23-mel-cepstra representation is used for each type of acoustic signal. The experiment setup is comparable to that used in [114]. The glottal flow derivative estimates are obtained via closed-phase, pitch-synchronous inverse filtering of speech signals. In the process of blind deconvolution, the vocal tract frequency response is modeled with 14 coefficients obtained through a covariance based linear prediction analysis. In order to enable a fair comparison between the two glottal models, the LF synthesis of voice source signals is based on the parameters obtained via the Characteristic Glottal Pulse Waveform Parameterization. Note that the Characteristic Glottal Pulse Waveform Parameterization and Modeling is performed for the following parameters $L=120$, $T_N=100$, and $\delta=10$. The results of the experiment are summarized in Table 6.2.

Table 6.2:
The average speaker identification rate for speech signals, voice source estimates, LF-based voice source reconstructions, and CGPWPM-based voice source reconstructions

Identification signal	**Male [%]**	**Female [%]**	**Average [%]**
Speech	100.0 %	100.0 %	100.0 %
Voice source estimate	95.6 %	94.4 %	95.0 %
Synthetic voice source - LF	54.4 %	53.3 %	53.9 %
Synthetic voice source - CGPWPM	65.6 %	62.2 %	63.9 %

For us, the most significant result is that the average identification rate obtained for CGPWPM-based voice source synthesis is 18.6 % higher than the identification rate corresponding to the Liljencrants-Fant's glottal pulse model. Liljencrants-Fant's model can not capture any of the fine structural elements of the glottal derivative waveforms and at best, it is only able to describe a general shape of the glottal flow derivative realizations. Thus, the average speaker identification rate for LF model was the lowest among the considered acoustic waveforms. Nevertheless, the overall results demonstrate that the voice source signal contains a considerable amount of speaker-dependant information. We used the word "considerable" rather loosely here. We certainly do not mean to imply that the voice source signal has a comparable level of speaker-dependant information as the vocal tract features. Instead, the speaker identification rates are high enough to illustrate the importance of adequate voice source representation in the source-filter based speech synthesis or any other relevant applications.

The substantial differences in the identification rates between the CGPWPM-based voice source reconstruction and the actual voice source estimates indicate that there is still a large space for improvement in the modeling of voice source signal. However, the gap might not be as large as these experiments would suggest. Baring in mind that for speech signals, 100 % identification rates are realized even for much larger databases than the one considered here. As such, it is entirely possible that the high identification rate obtained on the voice source estimates is to a large extent a tribute to the imperfect voice source deconvolution from vocal tract. The vocal tract artifacts in the voice source signal can give rise to significant improvements in the identification rates. On average, a slightly higher identification rates are achieved for male speakers than for female speakers. However, given that the discrepancy between the female and male identification rates is only marginal and that the results are obtained on a relatively small dataset, we do not think that we are in a position to form a valid judgment on gender dependency of speaker identification rates.

Having established that CGPWPM is perceptually a more acceptable glottal flow derivative model, and having demonstrated that it is able to retain more of the speaker-dependant information from the voice source estimates, we will proceed with the formal evaluation of CGPWPM-based speech synthesis.

6.4 Speech quality assessment for CGPWPM via MOS and DMOS

In this section we will evaluate the absolute perceptual acceptability and the quality of the CGPWPM-synthesized speech. The performance evaluation is based on Mean Opinion scores (MOS) and Degradation Mean Opinion Scores (DMOS).

The speech quality assessment is performed on a subset of *WSJCAM0* database involving 10 male read speech sentences and 10 female read speech sentences, sampled at 10 *kHz*. We have also evaluated the quality of speech synthesis on a database of 5 male and 5 female pathological speech files. Pathological files are listed in Table 6.3. Note that the evaluation of pathological speech synthesis is constrained to DMOS results, as MOS values are inherently unreliable for pathological voices. The Characteristic Glottal Pulse Waveform Parameterization and Modeling is performed for the following parameter values; $L=120$, $T_N=100$ and $\delta=10$.

Table 6.3:
Pathological Speech database

Index	Diagnosis	Gender
1	True vocal cords (TVC) contact ulcer	Male
2	Hoarse unilateral TVC carcinoma	Male
3	Vocal fry	Male
4	Bilateral Paralysis of TVC	Male
5	Hoarse	Male
6	Left TVC unilateral paralysis	Female
7	Pathologically breathy	Female
8	Hyper functional	Female
9	Enlarged vocalus muscle	Female
10	Right TVC unilateral paralysis	Female

Mean Opinion Score

Mean Opinion Score (MOS) is the most widely used speech quality assessment method. It is based on the Absolute Category Rating (ACR). Listeners are asked to grade the overall quality of the acoustic signals using the five categories shown in Table 6.4. The MOS value is defined as the mean of rating values obtained from a group of listeners. Note that the listeners are not presented with the reference signals and the speech quality assessment is based exclusively on the listeners' perceptual impression of the speech quality. Since the individual scales of goodness are inherently varied [143], the unconstrained subjective assessment is prone to the listeners' bias. This bias can be minimized by collecting the results from a large number of listeners. As such, we have asked 40 listeners to participate in the assessment of CGPWPM-based speech synthesis‡.

Degradation Mean Opinion Score

In Degradation Mean Opinion Score (DMOS) evaluation, listeners are asked to grade the level of degradation between the synthetic speech tokens and the corresponding original (reference) signals. As such, DMOS belongs to a group of Degradation Category Rating (DCR) measures. Unlike MOS results, DMOS results are not necessarily linked to the listeners' absolute acceptability of synthetic speech, but instead, they reflect the quality of speech reproduction. Thorpe and Shelton compared the MOS results with the DMOS results obtained for eight codecs with the dynamic background noise [136]. They have concluded that DMOS values are particularly useful in the instances when MOS results have a compressed range or/and are near the floor or ceiling value. For the same reason, CGPWPM is evaluated using the MOS and DMOS measures in conjunction. DMOS rating system is illustrated in Table 6.5.

Table 6.4:
MOS and the corresponding Speech Quality

Rating	Speech Quality Description
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Table 6.5:
DMOS and the corresponding Degradation Levels

Rating	Degradation Level Description
1	Very annoying
2	Annoying
3	Slightly annoying
4	Audible but not annoying
5	Inaudible

Results

Figure 6.7 shows the Mean Opinion Scores and Degradation Mean Opinion Scores for each of the 20 speech files in the test database (healthy voice). The average results are presented in Table 6.6. MOS values reveal that CGPWPM-based speech synthesis rates highly on the

‡ This particular number of listeners is recommended by ITU-T [76].

scale of absolute perceptual acceptability, while DMOS values suggest that the speech is faithfully reconstructed on consistent basis. The average MOS and DMOS values across the database are 4.45 and 4.55, respectively. Not a single speech reconstruction exhibits *annoying* degradation levels and only one speech file was judge below the second highest level of absolute perceptual acceptability. Overall, we can conclude that CGPWPM-based speech synthesis offers a high quality performance that is consistent across speakers and gender. Note that the MOS and DMOS results for LF-based speech synthesis are not presented as they were not comparable to those of CGPWPM-based speech synthesis, especially in terms of DMOS evaluation.

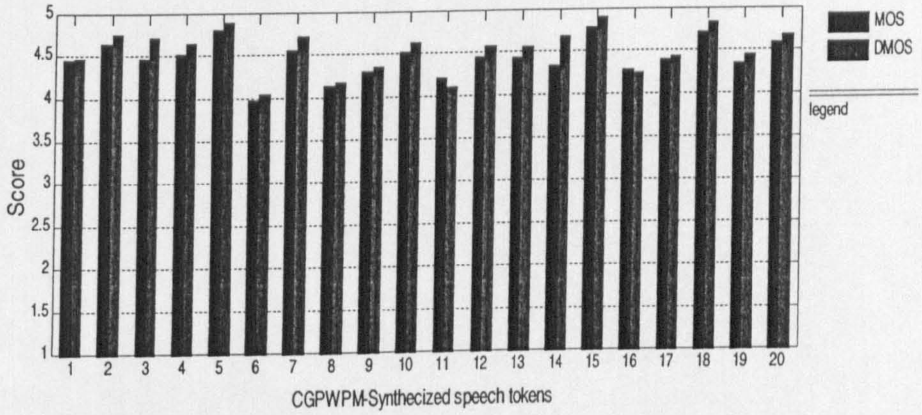


Figure 6.7: Mean Opinion Scores and Degradation Mean Opinion Scores for CGPWPM-Synthesized Speech. Speech tokens labeled 1-10 correspond to

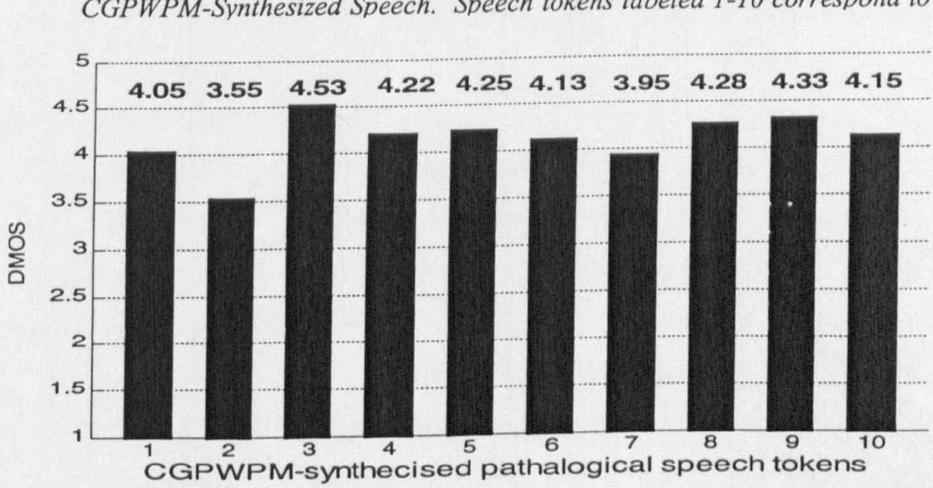


Figure 6.8: Degradation Mean Opinion Scores for CGPWPM-Synthesized Pathological Speech.

Table 6.6:
Gender dependant MOS and DMOS values for the normal and pathological speech

CGPWPM synthesis	MOS	DMOS
Male speech	4.44	4.54
Females speech	4.46	4.55
Pathological speech male	NA	4.12
Pathological speech female	NA	4.17

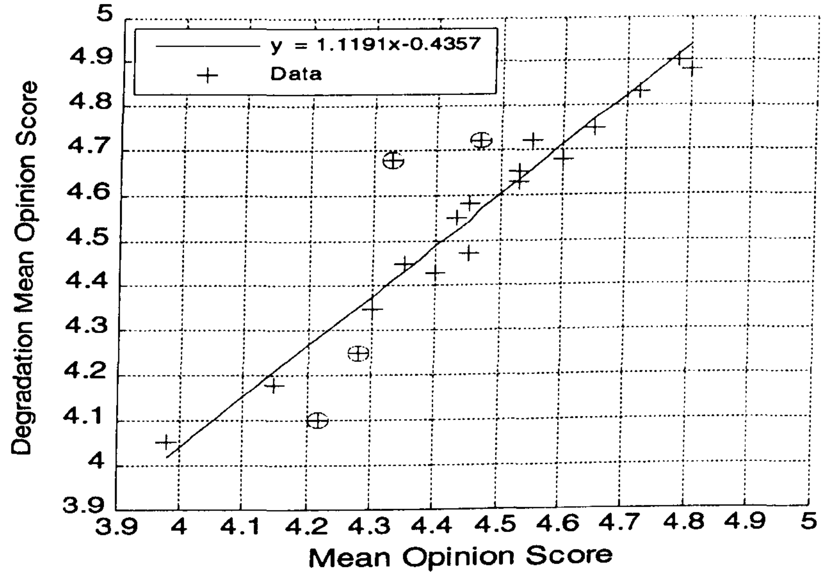


Figure 6.9 Scatter plot of DMOS vs. MOS and the optimal linear fit obtained via regression analysis; the outliers are circled.

The relationship between DMOS and MOS results is further examined in Figure 6.9, where a DMOS vs. MOS scatter plot and the optimal linear fit obtained via regression analysis are displayed. With respect to the outliers above the regression fit (speech tokens indexed as 3 and 14 in Figure 6.7), the interviews[§] that were conducted after the listening test reveal that a vast majority of listeners would rate the original speech tokens between *good* and *excellent* quality. This explains why in spite of high quality reconstructions, lower MOS values were reported. On the other hand, the outliers below the regression fit (indexed as 11 and 16 in Figure 6.7) exhibit substantially lower DMOS than MOS values. On closer examination, we

[§] The entire speech data base was made available to listeners for more thorough examination.

have found that the two synthetic speech tokens are perceptually more acceptable than their natural counterparts, due to a higher degree of regularity in the CGPWPM-synthesized voice source signal. Note that in the process of voice source synthesis, each glottal pulse waveform is represented by CGPW that is temporally warped in a non-linear manner according to voice source parameters. Since the warping function is constrained by a range of conditions, such as: monotonicity, limited extent of deviation from the diagonal form, etc., a relatively high level of regularity in the voice source signal emerges as an inherent property of CGPWPM-based synthesis. Some listeners perceived this type of speech enhancement as imperfect speech reconstruction, and correspondingly, they have reported lower DMOS scores. In the light of this investigation, we believe that the four outliers do not accurately reflect CGPWPM performance. When they are removed, the Pearson product-moment correlation coefficient, defined in (11), value increases, from 0.929 to 0.988. The latter value indicates a strong DMOS *vs.* MOS correlation, and correspondingly it validates the overall quality of experimental results.

$$r = \frac{N\sum_{i=1}^{N} X(i)Y(i) - \left(\sum_{i=1}^{N} X(i)\right)\left(\sum_{i=1}^{N} Y(i)\right)}{\left[\left(N\sum_{i=1}^{N} X(i)^2 - \left(\sum_{i=1}^{N} X(i)\right)^2\right)\left(N\sum_{i=1}^{N} Y(i)^2 - \left(\sum_{i=1}^{N} Y(i)\right)^2\right)\right]^{1/2}} \tag{6.2}$$

The DMOS results corresponding to the pathological speech database are presented in Table 6.6 and Figure 6.8. The level of pathological speech reconstruction was rated between *slightly annoying* and *audible but not annoying*. In most cases, CGPWPM enhanced the quality of pathological speech. Nevertheless, there were cases, such as *Hoarse unilateral TVC carcinoma*, where CGPWPM could not account for the extent of variation in the acoustic characteristics of the glottal flow derivative waveform realizations. We believe that at least two Characteristic Glottal Pulse Waveforms would be necessary to be able to adequately represent this particular voice source signal. Overall, DMOS results for pathological voices are very high, suggesting that CGPWPM is adaptable enough to cater for very complex and structurally rich forms of glottal flow derivative realizations.

Having described CGPWPM as a high quality voice source parameterization and synthesis system, in the following section, we will demonstrate that the proposed method can also provide an effective platform for voice source conversion and voice quality control.

6.5 Voice quality conversion

High quality voice source modification is a subject of considerable importance in a range of applications, such as: text-to-speech synthesis, psychoacoustics experiments, speaker normalization-based speech recognition, etc. In this section, we describe a technique, based on the Characteristic Glottal Pulse Waveform Parameterization and Modeling, for automatic conversion of one speaker's voice quality to another's. CGPWPM is used in the analysis of the *source* and *target* signals, and it is subsequently used to synthesize the modified speech. The text-independent voice quality conversion is accomplished by representing the voice quality as a multi-dimensional space that can be modified to produce the desired perceptual effects. The multi-dimensional space encodes the glottal pulse waveform and aspiration noise properties, the average frequency of vocal fold vibrations and aperiodicity features, namely shimmer and jitter.

Let us remind that the voice source synthesis via CGPWPM is defined for the following parameters, Characteristic Glottal Pulse Waveform, aspiration noise envelope, GM parameter trajectories, gain contour, pitch contour and the corresponding glottal closure instants.

In the proposed method of voice quality conversion, the following set of parameters are required from the *target* speaker: aspiration noise envelope, Characteristic Glottal Pulse Waveform and its parameters, the average frequency of vocal fold vibrations, and finally, the perturbation coefficient values for the pitch and glottal excitation strength contours.

Conversion of the glottal pulse waveforms and aspiration noise properties

Since, the temporal relationship between the *glottal matrix* (GM) parameters and the parameters describing the CGPW is related through a surface of alignment functions, a new set of *glottal matrix* parameter trajectories for the modified voice source signal can be obtained via the following algorithm:

$$P_i^M = \hat{W}_i^S(P_{CGPW}^T) \qquad P_{CGPW}^T = \{\delta+1,\ t_c,\ t_o,\ t_m,\ t_p,\ t_e\} \tag{6.3}$$

where P_i^M denotes the GM parameters for the i^{th} glottal pulse in the modified *glottal matrix.* $\hat{W}_i^S$ refers to the adjusted *source* alignment function relating CGPW to the i^{th} glottal pulse in the *source glottal matrix.* P_{CGPW}^T corresponds to the parametric description of *target's* CGPW. As in the voice source synthesis, P_i^M values are obtained through the monotone piecewise cubic interpolation of alignment functions.

An example of glottal waveform conversion is given in Figure 6.10. In this example the *source* corresponds to a male voice with *modal* phonation. A male voice of *tense* phonation is used as the *target.* The voice quality parameters for the *source* and *target* speaker are presented in Table 6.7, under M_1 and T_1, respectively. Both, the *source* and the *target* data correspond to the utterance "We were away a year ago". In order to enable an effective visual comparison, the trajectories of the *target* speaker are uniformly compressed such that the beginning and the end points of the *target* trajectories precisely align with those of the *source.*

Figure 6.10 shows that the modified GM parameters are actually somewhere in-between those of *source* and *target* speakers. To be more precise, the average values of the modified GM parameters correspond to the *target* speaker, whereas the nature of temporal evolution is retained from the *source.* Based on our experience, we believe that the temporal behavior of the voise sorce signal, with the respect to the general shape of glottal flow derivative waveforms, is at least to some extent related to the linguistic layer of speech communication. As such, we have purposefully aimed to preserve this aspect of a voice source signal from the *source* speaker. In addition, our informal listening tests have shown that when the GM parameter trajectories are obtained in this manner, as opposed to using the GM parameter trajectories directly from the *target* speaker, a higher quality speech is obtained.

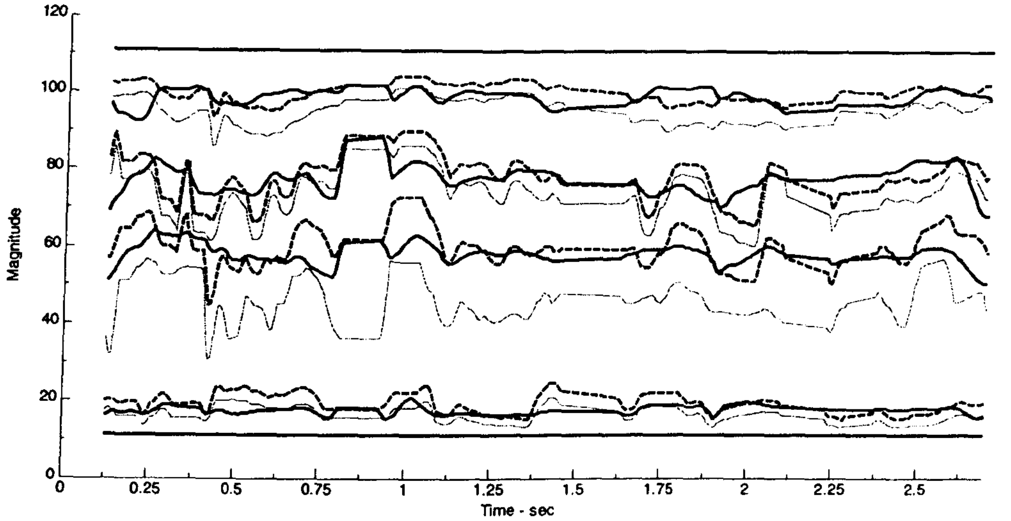


Figure 6.10: Glottal waveform conversion example; a) source: male modal voice (thin line); b) target: male tense: (thick line); c) modified: source to target (thick dashed line); The trajectories correspond to GM parameters: in increasing order, $P = \{\delta+1, t_o, t_o, t_m, t_p, t_e\}$. Both, the source and target data correspond to the utterance "We were away a year ago."

Conversion of glottal excitation strength and frequency of vibration

When segment-by segment changes have been imposed on the voice source parameters, the perception of "business" or distortion can arise in the LPC speech as a result of discontinuities between the successive speech segments. In order to reduce the effect of these discontinuities, especially for the *voiced/unvoiced* transitions, a modified excitation strength, E_e contour is obtained, such that the original (*source*) energy envelope of the speech waveform is preserved.

The pitch trajectory modification is performed by a simple procedure described by Childers in [24]. The modified pitch period contour, *T* is essentially obtained from the original *GCI* instants. The original contour is shifted upwards or downwards to satisfy the average pitch value of the *target* specifications. For each *voiced* segment a new vector of glottal closure instants is obtained as:

$$GCI_i = GCI_{i-1} + T(GCI_{i-1}) \tag{6.4}$$

where, GCI_i denotes the i^{th} instant of glottal closure. *T* corresponds to the modified pitch period envelope defined for each sample over the duration of the voice source segment. Note that the GCI_l value is initialized to the starting point of the *voiced* segment.

Conversion of aperiodicity features: shimmer jitter

The vocal tremor measures, jitter and shimmer are added to the pitch period and excitation strength contours, respectively, in form of a random Gaussian white noise with the standard deviation equal to the specified *target* perturbation coefficients. Prior to adding vocal tremor, the respective contours are preprocessed with the 5th order median filter to remove the turbulent components belonging to the *source* speaker. Note that the perturbation of glottal shape parameter R_d, is not included in this voice conversion method as that would require prediction of voice source parameters from R_d values, which is not always accurate.

6.5.1 Verification experiment

In the following experiment we will evaluate the perceptual effectiveness of the proposed voice conversion method. The experiment is performed on a database of 6 female and 6 male speakers. Each gender is represented by 2 speakers of *modal* phonation, 2 speakers of *lax* phonation, and 2 speakers of *tense* phonation. All of the acoustic stimuli correspond to the utterance: “We were away a year ago”. Parametric voice quality description for the female and male speakers is provided in Table 6.7 and Table 6.8, respectively. Note that R_d

perturbation values are included in the voice quality description for the sake of completeness, only. They are not used in this experiment. The test database is subsequently expended to include the modified stimuli. The *modal* voices were used as the *sources*, while *lax* and *tense* voices were used as the *targets* of voice quality conversion. In Table 6.7 and Table 6.8, M_i denotes *modal* phonation, where $i \in \{1,2\}$ corresponds to speaker's index. The *tense* and *lax* voices are denoted as T_i and L_i, respectively. We will represent the modified stimuli as $\tilde{T}_{ij}$ and $\tilde{L}_{ij}$, where i corresponds to the *source* speaker's index (*modal* phonation always) and j denotes the *target* speaker's index (*tense* or *lax* phonation). As an example, $\tilde{T}_{21}$ denotes the acoustic stimulus that is produced when M_2 voice is modified according to the voice quality description of T_1 voice.

Table 6.7:
Voice quality correlates evaluated for 6 female speakers

	R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	SNR [dB]	RDPQ [%]	Shimmer [%]	Jitter [%]	F0 [Hz]
M_1	2.43	34.96	79.23	1.06	21.05	0.55	1.96	0.22	207.04
M_2	2.23	37.54	81.14	1.15	19.45	0.73	1.73	0.18	227.27
T_1	1.62	32.31	61.93	0.72	19.55	1.12	1.99	1.13	172.12
T_2	1.28	27.88	62.00	0.61	20.6	0.81	2.91	1.20	196.85
L_1	2.59	50.92	90.31	1.80	21.02	0.48	1.85	0.19	204.75
L_2	4.21	42.06	80.43	1.47	20.08	0.72	1.92	0.82	247.95

Table 6.8:
Voice quality correlates evaluated for 6 male speakers

	R_a [10^{-2}]	R_k [10^{-2}]	R_o [10^{-2}]	R_d	SNR [dB]	RDPQ [%]	Shimmer [%]	Jitter [%]	F0 [Hz]
M_1	2.46	34.60	66.19	0.92	18	0.5	3.50	0.20	129.87
M_2	3.81	33.71	65.12	0.99	17	0.67	1.32	0.19	123.46
T_1	1.65	26.67	51.30	0.53	14	2.4	3.42	0.25	89.69
T_2	1.17	25.31	65.92	0.57	15.4	3.1	3.60	0.35	98.91
L_1	3.52	42.95	75.41	1.37	17.5	0.7	0.88	0.41	151.52
L_2	3.26	47.63	78.77	1.55	19	1.27	2.47	0.58	141.64

The verification experiment is based on the subjective triadic comparison listening tests. 10 listeners have participated in the experiment. The experiment is conducted under two settings, setting (I) and setting (II). In setting (I), the first two entries of a triplet {A, B, X} consist of one *lax* and one *tense* voice stimuli. The final entry could contain any voice stimuli other than *modal* voice and the stimuli that are already present in the triplet. In setting (II), the triplets consist of either *tense* and *modal* voices or *lax* and *modal* voices. As such, setting (II) examines smaller perceptual distances, while setting (I) examines larger perceptual distances. However, it is important to note that the perceptual distances between *lax* and *tense* voices are still small in relation to the size of the voice quality continuum, see Figure 5.42. Each triplet is presented twice, as $\{A, B, X\}$ and $\{B, A, X\}$. The order of triplets is randomized and presented in a similar fashion as in the *A/B* preference test described in Section 6.3.1. The listeners were asked to report which one of the two acoustic stimuli, *A* or *B*, has a perceptually closer vocal texture to that of stimulus *X*. The majority of listeners had at least some experience with the subject of voice quality. Those that did not have any experience in this subject were briefed prior to the experiment. With the respect to each triplet, the results of the listening test are rated as a percentage, representing the number of times that the voice qualities were correctly matched in relation to the number of times the triplet was presented. The results are referred to as *identification rates*. The Characteristic Glottal Pulse Waveform Parameterization and Modeling is performed for the following parameter values: $L=120$, $T_N=100$ and $\delta=10$.

6.5.2 Results and discussion

Setting (I)

For each gender, there are 180 possible triplets under setting (I). The individual triplets are shown in Table 6.9. Generally, the listeners have performed the task with ease. Almost every listener was able to correctly distinguish between *lax* and *tense* phonations, across the triplets and gender. With the respect to the female test data, the identification error was 0.069 %, while for the male test data, it was only slightly higher 0.111 %. Since the identification errors are quite trivial, and apparently random, we believe that they are a result of listening fatigue.

Setting (II): Female speakers

The results of the listening test for female test data, under Setting (II), are shown in Table 6.10. The table is organized in the following manner. The triplets containing *tense* voices are shown on the right hand side of the table, whereas the triplets containing *lax* voices are shown on the left hand side. The triplets are grouped according to the number of modified (synthetic) stimuli present in each triplet. Before we analyze the results of the listening test, it is important to note that the results do not only reflect the voice conversion performance, but also the abilities of the listeners to differentiate between the voice quality types and the perceptual voice quality distances.

For female test data, the overall identification rate is 88.54 %. The average identification rates for *tense* triplets and *lax* triplets are 88.47 % and 88.61 %, respectively. The average identification rate for the triplets containing only natural voice stimuli is 90.63 %. On the other hand, the average identification rates for the triplets containing one and two modified voice stimuli are 89.50 % and 86.25 %, respectively. The small identification rate differences between the triplets containing only natural speech stimuli and those that contain at least one synthetic speech stimuli reflect two important characteristics of CGPWPM-based voice quality conversion. The proposed voice conversion method is evidently able to achieve the desired perceptual effects, and secondly the synthetic speech remains as intelligible and human sounding as the natural speech. The quality of synthetic speech is an important factor in any experiment that is based on the subjective listening tests, especially when the experiment involves the entire speech sentences and not just small speech segments or sustained vowels. Voice quality analysis and modification is performed in a controlled and systematic manner, and as such, no significant, perceptually undesired, artifacts are induced. However, we have to remind that the spectral characteristics of the vocal tract filter are also correlates of voice quality. As such, we believe that the exclusion of vocal tract filter in the voice quality analysis and conversion is the primary cause of the slight performance difference between the natural and the synthetic speech stimuli.

Upon further examination we have established that the identification rates with the respect to each triplet exhibit strong correlation with the R_d distances between the individual stimuli that constitute the triplet. To be more exact, the identification rate is largely dependant on the minimum R_d distance between the two voice quality types that need to be differentiated to ensure correct identification.

For female test data, the two closest stimuli according to the glottal shape parameter are M_2 and L_2 with dR_d = 0.32. The average identification rate for the triplets containing M_2 voice and either L_2 voice or its corresponding synthetic versions $\tilde{L}_{12}$ or $\tilde{L}_{22}$ is ID = 82.67%. The second closest stimuli according to the glottal shape parameter are M_1 and T_1 with dR_d = 0.34. The average identification rate for the triplets containing M_1 voice and either T_1 voice or its corresponding synthetic versions $\tilde{T}_{11}$ or $\tilde{T}_{21}$ is ID = 86.00%.

The two furthest stimuli according to the glottal shape parameter are M_1 and L_1 with dR_d = 0.74. The average identification rate for the triplets containing M_1 and either L_1 or its corresponding synthetic versions $\tilde{L}_{11}$ or $\tilde{L}_{21}$ is ID = 92.00%. As far as the *tense* voice is concerned the furthest two stimuli according to the glottal shape parameter are M_2 and T_2 with dR_d = 0.54. The average identification rate for the triplets containing M_1 and either L_1 or its corresponding synthetic versions $\tilde{T}_{12}$ or $\tilde{T}_{22}$ is ID =90.00%.

As a final comment we want to say, that out of all the synthetic stimuli, only those that were converted according to T_2 specifications exhibited at times slight "unnaturalness". After some analysis we have established that the main cause of "unnaturalness" is the relatively large shimmer value that describes T_2 voice. In order to support our earlier argument that the quality of synthetic speech affects the results of the subjective listening tests, we have compared the average identification rates for the triplets containing $\tilde{T}_{12}$ and/or $\tilde{T}_{22}$ (ID = 86.75 %) and those containing $\tilde{T}_{11}$ and/or $\tilde{T}_{21}$ (ID = 88.00 %). Even though T_2 is perceptually further from *modal* phonation according to the glottal shape parameter, the triplets $\tilde{T}_{11}$ and $\tilde{T}_{21}$ attain higher identification rates because they do not have any undesired perceptual artifacts. We believe that a better vocal tremor model, such as the one developed by Kreiman *et al.* in [89], would improve the quality of modified speech.

Setting (II): Male speakers

The results of the listening test for male test data, under Setting (II), are shown in Table 6.11. These results essentially draw the same conclusions as those of female test data. As such, only a brief summary of the results is presented in order to provide further support for the arguments made in the previous section. We will only note that none of the modified stimuli have exhibited any undesired perceptual artifacts, and they have all sounded perfectly natural and intelligible. For male speakers, the overall identification rate is 88.68 %, 87.92 % for *tense* triplets and 89.44 % for *lax* triplets. The average identification rate for the triplets containing only natural voice stimuli is 89.38 %. On the other hand, the average identification rates for the triplets containing one and two modified voice stimuli are 89.13 % and 87.71 %, respectively.

For male test data, the two closest stimuli according to the glottal shape parameter are M_1 and T_2 with $dR_d = 0.35$. The average identification rate for the triplets containing M_1 and either T_2 or its corresponding synthetic versions $\tilde{T}_{12}$ or $\tilde{T}_{22}$ is ID = 85.67%. With the respect to *lax* voices, two closest stimuli according to the glottal shape parameter are M_2 and L_1 with dR_d = 0.38. The average identification rate for the triplets containing M_2 and either L_1 or its corresponding synthetic versions $\tilde{L}_{11}$ or $\tilde{L}_{21}$ is $ID = 84.67\%$.

The two furthest stimuli according to the glottal shape parameter are M_1 and L_2 with dR_d = 0.63. The average identification rate for the triplets containing M_1 and either L$_2$ or its corresponding synthetic versions $\tilde{L}_{12}$ or $\tilde{L}_{22}$ is ID = 93.67 %. With the respect to *tense* voices, the two furthest stimuli according to the glottal shape parameter are M_2 and T_1 with $dR_d = 0.46$. The average identification rate for the triplets containing M_2 and either T_1 or its corresponding synthetic versions $\tilde{T}_{11}$ or $\tilde{T}_{21}$ is 89.33 %.

Table 6.9:
Triplets used for Setting (I) of the voice quality conversion experiment. All of the triplets consist of Tense and Lax voice stimuli.

Triplets								
$T_1L_1L_2$	$T_1\tilde{L}_{11}L_2$	$T_1\tilde{L}_{21}L_2$	$\tilde{T}_{11}L_1L_2$	$\tilde{T}_{11}\tilde{L}_{11}L_2$	$\tilde{T}_{11}\tilde{L}_{21}L_2$	$\tilde{T}_{21}L_1L_2$	$\tilde{T}_{21}\tilde{L}_{11}L_2$	$\tilde{T}_{21}\tilde{L}_{21}L_2$
$T_1L_1\tilde{L}_{12}$	$T_1\tilde{L}_{11}\tilde{L}_{12}$	$T_1\tilde{L}_{21}\tilde{L}_{12}$	$\tilde{T}_{11}L_1\tilde{L}_{12}$	$\tilde{T}_{11}\tilde{L}_{11}\tilde{L}_{12}$	$\tilde{T}_{11}\tilde{L}_{21}\tilde{L}_{12}$	$\tilde{T}_{21}L_1\tilde{L}_{12}$	$\tilde{T}_{21}\tilde{L}_{11}\tilde{L}_{12}$	$\tilde{T}_{21}\tilde{L}_{21}\tilde{L}_{12}$
$T_1L_1\tilde{L}_{22}$	$T_1\tilde{L}_{11}\tilde{L}_{22}$	$T_1\tilde{L}_{21}\tilde{L}_{22}$	$\tilde{T}_{11}L_1\tilde{L}_{22}$	$\tilde{T}_{11}\tilde{L}_{11}\tilde{L}_{22}$	$\tilde{T}_{11}\tilde{L}_{21}\tilde{L}_{22}$	$\tilde{T}_{21}L_1\tilde{L}_{22}$	$\tilde{T}_{21}\tilde{L}_{11}\tilde{L}_{22}$	$\tilde{T}_{21}\tilde{L}_{21}\tilde{L}_{22}$
$T_1L_2\tilde{L}_{12}$	$T_1L_1\tilde{L}_{21}$	$T_1L_1\tilde{L}_{11}$	$\tilde{T}_{11}L_2\tilde{L}_{12}$	$\tilde{T}_{11}L_1\tilde{L}_{21}$	$\tilde{T}_{11}L_1\tilde{L}_{11}$	$\tilde{T}_{21}L_2\tilde{L}_{12}$	$\tilde{T}_{21}L_1\tilde{L}_{21}$	$\tilde{T}_{21}L_1\tilde{L}_{11}$
$T_1L_2\tilde{L}_{22}$	$T_1\tilde{L}_{11}\tilde{L}_{21}$	$T_1\tilde{L}_{12}\tilde{L}_{22}$	$\tilde{T}_{11}L_2\tilde{L}_{22}$	$\tilde{T}_{11}\tilde{L}_{11}\tilde{L}_{21}$	$\tilde{T}_{11}\tilde{L}_{12}\tilde{L}_{22}$	$\tilde{T}_{21}L_2\tilde{L}_{22}$	$\tilde{T}_{21}\tilde{L}_{11}\tilde{L}_{21}$	$\tilde{T}_{21}\tilde{L}_{12}\tilde{L}_{22}$
$T_2L_2L_1$	$T_2\tilde{L}_{12}L_1$	$T_2\tilde{L}_{22}L_1$	$\tilde{T}_{12}L_2L_1$	$\tilde{T}_{12}\tilde{L}_{12}L_1$	$\tilde{T}_{12}\tilde{L}_{22}L_1$	$\tilde{T}_{22}L_2L_1$	$\tilde{T}_{22}\tilde{L}_{12}L_1$	$\tilde{T}_{22}\tilde{L}_{22}L_1$
$T_2L_2\tilde{L}_{11}$	$T_2\tilde{L}_{12}\tilde{L}_{11}$	$T_2\tilde{L}_{22}\tilde{L}_{11}$	$\tilde{T}_{12}L_2\tilde{L}_{11}$	$\tilde{T}_{12}\tilde{L}_{12}\tilde{L}_{11}$	$\tilde{T}_{12}\tilde{L}_{22}\tilde{L}_{11}$	$\tilde{T}_{22}L_2\tilde{L}_{11}$	$\tilde{T}_{22}\tilde{L}_{12}\tilde{L}_{11}$	$\tilde{T}_{22}\tilde{L}_{22}\tilde{L}_{11}$
$T_2L_2\tilde{L}_{21}$	$T_2\tilde{L}_{12}\tilde{L}_{21}$	$T_2\tilde{L}_{22}\tilde{L}_{21}$	$\tilde{T}_{12}L_2\tilde{L}_{21}$	$\tilde{T}_{12}\tilde{L}_{12}\tilde{L}_{21}$	$\tilde{T}_{12}\tilde{L}_{22}\tilde{L}_{21}$	$\tilde{T}_{22}L_2\tilde{L}_{21}$	$\tilde{T}_{22}\tilde{L}_{12}\tilde{L}_{21}$	$\tilde{T}_{22}\tilde{L}_{22}\tilde{L}_{21}$
$T_2L_2\tilde{L}_{12}$	$T_2L_1\tilde{L}_{21}$	$T_2L_1\tilde{L}_{11}$	$\tilde{T}_{12}L_2\tilde{L}_{12}$	$\tilde{T}_{12}L_1\tilde{L}_{21}$	$\tilde{T}_{12}L_1\tilde{L}_{11}$	$\tilde{T}_{22}L_2\tilde{L}_{12}$	$\tilde{T}_{22}L_1\tilde{L}_{21}$	$\tilde{T}_{22}L_1\tilde{L}_{11}$
$T_2L_2\tilde{L}_{22}$	$T_2\tilde{L}_{11}\tilde{L}_{21}$	$T_2\tilde{L}_{12}\tilde{L}_{22}$	$\tilde{T}_{12}L_2\tilde{L}_{22}$	$\tilde{T}_{12}\tilde{L}_{11}\tilde{L}_{21}$	$\tilde{T}_{12}\tilde{L}_{12}\tilde{L}_{22}$	$\tilde{T}_{22}L_2\tilde{L}_{22}$	$\tilde{T}_{22}\tilde{L}_{11}\tilde{L}_{21}$	$\tilde{T}_{22}\tilde{L}_{12}\tilde{L}_{22}$
$L_1T_1T_2$	$L_1\tilde{T}_{11}T_2$	$L_1\tilde{T}_{21}T_2$	$\tilde{L}_{11}T_1T_2$	$\tilde{L}_{11}\tilde{T}_{11}T_2$	$\tilde{L}_{11}\tilde{T}_{21}T_2$	$\tilde{L}_{21}T_1T_2$	$\tilde{L}_{21}\tilde{T}_{11}T_2$	$\tilde{L}_{21}\tilde{T}_{21}T_2$
$L_1T_1\tilde{T}_{12}$	$L_1\tilde{T}_{11}\tilde{T}_{12}$	$L_1\tilde{T}_{21}\tilde{T}_{12}$	$\tilde{L}_{11}T_1\tilde{T}_{12}$	$\tilde{L}_{11}\tilde{T}_{11}\tilde{T}_{12}$	$\tilde{L}_{11}\tilde{T}_{21}\tilde{T}_{12}$	$\tilde{L}_{21}T_1\tilde{T}_{12}$	$\tilde{L}_{21}\tilde{T}_{11}\tilde{T}_{12}$	$\tilde{L}_{21}\tilde{T}_{21}\tilde{T}_{12}$
$L_1T_1\tilde{T}_{22}$	$L_1\tilde{T}_{11}\tilde{T}_{22}$	$L_1\tilde{T}_{21}\tilde{T}_{22}$	$\tilde{L}_{11}T_1\tilde{T}_{22}$	$\tilde{L}_{11}\tilde{T}_{11}\tilde{T}_{22}$	$\tilde{L}_{11}\tilde{T}_{21}\tilde{T}_{22}$	$\tilde{L}_{21}T_1\tilde{T}_{22}$	$\tilde{L}_{21}\tilde{T}_{11}\tilde{T}_{22}$	$\tilde{L}_{21}\tilde{T}_{21}\tilde{T}_{22}$
$L_1T_2\tilde{T}_{12}$	$L_1T_1\tilde{T}_{21}$	$L_1T_1\tilde{T}_{11}$	$\tilde{L}_{11}T_2\tilde{T}_{12}$	$\tilde{L}_{11}T_1\tilde{T}_{21}$	$\tilde{L}_{11}T_1\tilde{T}_{11}$	$\tilde{L}_{21}T_2\tilde{T}_{12}$	$\tilde{L}_{21}T_1\tilde{T}_{21}$	$\tilde{L}_{21}T_1\tilde{T}_{11}$
$L_1T_2\tilde{T}_{22}$	$L_1\tilde{T}_{11}\tilde{T}_{21}$	$L_1\tilde{T}_{12}\tilde{T}_{22}$	$\tilde{L}_{11}T_2\tilde{T}_{22}$	$\tilde{L}_{11}\tilde{T}_{11}\tilde{T}_{21}$	$\tilde{L}_{11}\tilde{T}_{12}\tilde{T}_{22}$	$\tilde{L}_{21}T_2\tilde{T}_{22}$	$\tilde{L}_{21}\tilde{T}_{11}\tilde{T}_{21}$	$\tilde{L}_{21}\tilde{T}_{12}\tilde{T}_{22}$
$L_2T_2T_1$	$L_2\tilde{T}_{12}T_1$	$L_2\tilde{T}_{22}T_1$	$\tilde{L}_{12}T_2T_1$	$\tilde{L}_{12}\tilde{T}_{12}T_1$	$\tilde{L}_{12}\tilde{T}_{22}T_1$	$\tilde{L}_{22}T_2T_1$	$\tilde{L}_{22}\tilde{T}_{12}T_1$	$\tilde{L}_{22}\tilde{T}_{22}T_1$
$L_2T_2\tilde{T}_{11}$	$L_2\tilde{T}_{12}\tilde{T}_{11}$	$L_2\tilde{T}_{22}\tilde{T}_{11}$	$\tilde{L}_{12}T_2\tilde{T}_{11}$	$\tilde{L}_{12}\tilde{T}_{12}\tilde{T}_{11}$	$\tilde{L}_{12}\tilde{T}_{22}\tilde{T}_{11}$	$\tilde{L}_{22}T_2\tilde{T}_{11}$	$\tilde{L}_{22}\tilde{T}_{12}\tilde{T}_{11}$	$\tilde{L}_{22}\tilde{T}_{22}\tilde{T}_{11}$
$L_2T_2\tilde{T}_{21}$	$L_2\tilde{T}_{12}\tilde{T}_{21}$	$L_2\tilde{T}_{22}\tilde{T}_{21}$	$\tilde{L}_{12}T_2\tilde{T}_{21}$	$\tilde{L}_{12}\tilde{T}_{12}\tilde{T}_{21}$	$\tilde{L}_{12}\tilde{T}_{22}\tilde{T}_{21}$	$\tilde{L}_{22}T_2\tilde{T}_{21}$	$\tilde{L}_{22}\tilde{T}_{12}\tilde{T}_{21}$	$\tilde{L}_{22}\tilde{T}_{22}\tilde{T}_{21}$
$L_2T_2\tilde{T}_{12}$	$L_2T_1\tilde{T}_{21}$	$L_2T_1\tilde{T}_{11}$	$\tilde{L}_{12}T_2\tilde{T}_{12}$	$\tilde{L}_{12}T_1\tilde{T}_{21}$	$\tilde{L}_{12}T_1\tilde{T}_{11}$	$\tilde{L}_{22}T_2\tilde{T}_{12}$	$\tilde{L}_{22}T_1\tilde{T}_{21}$	$\tilde{L}_{22}T_1\tilde{T}_{11}$
$L_2T_2\tilde{T}_{22}$	$L_2\tilde{T}_{11}\tilde{T}_{21}$	$L_2\tilde{T}_{12}\tilde{T}_{22}$	$\tilde{L}_{12}T_2\tilde{T}_{22}$	$\tilde{L}_{12}\tilde{T}_{11}\tilde{T}_{21}$	$\tilde{L}_{12}\tilde{T}_{12}\tilde{T}_{22}$	$\tilde{L}_{22}T_2\tilde{T}_{22}$	$\tilde{L}_{22}\tilde{T}_{11}\tilde{T}_{21}$	$\tilde{L}_{22}\tilde{T}_{12}\tilde{T}_{22}$

Table 6.10:
Identification rates for female test-data under setting (II)

Triplets	ID rate [%]	Triplets	ID rate [%]	Triplets	ID rate [%]	Triplets	ID rate [%]
No. synthetic stimuli: 1		*No. synthetic stimuli: 0*		*No. synthetic stimuli: 1*		*No. synthetic stimuli: 0*	
$M_1\tilde{T}_{11}T_1$	90	$M_1T_2T_1$	85	$M_1\tilde{L}_{11}L_1$	100	$M_1L_2L_1$	95
$M_1\tilde{T}_{12}T_1$	85	$M_2T_2T_1$	90	$M_1\tilde{L}_{12}L_1$	85	$M_2L_2L_1$	85
$M_1\tilde{T}_{21}T_1$	90	$T_1M_1M_2$	90	$M_1\tilde{L}_{21}L_1$	100	$L_1M_1M_2$	100
$M_1\tilde{T}_{22}T_1$	80	$T_2M_1M_2$	100	$M_1\tilde{L}_{22}L_1$	85	$L_2M_1M_2$	80
$M_1\tilde{T}_{11}T_2$	90	Group Average	91.25	$M_1\tilde{L}_{11}L_2$	85	Group Average:	90.00
$M_1\tilde{T}_{12}T_2$	85	*No. synthetic stimuli:*	2	$M_1\tilde{L}_{12}L_2$	90	*No. synthetic stimuli:* 2	
$M_1\tilde{T}_{21}T_2$	85			$M_1\tilde{L}_{21}L_2$	85		
$M_1\tilde{T}_{22}T_2$	90			$M_1\tilde{L}_{22}L_2$	90		
$M_2\tilde{T}_{11}T_1$	95	$M_1\tilde{T}_{12}\tilde{T}_{11}$	80	$M_2\tilde{L}_{11}L_1$	100	$M_1\tilde{L}_{12}\tilde{L}_{11}$	85
$M_2\tilde{T}_{12}T_1$	90	$M_1\tilde{T}_{21}\tilde{T}_{11}$	90	$M_2\tilde{L}_{12}L_1$	80	$M_1\tilde{L}_{21}\tilde{L}_{11}$	100
$M_2\tilde{T}_{21}T_1$	95	$M_1\tilde{T}_{22}\tilde{T}_{11}$	85	$M_2\tilde{L}_{21}L_1$	100	$M_1\tilde{L}_{22}\tilde{L}_{11}$	85
$M_2\tilde{T}_{22}T_1$	85	$M_1\tilde{T}_{21}\tilde{T}_{12}$	85	$M_2\tilde{L}_{22}L_1$	80	$M_1\tilde{L}_{21}\tilde{L}_{12}$	90
$M_2\tilde{T}_{11}T_2$	90	$M_1\tilde{T}_{22}\tilde{T}_{12}$	85	$M_2\tilde{L}_{11}L_2$	85	$M_1\tilde{L}_{22}\tilde{L}_{12}$	90
$M_2\tilde{T}_{12}T_2$	100	$M_1\tilde{T}_{22}\tilde{T}_{21}$	85	$M_2\tilde{L}_{12}L_2$	90	$M_1\tilde{L}_{22}\tilde{L}_{21}$	85
$M_2\tilde{T}_{21}T_2$	95	$M_2\tilde{T}_{12}\tilde{T}_{11}$	80	$M_2\tilde{L}_{21}L_2$	80	$M_2\tilde{L}_{12}\tilde{L}_{11}$	80
$M_2\tilde{T}_{22}T_2$	95	$M_2\tilde{T}_{21}\tilde{T}_{11}$	95	$M_2\tilde{L}_{22}L_2$	85	$M_2\tilde{L}_{21}\tilde{L}_{11}$	100
$\tilde{T}_{11}M_1M_2$	85	$M_2\tilde{T}_{22}\tilde{T}_{11}$	85	$\tilde{L}_{11}M_1M_2$	100	$M_2\tilde{L}_{22}\tilde{L}_{11}$	80
$\tilde{T}_{21}M_1M_2$	85	$M_2\tilde{T}_{21}\tilde{T}_{12}$	90	$\tilde{L}_{21}M_1M_2$	100	$M_2\tilde{L}_{21}\tilde{L}_{12}$	80
$\tilde{T}_{12}M_1M_2$	90	$M_2\tilde{T}_{22}\tilde{T}_{12}$	80	$\tilde{L}_{12}M_1M_2$	85	$M_2\tilde{L}_{22}\tilde{L}_{12}$	90
$\tilde{T}_{22}M_1M_2$	95	$M_2\tilde{T}_{22}\tilde{T}_{21}$	85	$\tilde{L}_{22}M_1M_2$	80	$M_2\tilde{L}_{22}\tilde{L}_{21}$	80
Group average	89.75	Group average	85.42	Group average	89.25	Group average	87.08

Table 6.11:
Identification rates for male test-data under setting (II)

Triplets	ID rate [%]	Triplets	ID rate [%]	Triplets	ID rate [%]	Triplets	ID rate [%]
No. synthetic stimuli: 1		*No. synthetic stimuli: 0*		*No. synthetic stimuli: 1*		*No. synthetic stimuli: 0*	
$M_1\tilde{T}_{11}T_1$	85	$M_1T_2T_1$	85	$M_1\tilde{L}_{11}L_1$	90	$M_1L_2L_1$	90
$M_1\tilde{T}_{12}T_1$	90	$M_2T_2T_1$	90	$M_1\tilde{L}_{12}L_1$	90	$M_2L_2L_1$	85
$M_1\tilde{T}_{21}T_1$	90	$T_1M_1M_2$	90	$M_1\tilde{L}_{21}L_1$	85	$L_1M_1M_2$	90
$M_1\tilde{T}_{22}T_1$	80	$T_2M_1M_2$	90	$M_1\tilde{L}_{22}L_1$	90	$L_2M_1M_2$	95
$M_1\tilde{T}_{11}T_2$	85	Group Average	88.75	$M_1\tilde{L}_{11}L_2$	95	Group Average:	90.00
$M_1\tilde{T}_{12}T_2$	90	*No. synthetic stimuli:*	2	$M_1\tilde{L}_{12}L_2$	95	*No. synthetic stimuli:*	2
$M_1\tilde{T}_{21}T_2$	85			$M_1\tilde{L}_{21}L_2$	95		
$M_1\tilde{T}_{22}T_2$	85			$M_1\tilde{L}_{22}L_2$	100		
$M_2\tilde{T}_{11}T_1$	100	$M_1\tilde{T}_{12}\tilde{T}_{11}$	85	$M_2\tilde{L}_{11}L_1$	85	$M_1\tilde{L}_{12}\tilde{L}_{11}$	90
$M_2\tilde{T}_{12}T_1$	90	$M_1\tilde{T}_{21}\tilde{T}_{11}$	95	$M_2\tilde{L}_{12}L_1$	90	$M_1\tilde{L}_{21}\tilde{L}_{11}$	90
$M_2\tilde{T}_{21}T_1$	100	$M_1\tilde{T}_{22}\tilde{T}_{11}$	85	$M_2\tilde{L}_{21}L_1$	85	$M_1\tilde{L}_{22}\tilde{L}_{11}$	85
$M_2\tilde{T}_{22}T_1$	90	$M_1\tilde{T}_{21}\tilde{T}_{12}$	85	$M_2\tilde{L}_{22}L_1$	85	$M_1\tilde{L}_{21}\tilde{L}_{12}$	90
$M_2\tilde{T}_{11}T_2$	90	$M_1\tilde{T}_{22}\tilde{T}_{12}$	85	$M_2\tilde{L}_{11}L_2$	80	$M_1\tilde{L}_{22}\tilde{L}_{12}$	100
$M_2\tilde{T}_{12}T_2$	90	$M_1\tilde{T}_{22}\tilde{T}_{21}$	90	$M_2\tilde{L}_{12}L_2$	95	$M_1\tilde{L}_{22}\tilde{L}_{21}$	95
$M_2\tilde{T}_{21}T_2$	85	$M_2\tilde{T}_{12}\tilde{T}_{11}$	85	$M_2\tilde{L}_{21}L_2$	85	$M_2\tilde{L}_{12}\tilde{L}_{11}$	80
$M_2\tilde{T}_{22}T_2$	90	$M_2\tilde{T}_{21}\tilde{T}_{11}$	90	$M_2\tilde{L}_{22}L_2$	90	$M_2\tilde{L}_{21}\tilde{L}_{11}$	85
$\tilde{T}_{11}M_1M_2$	85	$M_2\tilde{T}_{22}\tilde{T}_{11}$	85	$\tilde{L}_{11}M_1M_2$	85	$M_2\tilde{L}_{22}\tilde{L}_{11}$	85
$\tilde{T}_{21}M_1M_2$	90	$M_2\tilde{T}_{21}\tilde{T}_{12}$	85	$\tilde{L}_{21}M_1M_2$	85	$M_2\tilde{L}_{21}\tilde{L}_{12}$	85
$\tilde{T}_{12}M_1M_2$	85	$M_2\tilde{T}_{22}\tilde{T}_{12}$	90	$\tilde{L}_{12}M_1M_2$	100	$M_2\tilde{L}_{22}\tilde{L}_{12}$	95
$\tilde{T}_{22}M_1M_2$	80	$M_2\tilde{T}_{22}\tilde{T}_{21}$	85	$\tilde{L}_{22}M_1M_2$	95	$M_2\tilde{L}_{22}\tilde{L}_{21}$	80
Group average	88.25	Group average	87.08	Group average	90.00	Group average	88.33

6.6 Conclusion

In this chapter, the voice source reconstruction aspect of the Characteristic Glottal Pulse Waveform Parameterization and Modeling system is presented. In comparative evaluation of the CGPWPM and Liljencrants-Fant's model of glottal flow derivative waveform, we have demonstrated that the LF model does not provide enough degrees of freedom to adequately represent the structurally more complex examples of voice source realizations. On the other hand, the DMOS results obtained for the CGPWPM-based synthesis of pathological voices are very high, suggesting that CGPWPM is adaptable enough to cater for very complex and structurally rich forms of glottal flow derivative realizations. With the respect to healthy, *modal* phonation, the MOS values reveal that the CGPWPM-based speech synthesis rates highly on the scale of absolute perceptual acceptability, while the corresponding DMOS values suggest that the speech is faithfully reconstructed on consistent basis. The average MOS and DMOS values across the test database of 40 speakers are 4.45 and 4.55, respectively. Not a single speech reconstruction exhibited *annoying* degradation levels and only one speech file was judge below the second highest level of absolute perceptual acceptability. Overall, we can conclude that CGPWPM-based speech synthesis offers a high quality performance that is consistent across speakers and gender. The results of the speaker identification experiments have established that the fine structural elements of the glottal flow derivative waveforms, which are modeled through CGPWPM, contain a notable level of speaker-dependant information. The average speaker identification rate obtained for CGPWPM-based voice source synthesis is 18.6 % higher than the identification rate corresponding to the Liljencrants-Fant's glottal pulse model. Finally, the voice quality conversion experiments have shown that CGPWMP is able to successfully modify the glottal shape parameters, aspiration noise characteristics, and the aperiodic features of voice source signals and consequentially, to achieve the desired perceptual effects.

Chapter 7:

Conclusion and Future Directions

Conclusion

In this thesis, we have developed a novel framework for voice source analysis, parameterization and reconstruction which we refer to as Characteristic Glottal Pulse Waveform Parameterization and Modeling (*CGPWPM*). The proposed method is not constrained to the idealized glottal waveform approximations (e.g. LF, KLGLOTT88), but instead, relies on the estimates of the Characteristic Glottal Pulse Waveform to obtain the voice source parameters. It uses a set of modified LF parameters and the DTW algorithm to track the nonlinear evolution of the Characteristic Glottal Pulse Waveform in time. The constrictions of the optimal non-linear time alignment that are employed by DTW algorithm are also extended to voice source parameterization and as such, pathological parameterization is precluded. The design of the method is motivated by the fact that the parameterization performance is linked to the extent of similarity between the glottal model and the analyzed voice source signal. CGPWPM provides the means to model both, the course and the fine structural elements of the glottal flow derivative realizations. Thus, the proposed method enables a study of voice source signal features and their temporal behavior that could not be efficiently or accurately represented by a poorly deterministic glottal flow derivative model. Another useful characteristic of this method is that the parameterization of consecutive glottal pulses across the voiced source signal is referenced to a parametric description of a single glottal flow derivative realization. As such, CGPWPM method can be used very effectively in a semi-automatic manner, as well as in the fully automatic mode. The results of the synthetic dataset based experiments have shown that the performance of the Characteristic Glottal Pulse Parameterization is virtually insensitive to the inaccuracies in the glottal closure instant estimates. On natural speech, CGPWP exhibits a robust performance even for the significant presence of disturbances in the voice source signal estimates. Overall, CGPWP exhibits a superior performance over the standard *fit estimation* and *direct estimation* methods. In comparative evaluation of the CGPWPM and Liljencrants-Fant's model of glottal flow derivative waveform, we have demonstrated that the LF model does not provide enough degrees of freedom to adequately represent the structurally more complex examples of voice source realizations. Furthermore, the results of the speaker identification experiments have established that the fine structural elements of the glottal flow derivative waveforms, which are modeled through CGPWPM, contain a notable level of speaker-dependant information. The average speaker identification rate obtained for CGPWPM-based voice source synthesis is 18.6 % higher than the identification rate corresponding to the Liljencrants-Fant's glottal pulse model. The results of the voice quality profiling experiment were in a general agreement with those obtained by Karlsson and Liljencrants, Childers and Lee, and van Dinther. We have used the voice quality profiling results to derive a surprisingly simple relationship between the glottal shape parameter R_d, and voice quality. We have also demonstrated that the glottal shape is for all practical considerations independent of both, the frequency of vocal fold oscillations and the glottal excitation strength.

CGPWPM is applied under the source-filter model of speech production to develop the speech synthesis and voice conversion methods. The quality of CGPWPM-based speech synthesis is formally evaluated via the Mean Opinion Scores and Degradation Mean Opinion Scores. On the other hand, triadic listening tests are used to evaluate the performance of a CGPWPM-based voice quality conversion method. DMOS results obtained for the CGPWPM-based synthesis of pathological voices are very high, suggesting that CGPWPM is adaptable enough to cater for very complex and structurally rich forms of glottal flow derivative realizations. With the respect to healthy, *modal* phonation, the MOS values reveal that the CGPWPM-based speech synthesis rates highly on the scale of absolute perceptual acceptability, while the corresponding DMOS values suggest that the speech is faithfully reconstructed on consistent basis. The average MOS and DMOS values across the test database of 40 speakers are 4.45 and 4.55, respectively. Not a single speech reconstruction exhibits *annoying* degradation levels and only one speech file was judge below the second highest level of absolute perceptual acceptability. Overall, we can conclude that CGPWPM-based speech synthesis offers a high quality performance that is consistent across speakers and gender. The voice quality conversion experiments have shown that CGPWMP is able to successfully modify the glottal shape parameters, aspiration noise characteristics, and the aperiodic features of voice source signals and consequentially, to achieve the desired perceptual effects.

In this thesis, we have also considered a group delay approach to GCI estimation. Specifically, *average group delay* and *energy weighted group delay* measures are discussed in detail. We have proposed a GCI estimation method based on a group delay algorithm and the translation-invariant hard-thresholding of LPC residue. Thresholding is performed with the 6-coefficient Coiflet filter and a primary resolution level-7. The performances of the two group delay measures and the proposed method are evaluated for a range of fixed and pitch-synchronous group delay window lengths. We have found that in comparison to the *energy weighted group delay* measure with a fixed group delay window, the pitch synchronous *energy weighted group delay* measure with the wavelet–denoised LPC residue improves the *identification rate* and *accuracy* by 6.57 % and 0.158 ms, respectively. This represents a considerable improvement in the performance, especially if we consider that 6.57 % increase in the *identification rate* corresponds to 82.33 % reduction in the number of unidentified glottal excitations. In large, these results reflect a superior denoising performance of the wavelet thresholding method. The proposed method is based on a study where we have aimed to develop an optimal wavelet thresholding strategy for the glottal volume velocity derivative signals. The following methods have been considered: Universal thresholding, SureShrink thresholding, Hybrid-Sure thresholding, Translation-Invariant thresholding, Hypothesis-Testing-based thresholding, Block thresholding, and Bayesian Adaptive Multi-resolution Smoother. We have systematically investigated the thresholding performance as a function of two thresholding parameters - the choice of wavelet basis function and the coarsest level of the wavelet decomposition. The main problem that we have encountered is the fact that the relationship between the thresholding parameters and the thresholding performance is highly non-linear. Furthermore, this relationship differs from one thresholding method to another. However, some rather crude trends were made apparent. Short wavelet filters, i.e. wavelets with a small number of vanishing moments, tend to have inadequate approximation properties, and as such, the quality of the reconstructed signal is often poor. On the other hand, more regular wavelets, corresponding to higher filter orders, have better decorrelating properties at the expense of temporal compactness. In the vast majority of the considered thresholding methods, a reasonable compromise between these effects is found to exist for some moderate filter length. A choice of decomposition level is also found to have a strong effect on the quality of the reconstructed signal. In most cases, if the decomposition level is too deep, the reconstructed signal is found to contain distortions around the instant of glottal closure (over-smoothing of glottal peak). The translation invariant thresholding performs better than other considered thresholding methods as it is able to minimize the thresholding artifacts associated with misalignments between the sharp changes in the signal and the features of the wavelet. When the TI-H thresholding is applied on a voice source estimate obtained from the natural speech via inverse filtering, the results were very pleasing. Turbulent components in the voice source estimate were almost completely removed. On the other hand, the rapidly varying components in the glottal flow derivative waveforms were clearly preserved.

Future Directions

One of author's interests is speech pathology diagnostics. Aperiodicity features and noise features are the two most important categories in describing the pathological features. Along with the standard aperiodicity features, such as: jitter, shimmer, glottal to noise excitation ratio, we have also used the perturbation of the glottal shape parameter in the voice quality analysis. Our informal tests show that glottal shape perturbation is a particularly important parameter in describing pathological voice. For instance, laryngeal cancer voice exhibits a particularly high degree of glottal shape perturbation, whereas a pathologically breathy voice commonly attains a high level of regularity in the temporal structure of the consecutive glottal flow derivative pulses. We also believe that the fine structural elements of the characteristic glottal pulse waveforms are, at least to some extent, correlated with the nature of voice pathology. It is important to stress that the performance of any measure used in the analysis of pathological speech is inherently sensitive to the accuracy levels of voice source parameterization. For example, the difficulties in tracking pitch contour can severely limit the ability of pitch perturbation measures to separate between the normal and pathological voices. In this thesis, we have shown that CGPWMP is robust enough to accurately parameterize and faithfully re-synthesize a range of pathological voices. As such, we believe that the proposed method could be applied to speech pathology diagnostics, e.g. dysphonia analysis and forensic applications

During the course of this research we have made some attempts to further tie up the formant modulation analysis with the Characteristic Glottal Pulse Waveform Parameterization and Modeling. We have gathered preliminary evidence that the onsets of formant modulation within a glottal cycle can be successfully tracked via CGPWPM, i.e. the estimate of the format modulation onset obtained for the Characteristic Glottal Pulse Waveform can be extended to the other *glottal matrix* pulses via DTW alignment functions. However, further research effort is required to conclusively establish the robustness and accuracy of this approach. In the immediate future we intend to employ CGPWPM in conjunction with the Hidden Markov Model to evaluate the extent of correlation between the temporal structure of voice source signals and the linguistic content, accent, etc... The results of this analysis might be particularly useful in text-to-speech research - this is a speech processing field that the author would like to further explore.

Although, in this thesis, we have evaluated speech quality for un-quantized parametric reconstruction, the proposed method can be readily applied to low bit rate high quality speech coding. The reliance of CGPWPM on the DTW algorithm is the principal disadvantage of this method. We are currently working on improving the computational efficiency during the voice source analysis and parameterization. Two separate approaches are being developed and the preliminary results are encouraging.

Appendix A

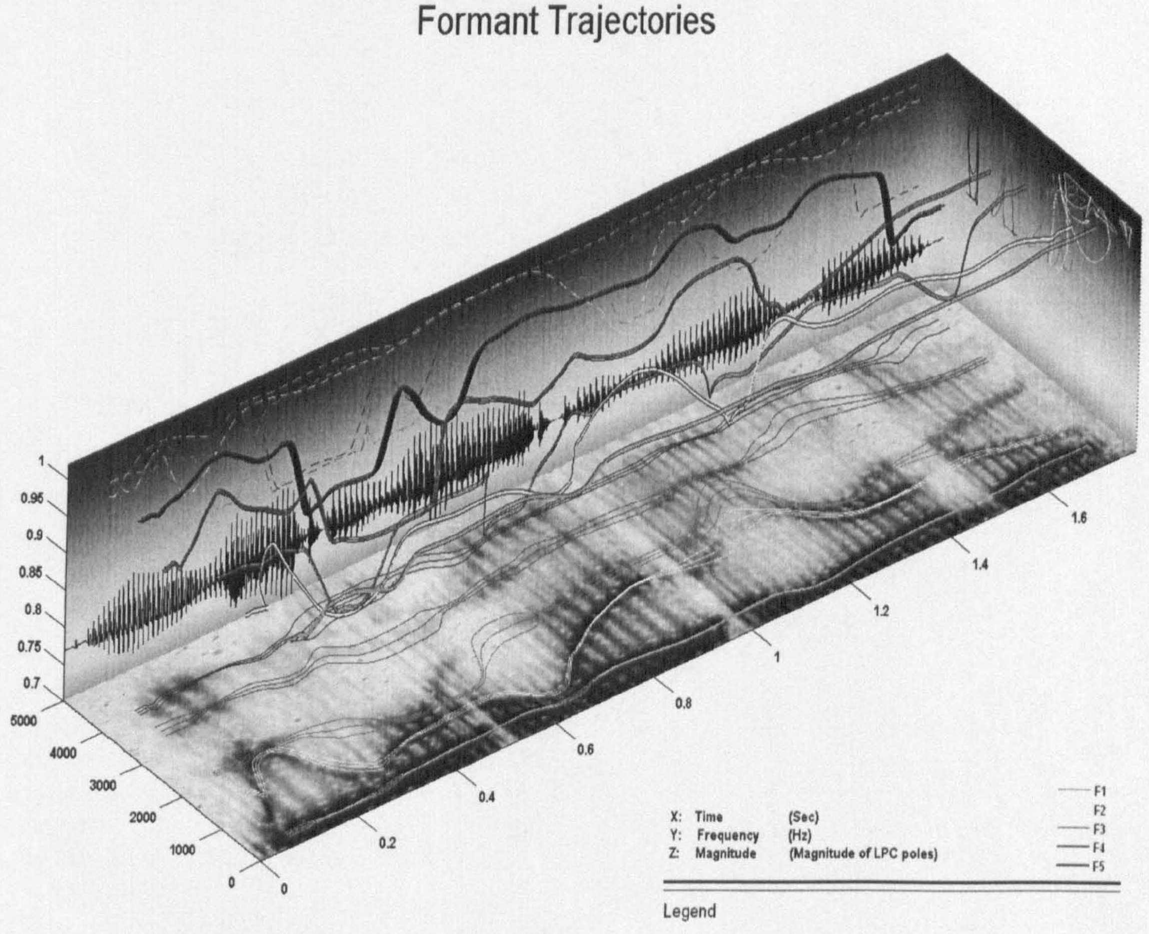


Figure A1: Formant trajectories over the utterance: "We were away a year ago" for a male speaker with modal voice. The trajectories are obtained via the closed-phase pitch synchronous inverse filtering,, 14^{th} order covariance based linear prediction analysis, and a Viterbi search algorithm. F_s =10 kHz

Appendix B

Translation Invariant Hard Thresholding Examples

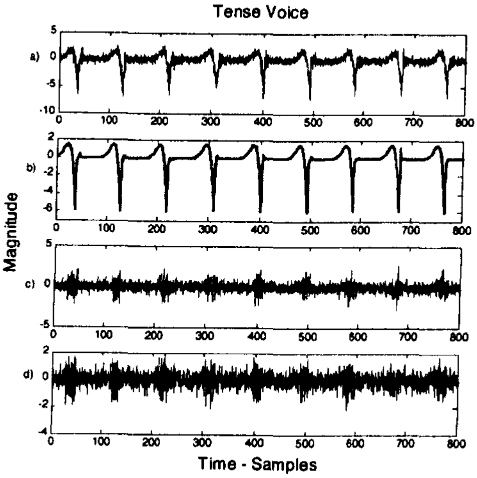


Figure B.1: Tense Voice: *Qualitative evaluation of the globally optimized TI-H performance: a) synthesized glottal excitation with SNR = 6dB; b) clean (dashed) and de-noised (solid) glottal excitation; c) synthesized turbulence noise; d) estimated noise.*

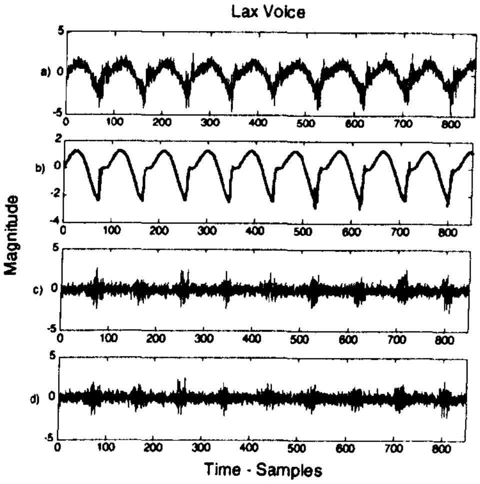


Figure B.2: Lax Voice: *Qualitative evaluation of the globally optimized TI-H performance: a) synthesized glottal excitation with SNR = 6dB; b) clean (dashed) and de-noised (solid) glottal excitation; c) synthesized turbulence noise; d) estimated noise.*

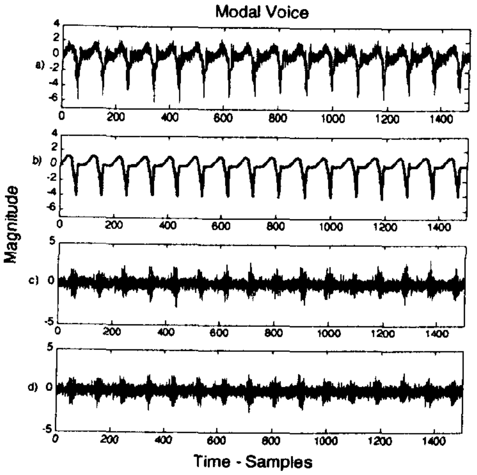


Figure B.3: Modal Voice*: Qualitative evaluation of the globally optimized TI-H performance: a) synthesized glottal excitation with SNR = 6dB; b) clean (dashed) and de-noised (solid) glottal excitation; c) synthesized turbulence noise; d) estimated noise.*

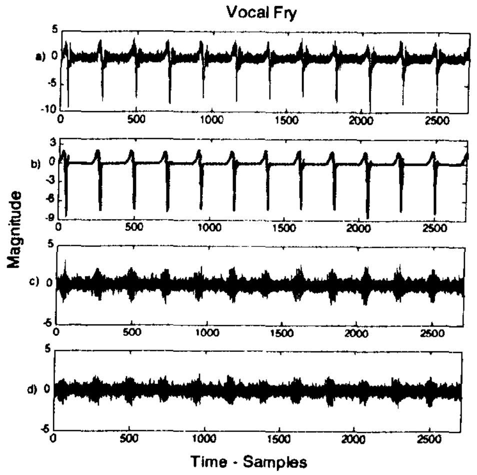


Figure B.4: Vocal Fry*: Qualitative evaluation of the globally optimized TI-H performance: a) synthesized glottal excitation with SNR = 6dB; b) clean (dashed) and de-noised (solid) glottal excitation; c) synthesized turbulence noise; d) estimated noise.*

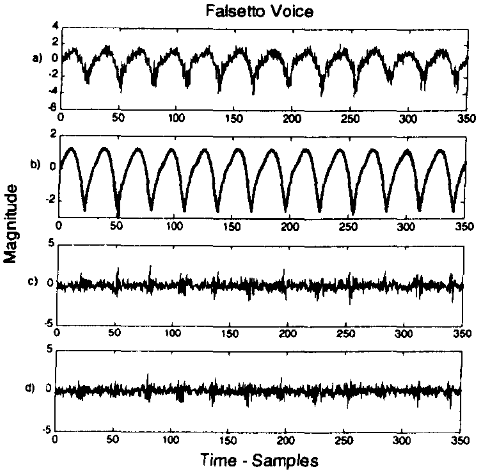


Figure B.5: Falsetto Voice*: Qualitative evaluation of the globally optimized TI-H performance: a) synthesized glottal excitation with SNR = 6dB; b) clean (dashed) and de-noised (solid) glottal excitation; c) synthesized turbulence noise; d) estimated noise.*

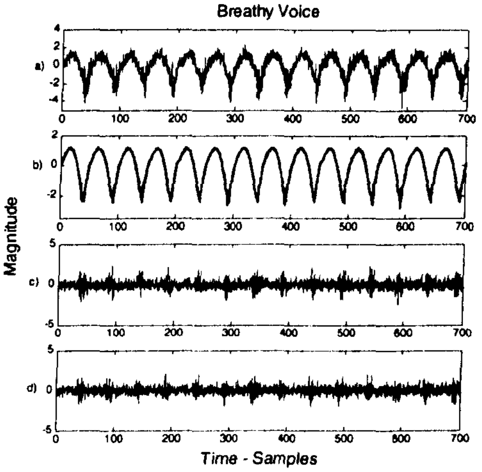


Figure B.6: Breathy Voice*: Qualitative evaluation of the globally optimized TI-H performance: a) synthesized glottal excitation with SNR = 6dB; b) clean (dashed) and de-noised (solid) glottal excitation; c) synthesized turbulence noise; d) estimated noise.*

Appendix C

Phase-Plane Plot Analysis

The phase-plane plots provide the means to examine the residual resonance characteristics, and thus, objectively asses the quality of the glottal flow derivative estimates. What follows is a brief review of phase-plane analysis.

Vocal tract can be modeled as a cascade of second order resonators. Let us consider the second order harmonic equation:

$$\frac{d^2x}{dt^2}+x=0 \tag{C.1}$$

Using the substitutions $x = x_1$ and $\frac{dx}{dt} = x_2$, eqn. C.1 can be rewritten in vector form as:

$$\begin{aligned}\frac{dx_1}{dt} &= x_2\\ \frac{dx_2}{dt} &= -x_1\end{aligned} \tag{C.2}$$

Eqn. C.2 can be solved by integrating the first-order differential equation $dx_2/dx_1 = -x_1/x_2$ to yield $x_1^2+x_2^2=K$, where K denotes a constant. Therefore, the solution to a harmonic equation is a combination of the periodic functions *cos(t)* and *sin(t).* Furthermore, the phase-plane plot of x vs. *dx/dt* belongs to a family of concentric circles. After a time period *T*, periodic solutions, x and *dx/dt*, resume their initial values and thus a periodic solution yields a closed loop in a phase plane.

We can extend this analysis to the vibratory motion of the vocal folds which has periodic solutions (with period T) of general form $dx/dt = f(t,x)$, such that $f(t,x) = f(t+T,x)$. Periodic solutions corresponding to the vibratory motion of vocal folds produce a single closed loop in the phase plane, although they are not exactly circular. If the vocal tracts resonances are present in the glottal flow estimates, than the phase plane loop will self–intersect to produce smaller loops within the large closed loop of driving solution.

The phase-plane plots can be used to evaluate the quality of voice source deconvolution from the vocal tract filter. A successful inverse filtering would remove all the vocal tract resonance information from the glottal waveform estimates. Therefore, the phase-plane plot of the voice source estimates should produce a single closed-loop with no self-intersection. If the phase-plane plot exhibits more than one loop or displays self-intersections, it would be an indication that the vocal tract resonances are present in the voice source estimate.

Bibliography

[1] P. Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," *Proc. Int. Conf. on Acoustic Speech Signal Process.*, San Francisco, USA, 2, 29-32., 1992.

[2] P. Alku, and E. Vilkman, "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," *J. Acoust. Soc. Am.* 98, 763-767, 1995.

[3] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication* 18, 131-138, 1996.

[4] T.V. Ananthapadmanabha and G. Fant, "Calculations of true glottal flow and its components", speech Communication, 1:167-184, 1982.

[5] T.V. Ananthapadmanabha, "Acoustic analysis of voice source dynamics", *Speech Transmiss. Lab. Q. Prog. Stat. Rep.*, 2-3, 1-24, 1984.

[6] A. Antoniadis, and J. Fan, "Regularization of wavelets approximations (with discussion)", *J. Am. Statist. Ass.*, 96, 2001.

[7] G. Bailley, and C. Benoit, "Talking Machines", *Elsevier Science Publishing*, North-Holland, Amsterdam, 1992.

[8] R.J. Baken, "Clinical Measurement of Speech and Voice", *College-Hill Press*, San Diego, CA., 1987.

[9] A. Barney, C. H. Shadle, and P.O.A.L. Devices, "Fluid Flow in a Dynamical Mechanical Model of the Vocal Folds and Tract: Measurements and Theory," *J. Acoustical Society of America*, vol. 105, no. 1., pp. 445-455, Jan. 1999.

[10] W. Becker, H.H. Naumann, C.R. Faltz, "Ear Nose and Throat Diseases", *Thieme Medical Publischers*, 2nd Edition, 1994.

[11] J.W. van den Berg, J.T. Zantema, P. Doorneball, "On the air resistance and the Bernoulli effect of the human larynx", J. Acoust. Soc. Am., 29(5):626-631, 1957.

[12] D.A. Berry, "Interpretation of biomechanical simulation of normal and chaotic

vocal folds oscillations with empirical Eigenfunctions", *J. Acoust. Soc. Am.*,95(6):3395-3604, June, 1994.

[13] D.A. Berry and I.R. Titze, "Normal modes in a continuum model of vocal fold tissues", *J. Acoust. Soc. Am.*, 100(5):3345-3354, Nov., 1996.

[14] T.T. Cai, and L.D. Brown, "Wavelet estimation for samples with random uniform design.", *Statist. Probab. Lett.*, 42, 313–321, 1999.

[15] T.T. Cai, and B.W. Silverman, "Incorporating information on neighboring coefficients into wavelet estimation", *Sankhya, Series A*, 63, 2001.

[16] R. Carlson, "Models of Speech Synthesis", *STL-QPSR, 1,1-14*, 1993.

[17] A. Ní Chasaide, and C. Gobl, "Linguistic and paralinguistic variation in the voice source," *Proc. Int. Conf. Spoken Language Process.*, Kobe, Jpn., 1, 85-88., 1990.

[18] A. Ní Chasaide, and C. Gobl, "Contextual variation of the vowel voice source as a function of adjacent consonants," *Language and Speech*, 36, 303-330, 1993.

[19] Y.M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1805-1815, December, 1989.

[20] D.G. Childers, and Lee, "Voice Quality factors: analysis, synthesis and perception.", *J Acoust. Soc. Am.*, 90: 2394-2410, 1991.

[21] D.G. Childers, and C. Wong, "Measuring and Modeling Vocal Tract Interaction", *IEEE Transactions on Biomedical Engineering*, 41, 663-671, 1994.

[22] D.G. Childers and T.H. Hu, "Speech synthesis by glottal excited linear prediction". *J. Acoust. Soc. Am.* 96(4), pp. 2026-36, 1994.

[23] 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.

[24] D.G. Childers, "Pitch Contour Modification", *Speech Processing and Synthesis Toolboxes*, John Wiley and Sons, pp.322-324, 2000.

[25] R.R. Coifman, and D.L. Donoho, "Translation-invariant de-noising", In Wavelets and Statistics", *Antoniadis, A. & Oppenheim, G. (Eds.), Lect. Notes Statist.*, 103, pp. 125–150, New York: SpringerVerlag, 1995.

[26] 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.

[27] K.E Cummings and M.A. Clements, "Analysis of Glottal Waveforms Across Stress Styles", *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 369-372,

Albuquerque, NM, 1990.

[28] V. Darsinos, D. Galanis, and G. Kokkinakis, “A method for fully automatic analysis and modelling of voice source characteristics”, *Proc. ESCA 4th European Conf. On Speech Communication and Technology*, Madrid, Spain, 1, 413-416., 1995.

[29] I. Daubechies, “Ten lectures on wavelets”, *CBMS, SIAM,* 61, 194-202, 1994.

[30] S.B. Davis, “Acoustic characteristics of laryngeal pathology”, *Speech Evaluation in Medicine*, J Darby, Ed. New York: Grune and Stratton, pp. 77-104, 1981.

[31] A. Dempster, N. Laird, and D. Rubin”, Maximum likelihood from incomplete data via the EM algorithm ”, *J. R. Soc.*, vol. 39, pp. 1-38, 1977.

[32] W. Ding, and H. Kasuya, “A novel approach to the estimation of voice source and vocal tract parameters from speech signals,” *Proc. Int. Conf. Spoken Language Process.*, Philadelphia, USA, 2, 1257-1260, 1996.

[33] R. van Dinther, “Perceptual aspects of voice source parameters”, *PhD thesis*, Technische Universiteit Eindhoven, pp. 51-55, 2003.

[34] R.P. Dixit, “On defining aspiration”, *Proceedings of the XIIIth International Conference of Linguistics,* Tokyo, Japan, pp. 606-610, 1988.

[35] D.L. Donoho, and I.M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage”, *Biometrika*, vol. 81, pp. 425-455, 1994.

[36] D.L. Donoho, “De-noising by soft-thresholding“, *IEEE Trans. Inform. Theory*, vol. 41, pp. 613-627, May 1995.

[37] D.L. Donoho and I.M. Johnstone, “Adapting to unknown smoothness via wavelet shrinkage”, *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200-1224, 1995.

[38] B. Doval, C. d’Alessandro, “Spectral correlates of glottal waveform models: an analytic study”, ICASSP-97, April 21-24, 1997.

[39] H. Dudley, “Synthesis Speech”, *Bell Labs. Record*, 15:98-102, 1936.

[40] J.A Edwards and J. A.S. Angus, “Using phase-plane plots to asses glottal inverse filtering”, Electronics letter, vol. 32, no. 3, pp. 192-193, 1996.

[41] M.J., Fadili, and E.T., Bullmore, “A comparative evaluation of wavelet based methods for hypothesis testing of brain activation maps,” *NeuroImage* 23, pp. 1112-1128, 2004.

[42] G. Fant, “Acoustic Theory of Speech Production”, The Hague: Mouton, 1960.

[43] G. Fant, J. Liljencrants and Q. Lin, “A four-parameter model of glottal flow”, *STL-QPSR 4*, pp. 1-13, 1985.

[44] G. Fant and Q. Lin, "Glottal source – vocal tract acoustic interaction", STL-QPSR, (1):13-27, 1987.

[45] G. Fant, "Some problems in voice source analysis", *Speech Communication*, 13:7-22, 1993.

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

[47] G. Fant, "The voice source in connected speech", *Speech Communication*, pp. 125-139, 1997.

[48] J.L. Flanagan, and L.L. Landgraf, "Self-oscillating source for vocal tract syntheciser", *IEEE Trans. Audio and Electronics*, AU-16:57-64, 1968.

[49] J.L Flanagan, "Speech Analysis, Synthesis and Perception", 3ed ed., New York: *Springer Verlag*, 1972.

[50] A. Fourcin, "Voice Quality Meassurement", Ch. 13, Kent R.D. and Ball M.J. ed., San Diago:: Singular Publishing Group, 2000.

[51] F.N. Fritsch, and R.E. Carlson, "Monotone Piecewise Cubic Interpolation," *SIAM J. Numerical Analysis*, Vol. 17, pp.238-246, 1980.

[52] H. Fujisaki, and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," *Proc. Int. Conf. on Acoustic Speech Signal Process.*, 4, Tokyo, Jpn., 1605-1608, 1986.

[53] H. Fujisaki and M. Ljungqvist, "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform", *In Proc. of IEEE Int. Conf. On Audio, Speech and Signal Processing (ICASSP-87)*, 637-640, 1987.

[54] K. Funaki, Y. Mitome, "A speech analysis method based on a glottal source model," *Proc. Int. Conf. Spoken Language Process*, Kobe, Jpn., 1, 45-48, 1990.

[55] H.Y. Gao, and A.G. Bruce, "WaveShrink with firm shrinkage", *Statist. Sinica*, 7, 855- 874, 1997.

[56] H.Y. Gao, "Wavelet shrinkage denoising using the nonnegative garrote", *J. Comp. Graph. Statist.*, 7, 469–488, 1998.

[57] J. Gauffin and J. Sundberg, "Data on the glottal voice source behavior in vowel production," *Speech Transmiss. Lab.* Q. Prog. Stat. Rep., 2-3, 61-70., 1980.

[58] C. Gobl, A. Ní Chasaide, "The role of voice quality in communicating emotion mood, attitude", *Speech Communication* 40, 189-212, 2003.

[59] M. Gordon, P. Ladefoged, "Phonation Types: a cross-linguistic overview", *Journal of Phonetics* 29,383-406, 2001.

[60] O. Gottesman, and A. Gersho, "Enhanced Waveform Interpolative Coding", *IEEE. Trans. Acoustics, Speech, and Signal Proc.*, vol. 9, no. 8, 2001

[61] P. Hall, and P. Patil, "Formulae for mean integrated square error of nonlinear wavelet-based density estimators", *Annals of Statistics*, 23, pp. 905-928, 1995.

[62] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech", *In Proc. International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, May pp. 238 – 241, 1989.

[63] W. Härdle, G. Kerkyacharian, D. Pikard, and A. Tsybakov, "A. Wavelets, Approximation, and Statistical Applications", *Lecture Notes in Statistics 129*, New York: Springer-Verlag, 1998.

[64] J. He, L. Liu and G. Palm, "On the Use of Features from Prediction Residual Signals in Speaker Identification", *Proc. Eurospeech95*, vol. 1, pp. 313 316, 1995.

[65] P. Hedelin, "A glottal LPC-vocoder", *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 161-64, 1984.

[66] D.J. Hermes, "Synthesis of breathy vowels: some research methods", *Speech Communication*, 10:497-502, 1991.

[67] H. Herzel, I. Steinecke, W. Mende, and K. Wermke, "Chaos and Bifurcations during Voiced Speech", chapter in Complexity, Chaos, and Biological Evolution, E. Moskilde, ed.,pp. 41-50, Plenum Press, NY, 1991.

[68] H. Herzel and C. Knudsen, "Bifurcation in vocal fold model", *Nonlinear Dynamics.*, 7, 53-64, 1995.

[69] S. Hiki, S. Imaizumi, M. Hirano, H. Matsushita, and Y. Kakita, "Acoustical analysis for voice disorders", *IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing*, 613-616, 1976.

[70] M. Hirano, "Morphological structure of the vocal cord as a vibrator and its variations." Folia Phoniatrica, Vol. 26, pp. 89-94, 1974.

[71] M. Hirano, "The laryngeal examination", *Speech Evaluation in Medicine*, J. Darby, Ed. New York: Grune and Stratton, 1981.

[72] M. Hirano, S. Kurita, T. Nakashima, "The structure of vocal folds", *In Vocal Fold Physiology*, K.N. Stevens, and M. Hirano, Eds. University of Tokyo Press, pp. 33-41, 1981.

[73] S. Hong, S. Kang, and S. Ann, "Voice parameter estimation using sequential SVD and wave shaping filter bank," *Proc. Int. Conf. Spoken Language Process.*, Yokohama, Jpn., 3, 1059-1062., 1994.

[74] Y., Hu, and P.C., Loizou, "Speech Enhancement Based on Wavelet thresholding Multitaper Spectrum," *IEEE Transactions on Speech and Audio Processing*, Vol.

12, no. 1, pp. 59-67, 2004.

[75] K. Ishizaka and J.L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords", *The Bell Syst. Tech. J.*, 51(6):1233-1268, July-August 1972.

[76] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality", Geneva, 1996.

[77] J. Jansen, B. Cranen, and Boves, L. "Modelling of source characteristics of speech sounds by means of the LF-model," *Proceedings of Eurospeech*, Genova, Italy, 1, 259-262, 1991.

[78] H.R. Javkin, N. Antonanzas-Barroso, I. Maddieson, "Digital inverse filtering for linguistic research", Journal of Speech and Hearing Research, vol. 30, pp. 122-129, 1987.

[79] J.J. Jiang, Y. Zhang, J. stern, "Modeling of chaotic vibrations in symmetric vocal folds", *J. Acoust. Soc. Am.* 110, 2120-2128, 2001.

[80] I.M Johnstone, and B.W. Silverman, "Wavelet threshold estimators for data with correlated noise", *J.R. Statist. Soc.*, vol. 59, pp. 319-351, 1997.

[81] I. Karlsson, "Voice source dynamics of female speakers," *Proc. Int. Conf. Spoken Language Process.*, Kobe, Jpn., 1, 69-72., 1990.

[82] I. Karlsson, "Analysis and synthesis of different voices with emphasis on female speech," *Ph.D. dissertation*, KTH, Stockholm., 1992.

[83] I. Karlsson, and J. Liljencrants, "Diverse voice qualities: models and data.", In STL-QPSR, volume 2/96, pp. 143-156, 1996.

[84] S. Kiritani, H. Imagawa, and H. Hirose, "Vocal cord vibrations and voice source characteristics - observations by a high speed digital recording", Proceedings of the International conf. on Spoken Language Processing, pp.61-64, Japan, 1990.

[85] G. Klasmeyer, W.F. Sendlmeier, "Voice and Emotional states", Voice Quality Measurement, *Singular Thompson Learning, Ch. 15*, pp. 339 -358, 2000.

[86] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am. 87(2)*, pp. 820-57, 1990.

[87] T. Koizumi, S. Taniguchi, and S. Hiromitsu., "Two-mass models of the vocal cords for natural sounding voice synthesis", *J. Acoust. Soc. Am.*, 82(4):1179-1192, 1987.

[88] J. Koreman, "Decoding linguistic information in the glottal airflow," *Ph.D. dissertation*, Univ. of Nijmegen., 1996.

[89] J. Kreiman, B. Gabelman, B.R. Geratt, "Perception of Vocal Tremor", *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 203-214, 2003.

[90] A.K. Krishnamurthy and D.G. Childers, "Two-channel speech analysis", IEEE Trans. Acoust., Speech, Signal Processing, vol. 34, pp. 730– 743, Aug 1986.

[91] B.J Kröger "A gestural production model and its application to reduction in German", *Phonetica*, 50, 213-233, 1993.

[92] G. de Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments", *J Speech Hear. Res.*, Vol. 38, pp. 794-811, 1995.

[93] J. Laver, "Principles of phonetics", *Oxford University Press.*, Oxford, UK., 1994.

[94] J. Liljencrants, "A translating and rotating mass model of the vocal folds", *STL-QPSR*, pp. 1-18, April 1991.

[95] Q.G. Lin, "Speech Production Theory and Articulatory Speech Synthesis", *PhD dissertation*, Royal Institute of Technology, Stockholm, Sweden, 1990.

[96] H.H. Lu, "Towards a high-quality singing synthesizer with vocal texture control", PhD thesis, *the department of Electrical Engineering*, Stanford University, pp.88-90, 2002.

[97] J.C. Lucero, "A theoretical study of hysteresis phenomenon at vocal fold oscillation onset-offset, *J. Acoust. Soc. Am.* 105, 423-431, 1999.

[98] J.C. Lucero, "Dynamics of the Vocal Fold Oscillation", *XXVII National Congress on applied and comp. mathematics – CNMAC*, Porto Alegre, 13-17 Sept., 2004.

[99] C. Ma, Y. Kamp, and L. F. Willems, "A frobenius norm approach to glottal closure detection from the speech signal", IEEE Trans. *Speech Audio Processing*, vol. 2, pp. 258 – 265, Apr 1994.

[100] S. Maeda, "A digital simulation method of the vocal tract system", *Speech Communication*, 1, 199-229, 1982.

[101] C. Maguire, P. de Chazal, R.B. Reilly, P. Lacy, " Automatic Classification of voice pathology using speech analysis", *World Congress on Biomedical Engineering and Medical Physics*, Sydney, August 2003.

[102] P. Mayer, R. Wilhelms, and H.W. Strube, "A quasi-articulatory speech synthesizer for German language running in real time", *J. Acoust. Soc. Am.*, 86(2):523-539, 1989.

[103] J.G. McKenna, "Automatic glottal closed-phase location and analysis by Kalman filtering", *In 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Blair Atholl, Aug 2001.

[104] D. Michealis, M. Frohlich, H.W. Strube, "Selection and combination of acoustic features for the description of pathologic voices", *J. Acoust. Soc. Am.*, Vol. 103, No 3, pp. 1628-1639, 1998.

[105] P.H. Milenkovic, "Voice source model for continuous control of pitch period", *J. Acoust. Soc. Am.* 93(2), pp. 1087-96, 1993.

[106] J.D. Miller, "Nature of the vocal cord wave", *J. Acoust. Soc. Am.,* 31: 667-677, 1959.

[107] J.D. Milton, "Head and Neck Rehabilitation Center", www.gbmc.org/voice/disorders.cfm.

[108] P.M. Morse, "Vibration of Sound", *McGraw-Hill,* New York, NY, 1948.

[109] T. Murry, E.T. Doherty, "Selected acoustic characteristics of pathological and normal speakers", J. Speech Hearing Res., vol. 23, pp. 361-359, 1980.

[110] P.S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," IEEE Trans. Speech Audio Processing, vol. 7, no. 6, pp. 609– 619, Nov 1999.

[111] J.L. Navarro-Messa, E. Lleida-Solano, and A. Moreno-Bilbao, "A new method for epoch detection based on the cohen's class of time frequency representations" , IEEE Signal Processing Lett., vol. 8, pp. 225 – 227, Aug 2001.

[112] W. Nowakowska, P. Zarnecki, "Dynamic model of the vocal tract in consonant nasalized articulation", *Archives of Acoustics,* 14, 67-96, 1989.

[113] R.T. Ogden, and E. Parzen, "Change-point approach to data analytic wavelet thresholding", *Statist. Comput.*, 6, 93–99, 1996.

[114] M. D. Plumpe, T.F. Quartiery, D.A. Reynolds, "Modeling of the Glottal Flow Derivative Waveform with Applications to Speaker identification", *IEEE Transactions on Speech and Audio Processing,* Vol. 7,no. 5, 1999.

[115] L. Rabiner, and B. Juang, "Fundamentals of Speech Recognition", *Englewood Cliffs, N.J. Prentice Hall,* 1993.

[116] E.L. Riegelsberger, and A.K. Krisnamurthy, "Glottal source estimation: methods of applying the LF-model to inverse filtering," *Proc. Int. Conf. on Acoustic Speech Signal Process, Minneapolis, USA,* 2, 542-545, 1993.

[117] M.R. Rothenberg, "A new inverse filtering technique for deriving the glottal airflow waveform during voicing", *J. Acoust. Soc. Am.,* 51: 1632-1645, 1973.

[118] M.R. Rothenberg, "Some relations between glottal airflow and vocal fold contact area", *ASHA Reports,* 11, 88-96, 1981.

[119] M.R. Rothenberg, "An interactive model for the voice source", STL-QPSR, Royal institute of Technology, Stockholm, Sweden, 4:1-17, 1981.

[120] C.A., Rosen, D., Anderson, T., Murrey, "Evaluating Hoarsness: Keeping Your Patient's Voice Healthy," American Family Physicians, 1998.

[121] H. Sakoe, and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE. Trans. Acoustics, Speech, and Signal Proc.*, Vol ASSP-26, 1978.

[122] J. Van Santen, R.W. Sproat, J. P. Olive, and J. Hirschberg, "Progress in Speech Synthesis", Springer, New York, NY, 1997.

[123] J. Schoentgen, "Dynamic Models of the Glottal Pulse," in *Levelsin Speech Communication: RelationsandInteractions*, a tribute to Max Wajskop, edited by C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen (Elsevier, Amsterdam, 249-266, 1995.

[124] R. Smits, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function", IEEE Trans. Speech Audio Processing, vol. 3, pp. 325–333, Sep 1995.

[125] K.H. Stevens, S. Kasowski, and G. Fant, "An electrical analogue of the vocal tract", *J. Acoust. Soc. Am.*, 25(4):734-742, 1953.

[126] K.N. Stevens, "Modelling affricate consonants", *Speech Communication*, 13:33-43, 1993.

[127] B. H. Story And I. R. Titze, "Voice Simulations with a body-cover model of vocal folds", *J. Acoust. Soc. Am.*, 97, 1249-1260, 1995.

[128] H. Strik, and L. Boves, "On the relation between voice source parameters and prosodic features in connected speech", *Speech Communication* 11, 167-174, 1992.

[129] H. Strik, and L. Boves, "Automatic estimation of voice source parameters" , *Proc. Int. Conf. Spoken Language Process., Yokohama,* Jpn., 1, 155-158, 1994.

[130] H. Strik, N. Oostdijk, C. Cucchiarini, "Testing two automatic methods for estimation of voice source parameters", Proceedings of the department of Language and Speech, Vol. 19, pp.105-127, Nijmegen, Netherlands, 1996.

[131] H. Strube, "Determination of the instant of glottal closure from the speech wave", J. *Acoustic Soc. America*, vol. 56, no. 5, pp. 1625–1629, 1974.

[132] Y. Stylianou, "Synchronization of speech frames based on phase data with application to concatenative speech synthesis," *in 6th European Conference on Speech Communication and Technology*, vol. 5, Budapest, pp. 2343–2346, Sep 1999.

[133] J.G. Švec, H.K, Schutte, F. Šram, "Variability of Vibration of Normal Vocal Folds as seen in Videokymography", *On Vibration Properties of Human Vocal Folds: Voice Registers, Bifurcations, Resonance Characteristics, Development and Application of Videokymography,* PhD thesis, University of Groningen, the Netherlands [ISBN: 90-367-1235-1], Ch. 8, pp. 91-93, 2000.

[134] H.M. Teager, S.M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract", chapter in *Speech Production and Speech Modeling,* W.J. Hardcastle and A. Marchal, eds., NATO Adv. Study Inst. Series D, vol. 55, Kluwer Academic Publishers, Boston, MA, pp. 241-261, 1990.

[135] P. Thévenaz, and H. Hügki, "Usefulness of the LPC-Residue in Text-Independent Speaker Verification", *Speech Communication*, vol. 17, no. 1-2, pp. 145 157, 1995.

[136] L. A. Thorpe and B. Shelton, "Subjective test methodology: MOS vs. DMOS in evaluation of speech coding algorithms," *IEEE Speech Coding Workshop*, pp.73-74 St. Adele, Quebec, Canada, 1993.

[137] I. R. Titze, "On the mechanics of vocal-fold vibration", *J. Acoust. Soc Am..,* 60(6):1366-1380, 1976.

[138] I.R. Titze, "The physics of small-amplitude oscillation of vocal cords", *J. Acoust. Soc. Am.*, 83, 1536-1552, 1988.

[139] I.R., Titze, B.H. Story, G.C. Burnett, J.F. Holzirichter, L.C Ng, and W.A. Lea, "Comparison between Electroglottography and electromagnetic glottography", The Journal of the Acoustic Soc. of America, vol. 107, no. 1, pp. 581-588, 2000.

[140] M.A. Trevisan, M.C. Eguia, and G.B. Mindlin, "Nonlinear aspects of analysis and synthesis of speech time series data", *Phys. Rev.,* 026216, E 63, 2001.

[141] R.N.J. Veldhuis, "A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation", *J. Acoust. Soc. Am.* 103, pp. 566-71, 1998.

[142] B. Vidakovic, and F. Ruggeri, "BAMS method: theory and simulations", *Discussion Paper, Institute of Statistics and Decision Sciences*, Duke University, USA, 2000.

[143] W.D. Voiers, "Methods of Predicting User Acceptance of Voice Communication Systems," Final Report, DCA100-74-C-0056, Jul. 1976.

[144] M.P. de Vries, H.K. Schutte, and G.J. Verkerke, "A new voice for the voiceless: Design and in-vitro testing of a voice producing element", *Wageningen: Ponsen and Looijen*, Ch. 2, 20-38, 2000.

[145] D.F., Walnut, "An Introduction to Wavelet Analysis," *Springer*, 2002.

[146] J.B. Weaver, X. Yansun, D.M. Healy Jr., L.D Cromwell, "Filtering noise from images with wavelet transforms", Magnetic Resonance in Medicine, 24, pp. 288-295, 1991.

[147] D.Y. Wong, J.D. Markel and A.H. Gray Jr., "Least-Squares Glottal Inverse Filtering from the acoustic Waveform", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no.4, pp. 350-355, Aug. 1979.

[148] B. Yegnanarayana and R. Veldhuis, "Extraction of Vocal-Tract System Characteristics from Speech Signals", *IEEE Trans. Speech and Audio Processing*, vol. 6, no.4, pp. 313-327, July 1998.

[149] B.D. Zangger, J. Sundberg, P.A. Lindestad, M. Thalen, "Vocal fold vibration and voice source aperiodicity in phonatorily distorted singing", *Speech, Music and Hearing*, TMH-QPSR, vol. 45: 87-91, 2003.

[150] X.P., Zhang and Z.Q., Luo, "A new time-scale adaptive denoising method based on wavelet shrinkage," *In. Proc. ICASSP99*, Phoenix, AZ, March, 1999.