University of Crete
Department of Computer Science

FO.R.T.H.
Institute of Computer Science

On the Inverse Filtering of Speech

(MSc. Thesis)

George P. Kafentzis

Heraklion
October 2010

Department of Computer Science
University of Crete

On the Inverse Filtering of Speech

Submitted to the
Department of Computer Science
in partial fulfillment of the requirements for the degree of
Master of Science

October 22, 2010

Author:

George P. Kafentzis
Department of Computer Science

Committee

Supervisor

Yannis Stylianou
Associate Professor

Member

Athanasios Mouchtaris
Assistant Professor

Member

Panagiotis Tsakalides
Professor

Accepted by:

Chairman of the
Graduate Studies Committee

Panos Trahanias
Professor

Heraklion, October 2010

ON THE INVERSE FILTERING OF SPEECH

by

George P. Kafentzis

A thesis submitted to the faculty of

University Of Crete

in partial fulfillment of the requirements for the degree of

Master of Science

Computer Science Department

University of Crete

September 2010

ABSTRACT

ON THE INVERSE FILTERING OF SPEECH

George P. Kafentzis

Computer Science Department

Master of Science

In all proposed source-filter models of speech production, Inverse Filtering (IF) is a well known technique for obtaining the glottal flow waveform, which acts as the source in the vocal tract system. The estimation of glottal flow is of high interest in a variety of speech areas, such as voice quality assessment, speech coding and synthesis as well as speech modifications. A major obstacle in comparing and/or suggesting improvements in the current state of the art approaches is simply the lack of real data concerning the glottal flow. In other words, the results obtained from various inverse filtering algorithms, cannot be directly evaluated because the actual glottal flow waveform is simply unknown. To this direction, suggestions on the use of synthetic speech that has been created using artificial glottal waveform are widely used in the literature. This kind of evaluation, however, is not truly objective because speech synthesis and IF are typically based on similar models of the human voice production

apparatus, in our case, the traditional source-filter model. This thesis presents three well-known IF methods based on Linear Prediction Analysis (LPA), and a new method, and its performance is compared to the others. The first one is based on the conventional autocorrelation LPA, and the second one on the conventional closed phase covariance LPA. The closed phase is identified using Plumpe and Quatieri's suggested method based on using statistics on the first formant frequencies during a pitch period. The third one is based the work of Alku et al, which proposed an IF method based on a Mathematically Constrained Closed Phase Covariance LPA, in which mathematical constraints are imposed on the conventional covariance analysis. This results in more realistic root locations of the model on the z-plane. Finally, Magi et al suggested a new method for extracting the vocal tract filter, called Stabilized Weighted LP Analysis (SWLP), in which a short time energy window controls the performance of the LP model. This method is suggested for IF due to its interesting property of applying emphasis on speech samples which typically occur during the closed phase region of the speech signal. This is expected to yield a more robust, in the acoustic sense, vocal tract filter estimate than the conventional autocorrelation LP. The three IF approaches along with the suggested new one are applied on a database of physically modeled speech signals. In this case, the glottal flow and the speech signal are available and direct evaluation of IF methods can be performed. Robust time and frequency parametrization measures are applied on both the actual glottal flow and the estimated ones, in order to evaluate the performance of the methods. These measures include the Normalized Amplitude Quotient (NAQ), the difference between the first two harmonics (H1-H2) of the glottal spectrum, and the Harmonic Richness Factor (HRF), along with the Signal to Reconstruction Error ratio (SRER).

Experiments conducted on physically modeled sustained vowels (/aa/, /ae/, /eh/, /ih/) of a wide range of frequencies (105 to 255 Hz) for both male and female speech. Glottal flow estimates were produced, using short time pitch synchronous analysis and synthesis for the covariance based methods, whereas for the autocorrelation methods, a long analysis window and a short synthesis window was used. The results and measures are compared and discussed, showing the prevalence of the covariance methods, but the suggested method typically produces better results than the conventional autocorrelation LP, according to our metrics.

ACKNOWLEDGMENTS

First, I would like to thank my supervisor, Professor Yannis Stylianou, for his continuous support in the M.Sc. program. I am really grateful for his advice, encouragement, guidance, motivation, and above all, trust and patience that he showed during the time we worked together. He gave me the chance to meet and work with exceptional people and scientists. He also taught me how to view things from different perspectives, how to do research, and how to be persistent when things are not going well. Finally, his help and care on matters beyond academia were really emotive, and for this I thank him twice.

I also thank Professor Paavo Alku, whom I had the opportunity to meet and work with in the Acoustics and Signal Processing Laboratory of the Aalto University of Technology, Helsinki, Finland, for his hospitality, kindness, support, and guidance, and for providing me the physically modeled speech database to work with in this Thesis.

In addition, I would like to thank Dr. Olivier Rosec, whom I had the opportunity to work with for four months in France Telecom R & D, Lannion, France. His continuous support during my stay there was very important for me.

I also thank my colleagues, past and present, with whom I shared my days at the University of Crete, and especially in the Multimedia Informatics Laboratory of the Computer Science Department. They made it a wonderful workplace and home for the past years. I would like to specially mention Christos and George Tzagkarakis, Pavlos Mattheakis, Maria Koutsogiannaki, Mairi Astrinaki, George Tzedakis, George Grekas, Christina Lionoudaki, Yannis Agiomyrgiannakis, Miltos Vasilakis, Yannis Pantazis, Andre Holzapfel, Maria Markaki, Nikos Pappas, Ilias Tsompanidis, Evgenios Kornaropoulos, and Dimitris Milioris, for helping me in so many ways and for being real friends.

Moreover, I would like to say ”thank you” to many others, including George Charitos, Kostas Perakakis, George Charitonidis, Pepi Katsigianni, Petros Androvitsaneas, Kostas Sipsas, Nasos Lagios, Andreas Vourdoulas, Panayotis Tripolitis, Eleni Choremioti, and Kostas Katsaros, for their care and friendship, even if most of them are away from Crete.

I am also grateful to the Institute of Computer Science (ICS) of the Foundation of Research and Technology - Hellas (F.O.R.T.H), for awarding me a fellowship during my graduate studies.

Many special thanks should go to Glykeria Stergioula, for consuming her time on reading and suggesting corrections on the grammar and syntax of this dissertation.

Last, but not least, I thank my family: my parents, Panayotis Kafentzis and Diamanto Tsirolia, for everything. I owe them the man I am today. My little sister Maria-Chrysanthi, for her patience, love, care, and support, in all these years we are together in Crete. My little brother Stelios, for his love and care, and for listening to his voice almost every night before going to bed.

Περίληψη

Σε όλα τα προτεινόμενα μοντέλα πηγής - φίλτρου της παραγωγής φωνής, το Αντίστροφο Φιλτράρισμα (ΑΦ) είναι μια γνωστή τεχνική για την απόκτηση της κυματομορφής της γλωττιδικής ροής, που λειτουργεί ως πηγή στο σύστημα της φωνητικής οδού. Η εκτίμηση της γλωττιδικής ροής είναι υψηλού ενδιαφέροντος σε μια ευρύτητα τομέων μελέτης της φωνής, όπως ο προσδιορισμός της ποιότητας φωνής, η κωδικοποίηση και η σύνθεση φωνής, καθώς επίσης και η τροποποίηση φωνής. Ένα μεγάλο εμπόδιο στη σύγκριση και/ή στην πρόταση βελτιώσεων όσον αφορά τις υπάρχουσες μεθόδους είναι η έλλειψη πραγματικών δεδομένων που αφορούν τη γλωττιδική ροή. Με άλλα λόγια, οι εκτιμώμενες κυματομορφές γλωττιδικής ροής από διάφορους αλγορίθμους ΑΦ, δεν μπορούν να αξιολογηθούν αντικειμενικά λόγω του ότι η πραγματική κυματομορφή γλωττιδικής ροής είναι άγνωστη. Προς αυτήν την κατεύθυνση, χρησιμοποιούνται συνθετικές κυματομορφές φωνής που έχουν δημιουργηθεί με συνθετικές κυματομορφές γλωττιδικής ροής. Όμως, αυτού του τύπου η αξιολόγηση δεν είναι πραγματικά αντικειμενική επειδή η σύνθεση φωνής και ΑΦ βασίζονται στο ίδιο μοντέλο της παραγωγής ανθρώπινης φωνής, δηλ. το γνωστό μοντέλο πηγής - φίλτρου. Σε αυτή τη διατριβή παρουσιάζονται τρεις γνωστές μέθοδοι ΑΦ βασισμένες στο μοντέλο της Γραμμικής Πρόβλεψης (ΓΠ) και μια νέα μέθοδος της οποίας η απόδοση ελέγχεται σε σχέση με τις υπόλοιπες. Η πρώτη βασίζεται στην κλασική ανάλυση ΓΠ με τη μέθοδο της αυτοσυσχέτισης και η δεύτερη στην κλασική ανάλυση ΓΠ στην κλειστή φάση της γλωττίδας με τη μέθοδο της συνδιασποράς. Η κλειστή φάση εκτιμάται με την προτεινόμενη από τους Quatier και Plumpe μέθοδο που βασίζεται σε στατιστικές πρώτης τάξης πάνω στην κίνηση της συχνότητας του πρώτου φορμαντ σε μια περίοδο. Επίσης, στην εργασία του Alku, προτάθηκε μια μέθοδος ΑΦ που βασίζεται στην ανάλυση ΓΠ με τη μέθοδο της συνδιασποράς σε κλειστή φάση με Μαθηματικούς Περιορισμούς, όπου στην κλασική ανάλυση ΓΠ συνδιασποράς εφαρμόζονται μαθηματικοί περιορισμοί που συντελούν σε πιο ρεαλιστικές

θέσεις των ριζών του μοντέλου στο μιγαδικό επίπεδο. Τέλος, στην εργασία του Magi, προτάθηκε η Ευσταθής Ανάλυση ΓΠ με Βάρη (Stabilised Weighted Linear Prediction), στην οποία ένα παράθυρο ενέργειας μικρής χρονικής διάρκειας ελέγχει την απόδοση του μοντέλου ΓΠ. Προτείνουμε τη χρήση της για ΑΦ λόγω των ιδιαίτερων ιδιοτήτων της στην απόδοση έμφασης στην κλειστή φάση της γλωττίδας, η οποία αναμένεται ότι οδηγεί σε εκτιμήσεις του φίλτρου της φωνητικής οδού που είναι πιο κοντά, με την ακουστική έννοια, στο πραγματικό φίλτρο της φωνητικής οδού. Αυτή η τεχνική, μαζί με τις δυο κλασικές και την ανάλυση ΓΠ με τη μέθοδο της συνδιασποράς με μαθηματικούς περιορισμούς, εφαρμόστηκαν σε μια βάση δεδομένων από σήματα φωνής που έχουν παραχθεί από φυσική μοντελοποίηση του συστήματος παραγωγής φωνής. Σε αυτήν την περίπτωση, η γλωττιδική ροή και το σήμα φωνής είναι διαθέσιμα και μπορεί να πραγματοποιηθεί αντικειμενική αξιολόγηση των μεθόδων ΑΦ. Εύρωστες μετρικές παραμετροποίησης χρησιμοποιήθηκαν τόσο στο πεδίο του χρόνου όσο και σε αυτό της συχνότητας, για να εκτιμηθεί η ομοιότητα των πραγματικών σημάτων με αυτές που παρήχθησαν από τις μεθόδους ΑΦ. Αυτές οι μετρικές περιλαμβάνουν τον Κανονικοποιημένο Λόγο Πλάτους (Normalized Amplitude Quotient - NAQ), τη διαφορά μεταξύ των δυο πρώτων αρμονικών του φάσματος της γλωττιδικής ροής, H1-H2, και τον Παράγοντα Αφθονίας Αρμονικών (Harmonic Richness Factor - HRF), μαζί με το λόγο Σήματος προς Σφάλμα Ανακατασκευής (Signal to Reconstruction Error ratio - SRER). Πειράματα διεξήχθησαν σε στάσιμα φωνήεντα (/αα/, /αε/, /ε/, /ι/) από την προαναφερθείσα βάση σε ένα εύρος συχνοτήτων (105 ως 255 Hz) για προσομοίωση τόσο της ανδρικής όσο και της γυναικείας φωνής. Η ανάλυση και η σύνθεση των παραγόμενων κυματομορφών γλωττιδικής ροής έγινε σύγχρονα με τις περιόδους του σήματος (pitch synchronously) με χρήση μικρού παραθύρου σύνθεσης και ανάλυσης για τις μεθόδους συνδιασποράς, και μεγάλου παράθυρου ανάλυσης και μικρού παραθύρου σύνθεσης για τις μεθόδους αυτοσυσχέτισης. Τα αποτελέσματα καταδεικνύουν την εν γένει κυριαρχία των μεθόδων συνδιασποράς αλλά η προτεινόμενη μέθοδος υπερέχει της κλασικής μεθόδου αυτοσυσχέτισης, σύμφωνα με τις μετρικές που χρησιμοποιήθηκαν.

Ευχαριστίες

Κατ΄ αρχάς, θα ήθελα να ευχαριστήσω τον επόπτη μου, Αναπληρωτή Καθηγητή Ιωάννη Στυλιανού, για τη συνεχή υποστήριξή του κατά τη διάρκεια του προγράμματος μεταπτυχιακών σπουδών. Είμαι ειλικρινά ευγνώμων για τις συμβουλές, την ενθάρρυνση, την καθοδήγηση, το κίνητρο, και πάνω απ΄ όλα, την εμπιστοσύνη και την υπομονή που έδειξε κατά το διάστημα που δουλέψαμε μαζί. Μου έδωσε την ευκαιρία να γνωρίσω και να συνεργαστώ με εξαιρετικούς ανθρώπους και επιστήμονες. Επίσης, μου έμαθε πώς να βλέπω τα πράγματα από διαφορετικές οπτικές γωνίες, πώς να κάνω έρευνα, και πώς να είμαι επίμονος όταν τα πράγματα δεν πηγαίνουν καλά. Τέλος, η βοήθεια και η έγνοια του για θέματα εκτός του ακαδημαϊκού περιβάλλοντος ήταν πραγματικά συγκινητική, και γι΄ αυτό τον ευχαριστώ διπλά.

Επίσης, θα ήθελα να ευχαριστήσω τον Καθηγητή Paavo Alku, τον οποίο είχα την ευκαιρία να γνωρίσω και να συνεργαστώ μαζί του στο εργαστήριο Ακουστικής και Επεξεργασίας Σήματος του Πολυτεχνείου Aalto, στο Helsinki της Φινλανδίας, για τη φιλοξενία, την ευγένεια, την υποστήριξη, και την καθοδήγησή του, και για την παροχή υλικού για τα πειράματα που παρουσιάζονται σε αυτό το σύγγραμμα.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Δρ. Olivier Rosec, τον οποίο είχα την ευκαιρία να γνωρίσω και να συνεργαστώ μαζί του για τέσσερις μήνες στις εγκαταστάσεις της France Telecom R & D, στη Lannion της Γαλλίας. Η συνεχής υποστήριξή του κατά τη διάρκεια της διαμονής μου εκεί ήταν πολύ σημαντική για μένα.

Επίσης, ευχαριστώ τους συναδέλφους μου, πρώην και νυν, με τους οποίους μοιράστηκα της μέρες μου στο Πανεπιστήμιο Κρήτης, και ειδικότερα στο Εργαστήριο Πολυμέσων του Τμήματος Επιστήμης Υπολογιστών. Η παρουσία τους το έκανε ένα υπέροχο εργασιακό περιβάλλον και ένα δεύτερο σπίτι για μένα, στα χρόνια που πέρασα εκεί. Θα ήθελα ιδιαίτερα να αναφέρω τους Χρήστο και Γιώργο Τζαγκαράκη, Παύλο Ματθαιάκη, Μαρία Κουτσογιαννάκη, Μαίρη Αστρινάκη, Γιώργο Τζεδάκη, Γιώργο Γρέκα, Χριστίνα Λιονουδάκη, Γιάννη Αγιομυργιαννάκη, Μίλτο Βασιλάκη, Γιάννη Πανταζή, Andre Holzapfel, Μαρία Μαρκάκη, Νίκο Παππά, Ηλία Τσομπανίδη, Ευγένιο Κορναρόπουλο, και Δημήτρη Μηλιώρη, για τη βοήθειά τους με χίλιους δυο τρόπους, και για το ότι μου στάθηκαν πραγματικοί φίλοι.

Ακόμα, θα ήθελα να πω ένα "ευχαριστώ"σε πολλούς ακόμα, στους οποίους συμπεριλαμβάνονται οι Γιώργος Χαρίτος, Κώστας Περακάκης, Γιώργος Χαριτωνίδης, Πέπη Κατσιγιάννη, Πέτρος Ανδροβιτσανέας, Κώστας Σιψάς, Νάσος Λάγιος, Ανδρέας Βούρδουλας, Παναγιώτης Τριπολίτης, Ελένη Χωρεμιώτη, και Κώστας Κατσάρος, για την έγνοια και τη φιλία τους, έστω κι αν οι περισσότεροι απ΄ αυτούς είναι μακριά απ΄ την Κρήτη.

Δε θα μπορούσα να ξεχάσω να ευχαριστήσω το Ινστιτούτο Πληροφορικής (ΙΠ) του Ιδρύματος Τεχνολογίας και Έρευνας (ΙΤΕ), για την υποτροφία που μου παρείχε κατά τη διάρκεια των σπουδών μου.

Επίσης, θα ήθελα να ευχαριστήσω ιδιαιτέρως τη Γλυκερία Στεργιούλα, για το χρόνο που αφιέρωσε στην ανάγνωση του συγγράμματος αυτού, και για τις διορθώσεις που πρότεινε στη γραμματική και στο συντακτικό του κειμένου.

Τελευταίους, αλλά όχι έσχατους, θα ήθελα να ευχαριστήσω την οικογένειά μου: τους γονείς μου, Παναγιώτη Καφεντζή και Διαμάντω Τσιρολιά, για τα πάντα. Τους χρωστάω αυτό που είμαι σήμερα. Την αδελφή μου, Μαρία-Χρυσάνθη, για την υπομονή, την αγάπη, την έγνοια, την ανεκτικότητα, και την υποστήριξή της, σε όλα αυτά τα χρόνια που είμαστε μαζί στην Κρήτη. Τον αδελφό μου, Στέλιο, για την αγάπη και την έγνοια του, και για το άκουσμα της φωνής του σχεδόν κάθε βράδυ πριν πέσω για ύπνο...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The source of human speech production system, called the *glottal flow*, poses an important role on the characteristics of speech and on several scientific fields. The glottal flow can be obtained from the speech signal using a technique called *inverse filtering* (IF). Inverse filtering is extensively used in basic research of speech production and its applications to speech analysis, synthesis, and modification. Also, an increased interest is recently risen in areas of speech science as environmental voice care, voice pathology detection, and analysis of the emotional content of speech.

Most of the suggested techniques in the literature are based on Linear Prediction (LP) Analysis [43]. The goal of this thesis is to evaluate the performance of a number of IF techniques, using robust time and frequency domain parametrization measures on a database of physically modeled speech signals. This chapter starts with a mathematical framework of the source-filter model of speech production. We then provide motivation for the importance of assessing the performance of IF techniques and this is followed by a description of previous efforts towards that direction. Finally, we discuss the contribution of this thesis and we provide an outline for the remainder of this thesis.

1.1 Background

A mathematical framework of the classic source-filter model of speech production model is presented here.

1.1.1 Source-Filter Production Model

Speech production can be considered as a linear filtering operation, which is time invariant over short time periods. The overall system of speech production can be divided into three parts: the vocal tract, with impulse response $v[n]$, the excitation signal $e[n]$, which is the input of the vocal tract, and the lip radiation $r[n]$:

$$s[n] = e[n] * v[n] * r[n]$$

where $s[n]$ is the speech pressure output, i.e. the speech signal, and * denotes convolution. For voiced speech, the excitation is a periodic series of pulses, whereas for unvoiced speech, the excitation has the properties of random noise. Figure 1.1 depicts this simple model.

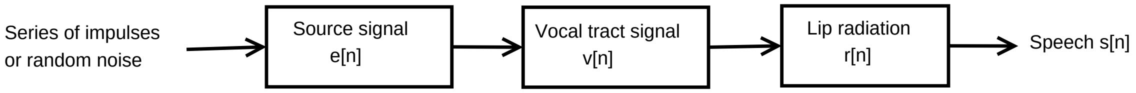

Figure 1.1 Simple Source-Filter Model.

In the Z-domain, the above equation can be written as:

$$S(z) = E(z)V(z)R(z)$$

It is shown that the vocal tract filter $V(z)$ can be written as an all-pole filter:

$$V(z) = \frac{1}{\prod_{k=1}^{p}(1 - c_k z^{-1})} = \frac{1}{\sum_{k=1}^{p}(\alpha_k z^{-k})},$$

where p is the number of poles of the filter.

1.1.2 Linear Prediction Analysis

A primary tool for inverse filtering speech waveforms is *Linear Prediction* (LP). LP is a very powerful modeling technique which may be applied to time series data. When applied to the speech signal, LP is used to produce an all-pole model of the system filter, $V(z)$, which turns out to be a model of the vocal tract and its resonances or formants. As it is previously mentioned, the input to such a model is either a series of pulses or white noise, for voiced speech and unvoiced speech respectively.

So, using LP we can estimate the vocal tract filter $v[n]$ from the speech signal, and then cancel its effect, thus resulting in the source waveform. In order to find the vocal tract response, we set up a least squares minimization problem where the error

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$$

is to be minimized, where a_k are the estimates of α_k. The minimization is performed over a region R. The total error is given by

$$E = \sum_{R} e^2[n]$$

The solutions of this minimization problem are called *linear prediction*, from the fact that a speech sample $x[n]$ can be written as a linear combination of p previous samples, that is, it can be predicted from p previous samples. Depending on the region selection, we are lead into two different techniques of linear prediction. These two techniques, as well as improvements on these, will be presented in a later chapter.

As is well known, using the method of least squares, this model has been successfully applied to a wide range of signals: deterministic, random, stationary, and non-stationary, including speech, where the method has been applied assuming local stationarity. LP has a number of advantages, including:

1. Mathematical tractability of the error measure
2. Stability of the model
3. Favorable computational characteristics of the resulting formulations
4. Spectral estimation properties
5. Wide applicability to a range of signal types

1.1.3 Inverse Filtering

The idea behind inverse filtering is to form a computational model for the vocal tract signal and then to cancel its effect from the speech waveform by filtering the speech signal through the inverse of the model. When we obtain an estimate, $\hat{V}(z)$, of the vocal tract filter $V(z)$, we can cancel its effect by removing the vocal tract response from the speech signal $s[n]$. After that, we have an estimate of the driving function, $\hat{e}[n]$, which is the combined signal of the glottal flow and the lip radiation. In frequency domain,

$$\hat{E}(z) = \frac{S(z)}{\hat{V}(z)}.$$

The above equation describes a process called *inverse filtering.* It is apparent that inverse filtering is greatly depended on the estimate of the vocal tract filter, $\hat{V}(z)$. The problem of robustly and accurately estimation of the vocal tract is often called *spectral estimation* in the literature.

A common procedure before linear prediction analysis and inverse filtering is *pre-emphasizing* the speech signal. Pre-emphasis is the process of filtering the speech signal with a single zero high-pass filter:

$$s_{emph} = s[n] - \beta_{emph} s[n-1],$$

where β_{emph} is the pre-emphasis coefficient and its values are between 0.9 and 0.99.

The reason for pre-emphasis is that the pre-emphasized spectrum is a closer representation of the vocal tract response, thus allowing linear prediction to better match the vocal tract response rather than the spectrum of the combined excitation and vocal tract.

Another common procedure before applying vocal tract LP analysis is applying a LP analysis of order one to acquire a preliminary estimate for the combined effects of the glottal flow and the lip radiation effect on the speech spectrum. Then, the estimated effects are cancelled from speech by invserse filtering with the corresponding filter.

1.1.4 The Glottal Flow

The goal of inverse filtering is the estimation of the voice source, e.g. the glottal flow. The glottal flow is the output of the glottis during voicing and thus it is a periodic signal. A period of the glottal flow can be divided in two main parts, corresponding to the state of the glottis:

1. The glottal *open phase*, where the glottis starts to open, reaches its full openess, and eventually closes again. The open phase can be further divided into the *opening phase*, where the glottal flow increases from baseline at time 0 to its maximum amplitude A_v, also called *amplitude of voicing* at time T_p, and the

closing phase, where the glottal flow decreases from A_v to a point at time T_e where the derivative reaches its negative extremum E_e. T_e is the so-called *glottal closing instant* (GCI) and E_e is called the *maximum excitation*.

2. The glottal *closed phase*, where the glottis is closed and no airflow pressure comes into the vocal tract. In normal voicing, after time T_e, the glottal flow derivative is continuous and exponentially returns to 0 at time T_c. This phase is called *return phase* and the exponential time constant is noted T_a.

Figure 1.2 illustrates this analysis on both the glottal flow and the glottal flow derivative waveform.

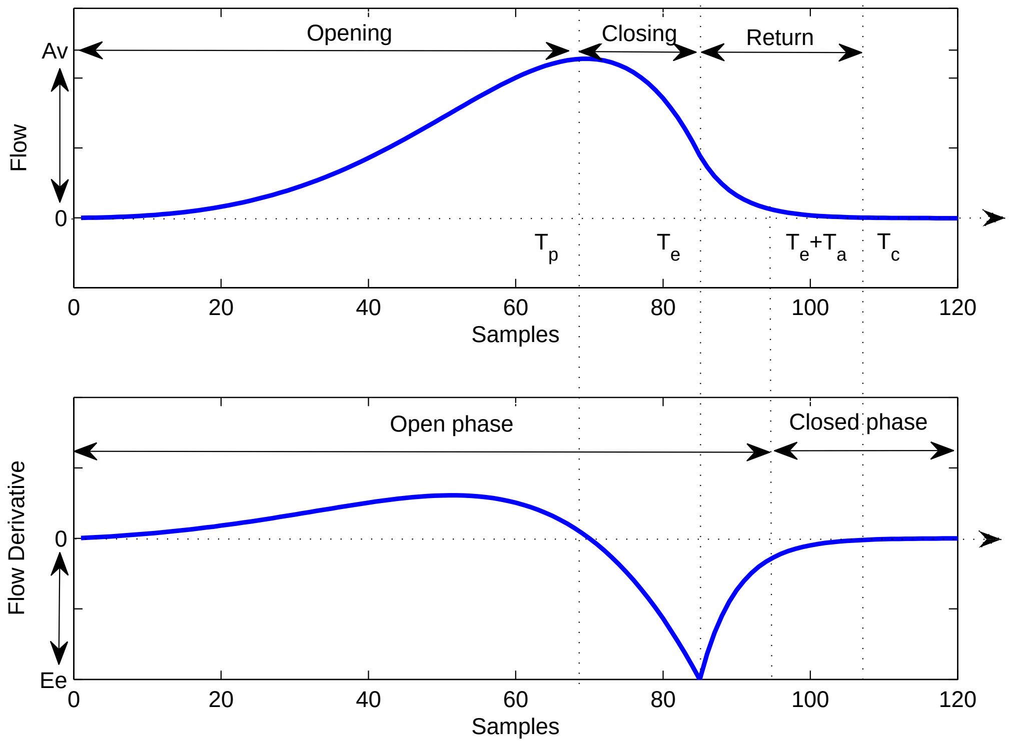

Figure 1.2 Phases of the glottal flow and its derivative.

1.1.5 Closed Phase Inverse Filtering

It is advantageous to restrict the linear prediction analysis region to the time interval where the glottis is closed. The reason is that during closed phase, there is no source/vocal tract interaction [18] and an accurate vocal tract estimate can be calculated. This estimate can then be used to inverse filter both the open and the closed phase.

The identification of the glottal closed phase interval from the speech signal is a difficult task. In the literature, there are several approaches in accurately determining the closed phase (and consequently, the glottal flow) in a non-invasive manner. Wong et al [71] proposed the use of the maximum determinant of the covariance matrix in order to find the closed phase interval. Ananthapadmanabha and Yegnanarayana [1] studied the linear prediction residual in order to find the closed phase. In [50], the authors proposed to detect discontinuities in frequency by confining the analysis around a single frequency. In this latter work, GCIs correspond to the positive zero-crossings of a filtered signal obtained by successive integrations of the speech waveform and followed by a mean removal operation. Childers et al have discussed two systems [16] [15] [37]. The first one is a two-channel analysis approach, using the electroglottographic (EGG) signal along with the speech signal. The second system uses weighting on speech samples based on the error in previous analysis windows. In [28], [17], and [71], it is suggested that an operator driven technique is required to estimate the closed phase interval. McKenna [46] suggested the use of Kalman filtering for closed phase identification. Plumpe et al [53] discussed the importance of an automatic system to find glottal opening and closure, and suggested the use of a sliding covariance analysis window for calculating the formant trajectories. Then, the closed phase interval was identified based on first order statistics on the motion of the first formant. This technique is used in this thesis whenever a closed phase covariance analysis is discussed, as it will be seen in later chapters.

1.1.6 Parametrization of the Source

Parametrization of the voice source has been the target of intensive research during the past few decades. The goal of parametrization is the expression of the most important features of the glottal flow (or its derivative) using few numerical values. In the past years, a large number of possible parametric representations of glottal flows given by inverse filtering have been proposed. Although quantitative signal processing measures such as the Signal to Reconstruction Error Ratio (SRER) can be applied on the glottal flow estimates, it is also desirable to apply measures that indicate the quality of IF based on the source-filter theory. Thus, we introduce here some of these metrics.

In time domain, there is a considerable amount of parametrization measures concerning the glottal flow (or its derivative) in the literature. This corresponds to quantifying the glottal flow using certain quotients between the closed phase, the opening phase, the closing phase of the glottal volume velocity waveform [31]. Time-based measures have also been computed using the first derivative of the glottal flow by applying, for example, the time difference between the beginning of the closing phase and the instant of the maximal negative peak [64]. The three most common time-based parameters are: (1) the open quotient (OQ), which is the ration of between the open phase of the glottal pulse and the length of the fundamental period; (2) the speed quotient (SQ), which is the ratio between glottal opening and closing phases; and (3) the closing quotient (CQ), which is the ratio between the glottal closing phase and the length of the fundamental period. However, the extraction of these time-based parameters is often problematic due to the presence of noise or formant

ripple in the estimated waveforms. In [7], a more robust time-domain parameter was introduced, called the *Normalized Amplitude Quotient - NAQ*, which is defined as the ratio between the maximum value of the glottal flow, f_{ac}, to the product of the minimum value of the glottal flow derivative, d_{peak}, and the length of the fundamental period, T. The NAQ's quality over its conventional counterpart, CQ, was demonstrated in both clean and noisy speech conditions. In addition, the calculation of NAQ is straightforward and can be computed automatically.
The NAQ is therefore selected as the time-domain parameter for IF evaluation in this thesis.

In addition, frequency-domain methods have been developed to quantify the voice source. These are typically based on measuring the decay of the voice source spectrum either from the spectral harmonics [33] [15], or from the pitch-synchronously computed spectrum [10]. This is justified by the fact that the harmonics in the speech signal below the first formant are often considered important for the perception of vocal quality [32]. It has also been found that the glottal spectra shows distinctive amplitude relationships between the fundamental and higher harmonics [15].
A well known technique for frequency domain parametrization is called Parabolic Spectral Parameter (PSP) [10], is based on fitting a parabolic function to a pitch synchronously computed spectrum of the estimated voice source. The PSP algorithm gives a single numerical value that describes how the spectral decay of an obtained glottal flow behaves with respect to theoretical bounds corresponding to maximal and minimal spectral tilting.
Another very common frequency domain parametrization measure is called H1H2, which is defined as the difference in dB between the amplitudes of the fundamental and the second harmonic of the source spectrum [68]. Finally, another parameter, the

Harmonic Richness Factor (HRF), is defined from the spectrum of the glottal flow as the difference in dB between the sum of the harmonic amplitudes above the fundamental frequency and the amplitude of the fundamental [15]. These two frequency fomain measures are selected for IF evaluation in this thesis. There are two reason for selecting these two parameters. First, both of them can be computed automatically, with low complexity, and without any user adjustments. Second, both of them are known to reflect the spectral decay of the glottal excitation. This means that if, for example, the glottal flow estimates appear to have the so-called “jags” [71], which are sharp negative peaks of the glottal flow near glottal closure, then this would be reflected into the H1H2 and HRF values.

Finally, one category of voice source parametrization methods is represented by techniques that fit certain predefined mathematical functions to the obtained glottal waveform. In noisy conditions, IF is known to perform poorly, so the glottal flow estimates (and their derivatives) would be severely distorted. This distortion can be alleviated if the mathematical function is firstly fit onto the distorted glottal flow derivative, and then calculate time-based or spectral measures using this mathematical model, instead of the original, distorted derivative. Among them, the Liljencrants-Fant (LF) [23] model is very popular. Parametrization methods for both time and frequency domain have been developed for the LF model [35].
LF-waveform parameterization is an intricate process, the complexities and implications of which are not always fully appreciated. In terms of ’accuracy’, the result of a parametric fit greatly depends on the optimization algorithm and the cost function to be minimized. In addition, the results are heavily influenced by the presence of random or systematic error in the signal to be fitted. Finally, if the model differs substantially from the process that generated the signal, the concept of accuracy of

the fit looses most of its meaning. A quick closer look at the LF model follows next.

The LF model

In a previous section, we saw that the speech signal can be represented as:

$$s[n] = e[n] * v[n] * r[n],$$

where $e[n]$ is the excitation signal, $v[n]$ is the vocal tract impulse response, $r[n]$ is the lip radiation, and * denotes convolution. The lip radiation can be modelled as a first order differentiator, $r[n] = \delta[n] - \delta[n-1]$.

Re-arranging the terms, we have:

$$s[n] = e[n] * v[n] * r[n] = e[n] * r[n] * v[n] = \hat{e}[n] * v[n],$$

where $\hat{e}[n]$ is the derivative of the excitation signal, which is the source signal. Since the source of the speech production system is the glottal flow, $\hat{e}[n]$ is the derivative of the glottal flow, the so-called *glottal flow derivative*. The glottal flow derivative is often called the *driving function*.

In the late 1980's, Liljencrants and Fant suggested a model for the derivative of the glottal flow, called the LF model [23]. The LF model is a four parameter model, although in [53], it is extended up to seven parameters, including time instants, for speaker identification applications. The LF model is described by the following equations:

$$x_{LF}(t) = \begin{cases} E_0 e^{at} \sin(\omega_g t), & 0 \leq t \leq T_e, \\ -\dfrac{E_0}{\beta T_a}(e^{-\beta(t-T_e)} - e^{-\beta(T_c-T_e)}), & T_e \leq t \leq T_c \\ 0, & elsewhere \end{cases},$$

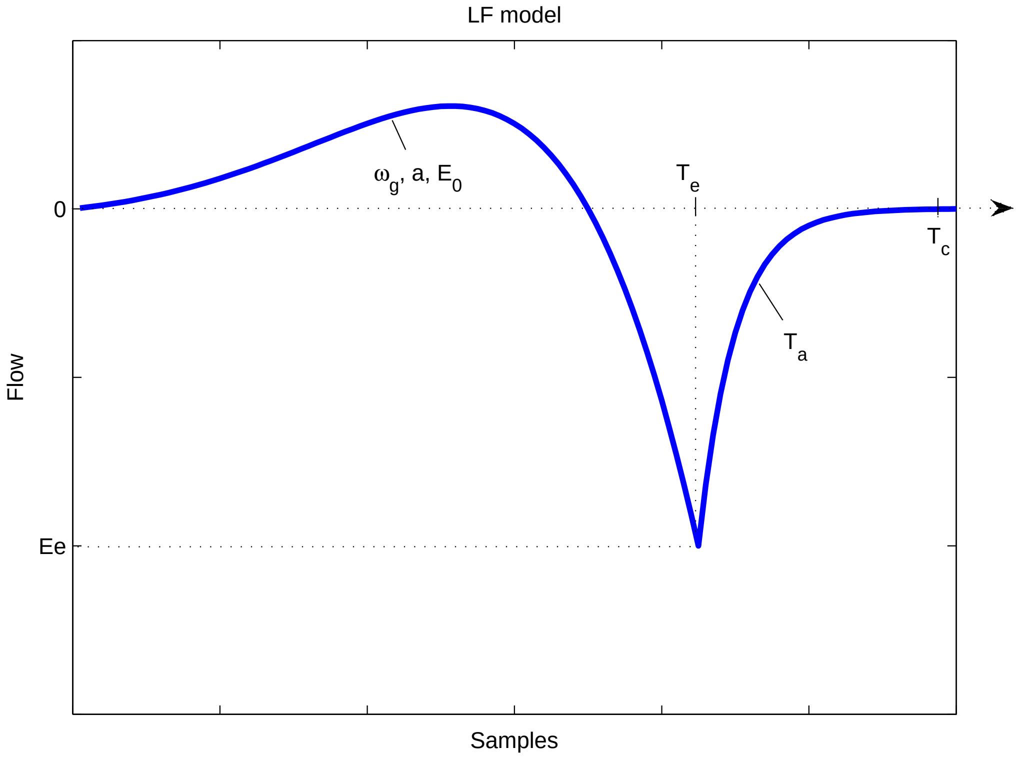

Figure 1.3 Liljencrants-Fant Model.

where T_c, T_e, T_a are illustrated on Figure 1.3, and represent the glottal closure, the glottal excitation, and the return phase. The four parameters of the model are α, ω_g, E_0, which describe the open phase, and T_a, which describes the return phase. The parameter β is dependent on T_a and although there is no closed form for the relationship of β and T_a, it can be assumed that $\beta \approx \frac{1}{T_a}$ for small values of β. Another assumption is that the glottal closure instant, T_c, can be considered to coincide with the glottal opening instant, T_o, of the next period. The four parameters do not include the timings of glottal opening, excitation, and closure, so these timing values must be given.

Due to the large dependence of E_0 on α [53], the parameter E_e, the value of the waveform at time T_e, can be estimated instead of E_0. To calculate E_0 from E_e the equation

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

is used.

The parameter T_a is the most important parameter in terms of human perception, as it controls the amount of spectral tilt present in the source. The return phase of the LF-model is equivalent to a first order lowpass filter [22] with a corner frequency of

$$F_a = \frac{1}{2\pi T_a}$$

The LF-model parameters and their significance is illustrated in Table 1.1.

LF model parameters	
E_e	The value of the waveform at time T_e
α	Determines the ratio of E_e to the height of the positive portion of the glottal flow derivative
ω_g	Determines the curvature of the left side of the glottal pulse.
T_a	An exponential time constant which determines how quickly the waveform returns to zero after time T_e

Table 1.1 The four parameters of the Liljencrant-Fant model for the glottal flow derivative waveform.

1.2 Evaluation of Inverse Filtering Methods

A major obstacle both in developing new IF algorithms and in comparing existing methods is the complication of assessing the performance of an IF technique. This happens mostly because the signal to be estimated, the glottal flow, is unavailable. Therefore, when IF is used to estimate the glottal flow of natural speech, it is actually never possible to assess in detail how closely the estimated waveform corrseponds to the true glottal flow generated by the vibrating vocal folds.

However, it is possible to use synthesized speech that has been created using artificial glottal flow waveforms to assess the accuracy and efficiency of an IF technique. The success of the algorithm can be judged by quantifying the error between the known input waveform and the version recovered by the algorithm. Although this approach is typically used in the literature [5], [6], [45], [65], [70], this kind of evaluation is not truly objective because speech synthesis and IF analysis are based on similar models of the human voice production system, such as the source-filter model. An improvement would be to provide synthesized speech generated by a more sophisticated articulatory model [34], which allows source-tract interaction [47] [8] [9]. A similar approach will be followed in this thesis, as it will be presented later.

Once the algorithm has been verified and is being used for inverse filtering real speech samples, there are two possible approaches to evaluate the results. One is to compare the waveforms obtained with those obtained by earlier methods. As, typically, the aim of this is to establish that the new method is superior, the objectivity of this approach is also doubtful. This approach can be made most objective when methods are compared using synthetic speech and results can be compared with the original source, as in [6]. In many works, no comparisons are made, a stance which is not wholly unjustified because there is not enough data available to say which are the correct glottal flow waveforms.

On the other hand, using two different methods to extract the glottal flow could be an effective way to confirm the appearance of the waveform as correct. If new techniques for glottal inverse filtering produce waveforms which 'look like' the other waveforms that have been produced before, then they are evaluated as better than those which do not: examples of the latter include [4], [19].

In the direction of evaluation of IF methods, a physiological-based model of vocal folds and vocal tract is used in order to evaluate different IF methods. In this model, time-varying waveforms of the glottal flow and radiated acoustic pressure are simulated. A detailed description of this model follows next.

1.3 Physical Modeling of Speech Signals

A computational model of the vocal folds and acoustic wave propagation generated the sound pressure and glottal flow waveforms used in this thesis. In detail, self sustained vocal fold vibration was simulated with three masses coupled to one another through stiffness and damping elements [60]. A schematic diagram depicting this model is shown in Figure 1.4, where the arrangement of the masses was designed to emulate the body-cover structure of the vocal folds [29].

The input parameters of this model are the lung pressure, prephonatory glottal half-width (adduction), resulting vocal fold length and thickness, and normalized activation levels of the cricothyroid (CT) and thyroarytenoid (TA) muscles. These values were transformed into mechanical parameters of the model, such as mass, stiffness,

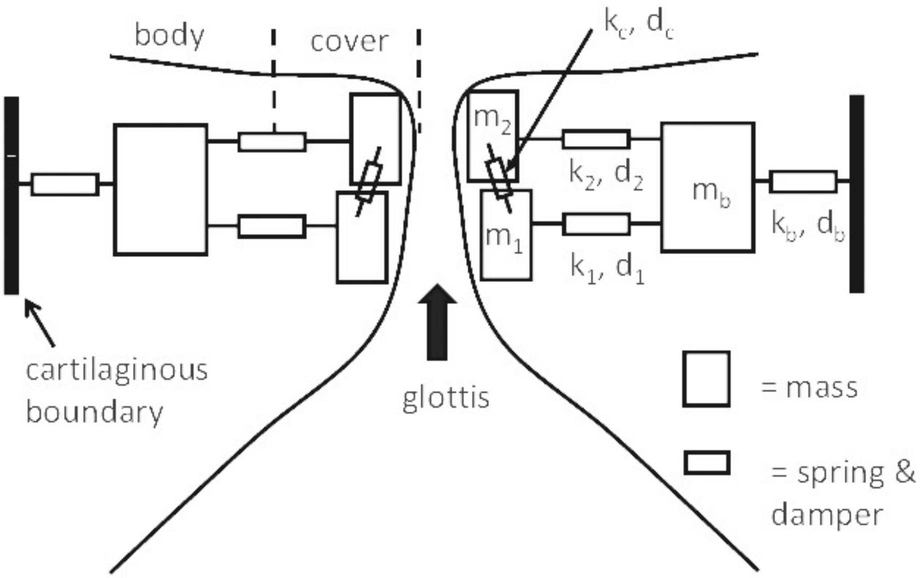

Figure 1.4 Schematic diagram of the lumped-element vocal fold model. The cover-body structure of each vocal fold is represented by three masses that are coupled to each other by spring and damping elements. Bilateral symmetry was assumed for all simulations.

and damping, acording to [67]. In [66], through aerodynamic and acoustic considerations, the vocal fold model was coupled to the pressures and air flows in the trachea and vocal tract. Bilateral symmetry was assumed for all simulations such that identical vibrations occur within both the left and right folds. The modification of the resting vocal fold length and activation levels of CT and TA muscles resulted into eight different fundamental frequency values $(105, 115, 130, 145, 205, 210, 230,$ and 255 Hz). These values roughly approximate the range of the fundamental frequency in adult male and female speech [30]. The input parameters for all nine cases are shown in Table 1.2.

Acoustic wave propagation in both the trachea and vocal tract was computed in time synchrony with the vocal fold model. This was performed with a wave-reflection

Parameter value	105	115	130	145	205	210	230	255
a_{CT}	0.1	0.4	0.1	0.4	0.2	0.3	0.3	0.4
a_{TA}	0.1	0.1	0.4	0.4	0.2	0.2	0.3	0.4
$L_0(cm)$	1.6	1.6	1.6	1.6	0.9	0.9	0.9	0.9
$T_0(cm)$	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
$\xi_{01}(cm)$	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
$\xi_{02}(cm)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$P_L(dyn/cm^2)$	7840	7840	7840	7840	7840	7840	7840	7840

Table 1.2 Input parameters for the vocal fold model used to generate the nine different fundamental frequencies. Notation is identical to that used in Titze and Story (2002). The a_{CT} and a_{TA} are normalized activation levels (can range from 0 to 1) of the CT and TA muscles, respectively. L_o and T_o are the resting length and thickness of the vocal folds, respectively. $\xi 01$ and $\xi 02$ are the prephonatory glottal half-widths at the inferior and superior edges of vocal folds, respectively, and P_L is the respiratory pressure applied at the entrance of the trachea. The value of P_L shown in the table is equivalent to a pressure of 8 cm H_2O.

approach [63] [39], where the area function of the vocal tract and trachea were discretized into short cylindrical sections or tubelets. Reflection and transmission coefficients were calculated at the junctions of consecutive tubelets, at each time sample. From these, pressure and volume velocity were then computed to propagate the acoustic waves through the system. The glottal flow was determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis, as described by Titze [66]. At the lip termination, the forward and backward traveling pressure wave components were subjected to a radiation load modeled as a resistance in parallel with an inductance, as suggested by Flanagan [25], intended to approximate a piston in an infinite plane baffle. The output pressure is assumed to be representative of the pressure radiated at the lips. To the extent that the piston-in-a-baffle reasonably approximates the radiaton load, the calculated output pressure can also be assumed to be representative of the pressure that would be transduced by a microphone in a non-reflective environment. The specific implementation of the vocal tract model used for in this thesis was presented in Story [59], and included energy losses due to viscosity, yielding walls, heat conduction, as well as radiation at the lips.

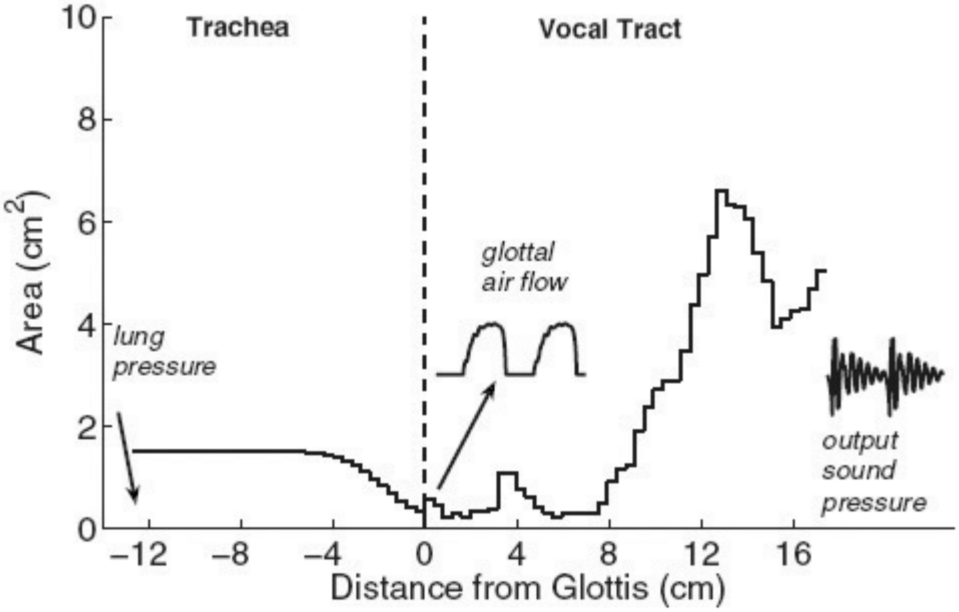

Figure 1.5 Area function representation of the trachea and vocal tract used to simulate the male "a"vowel. The vocal fold model of Fig. 1.4 would be located at the 0 cm point indicated by the dashed vertical line. Examples of the glottal flow and output pressure waveforms are shown near the locations at which they would be generated.

In this thesis, glottal flow and speech pressure waveforms were generated for 4 vowels (/aa/, /ae/, /eh/, and /ih/) in both male and female configurations. The area functions were taken from those reported for an adult male in Story et al [61]. For the simulations of male speech, the area functions were used directly with the exception of the vocal tract length, which was normalized to 17.46 cm. For female speech simulations, the same male area factors were based on those reported in Fitch and Giedd [24]. The trachea in all cases was the same as that shown in Figure 1.5.

In summary, the model is a simplified but physically motivated representation of a speaker. It generates both the signal on which inverse filtering is typically performed (microphone signal) and the signal that is seeked to be determined (glottal flow). This provides an idealized test case for inverse filtering algorithms.

1.4 Thesis Contribution

In this thesis, a physiological-based model of the vocal folds and vocal tract is used in order to evaluate different IF methods. In this model, time-varying waveforms of the glottal flow and radiated acoustic pressure are simulated. By using this simulated speech pressure waveform as an input to an IF method, it is possible to determnine how close the obtained estimate of the glottal flow is to the simulated glottal flow. This approach differs from using synthetic speech excited by an artificial glottal flow because the glottal flow waveform results from the interaction of the self-sustained oscillations of the vocal folds with subglottal and supraglottal pressures, as it would happen during real speech production. Therefore, this model generates a glottal flow waveform that is expected to provide a more firm and realistic test of the IF method than a parametric flow model, where no source-tract interaction is taken into account.

In order to evaluate the different IF techniques, robust time and frequency domain parametrization measures are typically used. In this way, the most important features of the glottal flow (or its derivative) estimates are expressed using a few numerical values. The parametrization measures used in this thesis are the Normalized Amplitude Quotient (NAQ), the Signal to Reconstruction Error ratio (SRER), the difference in dB between the first two harmonics of the glottal spectrum (H1-H2), and the Harmonic Richness Factor (HRF), which were previously discussed.

1.5 Thesis Organization

In this chapter, the basic concepts of source-filter voice production system were introduced, and their relation to the inverse filtering process. The problem of evaluation of IF techniques was illustrated, and a database of physically modeled speech pressure signals was delineated. This database will greatly help in comparing and evaluating different methods of IF. In Chapter 2, we will discuss the different vocal tract filter estimation methods and their properties, as well as the inverse filtering procedure followed for each method. Chapter 3 covers the results of each IF method, using the measures introduced in this chapter, and finally Chapter 4 concludes the thesis and discusses ideas for future directions in related research.

Chapter 2

Vocal Tract Filter Estimation with Linear Prediction

As discussed in Chapter 1, most of the IF approaches suggest the use of Linear Prediction Analysis in order to estimate the vocal tract filter. In this chapter, we firstly discuss the linear prediction techniques that are used througout the rest of the thesis. Next, we describe the inverse filtering procedure that we follow to estimate the glottal flow waveforms.

2.1 Autocorrelation based methods

2.1.1 Classic Linear Prediction - Autocorrelation Method

If we assume that the speech signal is zero outside an interval $0 \leq n \leq N-1$, then the error signal $e[n]$ will be nonzero in the interval $0 \leq n \leq N+p-1$. That interval is the region R. Since we are trying to predict nonzero samples from zero samples at the beginning of the interval, this will result in large errors at that point. This is also valid for the end of the interval, where zero samples are predicted from nonzero ones. For this reason, a tapered window (e.g., Hamming) is often used. The forementioned assumptions result in the *autocorrelation method* of linear prediction. It can be formulated as follows:

By using Hamming window $w[n]$ with length N, we get a windowed speech segment $s_N[n]$, where $s_N[n] = s[n]w[n]$. Then the mean squared prediction error is defined as:

$$E_N = \sum_{n=-\infty}^{\infty} e^2[n] = \sum_{n=-\infty}^{\infty} (s_N[n] - \sum_{k=1}^{p} a_k s_N[n-k])^2. \tag{2.1}$$

The values of a_k that minimize E_N are found by assigning the partial derivatives of E_N with respect to a_k to zeros. This follows in the following p equations with p unknown variables a_k:

$$\sum_{k=1}^{p} a_k \sum_{n=-\infty}^{\infty} s_N[n-i] s_N[n-k] = \sum_{n=-\infty}^{\infty} s_N[n-i] s_N[n], \quad 1 \leq \ i \leq p. \tag{2.2}$$

Noticing that the windowed speech signal $s_N[n] = 0$ outside the window $w[n]$, and by introducing the autocorrelation function

$$R(i) = \sum_{n=i}^{N-1} s_N[n] s_N[n-i], \quad 0 \leq i \leq p, \tag{2.3}$$

equation X becomes

$$\sum_{k=1}^{p} R(|i-k|) a_k = R(i), \quad 1 \leq i \leq p. \tag{2.4}$$

Using matrix notation, the latter equation can be written as

$$\mathbf{\Phi} \vec{a} = \vec{r}, \tag{2.5}$$

where matrix Φ is called the *autocorrelation matrix* and its elements are given by $\Phi_{i,j} = R(|i-j|), \quad 1 \leq i, j \leq p$. The other two vectors are given by

$$\vec{a} = [a_1, a_2, a_3, ..., a_p]^T \text{ and } \vec{r} = [R(1), R(2), R(3), ..., R(p)]^T.$$

The main advantage of the autocorrelation method is that it always produces a stable filter with a reasonable computational load and that there are fast algorithms, such as *the Levinson-Durbin recursion algorithm* [55], for solving the matrix system. The effects of the problems at the beggining and end of the analysis region can be reduced using a non-rectangular window, such as a Hamming window.

2.1.2 Stabilized Weighted Linear Prediction

Stabilized Weighted Linear Prediction (SWLP), introduced by Magi et al [42], is an all-pole modeling method based on Weighted Linear Prediction (WLP) [41]. WLP uses time domain weighting of the square of the prediction error signal. This temporal weighting emphasizes the speech samples which have a high signal-to-noise ratio, and thus it has been shown that WLP improves the spectral envelope estimation of noisy speech in comparison to the conventional LP analysis. Moreover, in contrast to other robust methods of LP [40] [73], the WLP filter parameters can be calculated without any iterative update. A problem is that the WLP filter is not guaranteed to be stable. This drawback is dissolved through SWLP, where the weighting is such that the all-pole model is always stable.

The formulation of WLP is presented next, as well as the SWLP algorithm that ensures stability of the all-pole model.

Weighted Linear Prediction

As in conventional LP, sample $x[n]$ is estimated by a linear combination of the past p samples:

$$\hat{x}[n] = -\sum_{i=1}^{p} a_i x[n-i], \tag{2.6}$$

where the coefficients $a_i \in \Re$. The prediction error $e_n(\mathbf{a})$, the residual, is defined as

$$e_n(\mathbf{a}) = x[n] - \hat{x}[n] = x[n] + \sum_{i=1}^{p} a_i x[n-i] = \mathbf{a}^T\mathbf{x}[n], \tag{2.7}$$

where $\mathbf{a} = [a_0 a_1 ... a_p]^T$ with $a_0 = 1$ and $\mathbf{x}[n] = [x[n] ... x[n-p]]^T$. The prediction error energy $E(\mathbf{a})$ in the WLP method is

$$E(\mathbf{a}) = \sum_{n=1}^{N+p} (e_n(\mathbf{a}))^2 w_n = \mathbf{a}^T \big(\sum_{n=1}^{N+p} w_n \mathbf{x}[n]\mathbf{x}^T[n])\mathbf{a} = \mathbf{a}^T\mathbf{R}\mathbf{a}, \tag{2.8}$$

where w_n is the weight imposed on sample n, N is the length of the signal $x[n]$, and $\mathbf{R} = \sum_{n=1}^{N+p} w_n \mathbf{x}[n]\mathbf{x}^T[n]$. This problem is a constrained minimization problem,

$$\text{minimize } E(\mathbf{a}) \text{ subject to } \mathbf{a}^T\mathbf{u} = 1,$$

where $\mathbf{u}$ is the vector defined as $\mathbf{u} = [1\ 0\ ...\ 0]^T$. It can be seen that the autocorrelation matrix $\mathbf{R}$ is weighted, in difference to the conventional LP analysis.

Matrix $\mathbf{R}$ is symmetric but not Toeplitz, due to the weighting process. However, it is positive definite, and this makes the minimization problem convex. Using Lagrange multipliers, it can be shown that $\mathbf{a}$ satisfies the linear equation

$$\mathbf{R}\mathbf{a} = \sigma^2\mathbf{u}, \tag{2.9}$$

where $\sigma^2 = \mathbf{a}^T\mathbf{R}\mathbf{a}$ is the error energy. Finally, the WLP all-pole model is obtained as $H(z) = 1/A(z)$, where $A(z)$ is the Z-transform of vector $\mathbf{a}$.

Weighting function

The time domain weighting function w_n is the key point of both WLP and SWLP. In [41], the weighting function was chosen to be the Short-Time Energy (STE)

$$w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2, \tag{2.10}$$

where M is the length of the STE window. The use of STE window can be justified as following. The STE function emphasizes the speech samples of large amplitude. It is well-known that applying LP analysis on speech samples that belong to the glottal closed phase interval will generally result in a more robust spectral representation of the vocal tract. So, by emphasizing on these samples that occur during the glottal closed phase, it is likely to yield more robust acoustical cues for the formants. In Figure 2.1, the focus on the glottal closed phase of the STE weighting function is illustrated on a clean vowel.

Stability

However, the WLP method with the STE window does not ensure stability of the all-pole model. Therefore, in [42], a formula for a generalized weighting function to be used in WLP is developed in order to guarantee stability. The autocorrelation matrix $\mathbf{R}$ in Eq. X can be expressed as

$$\mathbf{R} = \mathbf{Y}^T\mathbf{Y}, \tag{2.11}$$

where

$$\mathbf{Y} = [\mathbf{y_0}\,\mathbf{y_1}\cdots\mathbf{y_p}] \in \Re^{(N+p)x(p+1)}$$

and

$$\mathbf{y}_0 = [\sqrt{w_1}x[1]\cdots\sqrt{w_N}x[N]\ \ 0\cdots 0]^T.$$

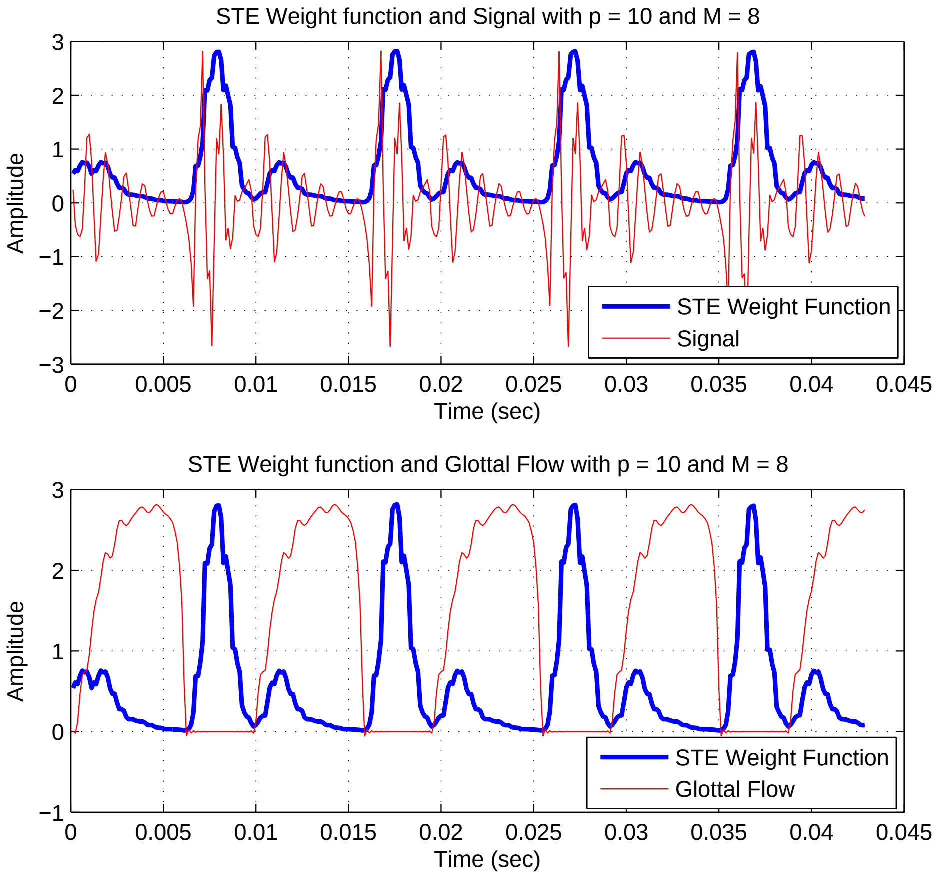

Figure 2.1 Upper panel: time domain waveforms of speech (vowel /a/ produced by male speaker) and short-time energy (STE) weight function (M=8).
Lower panel: Glottal flow waveform of the vowel /a/ together with the STE weight function (M=8).

The column vectors are given by

$$\mathbf{y}_{k+1} = \mathbf{B}\mathbf{y}_k, \; k = 0, 1, \cdots, p-1, \tag{2.12}$$

where

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \sqrt{w_2/w_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{w_3/w_2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{w_{N+p}/w_{N+p-1}} & 0 \end{bmatrix}$$

So, before forming the matrix $\mathbf{Y}$, the elements of the secondary diagonal of the matrix $\mathbf{B}$ are defined for all $i = 1, \cdots, N+p-1$ as

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \leq w_{i+1} \\ 1, & \text{if } w_i > w_{i+1} \end{cases}$$

Finally, the WLP method computed using matrix $\mathbf{B}$ is called the *Stabilized Weighted Linear Prediction* model, and the stability of the all-pole filter is ensured. For more information on the stability of SWLP, see [42].

2.2 Covariance based methods

2.2.1 Classic Linear Prediction - Covariance Method

If we assume that the speech signal is zero outside an interval $p \leq n \leq N-1$, thus p samples outside the region are available, then the mean squared prediction error is given by

$$E_N = \sum_{m=0}^{N-1} e^2[m] \tag{2.13}$$

where $e[m] = s[m] - \sum_{k=1}^{p} a_k s[m-k], \quad 0 \leq m \leq N-1$ and the interval $[0, N-1]$ is the prediction error interval.

Using a similar approach with the autocorrelation method in minimizing the prediction error, we result in the *covariance method* of linear prediction, which is given by the following equation:

$$\mathbf{\Phi}\vec{a} = \vec{\psi}, \tag{2.14}$$

where the elements of the *covariance matrix* $\mathbf{\Phi}$ are given by

$$\phi_{i,j} = \sum_{n=0}^{N-1} s[n-i]s[n-j], \tag{2.15}$$

where $1 \leq i \leq p$ and $1 \leq j \leq p$. The other two vectors are given by

$$\vec{a} = [a_1, a_2, a_3, ..., a_p]^T \quad , \vec{\psi} = [\phi_{0,1}, \phi_{0,2}, \phi_{0,3}, ..., \phi_{0,p}]^T. \tag{2.16}$$

Matrix Φ has the properties of a covariance matrix and thus the system can be efficiently solved using Cholesky Decomposition.

The main advantage of the covariance method is that it always results in the correct solution for any window length greater than p. Also, if the boundaries of the window are handled properly, a rectangular window can be used. Finally, the main disadvantage of this method is that stability of the filter is not always guaranteed. It should also be noted that in high pitched speakers, the closed phase interval is typically too short, and thus the estimation of the vocal tract filter is not accurate. This yields severe distortions on the estimated glottal flow, such as increased ripple in its closed phase interval.

As it was mentioned in Chapter 1, it is advantageous to restrict the analysis region in the closed phase interval of the glottal flow waveform. This approach is called *Closed Phase (CP) Covariance LP Analysis* and it is the covariance technique that is tested in this thesis.

2.2.2 Constrained Closed Phase Covariance Linear Prediction

Problems with conventional CP covariance analysis

The classic CP covariance LP analysis described in the previous section suffers from certain shortcomings. Several previous studies [38] [69] [72] [57] indicate that the CP analysis is very sensitive to the position of the covariance frame, thus giving glottal flow estimated that vary greatly. This is understandable if we consider that CP length is typically short, and the amount of data used to define the parametric model of the vocal tract is sparse. If the position of this frame is misaligned, then this results in poor modeling of the vocal tract resonances, and thus severe distortion of the glottal flow estimates. The problem is more apparent in high pitch speakers, where the length of the CP interval is very short. This type of distortion is demonstrated in Figure 2.2.

In this figure, three glottal flow estimates are shown on the left, which were inverse filtered from the same token of a male subject uttering the vowel [a], by using a minor change in the position of the covariance frame. This example shows how a minor change in the position of the covariance frame has resulted in a major change in the glottal flow estimates. On the right of this figure, the corresponding pole-root locations of the glottal flow estimates are shown. It is interesting to notice that in Figs.

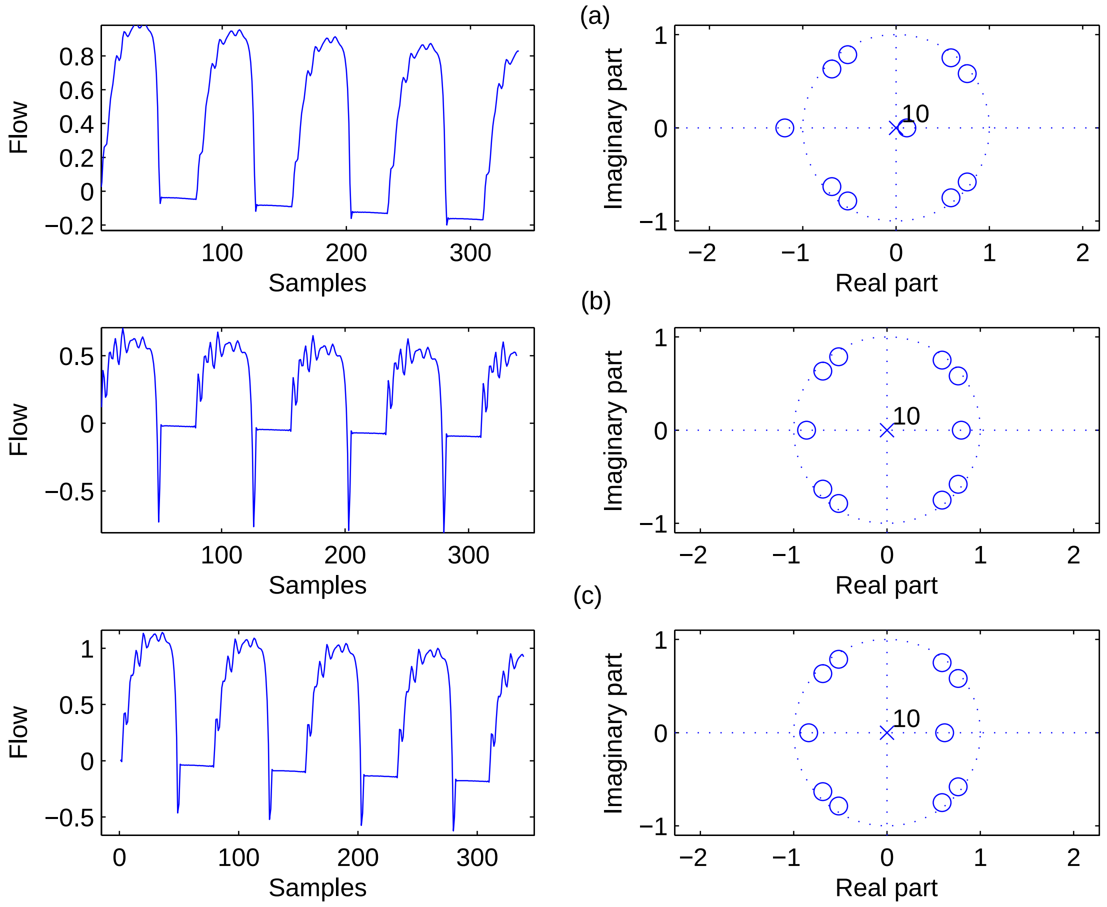

Figure 2.2 Covariance Frame Misalignment and Glottal Flow Distortion.

2.2(a), 2.2(b), and 2.2(c), the inverse filters have one root on the positive real axis in the z-domain. The effect of an inverse filter root which is located on the positive real axis has the properties of a first order differentiator, when the root approaches the unit circle, and a similar effect is also produced by a pair of complex conjugate roots at low frequencies. Thus, the glottal flow estimate in such cases becomes similar to a time-derivative of the flow candidate given by an inverse filter with no such roots or roots located in a more neutral position near the origin of the z-plane. This distortion is more apparent at the time instants where the glottal flow changes more rapidly, that is, near glottal closure. As it can be seen in Figs. 2.2(b), 2.2(c), this distortion is typically seen as sharp negative peaks, called ”jags” by Wong et al [71], at the time instants of glottal closure.

The theory of source-filter speech production suggests that poles of the vocal tract for non-nasalized voiced sounds occur at complex conjugate pairs and the low frequency emphasis of the spectrum results from the glottal source. However, as it can be seen in Figs. 2.2(b), 2.2(c), the estimated vocal tract model has roots on the positive real axis or at low frequencies, and thus the amplitude spectrum shows boosting of low frequencies, which comes in contrast with Fant’s suggested theory. Hence, it can be argued that among the three vocal tract models, the one depicted in Fig. 2.2(a) is the one that is more close to represent an amplitude spectrum of an all-pole vocal tract of a vowel. Also, the removal of the roots of the vocal tract model located on the real axis results in glottal flow estimates which are less dependent on the covariance frame location [71] [13].

Another source of distortion in conventional CP covariance analysis is the fact that the inverse filter might be non-minimum phase, that is, the filter has roots outside the unit circle in the z-domain. From a stability point of view, this is not a problem, since the IF is computed using a FIR filter, and thus non-minimum phase filters do not cause stability problems. However, the use of non-minimum phase filters does cause other kinds of distortion. According to the source-filter theory of speech production, the glottal flow is filtered through the vocal tract, which is considered to be a stable all-pole system for vowels and liquids. In the z-domain, this system must have all its poles inside the unit circle. Its inverse filter cancels the vocal tract contribution by mapping each pole of the vocal tract into a zero of the IF filter inside the unit circle, this resulting in a minimum phase filter. However, it is well-known in the theory of digital signal processing that a zero of a FIR filter can be *mirrored*, that is, it can be replaced by its mirror image partner. A zero at $z = z_0$ can be replaced by a zero at $z = 1/z_0^*$, without changing the shape of the amplitude spectrum of the filter. Hence, from an inverse filtering point of view, there are several inverse filters: one of them is minimum phase and the others are non-minimum phase, and all of them are considered equal. These filters, however, are different in terms of phase characteristics. This difference can cause severe distortion, and it is particularly strong in cases where zeros of the inverse filter located in the vicinity of the lowest two formants are moved from inside to outside the unit circle.

Figure 2.3 demonstrates this distortion caused by IF a vowel waveform with a minimum and a non-minimum phase FIR filter. In Fig. 2.3(a), the glottal flow estimated with a minimum phase filter is shown on the left, and the corresponding z-plane representation is shown on the right. In Fig. 2.3(b), the complex conjugate root pair that models the first formant is replaced by its counterpart outside the unit circle. Even though this slight modification only changed the root radius by 0.04, the distortion caused in the glottal flow estimate is severe and is displayed as increased ripple during the closed phase interval of the glottal cycle, as shown in the left panel of Fig. 2.3(b).

Mathematically Constrained Linear Prediction

The concept of mathematically constrained linear prediction is the modification of the conventional covariance analysis in order to reduce the distortion which is caused by unrealistic vocal tract model roots location. This is achieved by not allowing the

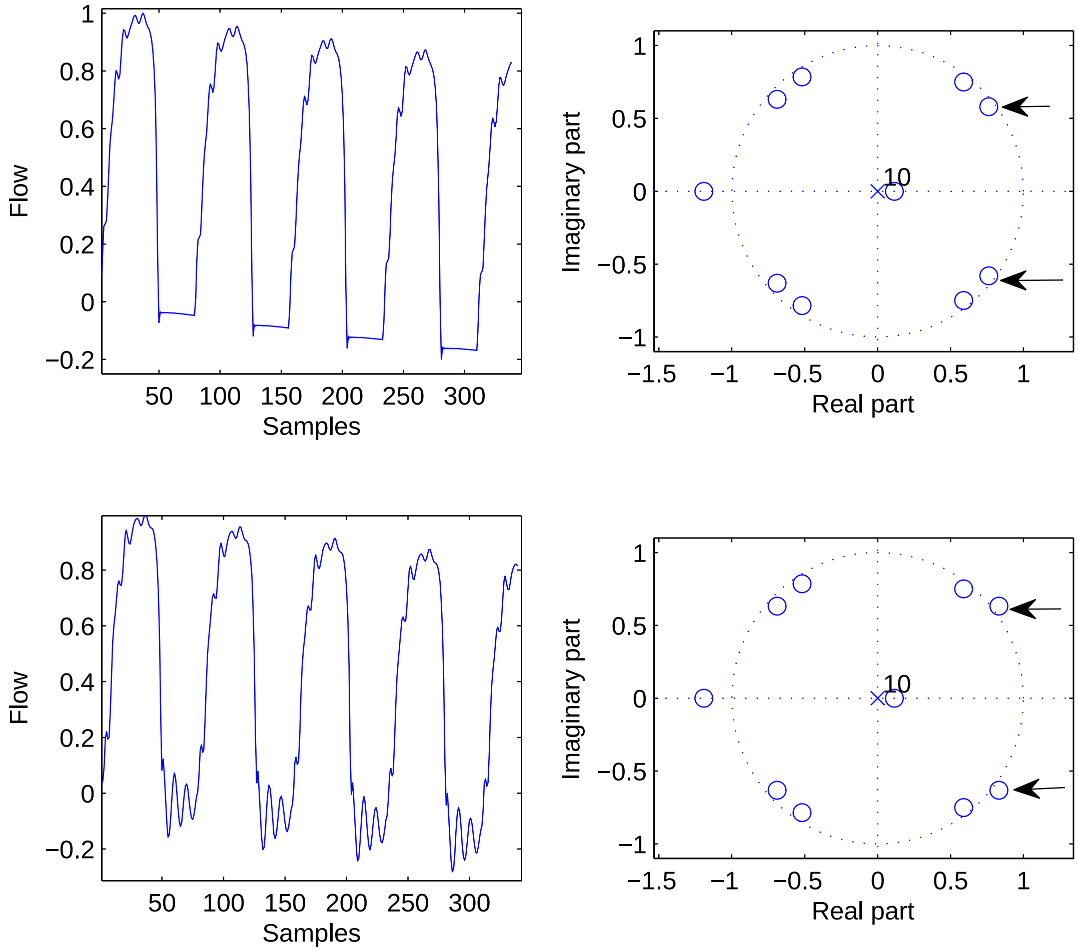

Figure 2.3 Distortion caused by non-minimum phase filter.

mean square error (MSE) criterion to locate the roots freely on the z-domain, thus providing certain restrictions in the predictor structure, which results in more realistic root locations, from the viewpoint of the acoustic theory described by Fant.

In order to implement such restrictions in the MSE error, someone has to find a method to express the constraint in a form of a concise mathematical equation and apply the constraint in the MSE minimization problem. One such constraint can be expressed with the help of the DC gain of the LP filter. The DC gain can be expressed as the sum of the predictor coefficients

$$V(e^{j0}) = \sum_{k=0}^{p} \alpha_k = l_{DC}. \tag{2.17}$$

where α_k are the filter coefficients of the constrained inverse filter and the l_{DC} is a pre-defined real value for the gain of the filter at DC. The reason for selecting the constraint on the DC gain is that, without it, it is possible that the amplitude response of the vocal tract model shows excessive boost at zero frequency. It is known from Fant's suggested source-filter theory, that the amplitude response of voiced sounds approaches unity at zero frequency [21]. Therefore, if a misplaced and short covariance frame occurs, it might even lead to an amplitude response with larger gain at zero frequency than at formants, which is a clear violation of the source-filter theory and its underlying acoustical theory of tube shapes. By imposing this constraint on the DC gain of the covariance linear prediction analysis, one might expect that the amplitude response of the resulting vocal tract estimates will better match Fant's source-filter theory.

However, it should be noted that this method still leaves the determination of the exact z-domain root locations of the vocal tract model to the MSE criterion, and does not bias the root locations prior the optimization. The formulation of the DC-constrained LP follows.

A mathematically straightforward way to implement such a restriction in the conventional LP, as it is described in a previous section, is to set a certain pre-defined value for the frequency response of the all-pole model at zero frequency, as it is written in Eq. 2.17. Using matrix notation, the DC-constrained minimization problem can now be formulated as follows:

$$\text{minimize } \mathbf{a}^T \mathbf{\Phi} \mathbf{a} \text{ subject to } \mathbf{\Gamma}^T \mathbf{a} = \mathbf{b},$$

where $\mathbf{a} = [a_0, \cdots, a_p]^T$ is the filter coefficient vector with $a_0 = 1$, $\mathbf{b} = [1,\ l_{DC}]^T$, and $\mathbf{\Gamma}$ is a $(p+1)$ by 2 constraint matrix defined as

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \tag{2.18}$$

The $\mathbf{\Gamma}$ matrix is positive definite. Thus, the minimization problem is convex. The Lagrange multiplier method is suitable for efficient solution of this problem. The objective function is defined as:

$$\eta(\alpha, \mathbf{g}) = \alpha^T \mathbf{\Phi} \alpha - 2\mathbf{g}^T (\mathbf{\Gamma}^T \alpha - \mathbf{b}), \tag{2.19}$$

where $\mathbf{g} = [g_1\, g_2]^T > 0$ is the Lagrange multiplier vector. The above equation can be minimized by setting the derivative with respect to vector α to zero. Thus, by taking into account that matrix $\mathbf{\Phi}$ is symmetric, we have:

$$\frac{\partial \eta}{\partial \alpha}(\alpha, \mathbf{g}) = \alpha^T (\mathbf{\Phi}^T + \mathbf{\Phi}) - 2\mathbf{g}^T \mathbf{\Gamma}^T = 2\alpha^T \mathbf{\Phi} - 2\mathbf{g}^T \mathbf{\Gamma}^T = 2(\mathbf{\Phi}\alpha - \mathbf{\Gamma}\mathbf{g}) = 0. \tag{2.20}$$

Vector $\mathbf{c}$ can be solved from the group of equations

$$\mathbf{\Phi}\alpha - \mathbf{\Gamma}\mathbf{g} = 0 \tag{2.21}$$

$$\mathbf{\Gamma}^T \alpha - \mathbf{b} = 0 \tag{2.22}$$

which finally gives that the optimal coefficients of the constrained inverse filter are:

$$\alpha = \mathbf{\Phi}^{-1} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{\Phi}^{-1} \mathbf{\Gamma})^{-1} \mathbf{b} \tag{2.23}$$

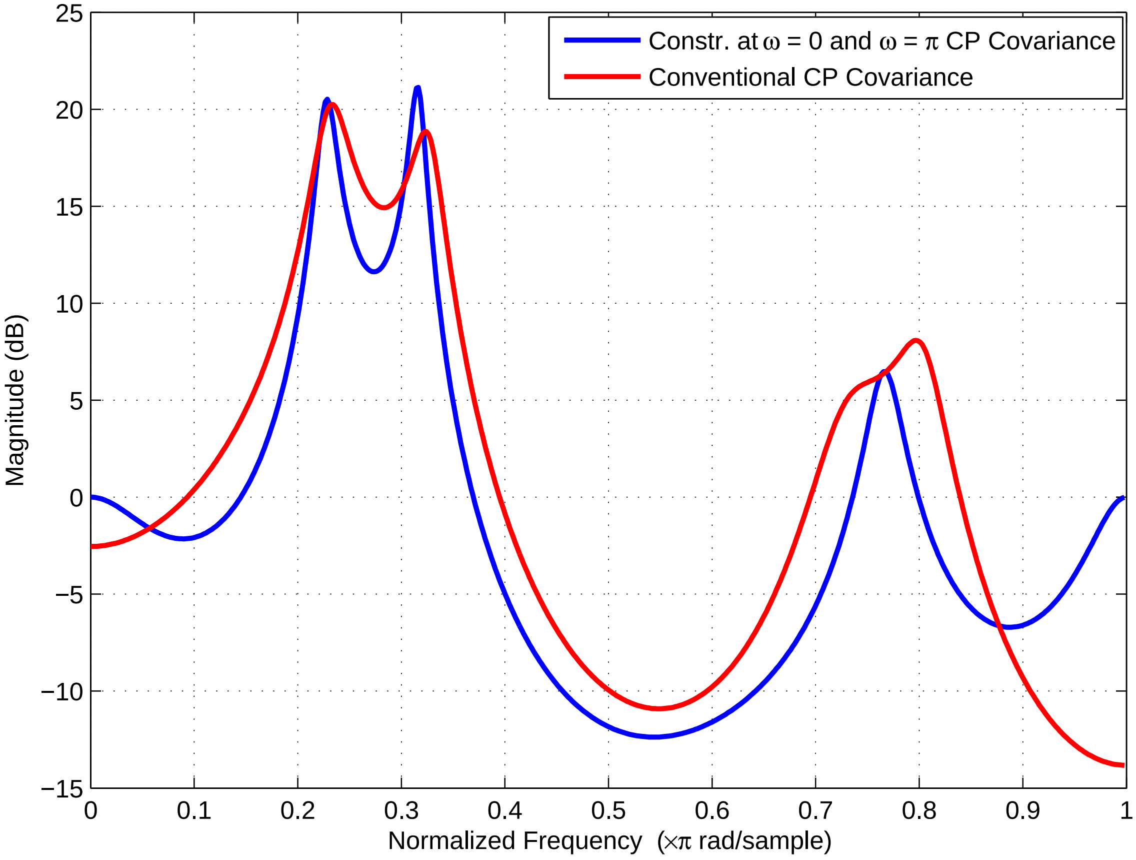

Figure 2.4 Examples of all-pole spectra computed in the closed phase covariance analysis by the conventional LP and by the DC-π constrained LP.

In a similar manner, a constrained at $\omega = \pi$ can be formed. By denoting the transfer function of a pth order π-constrained inverse filter $D(z)$, the following equation can be written:

$$D(z) = \sum_{k=0}^{p} d_k z^{-k} \Rightarrow D(e^{j\pi}) = \sum_{k=0}^{p} d_k (-1)^k = l_\pi. \tag{2.24}$$

The π-constrained minimization problem can now be formulated:

$$\text{minimize } \mathbf{d}^T \mathbf{\Phi} \mathbf{d} \text{ subject to } \mathbf{\Gamma}^T \mathbf{d} = \mathbf{e},$$

where $\mathbf{d} = [d_0 \cdots d_p]^T$ is the filter coefficient vector with $d_0 = 1$, $\mathbf{e} = [1 \; l_\pi]^T$, and Γ

the new $(p+1)$ by 2 constraint matrix is defined as:

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 0 & 1 \\ 0 & -1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \tag{2.25}$$

It is also possible to assign a third constraint by imposing simultaneously that the first inverse filter coefficient is equal to unity and that the filter gain at both $\omega = 0$ and $\omega = \pi$ are equal to l_{DC} and l_π respectively. Then, the constraint equation becomes $\mathbf{\Gamma}^T\mathbf{v} = \mathbf{h}$, where $\mathbf{v} = [v_0 \cdots v_p]^T$ is the filter coefficient vector with $v_0 = 1$, $\mathbf{h} = [1\, l_{DC}\, l_\pi]$ and the resulting $(p+1)$ by 3 constraint matrix is defined as:

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \\ \vdots & \vdots & \\ 0 & 1 & 1 \end{bmatrix} \tag{2.26}$$

An example of the vocal tract filter estimations during a closed phase region for the Constrained Covariance method and the conventional one is depicted in Figure 2.4.

2.3 Inverse Filtering Procedure

The inverse filtering procedure that is followed in this thesis is illustrated in Figure 2.5.

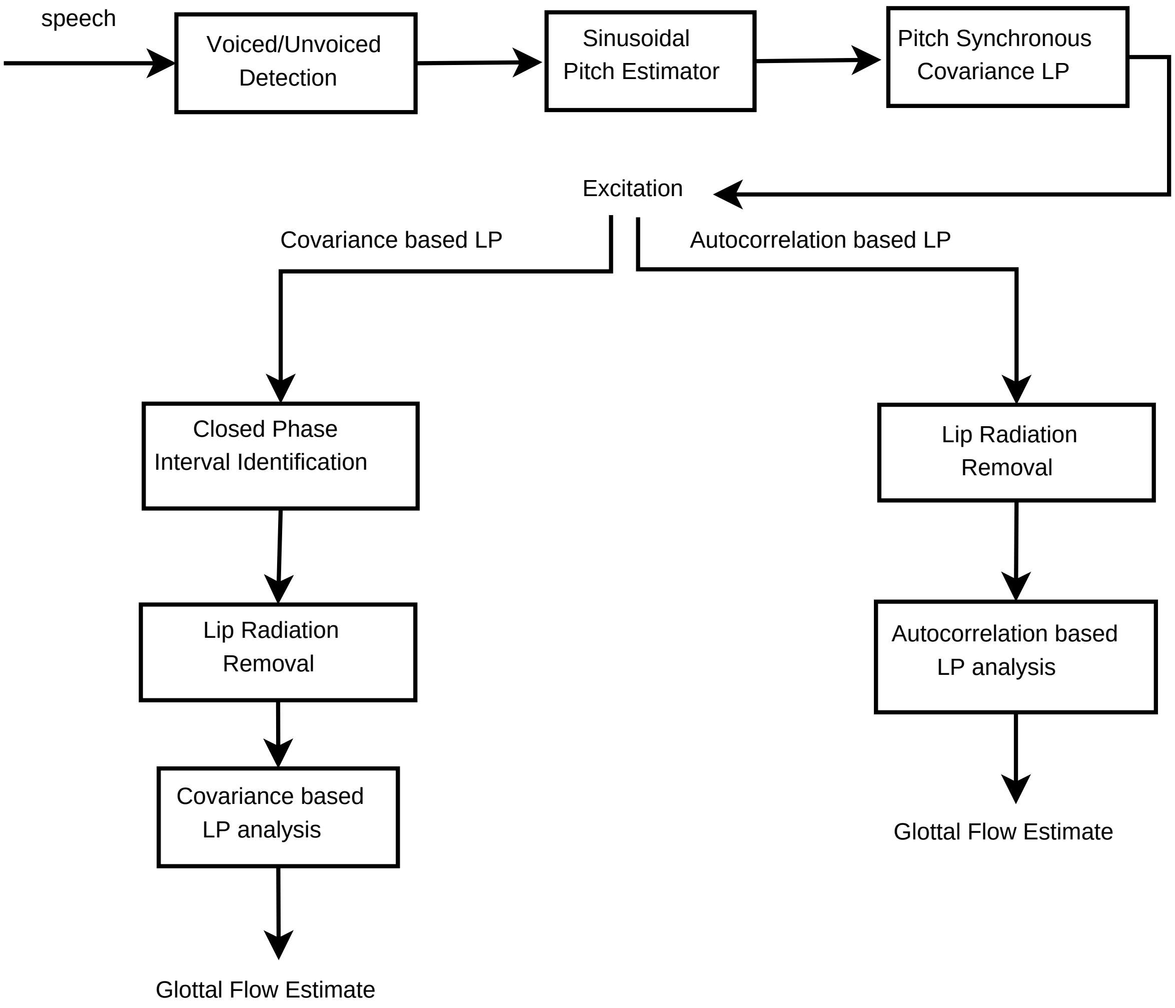
speech
Voiced/Unvoiced Detection
Sinusoidal Pitch Estimator
Pitch Synchronous Covariance LP
Excitation
Covariance based LP
Autocorrelation based LP
Closed Phase Interval Identification
Lip Radiation Removal
Covariance based LP analysis
Glottal Flow Estimate
Lip Radiation Removal
Autocorrelation based LP analysis
Glottal Flow Estimate

Figure 2.5 Inverse Filtering Procedure.

Before processing, the speech and airflow signals are downsampled from 44 kHz to 8 kHz. Care is taken in order to avoid aliasing before downsampling. This is achieved by using a FIR filter with a cutoff frequency at 4 kHz.

An analytic description of each sub-system follows next.

2.3.1 VUS detector

As first, the speech waveforms generated by physical modeling of the vocal tract are passed through a voiced/unvoiced/silence (VUS) detection algorithm in order to cut any silent or unvoiced regions. Any VUS detection algorithm can be used at this point, but one based on energy and zero crossings is preferred here due to its simplicity, low complexity, and speed. The analysis is performed small segments of speech, with 30 ms of duration and a 15 ms of overlap between successive segments.

2.3.2 Pitch Estimation

Afterwards, an estimate of the pitch of the voiced parts of the waveform is obtained. Although any pitch estimator can be used in this part, the sinusoidal pitch estimator [44] is used here. Pitch estimates are calculated on a speech frame of 30 ms with a 15 ms overlap.

2.3.3 Excitation Instants Detection

According to the pitch estimates generated by the previous algorithm, a pitch-synchronous covariance-based LP analysis on the waveform is performed, with an order of $p = 10$, for speech signals with sampling frequency $F_s = 8$ kHz. The purpose of this analysis is not an accurate estimate of the vocal tract or the glottal flow in each frame, but a rough approximation of the excitation and the glottal excitation instants, which indicate glottal closure. The reason for this analysis is the identification of pitch periods throughout the whole speech signal and it is an initial step for CP covariance analysis, as it will be shown later.

After this point, the analysis differs depending on the LP method that is used.

2.3.4 Covariance-based LP Analysis

For CP covariance-based methods, it is necessary to identify the closed phase interval. This is achieved by using the sliding covariance analysis suggested by Plumpe et al [53], which provides a robust method to extract the glottal closed phase interval out of the speech signal. Specifically, this method of glottal closed phase estimation relies on a sliding covariance analysis and uses vocal tract formant modulation which is predicted by Ananthapadmanabha and Fant [2] to vary more slowly in the glottal closed phase than in its open phase and to respond quickly to a change in glottal area. A ”stationary” region of formant modulation gives a closed-phase time interval, over which we estimate the vocal tract transfer function. A stationary region is present even when the vocal folds remain partly open [53]. For high pitch speakers, where the closed phase samples are less than twice the order of the LP analysis, a fixed length of $N_{CP} + order$ closed phase samples was used.

Then, before inverse filtering, the lip radiation is cancelled using a first order all-pole filter with its pole at $z = 0.999$. Having the closed phase intervals for each period, a covariance-based LP analysis with order $p = 10$ is set up on the corresponding speech samples. Finally, the vocal tract estimate that is obtained inverse filters pitch synchronously a speech segment consisting of two pitch periods, and thus the glottal flow is obtained.

2.3.5 Autocorrelation-based LP Analysis

For autocorrelation based methods, the analysis is simpler. The lip radiation is cancelled using a first order all-pole filter with its pole at $z = 0.999$. Next, using the glottal pulses identified in a previous step, a pitch-synchronous LP analysis with order $p = 10$ is performed over a speech segment of 250 ms. The vocal tract contribution is then cancelled by inverse filtering over a region of two pitch periods, with an overlap of one pitch period.

The overall glottal flow is synthesized using the well-known Overlap-Add method (OLA) [55].

2.4 Summary

In this chapter, four LP-based spectral estimation techniques were discussed and applied to an inverse filtering process; two of them are well-kwown and widely used: the conventional autocorrelation and covariance LP. The other two were recently developed and they are referred to as Stabilised Weighted Linear Prediction and Constrained Closed Phase Covariance LP. SWLP is suggested for IF since it has the interesting property of emphasizing the speech samples that typically occur during the closed phase region. Finally, an automatic system for inverse filtering was described.

Chapter 3

Results

In this chapter, we firstly present glottal flow estimates of several vowels of different frequencies, using the IF techniques described in Chapter 2. Then, parametrization measures in time and frequency domain are applied on the estimates, in order to compare and evaluate the performance of different IF techniques. Finally, comments are presented on both the shape of the glottal flow estimates and the parametrization measures.

3.1 Glottal Flow Estimates Examples

The database consists of four physically modeled speech waveforms, all vowels (/aa/, /ae/, /eh/, /ih/), each one in eight different frequencies $(105, 115, 130, 145, 205, 210, 230,$ and 255 Hz). That is a total of 32 speech signals. For brevity, we present here characteristic examples for each of the five vowels. In each figure (from left to right), we show first the actual glottal flow waveform, which is encapsulated by a thick-line box. Then, it follows the estimates using the autocorrelation-based LPC IF technique, the Stabilized Weighted LPC IF technique with parameter $M = 8$, the Stabilized Weighted LPC IF technique with parameter $M = 24$, the covariance-based LPC IF technique, and finally the constrained covariance-based LPC IF technique.

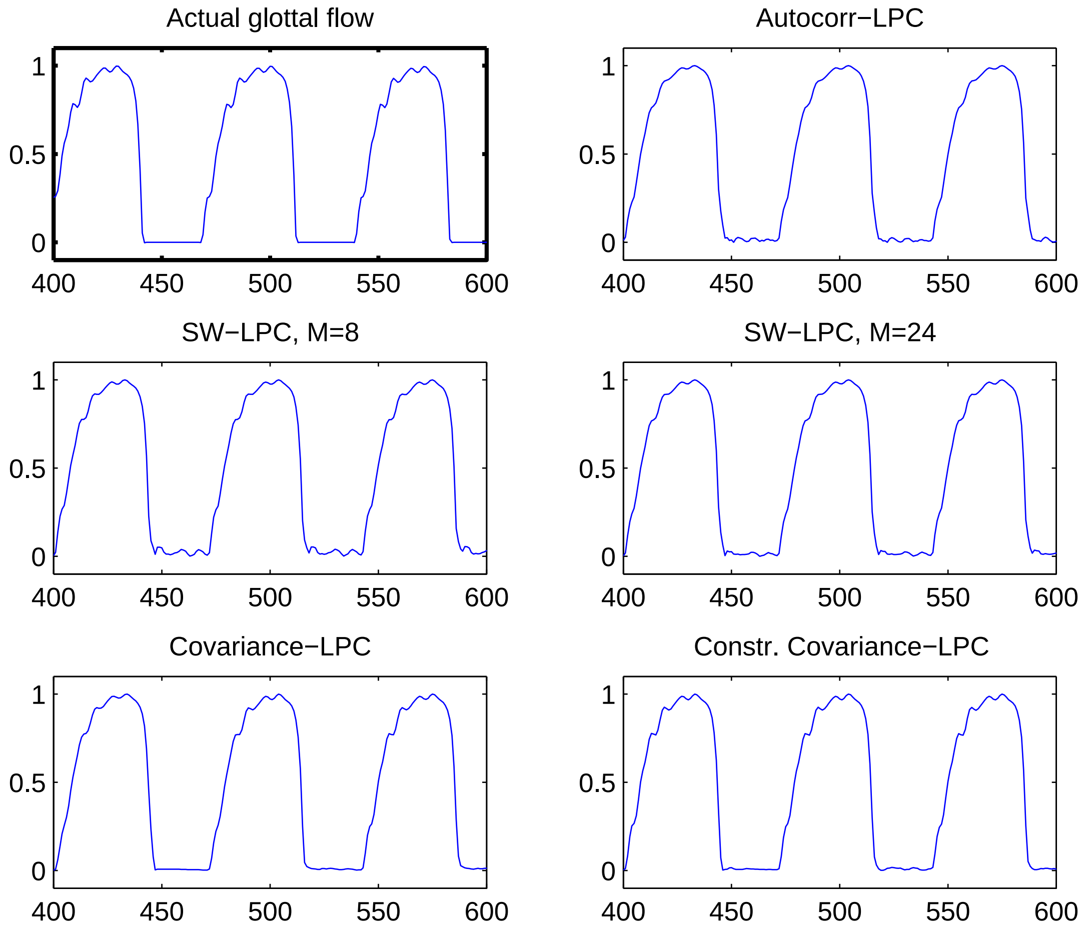

Figure 3.1 Glottal flow estimates for vowel /aa/, fundamental frequency of 105 Hz.

As it can be seen in the figures, the estimated glottal flow waveforms are, in general, close to the original one. However, the best fit is achieved by the CP covariance methods, because of a more accurate extraction of the vocal tract filter during the closed phase interval. Also, the $SWLP_{24}$ waveform, which corresponds to the SWLP

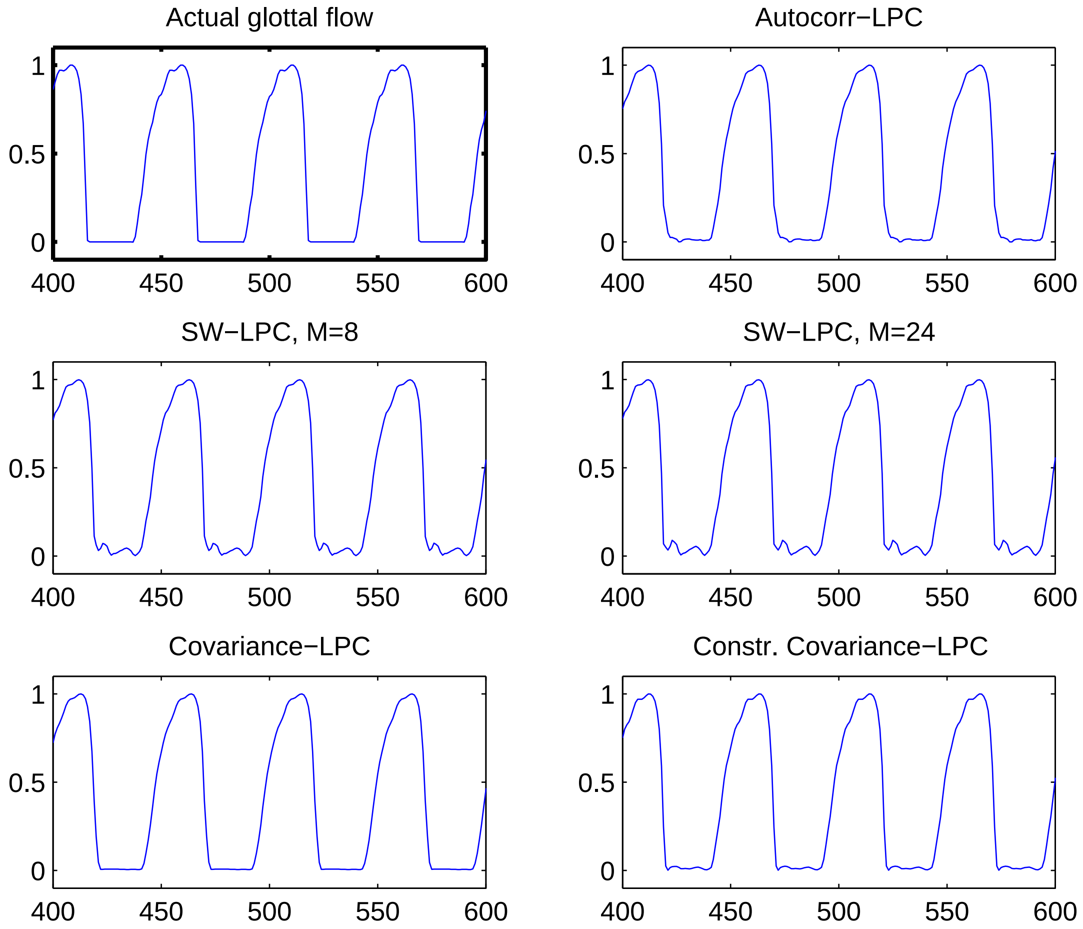

Figure 3.2 Glottal flow estimates for vowel /aa/, fundamental frequency of 145 Hz.

with $M = 24$, is closer to the original one, compared to the autocorrelation method of LP.

3.2 Parametrization Results

In order to parametrize our results, the previously discussed time and frequency domain measures were applied on the glottal flow estimates. In this chapter, the IF methods are abbreviated as: $WLPC_8$ stands for *Stabilized Weighted Linear Predic-*

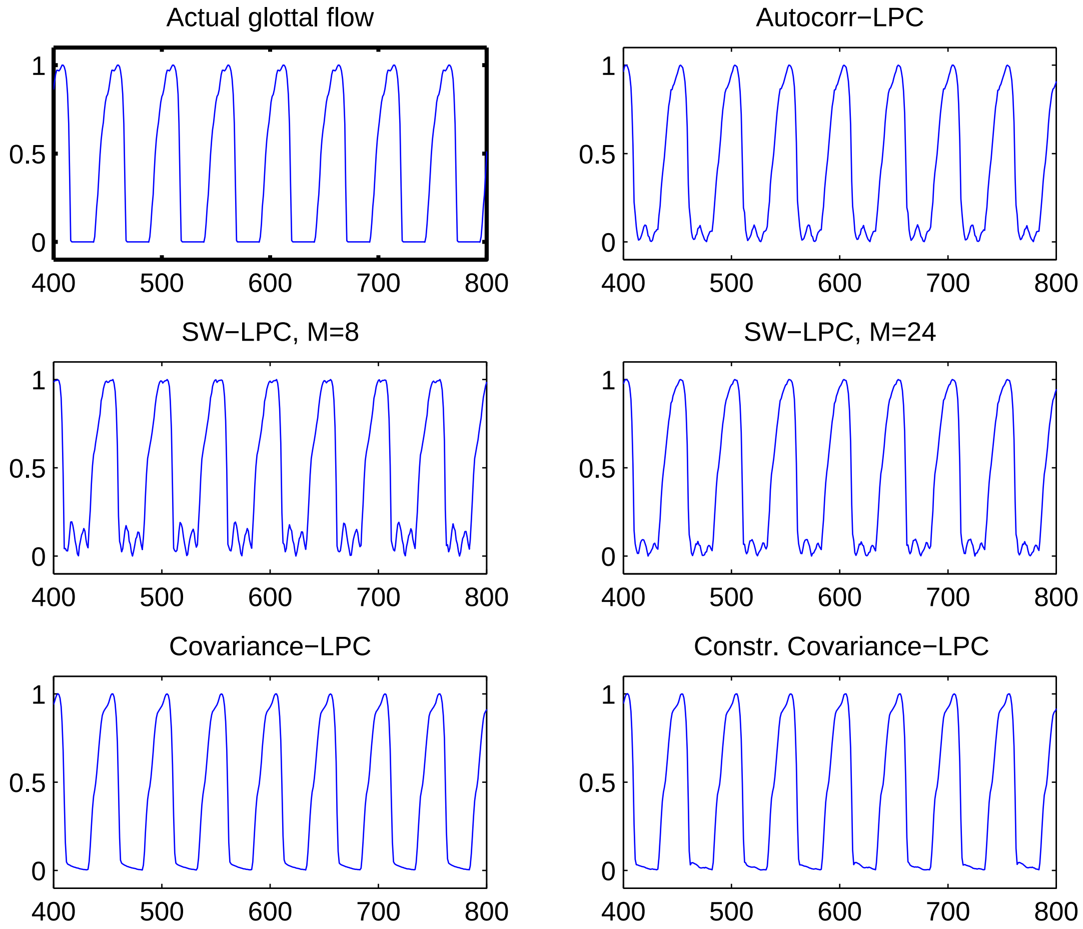

Figure 3.3 Glottal flow estimates for vowel /ae/, fundamental frequency of 145 Hz.

tion with STE window of length $M = 8$, $WLPC_{24}$ stands for *Stabilized Weighted Linear Prediction* with STE window of length $M = 24$, *LPC* stands for the conventional *autocorrelation Linear Prediction*, *CovLPC* stands for the conventional *CP covariance Linear Prediction* and *CLPC* stands for the *Constrained CP covariance Linear Prediction*.

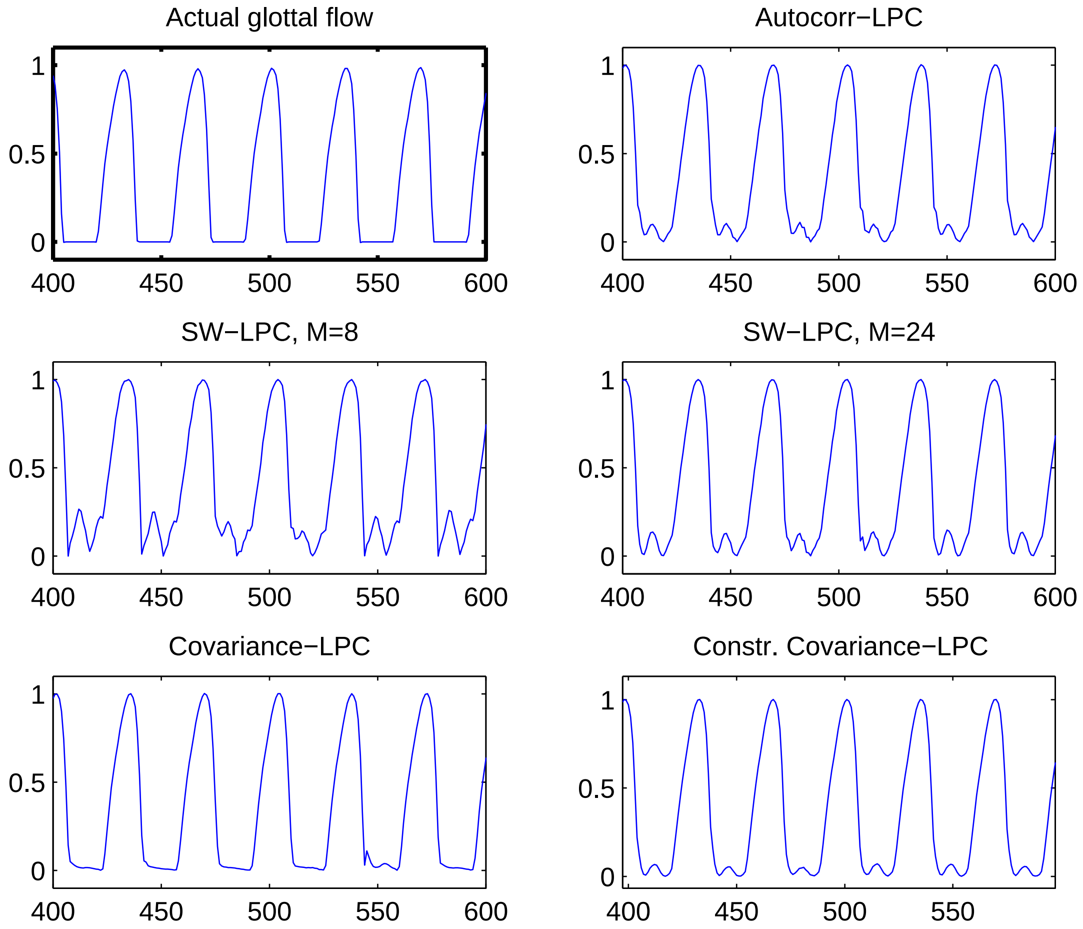

Figure 3.4 Glottal flow estimates for vowel /ae/, fundamental frequency of 210 Hz.

3.2.1 Signal to Reconstruction Error Ratio - SRER

A typical measure in signal processing algorithms is the Signal to Reconstruction Error Ratio (SRER). SRER is defined as:

$$SRER = 10log_{10}\left(\frac{\sigma_{s[n]}}{\sigma_{e[n]}}\right) \tag{3.1}$$

where $s[n]$ is the obtained glottal flow in our case, $e[n] = s[n] - \hat{s}[n]$ is the error between the original and the obtained glottal flow, and σ denotes the corresponding

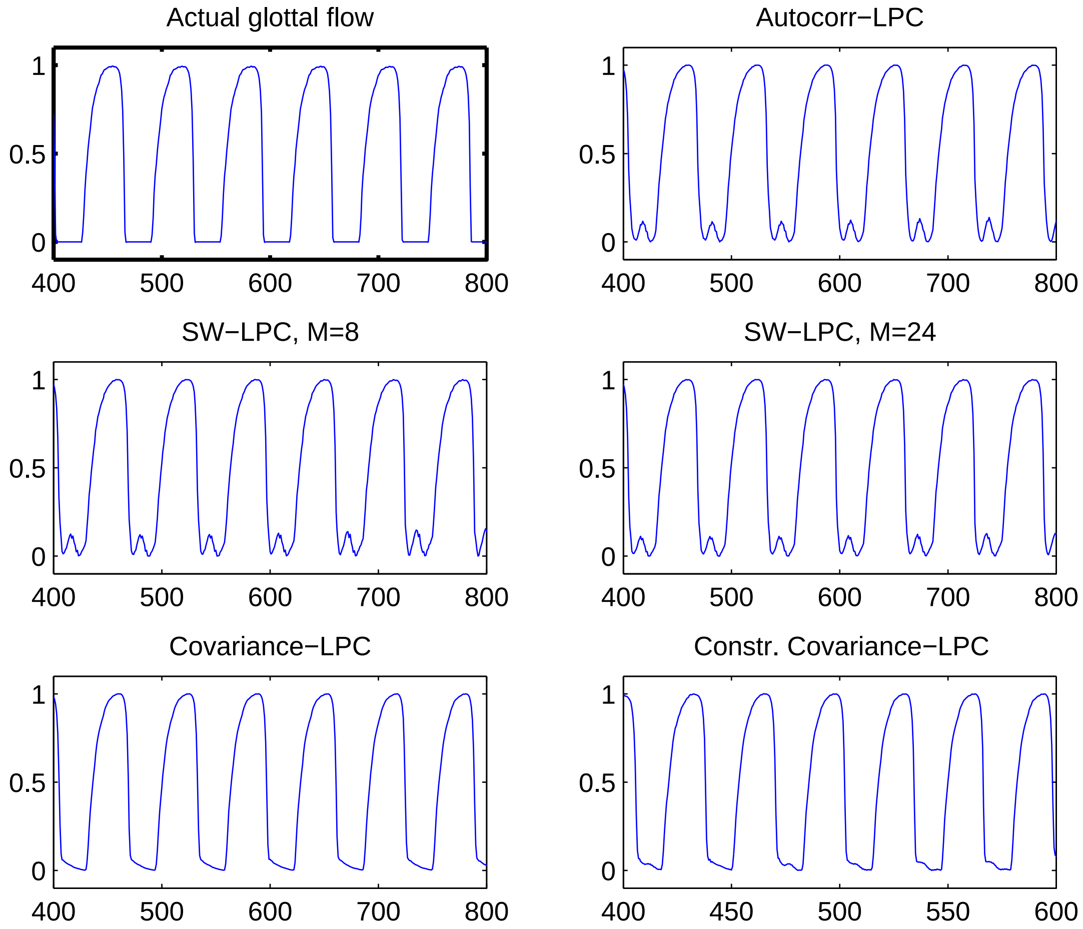

Figure 3.5 Glottal flow estimates for vowel /ih/, fundamental frequency of 115 Hz.

standard deviation. SRER is a means to estimate how well the estimated glottal flow ”fits” on the original one. In other words, it is an index of the quality of the pitch synchronously resynthesized glottal flow waveform.

Table 3.1 shows the mean and the standard deviation of SRER for each IF method per vowel.

In table 3.1, it is clearly seen that the CP covariance methods outperform the autocorrelation methods in terms of the SRER. However, CP covariance methods shows a small decrease in performance for the /ih/ vowel. This can be justified by the fact

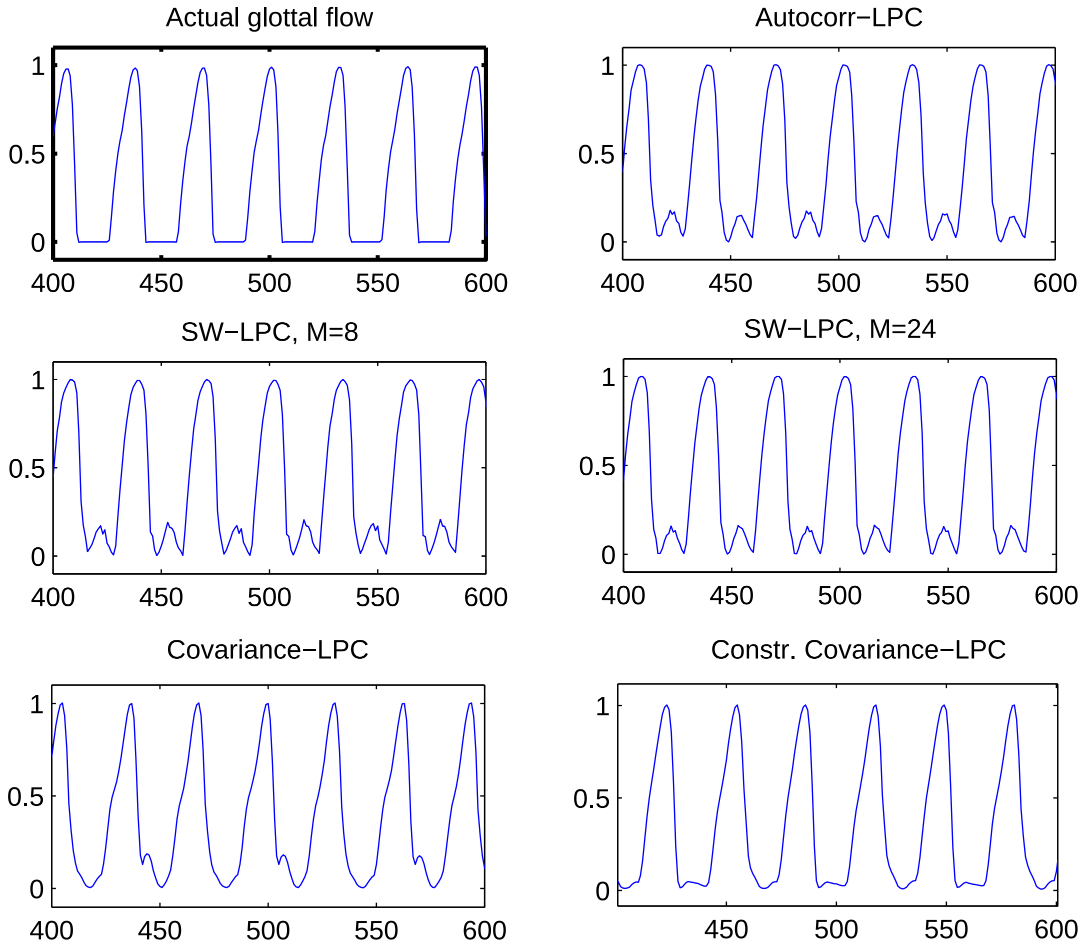

Figure 3.6 Glottal flow estimates for vowel /ih/, fundamental frequency of 230 Hz.

that the first formant for the /ih/ vowel is typically low, and LP is known to perform poor in such cases. This, along with a possible misalignment of the CP interval in some frames, and the short length of the CP interval for high pitch speakers, causes severe distortion in certain parts of the synthesized glottal flow, and thus leads to the decreased SRER value. A clearer picture could be illustrated using a segmental SRER, while discarding frames where the analysis had poor estimation results.

In table 3.2, the SRER mean and standard deviation for each IF method per frequency is illustrated. Here, it is evident that the performance of all IF methods are

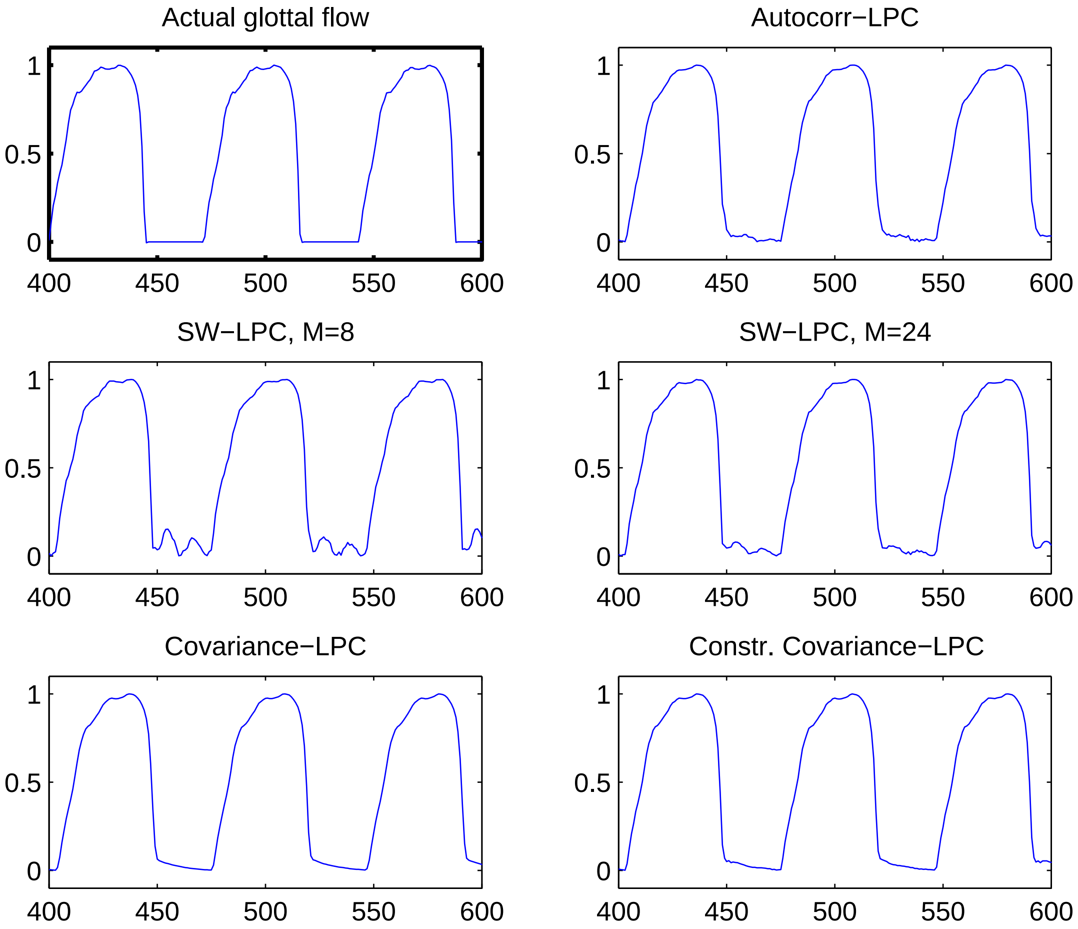

Figure 3.7 Glottal flow estimates for vowel /eh/, fundamental frequency of 105 Hz.

worse when the pitch increases, and especially in the CP covariance methods, where the covariance frame becomes too short. In almost all frequencies, $SWLP_{24}$ SRERs are higher than conventional autocorrelation LP.

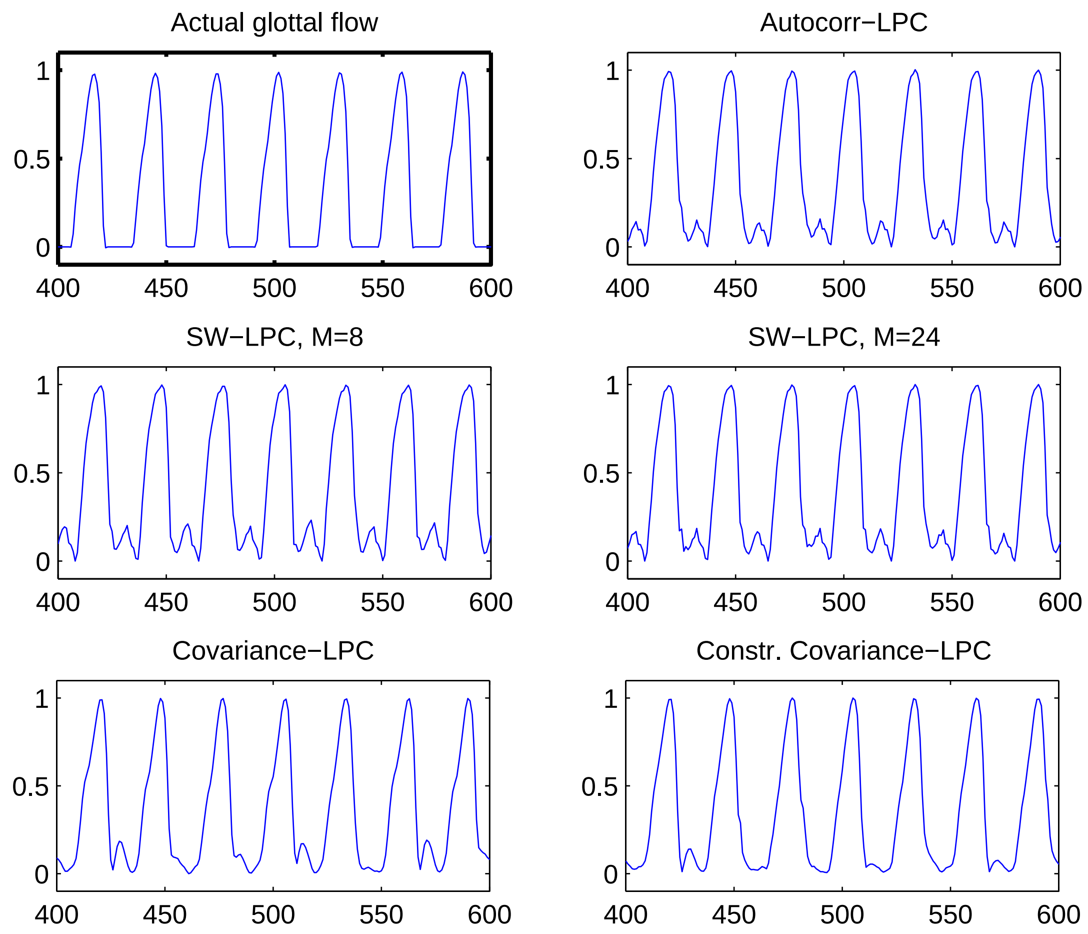

Figure 3.8 Glottal flow estimates for vowel /eh/, fundamental frequency of 255 Hz.

3.2.2 Normalized Amplitude Quotient - NAQ

For the synthesized glottal flows, the NAQ was estimated for each cycle. In order to compare the NAQ values of the original and the estimated glottal flows, the *ratio* of the NAQ of the actual glottal flow to the estimated one is formed, called NAQ_{rat}. The ratio should be equal to unity when the estimation of the glottal flow has succeeded perfectly.

Table 3.3 shows the mean and the standard deviation of NAQ_{rat} for each IF method

Vowel	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
/aa/	33.5(±2.0)	39.7(±4.5)	36.2(±5.7)	34.9(±9.2)	42.1(±13.1)
/ae/	32.6(±2.8)	38.3(±3.3)	36.9(±5.2)	37.5(±8.4)	46.1(±6.3)
/eh/	34.0(±1.9)	38.4(±4.2)	34.0(±4.0)	34.5(±8.2)	43.4(±5.6)
/ih/	37.3(±1.6)	37.6(±3.1)	33.3(±4.6)	32.2(±9.6)	38.8(±8.0)

Table 3.1 In this table, the mean and the standard deviation of the SRER value for each vowel (all 8 frequencies) is illustrated. All values are in dB.

Frequency	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
$F_0 = 105$ Hz	37.1(±0.8)	42.8(±2.1)	40.7(±3.7)	41.1(±2.3)	51.5(±4.4)
$F_0 = 115$ Hz	35.4(±1.9)	40.1(±1.8)	37.7(±2.8)	41.5(±2.3)	50.0(±2.4)
$F_0 = 130$ Hz	34.1(±1.0)	39.2(±4.0)	38.7(±3.4)	42.4(±2.1)	48.8(±3.8)
$F_0 = 145$ Hz	34.0(±2.1)	37.1(±0.7)	37.5(±3.0)	40.7(±1.8)	49.3(±6.2)
$F_0 = 205$ Hz	33.5(±3.8)	38.4(±1.8)	32.7(±2.5)	31.4(±7.3)	36.4(±2.6)
$F_0 = 210$ Hz	34.1(±3.6)	37.2(±3.2)	32.4(±4.0)	28.2(±5.3)	36.3(±6.7)
$F_0 = 230$ Hz	33.2(±3.5)	38.8(±3.9)	30.9(±4.0)	32.3(±9.4)	33.5(±7.3)
$F_0 = 255$ Hz	33.3(±3.0)	34.2(±5.7)	30.0(±4.2)	20.7(±2.8)	35.0(±5.0)

Table 3.2 In this table, the mean and the standard deviation of the SRER value for each frequency (all 4 vowels) is illustrated. All values are in dB.

per vowel.

In table 3.3, it is evident that SWLP outperforms both the conventional autocorrelation LP and the CP covariance LP. However, an interesting point in this table is the fact that the CP covariance waveforms, especially the conventional one, have NAQ_{rat} values that are higher than the others. This comes in contrast with the SRER index, which clearly showed that, in general, the CP covariance glottal flows are closer to the original glottal flow than the other estimated flows. Since NAQ is a measure that depends only on the amplitude values of the waveform and its derivative, even a

Vowel	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
/aa/	1.06(±0.12)	1.10(±0.12)	1.23(±0.13)	1.27(±0.18)	1.21(±0.12)
/ae/	0.99(±0.11)	1.09(±0.11)	1.25(±0.13)	1.23(±0.10)	1.18(±0.08)
/eh/	1.05(±0.12)	1.13(±0.12)	1.31(±0.14)	1.25(±0.13)	1.18(±0.11)
/ih/	1.10(±0.12)	1.14(±0.12)	1.32(±0.14)	1.24(±0.16)	1.17(±0.11)

Table 3.3 In this table, the mean and the standard deviation of the ratio, denoted as NAQ_{rat}, between the NAQ of the actual glottal flow to the NAQ of the estimated one, for each vowel (all 8 frequencies) is illustrated.

small distortion due to an incomplete estimation of the vocal tract filter in the closed phase might result in sever amplitude changes on the glottal flow. That is the reason why NAQ is usually not preffered in the literature when the CP covariance method is examined [8].

In table 3.4, the NAQ_{rat} values per frequency are illustrated. The results show no significant change depending on the frequency.

Frequency	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
$F_0 = 105$ Hz	1.03(±0.12)	1.12(±0.13)	1.28(±0.14)	1.33(±0.11)	1.20(±0.10)
$F_0 = 115$ Hz	1.07(±0.13)	1.14(±0.14)	1.36(±0.16)	1.33(±0.13)	1.22(±0.12)
$F_0 = 130$ Hz	1.03(±0.11)	1.10(±0.12)	1.27(±0.14)	1.26(±0.12)	1.19(±0.11)
$F_0 = 145$ Hz	1.01(±0.10)	1.05(±0.10)	1.28(±0.12)	1.28(±0.10)	1.15(±0.10)
$F_0 = 205$ Hz	1.03(±0.11)	1.11(±0.10)	1.23(±0.12)	1.20(±0.22)	1.17(±0.09)
$F_0 = 210$ Hz	1.06(±0.13)	1.15(±0.12)	1.23(±0.14)	1.20(±0.22)	1.14(±0.16)
$F_0 = 230$ Hz	1.10(±0.11)	1.13(±0.12)	1.30(±0.14)	1.18(±0.13)	1.16(±0.10)
$F_0 = 255$ Hz	1.09(±0.11)	1.10(±0.11)	1.28(±0.13)	1.26(±0.25)	1.24(±0.07)

Table 3.4 In this table, the mean and the standard deviation of the ratio, denoted as NAQ_{rat}, between the NAQ of the actual glottal flow to the NAQ of the estimated one, for each frequency (all 4 vowels), is illustrated.

3.2.3 H1-H2

For the synthesized glottal flows, the H1-H2 parameter was estimated for a speech frame of five periods. In order to compare the H1-H2 values of the original and the estimated glottal flows, the *difference* of the H1-H2 of the actual glottal flow to the estimated one is formed, called $H1 - H2_{dif}$. The difference should be equal to zero when the estimation of the glottal flow has succeeded perfectly. H1-H2 is an index for the spectral tilt of the glottal spectrum.

Table 3.5 shows the mean and the standard deviation of $H1 - H2_{dif}$ for each IF method per vowel.

Vowel	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
/aa/	0.61(±0.03)	0.20(±0.03)	0.54(±0.03)	0.20(±0.15)	0.08(±0.08)
/ae/	0.19(±0.03)	0.18(±0.02)	0.38(±0.01)	0.18(±0.12)	0.07(±0.04)
/eh/	0.23(±0.03)	0.32(±0.02)	0.44(±0.01)	0.38(±0.15)	0.15(±0.05)
/ih/	-0.16(±0.01)	-0.029(±0.01)	0.26(±0.01)	1.10(±0.34)	0.27(±0.15)

Table 3.5 In this table, the mean and the standard deviation of the difference in dB, denoted as $H1 - H2_{dif}$, between the H1-H2 of the actual glottal flow and the H1-H2 of the estimated one, for each vowel (all 8 frequencies) is illustrated.

The data on Table 3.5 shows that the performance of CP covariance methods outperforms the autocorrelation methods, except for the /ih/ vowel, where the problem of estimating the formants and the length of the CP length affects the resulting waveforms. This is consistent with the theory of CP analysis. However, it is not clear whether SWLP with $M = 8$ or $M = 24$ performs better with this criterion. In any case, they both outperform conventional autocorrelation LP.

In table 3.6, the $H1 - H2_{dif}$ values per frequency is illustrated. Here, the $H1 -$

$H2_{dif}$ metric performs worse when the frequency increases, which is expected.

Vowel	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
$F_0 = 105$ Hz	0.13(±0.03)	0.18(±0.01)	0.20(±0.01)	0.20(±0.04)	0.16(±0.03)
$F_0 = 115$ Hz	-0.11(±0.04)	-0.02(±0.02)	-0.03(±0.02)	0.14(±0.04)	0.07(±0.04)
$F_0 = 130$ Hz	0.05(±0.09)	0.09(±0.01)	0.15(±0.01)	0.19(±0.03)	0.11(±0.02)
$F_0 = 145$ Hz	-0.12(±0.01)	-0.05(±0.01)	-0.07(±0.01)	0.15(±0.03)	0.14(±0.04)
$F_0 = 205$ Hz	0.35(±0.01)	0.25(±0.01)	0.53(±0.01)	0.35(±0.19)	0.11(±0.08)
$F_0 = 210$ Hz	0.54(±0.05)	0.36(±0.03)	0.62(±0.04)	0.56(±0.31)	0.18(±0.14)
$F_0 = 230$ Hz	0.33(±0.03)	0.23(±0.02)	0.52(±0.01)	0.60(±0.31)	0.24(±0.21)
$F_0 = 255$ Hz	0.38(±0.02)	0.28(±0.02)	1.32(±0.01)	2.81(±0.56)	0.35(±0.06)

Table 3.6 In this table, the mean and the standard deviation of the difference in dB, denoted as $H1 - H2_{dif}$, between the H1-H2 of the actual glottal flow and the H1-H2 of the estimated one, for each frequency (all 4 vowels) is illustrated.

3.2.4 Harmonic Richness Factor - HRF

For the synthesized glottal flows, the HRF parameter was estimated for a frame of five glottal periods. HRF is defined as the ratio, in dB, of the sum of the harmonic amplitudes above the fundamental to the amplitude of the fundamental:

$$HRF_N = \frac{\sum_{n\geq 2}^{N} H_n}{H_1} \tag{3.2}$$

where H_1 is the amplitude of the fundamental and H_n are the amplitudes of the higher harmonics. In order to compare the HRF values of the original and the estimated glottal flows, the *difference* of the HRF of the actual glottal flow to the estimated one is formed, called HRF_{dif}. The difference should be equal to zero when the estimation of the glottal flow has succeeded. The HRF criterion is an extension of the H1-H2, illustrating the relationship of the fundamental with the higher harmonics. For the calculation of the HRF, the first eight ($N = 8$) harmonics were included.

Vowel	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
/aa/	-0.23(±0.02)	-0.16(±0.02)	-0.45(±0.04)	-0.14(±0.16)	-0.19(±0.10)
/ae/	0.12(±0.01)	-0.19(±0.01)	-0.44(±0.01)	-0.04(±0.08)	-0.18(±0.05)
/eh/	0.01(±0.02)	-0.24(±0.02)	-0.46(±0.02)	-0.02(±0.15)	-0.08(±0.05)
/ih/	-0.23(±0.01)	-0.29(±0.01)	-0.53(±0.02)	0.09(±0.15)	-0.03(±0.09)

Table 3.7 In this table, the mean and the standard deviation of the difference in dB, denoted as HRF_{dif}, between the HRF of the actual glottal flow and the HRF of the estimated one, for each vowel (all 8 frequencies) is illustrated.

Table 3.7 shows the mean and the standard deviation of HRF_{dif} for each IF method per vowel. It is evident that CP covariance techniques outperforms the autocorrelation methods, thus proving that the vocal tract estimation in higher formant regions is more accurate when an accurate CP analysis is applied on the speech signal. In addition, the SWLP method provides better results than the conventional autocorrelation LP.

In table 3.8, the HRF_{dif} values per frequency is illustrated. It is interesting to note that, except for the constrained CP covariance method, all other methods perform worse when the frequency increases.

3.3 Summary

In this chapter, the resulting waveforms of the IF techniques were parametrized using well known measures in time and frequency domain. The prevalence of the covariance methods is depicted, although the required analysis for this method has certain issues,

Frequency	$WLPC_8$	$WLPC_{24}$	LPC	CovLPC	CLPC
$F_0 = 105$ Hz	-0.024(±0.003)	-0.119(±0.003)	-0.180(±0.003)	-0.112(±0.03)	-0.098(±0.03)
$F_0 = 115$ Hz	-0.033(±0.006)	-0.096(±0.006)	-0.195(±0.008)	-0.129(±0.025)	-0.093(±0.03)
$F_0 = 130$ Hz	0.041(±0.008)	-0.118(±0.007)	-0.297(±0.005)	-0.133(±0.036)	-0.107(±0.032)
$F_0 = 145$ Hz	0.019(±0.004)	-0.122(±0.004)	-0.382(±0.007)	-0.165(±0.028)	-0.078(±0.025)
$F_0 = 205$ Hz	-0.086(±0.017)	-0.288(±0.017)	-0.630(±0.019)	0.106(±0.243)	-0.154(±0.108)
$F_0 = 210$ Hz	-0.224(±0.032)	-0.427(±0.042)	-0.669(±0.061)	-0.118(±0.258)	-0.143(±0.140)
$F_0 = 230$ Hz	-0.202(±0.037)	-0.335(±0.034)	-0.712(±0.041)	0.072(±0.223)	-0.015(±0.167)
$F_0 = 255$ Hz	-0.152(±0.033)	-0.276(±0.042)	-0.711(±0.046)	0.272(±0.267)	-0.288(±0.097)

Table 3.8 In this table, the mean and the standard deviation of the difference in dB, denoted as HRF_{dif}, between the HRF of the actual glottal flow and the HRF of the estimated one, for each frequency (all 4 vowels) is illustrated.

such as the accurate identification of the closed phase region. However, it is interesting that the IF method based on SWLP with $M = 24$ performs better, in general and given a long enough analysis window, than the conventional autocorrelation method.

Chapter 4

Conclusions

4.1 Summary of Findings

In this thesis, an evaluation of inverse filtering techniques for reliably estimating the glottal flow waveform directly from speech, and to quantify the quality of the estimated glottal flows using parametrization measures. The volume velocity airflow through the glottis, called the *glottal flow*, is the source for voiced speech. Previous studes have shown the importance of the glottal flow in several areas of speech sciences.

Four different techniques were examined in this thesis, all of them based on linear prediction analysis. Two of them are widely used in the literature, and they are based on autocorrelation and covariance method of linear prediction. The covariance analysis was restricted inside the closed phase region of the glottal cycle. Also, two recently developed LP techniques were introduced: Stabilized Weighted Linear Prediction and Constrained Closed Phase Covariance Linear Prediction. SWLP, which was recently suggested for robust vocal tract filter extraction in noisy conditions, computes the all-pole model by imposing temporal weighting of the square of the residual signal.

This is achieved by using STE as a weighting function. Thus, samples that fit the underlying speech production model well are emphasized. The performance of SWLP in inverse filtering depends on the length of the STE window. This property makes it interesting for IF, and it was our motivation for using it. The Constrained CP Covariance LP is a modified CP algorithm based on imposing certain predefined values on the gains of the vocal tract inverse filter at angular frequencies of 0 and/or π in optimizing filter coefficients. With these constraints, vocal tract models are less prone to show false low-frequency roots.

A major problem in assessing the performance of IF techniques is the lack of direct comparison of the estimated glottal flow waveforms with the actual glottal flow, since the latter is can be obtained only with special equipment and in an invasive manner. To this direction, the performance of the discussed techniques is evaluated on a database of speech signals which are produced by physical modeling of the human voice production system. In this way, both the glottal airflow signals and the speech pressure signals are available and direct comparisons can be made. Several frequencies of different vowels were produced and compared to the inverse filtered estimates. The glottal flows were estimated in a pitch synchronous manner using a system for inverse filtering. The system inverse filters each waveform using all diefferent LP methods, and synthesizes the glottal flow estimate. For CP analysis techniques, the CP interval was determined using a statistical technique, which eliminates dependence on a particular type of frequency modulation, and allows the algorithm to adapt to the degree of glottal closure. In order to compare the resulted glottal flow waveforms with the original one in a qualitative way, time and frequency domain parametrization measures were used. In time domain, the Normalized Amplitude Quotient and the Signal to Reconstruction Error ratio was used. In frequency domain, the difference in dB between the first two harmonics (H1-H2), and the Harmonic Richness Factor were used.

Inverse filtering experiments showed that SWLP outperforms the conventional autocorrelation LP for a large STE window in our IF system, and shows comparable performance to the covariance based methods. However, the prevalence of the covariance methods, when the CP region is accurately estimated, is obvious, as expected.

4.2 Suggestions for Future Work

The results of this work demonstrate the different LP-based vocal tract filter estimation techniques, and their effect on inverse filtering. It was shown that robust and accurate vocal tract filter estimation is crucial in the process of inverse filtering.

The use of more sophisticated all-pole models for inverse filtering could be used, such as Discrete All-Pole Modeling (DAP) [20]. DAP tries to fit the all-pole model using only the finite set of spectral locations that are related to the harmonic positions of the fundamental frequency. Through DAP, it is possible to obtain estimates of the formant frequencies that are less biased by the harmonic structure of the speech spectrum. The DAP model uses the discrete Itakura-Saito (IS) error measure and the optimisation criterion is derived in the frequency domain, where the error function reaches the minimum only when the model spectrum coincide on all discrete points. Moreover, the DAP method tries to maximise the error flatness. This could be useful fow high pitch speakers, as in female speech, where the bias of the low formant frequencies is more intense.

A further problem is that if the order of the original LP method is increased, then the corresponding envelope overestimates the original voiced speech power spectrum. This means that the LP envelope is resolving the harmonics and not the spectral envelope. The Minimum Variance Distortionless Response (MVDR) method [49] provides a smooth spectral envelope even when the model order is increased. In particular, if one chooses the proper order for the MVDR method, the all-pole envelope obtained models a set of spectral samples exactly.

In the same sense, nearly any vocal tract filter estimation technique can be used in order to examine how well the effect of the vocal tract filter can be cancelled on a speech signal.

Finally, since the CP covariance method is the prevalent method for inverse filtering and a major problem is the accurate identification of the closed phase, the presented database can be useful in this direction. The closed phases for each phoneme are available and a direct comparison of the performance of several CP identification algorithms can be made.

On the parametrization of the glottal flow, the LF model could be used to parametrize both the estimates and the original glottal flow. Since there can be either time or frequency domain LF-parametrization, a more robust framework of ”similarity” between glottal flow waveforms can be established, along with the measures described in this thesis. Also, the accuracy of quotients or spectral parameters is deteriorated when the glottal flows are severely affected by noise. The LF model can alleviate this distortion caused by instantaneous noisy peaks of the flow, especially in the case of NAQ, where the amplitude values of the glottal flow and the flow derivative may be distorted. Finally, both time and frequency domain metrics that were discussed here do not take into account the presence of a ripple component in the closed phase, an event that happens when there is incomplete cancelling of some of the higher formant by the inverse filter. Alternatively, this component might be explained by the existence of nonlinear coupling between the source and the tract, which cannot be taken into account in any analysis based on linear modeling of the voice production system.

In the direction of evaluating the performance of IF, it can be suggested that a more detailed simulation of the physical model of the human voice production system will reveal properties and details that are unknown so far. The database selected for the experiments in this thesis consists of signals that are produced by a simplified but physically-motivated representation of a speaker. A problem is that these vowels sound ”unnatural”. An interesting approach is the ”3D Vocal Tract Project” [74], which is a three-dimensional vocal tract model for articulatory and visual speech synthesis developed within CTT, the Centre for Speech Technology, KTH Royal Institute of Technology, Sweden. An ideal physical model would provide realistic waveforms in several midpoints of the vocal tract, such as near the glottis, inside the oral cavity, and at the lips.

Furthermore, since the area of Voice Care has shown increased interest in non invasive methods of extracting the glottal flow waveform, it would be interesting to implement a real-time system of glottal inverse filtering. Recently, a real time voice pathology detection system based on pitch estimation and short time jitter estimator was implemented [12]. This system was implemented in Pure Data (PD), which is a real-time graphical programming environment for audio and graphical processing. This worked is also based on short time speech analysis, as in the IF system described in this thesis, and showed that a real time approach in speech analysis is both efficient and accurate. In addition, the algorithms embedded in the IF system are low in computational cost, a fact that supports an effort for a real time approach.

Bibliography

[1] Ananthapadmanabha, T. and Yegnanarayana, B., 1979, "Epoch extraction from linear prediction residual for identification of closed glottis interval", IEEE Trans. Acoust., Speech and Signal Processing, vol. 27, no. 4, pp. 309-319.

[2] Ananthapadmanabha, T. and Fant, G, 1985, "Calculation of the true glottal flow and its components", STL-QPR, 1-30.

[3] Akande, O., and Murphy, P., 2005, "Estimation of the vocal tract transfer function with application to glottal wave analysis", Speech Commun. 46, 15-36

[4] Alkhairy, A, 1999, "An algorithm for glottal volume velocity estimation", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 1, 233-236.

[5] Alku, P., 1992, "An automatic method to estimate the time-based parameters of the glottal pulseform", Proc. IEEE, Int. Conf. Acoustics, Speech and Signal Proc. 2, 29-32.

[6] Alku, P., 1992, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", Speech Commun. 11, 109-118.

[7] Alku, P., Backstrom, T., and Vilkman, E., 2002, "Normalized Amplitude Quotient for parametrization of the glottal flow", J. Acoust. Soc. Am., 112, 701-710.

[8] Alku, P., Magi, C., Yrttiaho, S., Backstrom, T., and Story, B., 2009, ”Closed Phase Covariance Analysis based on Constrained Linear Prediction for Glottal Inverse Filtering”, J. Acoust. Soc. Am., 125(5), 3289-3305.

[9] Alku, P., Story, B., and Airas, M., 2006, ”Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production”, Folia Phoniatrica et Logopaedica, 58(2), 102-113.

[10] Alku, P., Strik, H., and Vilkman, E., 1997, ”Parabolic Spectral Parameter - A new method for quantification of the glottal flow”, Speech Communications, 22, 67-79.

[11] Alku, P., and Vilkman, E., 1996, ”A comparison of glottal voice source quantification parameters in breathy, normal, and pressed phonation of female and male speakers”, Folia Phoniatr. Logop. 48, 240-254.

[12] Astrinaki M., 2010, ”Real Time Voice Pathology Detection Using Autocorrelation Pitch Estimation and Short Time Jitter Estimator”, M.Sc. Thesis, Computer Science Department, University of Crete.

[13] Childers, D., and Ahn, C., 1995, ”Modeling the glottal volume velocity waveform for three voice types”, J. Acoust. Society of America, 97, 505-519.

[14] Childers, D., and Hu, H., 1994, ”Speech synthesis by glottal excited linear prediction”, J. Acoust. Society of America, 96, 2026-2036.

[15] Childers, D., Lee, C., 1991, ”Vocal Quality Factors: Analysis, synthesis, and perception”, J. Acoust. Society of America, 90, 2394-2410.

[16] Childers, D., Principe, J.C., and Ting, Y.T., 1995, "Adaptive WRLS-VFF for speech analysis", IEEE Transactions on Speech and Audio processing, 3(3), 209-213.

[17] Cummings, K. E., and Clements, M. A., 1995, "Analysis of the glottal excitation of emotionally stlyed and stressed speech", J. Acoust. Society of America, 98, 88-98.

[18] Childers, D., Wong, C.-F., 1994, "Measuring and modeling vocal source-tract interaction", IEEE Trans. Biomed. Eng. 41(7), 663-671.

[19] Deng, H., Beddoes, M. P., Ward, R. K., Hodgson, M, 2003, "Estimating the Glottal Waveform and the Vocal-Tract Filter from a Vowel Sound Signal", Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing, 1, 297-300.

[20] El-Jaroudi, A. and Makhoul, J., 1991, "Discrete all-pole modelling", IEEE Transactions on Signal Processing, 39(2), 411-423.

[21] Fant, G., 1970, *Acoustic Theory of Speech Production* (Mouton, The Hague).

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

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

[24] Fitch, T., and Giedd, J., 1999, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging", J. Acoust. Soc. America, 106, 1511-1522

[25] Flanagan, J., 1972, *Speech Analysis, Synthesis, and Perception*, (Springer, New York)

[26] Frohlich, M., Michaelis, D., and Strube, H., 2001, SIM - Simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals", J. Acoust. Soc. America, 110, 479-488.

[27] Fu, Q., and Murphy, P., 2006, "Robust glottal source estimation based on joint source-filter model optimization", IEEE Trans. Audio, Speech, Lang. Process., 14, 492-501.

[28] Gobl, C., 1988, "Voice source dynamics in connected speech", STL-QPSR, 1, 123-159.

[29] Hirano, M., 1974, "Morphological structure of the vocal cord as a vibrator and its variations", Folia Phoniatr. Logop., 26, 89-94.

[30] Hollien, H., Dew, D., and Philips, P., 1971, "Phonational frequency ranges of adults", J. Speech Hear. Res., 14, 755-760.

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

[32] Holmes, J.N., 1973, "The influence of glottal waveform on the naturalness of speech form a parallel formant synthesizer", IEEE Trans. Audio Electroacoust., AU-21, 298-305.

[33] Howell, P., and Williams, M., 1992, "Acoustic Analysis and perception of vowels in children's and teenagers' stuttered speech", J. Acoust. Soc. Am., 91, 1697-1706.

[34] Ishizaka, K., and Flanagan, J. L., 1972, ”Synthesis of voiced sounds from a two mass model of the vocal cords”, Bell Syst. Tech. J., 51, 1233-1268.

[35] Kane, J., Kane, M., Gobl1, C., 2010, ”A spectral LF model based approach to voice source parameterisation”, Interspeech, Kyoto, Japan.

[36] Klatt, D., and Klatt, L., 1990, ”Analysis, synthesis, and perception of voice quality variations among female and male talkers”, J. Acoust. Soc. America, 87, 820-857.

[37] Krishnamurthy, A., and Childers, D., 1986, ”Two-channel speech analysis”, IEEE Trans. Acoust. Speech, Signal Proc., 34, 730-743.

[38] Larar, J., Alsaka, Y., and Childers, D., 1985, ”Variability in closed phase analysis of speech”, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Tampa, FL, 1089-1092.

[39] Liljencrants, J., 1985, ”Speech synthesis with a reflection-type line analog”, DS dissertation, Department of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm, Sweden.

[40] Lim, J., Oppenheim, A., 1978, ”All-pole modeling of degraded speech”, IEEE Trans. Acoust. Speech and Signal Process., ASSP-26(3), 197-210.

[41] Ma, C., Kamp, Y., Willems, L., 1993, ”Robust signal selection for linear prediction analysis of voiced speech”, Speech Comm. 12 (1), 69-81.

[42] Magi, C., Pohjalainen, J., Backstrom, T., Alku, P., 2009, ”Stabilised Weighted Linear Prediction”, Speech Communication, doi:10.1016/j.specom.2008.12.005

[43] Makhoul, J., 1975, ”Linear Prediction: a tutorial review”, Proc. IEEE 63(4), 561-580.

[44] McAulay, R.J., Quatieri, T.F., 1990, ”Pitch estimation and voicing detection based on a sinusoidal speech model”, IEEE International Conference on Acoustics, Speech, and Signal Processing, (1), 249-252.

[45] Matausek, M. R., Batalov, V. S., 1980, ”A new approach to the determination of the glottal waveform”, IEEE Trans. Acoust., Speech, Signal Proc., 28, 616-622.

[46] McKenna, G., 2001, ”Automatic Glottal Closed-Phase Location and Analysis by Kalman Filtering”.

[47] Milenkovic, P., 1986, ”Glottal inverse filtering by joint estimation of an AR system with a linear input model”, IEEE Trans. Acoust. Speech, Signal Process., 34, 28-42.

[48] Miller, R., 1959, ”Nature of the vocal cord wave”, J. Acoust. Soc. Am., 31, 667-677.

[49] Murthi, M.N, and Rao, B.D., 2000, ”All-pole modelling of speech based on the minimum variance distortionless response spectrum”, IEEE Transactions on Speech and Audio Processing, 8(3), 221-239.

[50] Murty, K. and Yegnanarayana, B., 2008, ”Epoch Extraction From Speech Signals”, IEEE Trans. Audio Speech Lang. Processing, 16, 1602-1613.

[51] Naylor, P., Kounoudes, A., Gudnason, J., and Brookes, M., 2007, ”Estimation of glottal closure instants in voiced speech using the DYPSA algorithm”, IEEE Trans. Audio, Speech, Lang. Process., 15, 34-43.

[52] Oppenheim, A., and Schafer, R., 1989, *Discrete-Time Signal Processing*, (Prentice Hall, Englewood Cliffs, NJ).

[53] Plumpe, M., Quatieri, T., and Reynolds, D., 1999, ”Modeling of the glottal flow derivative waveform with application to speaker identification”, IEEE Trans. Speech Audio Process., 7, 569-586.

[54] Price, P., 1989, ”Male and female voice source characteristics: inverse filtering results”, Speech Commun., 8, 261-277.

[55] Quatieri, T., 2001, *Discrete Time Speech Signal Processing: Principles and Practice*, Prentice Hall, Englewood Cliffs, NJ.

[56] Rabiner, L., and Schafer, R., 1978, *Digital Processing of Speech Signals*, (Prentice Hall, Englewood Cliffs, NJ).

[57] Riegelsberger, E., and Krishnamurthy, A., 1993, ”Glottal source estimation: methods of applying the LF model to inverse filtering”, in Proceeding of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, MN, Vol. 2, pp. 542-545.

[58] Rothenburg, M., 1973, ”A new inverse filtering technique for deriving the glottal air flow waveform during voicing”, J. Acoust. Soc. Am., 53, 1632-1645.

[59] Story, B., 1995, Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract”, Ph.D. dissertation, University of Iowa.

[60] Story, B., and Titze, I., 1995, ”Voice simulation with a body-cover model of the vocal folds”, J. Acoust. Soc. Am., 97, 1249-1260.

[61] Story, B., Titze, I., and Hoffman, E., 1996, ”Vocal tract area functions from magnetic resonance imaging”, J. Acoust. Soc. Am., 100, 537-554.

[62] Strube, H., 1974, ”Determination of the instant of glottal closure from the speech wave”, J. Acoust. Soc. Am., 56, 1625-1629.

[63] Strube, H., 1982, ”Time-varying wave digital filters for modeling analog systems”, IEEE Trans. Acoust. Speech and Signal Processing, 30, 864-868.

[64] Sundberg, J., Titze, I., and Scherer, R., 1993, ”Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source”, J. Voice, 7, 15-29.

[65] Ting, Y. T., Childers, D. G., 1990, ”Speech Analysis using the Weighted Recursive Least Squares Algorithm with a Variable Forgetting Factor”, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 1, 389-392.

[66] Titze, I., 2002, ”Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model”, J. Acoust. Soc. Am., 111, 367-376.

[67] Titze, I., and Story, B., 2002, ”Rules for controlling low-dimensional vocal fold models with muscle activites”, J. Acoust. Soc. Am., 112, 1064-1076.

[68] Titze, I., and Sundberg, J., 1992, ”Vocal intensity in speakers and singers”, J. Acoust. Soc. Am., 107, 581-588.

[69] Veeneman, D., and BeMent, S., 1985, ”Automatic glottal inverse filtering from speech and electroglottographic signals”, IEEE Trans. Acoust. Speech, Signal Process., 33, 369-377.

[70] Walker, J, 2003, ”Application of the bispectrum to glottal pulse analysis”, Proc. NoLisp ’03.

[71] Wong, D., Markel, J., and Gray, A., Jr., 1979, ”Least Squares glottal inverse filtering form the acoustic speech waveform”, IEEE Trans. Acoust., Speech, Signal Process. 27, 350-355.

[72] Yegnanarayana, B., and Veldhuis, N., 1998, ”Extraction of vocal-tract system characteristics from speech signals”, IEEE Trans. Speech Audio Process., 6, 313-327.

[73] Zhao, Q., Shimamura, T., Suzuk, J., 1997, ”Linear Predictive Analysis of noisy speech”, Communications, Computers and Signal Processing, PACRIM’97, Victoria, Canada, August 20-22, (2), 585-588.

[74] ”A three-dimensional vocal tract model for articulatory and visual speech synthesis developed within CTT, the Centre for Speech Technology, KTH”: http://www.speech.kth.se/multimodal/vocaltract.html