Glottal Source Parametrisation by Multi-estimate Fusion

by

Haoxuan Li

A Dissertation submitted in partial fulfilment of the requirements

for the Degree of Doctor of Philosophy

Supervisors: Dr. Ronan Scaife and Dr. Darragh O'Brien

School of Electronic Engineering

Dublin City University

February, 2013

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of the degree of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ (Candidate) ID No.: _____

Date: _____ /_____ /_____

I would like to dedicate this thesis to my loving parents ...

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Ronan Scaife for the continuous support of my Ph.D research work. His expert advise and guidance helped me in all the time of study and completion of this thesis. Much thanks also goes to my co-supervisor Dr. Darragh O'Brien. His advise always gave me inspirations for the research and his feedback was extremely useful for the correction of the thesis. I am grateful to the technical staff at DCU, particularly Robert Clare for their assistance all the time. Thanks to Dr. Xiaojun Wang, who helped me to apply for this project.

I extend my thanks to my friends in the school of electronic engineering, Zhenhui, Longhao, Chen for supporting and encouraging me, especially during the time of pressure.

My lovely parents have always been a constant source of support during my four years Ph.D program. Thanks to my mother for her wise advise for applying project by the Chinese Scholarship Council as the source of my funding.

Last but not least, I acknowledge my beautiful and cute girlfriend, Zhuoran. During the hardest time of my work she has been, always, my pillar, my joy and my sunlight.

List of Publications

H. Li, R. Scaife and D. O'Brien, "LF model based glottal source parameter estimation by extended Kalman filtering", in Proceedings of the Irish Signals and Systems Conference, 2011.

H. Li, R. Scaife and D. O'Brien, "Comparison of Time- and Frequency-domain Based LF-model Fitting Methods for Voice Source Parameterisation", in Proceedings of the Irish Signals and Systems Conference, 2012.

H. Li, R. Scaife and D. O'Brien, "Automatic LF-model fitting to the glottal source waveform by extended Kalman filtering", in Proceedings of the 20th European Signal Processing Conference, 2012.

H. Li, D. O'Brien and R. Scaife, "Robust tracking of glottal LF-model parameters by multi-estimate fusion", in Proceedings of the 20th European Signal Processing Conference, 2012.

Contents

Contents **v**

List of Figures **x**

List of Tables **xiv**

1 Introduction **1**

1.1 Motivation of the thesis 1

1.2 Research Question and Hypothesis 3

1.3 Contributions of the thesis 4

1.4 Structure of the thesis 4

2 Background **7**

2.1 Introduction 7

2.2 Human Speech Production 8

2.2.1 Speech Anatomy 8

2.2.2 Speech Categorisation 8

2.2.3 Fundamental Frequency 10

2.2.4 Formants 11

2.3 The Source-Filter Model 11

2.3.1 Glottal Source Modelling 13

2.3.2 Vocal Tract Modelling 16

2.3.3 Lip Radiation Modelling 16

2.4 Conclusion 18

3 Glottal Waveform Extraction **19**
3.1 Introduction . . . 19
3.2 Glottal Inverse Filtering . . . 20
3.2.1 Closed Phase Inverse Filtering (CPIF) . . . 22
3.2.2 Iterative Adaptive Inverse Filtering (IAIF) . . . 25
3.2.3 Weighted Recursive Least Squares with Variable Forgetting Factor Analysis (WRLS-VFF) . . . 30
3.3 Performance Study . . . 34
3.4 Other Speech Decomposition Methods . . . 46
3.4.1 Mixed-phase Speech Decomposition . . . 46
3.4.2 Higher Order Statistics Approaches . . . 49
3.5 Conclusion . . . 50

4 Automatic Glottal LF-model Fitting **52**
4.1 Introduction . . . 52
4.2 Curve Fitting . . . 53
4.3 Automatic LF-model Fitting Related Work . . . 53
4.3.1 Review of the LF-model . . . 55
4.3.2 Time-domain LF-model Fitting . . . 56
4.3.2.1 Strik's Method . . . 57
4.3.2.2 Riegelsberger's Study . . . 59
4.3.2.3 Childers and Ahn's Method . . . 60
4.3.3 Frequency-domain LF-model Fitting . . . 60
4.3.4 Factors Affecting LF-model Fitting . . . 63
4.4 A New Time-domain LF-model Fitting Algorithm by Extended Kalman Filtering . . . 64
4.4.1 Extended Kalman Filtering (EKF) . . . 64
4.4.2 Discrete Time LF-model representation . . . 66
4.4.3 LF-model Shape-controlling Parameter Tracking by EKF . . . 67
4.4.3.1 Tracking for α by EKF . . . 67
4.4.3.2 Tracking for ε by EKF . . . 69
4.4.4 Algorithm Implementation . . . 71
4.5 Performance Study . . . 73

4.5.1 Comparison with a Standard Time-domain Method 73
4.5.1.1 Synthetic Speech 73
4.5.1.2 Real Speech 76
4.5.2 Comparison with a Modified Frequency-domain Method 78
4.5.2.1 Artificial Glottal Source 79
4.5.2.2 Real Glottal Source 81
4.6 Conclusion 84

5 A Multi-estimate Fusion Framework for Glottal Source Parameter Estimation 86
5.1 Introduction 86
5.2 Quantitative Data Fusion 87
5.2.1 An Introduction to Data Fusion 87
5.2.2 Two General Fusion Structure: State-Vector Fusion and Measurement Fusion 89
5.2.2.1 State-Vector Fusion 90
5.2.2.2 Measurement Fusion 91
5.2.3 A Basic Fusion Formula: Millman's Formula 92
5.2.4 A Data Fusion Tool: Kalman Filter 94
5.3 Glottal LF-model Parameter Multi-estimate Fusion 96
5.3.1 Multi-estimate Fusion Framework 96
5.3.1.1 Multiple Glottal Source Extractions 98
5.3.1.2 Multiple LF-model Fitting algorithms 98
5.3.1.3 Multiple Estimates Combination 99
5.3.1.4 Fused Estimates Smoothing 99
5.3.2 Advantages and Limitations 102
5.3.3 Factors Affecting the Performance of the Proposed Fusion Algorithm 103
5.4 Conclusion 104

6 Multi-estimate Fusion Evaluation 106
6.1 Introduction 106

6.2 Implementation of the Fusion Algorithm 107
6.3 Synchronisation by Glottal Closing Instants 110
6.4 Evaluation on Synthetic Voiced Segments 112
6.5 Evaluation on Utterance Recordings 114
6.6 Detailed Analysis of Performance Variation across Individual Algorithms 120
6.7 Evaluation on Hand-labelled Data 125
6.8 Extending the Fusion Framework 132
6.8.1 Adding FD-LF Fitting to the Fusion Framework 133
6.8.2 Adding JKN-LF Fitting to the Fusion Framework 137
6.9 Assessment of the Fusion Framework 140
6.10 Conclusion 141

7 Summary & Conclusions **143**
7.1 Introduction 143
7.2 Summary of the thesis 143
7.3 Contribution of the thesis 146
7.4 Further work suggestions 148

References **152**

Abstract

Glottal source information has been proven useful in many applications such as speech synthesis, speaker characterisation, voice transformation and pathological speech diagnosis. However, currently no single algorithm can extract reliable glottal source estimates across a wide range of speech signals. This thesis describes an investigation into glottal source parametrisation, including studies, proposals and evaluations on glottal waveform extraction, glottal source modelling by Liljencrants-Fant (LF) model fitting and a new multi-estimate fusion framework.

As one of the critical steps in voice source parametrisation, glottal waveform extraction techniques are reviewed. A performance study is carried out on three existing glottal inverse filtering approaches and results confirm that no single algorithm consistently outperforms others and provide a reliable and accurate estimate for different speech signals.

The next step is modelling the extracted glottal flow. To more accurately estimate the glottal source parameters, a new time-domain LF-model fitting algorithm by extended Kalman filter is proposed. The algorithm is evaluated by comparing it with a standard time-domain method and a spectral approach. Results show the proposed fitting method is superior to existing fitting methods.

To obtain accurate glottal source estimates for different speech signals, a multi-estimate (ME) fusion framework is proposed. In the framework different algorithms are applied in parallel to extract multiple sets of LF-model estimates which are then combined by quantitative data fusion. The ME fusion approach is implemented and tested in several ways.

The novel fusion framework is shown to be able to give more reliable glottal LF-model estimates than any single algorithm.

List of Figures

2.1 An overview of the human vocal system (from [Mannell, 2009]) . . 9
2.2 a) Voiced, b) Unvoiced and c) Plosive sounds 10
2.3 A typical glottal excitation pulse train 11
2.4 A typical vowel spectrum with labelled formant frequencies and bandwidths 12
2.5 Caption for LOF 12
2.6 Plot of an idealised glottal source (from [Taylor, 2009]) 14
2.7 A typical LF-model pulse and its parameters in time domain . . . 15
2.8 A four-formant, eight-pole model of vowel /a:/, top: spectrum plot and bottom: Z-plane plot. 17

3.1 The process of glottal inverse filtering in the frequency (upper plot) and time domains (lower plot) (from [Gobl, 2003]) 21
3.2 A block diagram of the ICPIF algorithm 24
3.3 Glottal component estimates and the a_1 values by iteration (from [Moore and Clements, 2004]) 26
3.4 Block diagram of the Iterative Adaptive Inverse Filtering (IAIF) algorithm. (from [Airas, 2008]) 28
3.5 An example of applying IAIF to a male speech segment. Top: spectra plot and bottom: poles in Z-plane 29
3.6 Applying WRLS-VFF analysis to a male speech segment /aa/ . . 32
3.7 Glottal inverse filtering by WRLS-VFF algorithm (from [Ting and Childers, 1990]) 33
3.8 High quality glottal flow and differentiated glottal flow waveform estimate (above), Phase-plane plots and the four GQMs (below) . 36

3.9 Poor quality glottal flow and differentiated glottal flow waveform estimate (above), Phase-plane plots and the four GQMs (below) . 37
3.10 Histograms of a segment of voiced speech (top) and the corresponding glottal estimate (bottom) 38
3.11 Waveform of male speech frame /æ/ and corresponding glottal estimates by the three inverse filtering algorithms. 40
3.12 Waveform of female speech frame /æ/ and corresponding glottal estimates by the three inverse filtering algorithms. 41
3.13 Waveform of male speech frame /ə/ and corresponding glottal estimates by the three inverse filtering algorithms. 42
3.14 Waveform of female speech frame /ə/ and corresponding glottal estimates by the three inverse filtering algorithms. 43
3.15 Waveform of male speech frame /ɔ/ and corresponding glottal estimates by the three inverse filtering algorithms. 44
3.16 Waveform of female speech frame /ɔ/ and corresponding glottal estimates by the three inverse filtering algorithms. 45
3.17 Waveforms of a voiced real speech frame and the corresponding glottal estimate by ZZT 48

4.1 Data set and curve fitting results (linear fitting) 54
4.2 Data set and curve fitting results (fifth order polynomial fitting) . 54
4.3 A typical LF-model pulse (bottom) and its undifferentiated waveform (top) 55
4.4 Flow chart of the spectral LF-fitting algorithm [Kane et al., 2010]. 62
4.5 Fitted LF-model open phase signals according to different α_0 values 69
4.6 MSE scores of estimated LF-model parameters for a) modal voice, b) vocal fry voice and c) breathy voice 75
4.7 Top: male speech waveform, Bottom: GFD signals and fitted LF-model pulses (LFP) from the two algorithms 77
4.8 Top: female speech waveform, Bottom: GFD signals and fitted LF-model pulses (LFP) from the two algorithms 77
4.9 Single pitch period of GFD and fitted LF waveforms for top: male and bottom: female 78

4.10 Artifical glottal source LF-model parameter true values and the estimates by EKFLF and MFDF 80
4.11 Real glottal source LF-model parameter hand-labelled values and the estimates by EKFLF and MFDF 82
4.12 An example where the LF-model is better fitted by EKFLF to the real glottal source 83
4.13 An example where the LF-model is better fitted by MFDF to the real glottal source 83

5.1 State-vector fusion structure 90
5.2 Measurement fusion structure 91
5.3 Kalman filter prediction and update equations 95
5.4 A general framework of the multi-estimate fusion algorithm 97

6.1 Implementation of multi-estimate fusion algorithm 108
6.2 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the male utterance “Robbery, bribery, fraud.” 116
6.3 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the female utterance “Robbery, bribery, fraud.” 116
6.4 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the male utterance “Shall I carry you.” 117
6.5 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the female utterance “Shall I carry you.” 117
6.6 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the male utterance “He did not rush in.” 118
6.7 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the female utterance “He did not rush in.” 118

6.8 Example 1: Waveform, spectrum, estimated GFD and poles by multiple algorithms of sound /ɔ/ in "robbery" by male speaker . . 122
6.9 Example 2: Waveform, spectrum, estimated GFD and poles by multiple algorithms of sound /i/ in "robbery" by male speaker . . 123
6.10 Example 3: Waveform, spectrum, estimated GFD and poles by multiple algorithms of sound /i/ in "carry" by female speaker . . 124
6.11 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for JK 127
6.12 Weights by different approaches across JK 127
6.13 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for JD 128
6.14 Weights by different approaches across JD 128
6.15 Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for LP 129
6.16 Weights by different approaches across LP 129

7.1 Implementation of the state-vector multi-estimate fusion algorithm 151

List of Tables

3.1 Equations of the WRLS-VFF algorithm 32
3.2 Glottal waveform quality measures by the three GIF methods for male speech frame /æ/ 40
3.3 Glottal waveform quality measures by the three GIF methods for female speech frame /æ/ 41
3.4 Glottal waveform quality measures by the three GIF methods for male speech frame /ə/ 42
3.5 Glottal waveform quality measures by the three GIF methods for female speech frame /ə/ 43
3.6 Glottal waveform quality measures by the three GIF methods for male speech frame /ɔ/ 44
3.7 Glottal waveform quality measures by the three GIF methods for female speech frame /ɔ/ 45

4.1 Implementation of the new time-domain LF-model fitting algorithm to the GFD train 72
4.2 LF-model parameters for three voice qualities 74
4.3 Formant frequencies and bandwidths of three vowels 76
4.4 MSE scores for real speech segments from two automatic time-domain LF-model fitting algorithms 77
4.5 Range of LF-model parameters 80
4.6 RMSE scores for the three LF-model parameters by EKFLF and MFDF applied to artificial speech data 80
4.7 RMSE scores for the three LF-model parameters by EKFLF and MFDF applied to real speech data 81

6.1 Asynchronous glottal estimates 110
6.2 Asynchronous glottal estimates 111
6.3 RMSE scores of LF-model parameters estimated by different algorithms for clean synthetic speech 113
6.4 RMSE scores to hand-labelled data for JK 130
6.5 Means of fitting error covariance and weight for JK 130
6.6 RMSE scores to hand-labelled data for JD 131
6.7 Means of fitting error covariance and weight for JD 131
6.8 RMSE scores to hand-labelled data for LP 132
6.9 Means of fitting error covariance and weight for LP 132
6.10 RMSE to TCD-labelled data by different algorithms for JK 134
6.11 Means of fitting error covariance and weight for JK 134
6.12 RMSE scores to TCD-labelled data by different algorithms for JD 135
6.13 Means of fitting error covariance and weight for JD 135
6.14 RMSE scores to TCD-labelled data by different algorithms for LP 136
6.15 Means of fitting error covariance and weight for LP 136
6.16 RMSE to TCD-labelled data by different algorithms for JK 137
6.17 Means of fitting error covariance and weight for JK 137
6.18 RMSE to TCD-labelled data by different algorithms for JD 138
6.19 Means of fitting error covariance and weight for JD 138
6.20 RMSE to TCD-labelled data by different algorithms for LP 139
6.21 Means of fitting error covariance and weight for LP 139

Chapter 1

Introduction

1.1 Motivation of the thesis

Glottal source parametrisation is useful in a number of speech applications such as:

• Glottal source parameters can be used to produce more realistic source signal then a pulse train, and when used by a HMM-based speech synthesis system the naturalness of the corresponding synthetic speech can be improved [Cabral et al., 2007, 2008, 2011]

• Glottal flow model parameters can be applied to speaker identification [Plumpe et al., 1999]. It is shown that the glottal features contain significant speaker dependent information by utilising a Gaussian mixture model speaker identification system to a large TIMIT database subset.

• Prosody analysis and modification needs not only intonation information but also voice source control rules for quality improvement [Strik and Boves, 1992], where we need to understand the relation between glottal source parameters and voice quality.

• Voice source features are established to be useful to detect speech pathologies [Drugman et al., 2009c; Dubuisson et al., 2009]. The relevancy of the glottal source-based features, speech signal-based features and prosodic features is assessed and it is demonstrated the glottal features such as the glottal formant, glottal open quotient and amplitude of the main excitation can be incorporated

to detect pathological voice.

For its usefulness, over recent decades, much research effort has been devoted to estimating the glottal source from the speech waveform signals. Generally, speech is considered to be produced by a source-filter model, where source is the glottal source and filter is the vocal tract effect. Thus, to track the glottal source parameters we need to decompose speech into its two components. The most widely used method is glottal inverse filtering (GIF) [Wong et al., 1979; Alku and Vilkman, 1994], which is to remove the vocal tract effect from the speech signal to yield the glottal source waveform. Subsequently the source component parameters are obtained by fitting a parametric model [Fant et al., 1985] to the glottal waveform [Strik et al., 1993].

The diversity and complexity of human speech (and extraneous factors such as recording device characteristics and ambient noise) pose significant challenges to any single glottal source parametrisation algorithm. Currently, no individual algorithm performs the best for all kinds of speech signals. The Iterative Adaptive Inverse Filtering (IAIF) algorithm [Alku, 1992] has its limitation in estimating the glottal flow for speech with low frequency of the first formant. Closed Phase Inverse Filtering (CPIF) [Wong et al., 1979] may generate inaccurate glottal estimate when the analysed speech signal has short or no duration of the closed phase. Zero of Z-transform analysis [Bozkurt, 2005] decomposes the speech into maximum- and minimum-phase components, where the maximum-phase component represents the glottal source, however, the return phase information cannot be extracted. In addition, there is still no scientific tool to analyse and extract the glottal source parameters for arbitrary input speech signals, which is useful to study the variation features of the source parameters across different speakers and sounds. Also, in the area of glottal source parametrisation, based on no a priori information of the real glottal component it is difficult to make comparisons between different algorithms.

Under these circumstances, a glottal source parametrisation algorithm is required, which should be able to

- accurately estimate the voice source parameters across a wide range of speech,
- extract all the useful glottal parameters from the original speech signal,

• present the trajectories of the source parameters for an arbitrary input speech signal, such as a word or an utterance and,

• offer useful information as references to study the performance of different algorithms.

The goal of this study is to attempt to find a such solution. To develop a completely new approach as defined above is a difficult task, which requires a comprehensive study of various features of a large number of speech signals. However, it is reasonable to develop a system which can intelligently combine estimates from multiple algorithms, by utilising the data fusion technique. Such a fusion algorithm is theoretically more reliable than individual algorithms for its ability to automatically lock to well performing local algorithms.

1.2 Research Question and Hypothesis

Multi-sensor data fusion (MSDF) aims to enhance the stability, accuracy and robustness of applications trying to track, identify and extract features of objects by collecting, filtering and fusing information from different sources. Generally, the source can be sensors for capturing static images, video and audio streams, object locations and so forth. MSDF combines multiple sets of data from different sensors to estimate one or more properties of the object, providing more robustness and accuracy than a single sensor.

Where no a priori information is available for selecting the optimal approach for analysing input speech signals, it may be more reasonable to apply a selection of different algorithms in parallel to extract multiple sets of source parameter estimates and to combine them by MSDF techniques. For voice source parameter estimation, different algorithms can be regarded as different sensors tracking the same set of source parameters. Accordingly, multiple sets of estimates will be obtained and if combined in an appropriate way, the reliability and accuracy of the estimates may be enhanced.

This idea of combining multiple sets of estimates for glottal source parametrisation has the potential to overcome the limitations of single algorithms and improve the accuracy of the estimated source parameters across different speech signals, which is crucially important for relevant applications. In addition, it will

give direction to future investigation for utilising more advanced techniques to further improve the performance of the fusion algorithm, and will be a useful reference for researchers working in the area. Thus, it is a worthy investigation.

1.3 Contributions of the thesis

This thesis contributes the following to the area of glottal source estimation:

- A comprehensive review and investigation of the performance of existing techniques for glottal source extraction.

- A review of existing approaches to fitting glottal Liljencrants-Fant (LF) model [Fant et al., 1985] to extracted glottal estimates.

- Proposal and evaluation of a new time-domain LF-model fitting algorithm, by extended Kalman filtering.

- Proposal of a novel general multi-estimate fusion framework for accurate glottal source estimation, which draws on different algorithms in parallel to obtain multiple sets of estimates and combine them in the fusion centre.

- Implementation and evaluation of the newly proposed fusion approach.

1.4 Structure of the thesis

The thesis is structured as follows:

Chapter 2 briefly describes the human speech anatomy and some important features of the speech signal. The source-filter model is presented which is widely used for speech signal generation and analysis. Also in this chapter, the glottal source model, vocal tract model and lip radiation model are introduced. The validity of the source-filter model is the basic assumption for all relevant techniques in the thesis.

Chapter 3 focuses on techniques for glottal source extraction. Accurate glottal source extraction is crucial before fitting the parametric model to the glottal estimates. Although a large number of glottal source extraction algorithm exist, less work has been carried out to study and compare the performance of different approaches. In this chapter, several effective glottal inverse filtering approaches are reviewed and the performance of each is studied by applying it to real speech signal segments. Results confirm that no single algorithm consistently outperforms others. In addition, some non-linear-prediction based techniques for speech decomposition are discussed.

Chapter 4 investigates model fitting methods and introduces a new approach for automatically fitting the LF-model (introduced in Chapter 2) to the glottal source component obtained by the approaches studied in Chapter 3. Firstly, curve fitting is briefly explained, which is the basis for the LF fitting method before several existing LF-model fitting algorithms are reviewed, including both time-domain and frequency-domain based methods. Subsequently, a new time-domain fitting approach by extended Kalman filter is proposed. The novel algorithm is compared with a standard time-domain method and a typical spectral fitting approach. Experimental results show the effectiveness of the new algorithm.

Chapter 5 firstly introduces some techniques relevant to quantitative data fusion including Millman's fusion formula, state-vector fusion, measurement fusion and Kalman filtering. Subsequently, a multi-estimate fusion framework for voice source parametrisation is proposed. The fusion framework operates across four stages: multiple glottal source estimation, multiple LF-model parameter estimation, fusion of multiple estimates and Kalman filter smoothing. The functions of each stage are discussed in detail, and the advantages and disadvantages of the algorithm are considered. Also, possible factors that may affect the performance of the fusion method are presented.

Chapter 6 tests the effectiveness of the proposed multi-estimate fusion algorithm by presenting several evaluations. Firstly, the fusion algorithm is implemented incorporating three inverse filtering approaches and one LF-model fitting

method. It is applied to synthetic speech signals and the results are compared with those of single algorithms. Subsequently, this implementation is tested with real speech utterances. Analysis is carried out by presenting several examples to illustrate the performance of each algorithm across various speech frames. In a further evaluation, the fusion algorithm is applied to to an all voiced utterance by difference speakers for which hand-labelled glottal source parameters exist. To test the effect of adding another algorithm to the existing framework, an additional LF fitting approach is integrated by two alternatives, where one is poorly performing and the other is well performing. It can be observed from the results that the fusion algorithm can generate acceptable estimates and overall it is more reliable than single algorithms, since the fusion method can automatically assign more weight on good estimate.

Chapter 7 presents conclusions. In this section, the main findings are summarised and the contributions of the thesis are presented. Finally, further investigations that would extend the work described here are suggested.

Chapter 2

Background

2.1 Introduction

For better understanding the extraction of glottal source parameter estimates, it is necessary to outline the human speech production system. Indeed, an understanding of the source-filter model of speech signal production is central to the thesis.

One of the most prominent differences between human beings and other animals is that we can speak. Speech is our most natural method of communication and it is an exclusive skill of human beings. We learn to speak from our parents and other individuals during infancy. We learn foreign languages because we want to communicate with people outside our countries and understand their cultures. Speech is also an important identifying feature because each of us has a different set of vocal organ parameters giving rise to unique pronunciation characteristics. Identification by voice has already been used in many areas such as criminal investigation and security systems. The actual human speech production system is very complex and even now has not been accurately modelled or fully understood. However, the basic mechanism for pronunciation is clear and we can mimic it with a relatively simple model.

In Section 2.2 we will firstly give an introduction to the basic aspects of speech production. Afterwards, a classical model, the source-filter model, which is simple but widely used by many speech processing applications is presented

and all components of the model are described in Section 2.3.

2.2 Human Speech Production

In this section we give an introduction to the human speech production system and describe some basic characteristics of speech signals. For further information see [Fant, 1970; Flanagan, 1972; Rabiner and Schafer, 1978; Deller et al., 1993].

2.2.1 Speech Anatomy

For the development of useful speech processing applications, it is necessary to understand how human speech is produced. The anatomy of our vocal mechanism determines the generation of different speech sounds. Presented in Fig. 2.1 is a schematic of the anatomy of speech production.

In general, a speech signal is an air pressure wave that travels from the speaker's mouth to the listener's ears. The main organs of the production system include the larynx, vocal tract and nasal tract. The vocal tract starts at the opening of vocal cords, which is the glottis, and ends at the lips. For male adults the average length of the vocal tract is approximated as 17 cm. The cross sectional area of the vocal tract depends on the positions of the tongue, lip, jaw and uvula; it varies from zero (fully closed) to about 20 cm^2. The nasal passage starts at the uvula and ends at the nostrils. The nasal passage and the vocal tract work together to generate the nasal sounds [Taylor, 2009].

The complete speech production system also includes the sub-glottal system which includes the lungs, bronchus and trachea. Air moves from the lungs along the vocal tract, impeded by constrictions at certain positions in the vocal tract, exiting as a speech sound wave at the lips.

2.2.2 Speech Categorisation

Speech can be divided into three categories according to different inputs to the vocal tract:

a) **Voiced Speech** When air flows across the glottis, if it makes the vocal

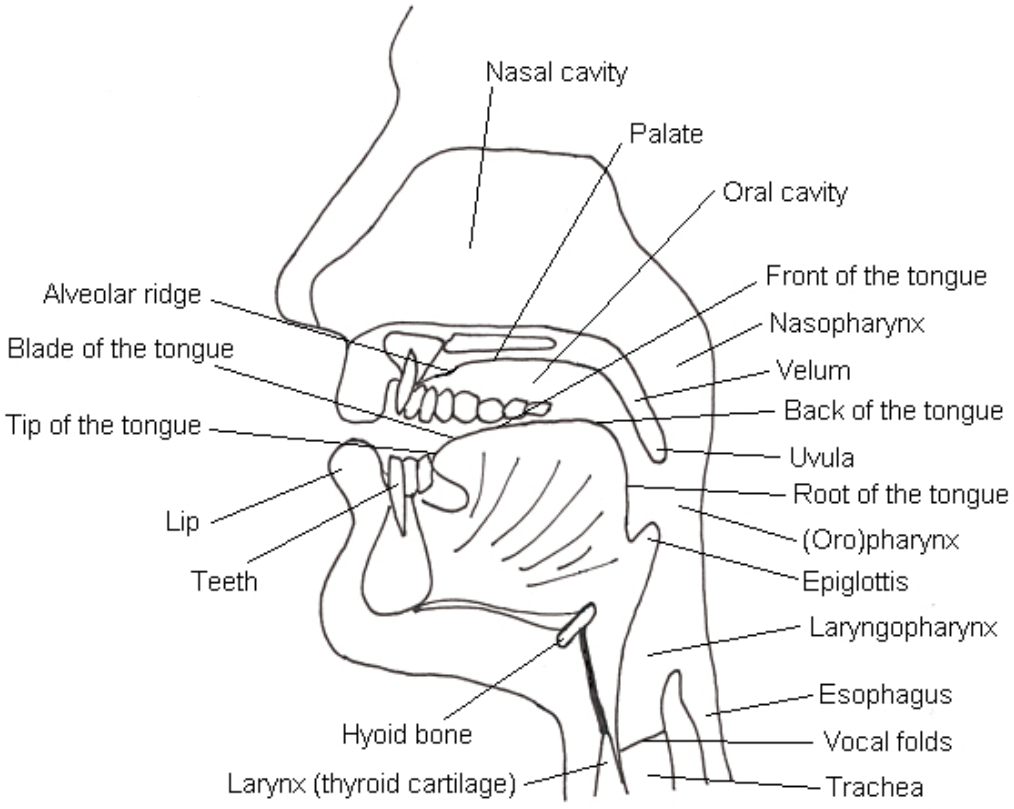

Figure 2.1: An overview of the human vocal system (from [Mannell, 2009])

folds vibrate repeatedly, the vocal tract is excited with a quasi-periodic input source and a voiced sound is generated.

b) **Unvoiced Speech** When air flows through the glottis, if the vocal folds shrink instead of vibrating, which makes the airflow pass through the constriction with a high velocity and produces a turbulent flow, an unvoiced sound is generated.

c) **Plosive Speech** If the vocal folds or lips are fully closed and the air pressure increases, a plosive speech sound is created after the abrupt release of the airflow. Fig. 2.2 shows example waveforms of the three kinds of speech sounds.

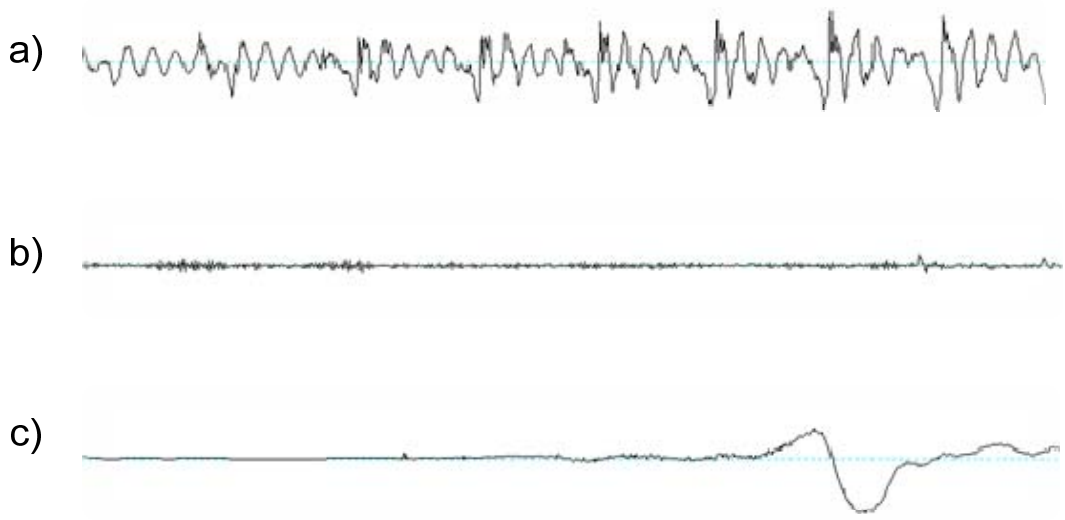

Figure 2.2: a) Voiced, b) Unvoiced and c) Plosive sounds

2.2.3 Fundamental Frequency

When a voiced sound is generated, airflow passes through the glottis and causes the vocal cords to vibrate, producing a quasi-periodic excitation pulse train. Such a typical glottal flow waveform is presented in Fig. 2.3. The period of the pulse train is generally represented by T_0, and its reciprocal is the fundamental frequency, usually represented by f_0.

f_0 is related to the length, thickness, tenacity of an individual's vocal cords. Generally, the fundamental frequency of an adult male speaker has a distribution

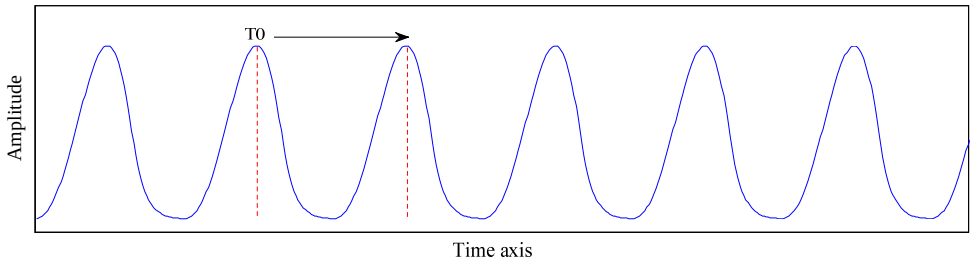

Figure 2.3: A typical glottal excitation pulse train

in the range from 50∼250Hz. For an adult female speaker or child, the range is 120∼500Hz [Deller et al., 1993].

2.2.4 Formants

The vocal tract can be considered as a tube of heterogeneous sections, and each resonance frequency of the tube is called a formant frequency (typically shortened to formant for convenience). Formants depend on the positions of the vocal organs, which means that formant frequencies are related to the shape of the vocal tract. Each vocal tract shape has its corresponding set of formant frequencies. Therefore, if the shape of the vocal tract changes, a different sound is generated, and the spectrum of the speech signal will also change. Formants are numbered from low-to-high frequencies by F_1, F_2, F_3, etc. In voiced speech, generally five formants can be distinguished, with the first three of vital importance in discriminating different speech sounds [Childers, 1999]. A typical vowel spectrum with labelled four formants is shown in Fig. 2.4.

2.3 The Source-Filter Model

According to [Fant, 1981], the human speech signal can be modelled by a source-filter model depicted in Fig. 2.5.

For voiced sounds, the source is defined as the glottal volume velocity signal through the glottis, which is represented by U_g. The vocal tract filter V has a particular configuration which filters the source. A lip radiation effect R com-

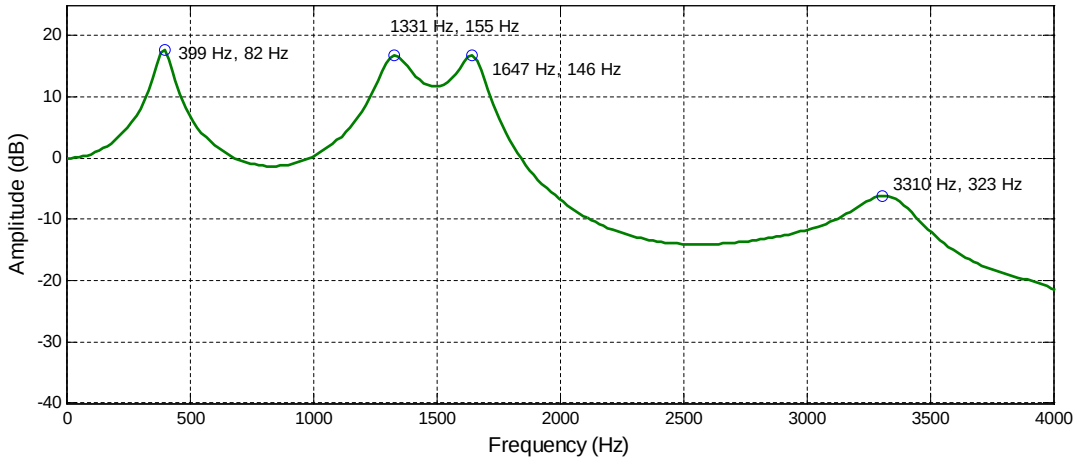

Figure 2.4: A typical vowel spectrum with labelled formant frequencies and bandwidths

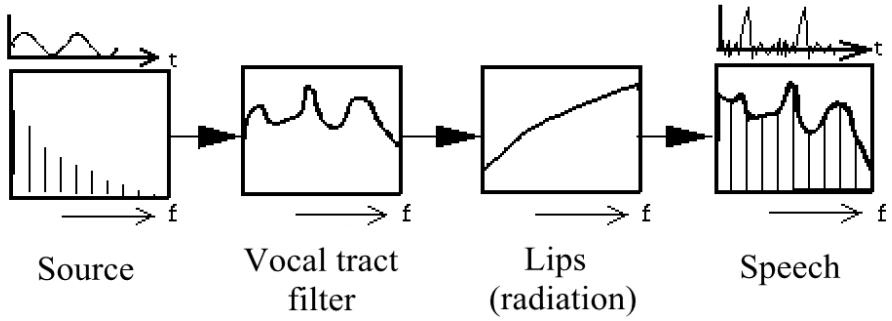

Figure 2.5: The Source-Filter model[a]

[a]obtained from https://www.msu.edu/course/asc/232/index.html

pletes the model. If we represent the output speech signal as S, the source-filter model can be defined by equation (2.1) (in the complex Z-domain):

$$S(z) = U_g(z)V(z)R(z) \tag{2.1}$$

In fact, the effect of the lip radiation can be combined with the glottal flow to give the differentiated glottal flow signal, also called glottal flow derivative U_d. So equation (2.1) can be re-written as:

$$S(z) = U_d(z)V(z) \tag{2.2}$$

One typical application of the source-filter model is the linear predictive coding (LPC) [Makhoul, 1975]. LPC utilises a simple pulse train as the glottal source, whose amplitude and F_0 are obtained by linear prediction analysis. The vocal tract and lip radiation effect are modelled by filters which are described below. Speech is generated by putting the pulse train through the vocal tract and lip radiation filters. The disadvantage of LPC is the lack of naturalness of the output signal resulting from its simplified glottal source representation. To improve the quality of synthetic speech, a more complex parametric model must be used to describe the shape of the voice source. Such a model is presented in the following section.

2.3.1 Glottal Source Modelling

In section 2.2, we described how the glottis produces different kinds of speech sounds. For voiced speech, the vocal cords vibrate one cycle after another, which gives rise to a quasi-periodic waveform. Fig. 2.6 shows a diagram of a typical glottal flow in the time domain.

Generally one pitch period of the glottal source has three phases: the open phase, the return phase and the closed phase. The open phase is so-named because of the opening of the glottis during this period due to the pressure coming from the lungs. The return phase starts at the glottal closing instant and ends when the glottis is just fully closed because of the vocal tract tension. During the closed phase, the vocal folds are closed before the next pitch cycle.

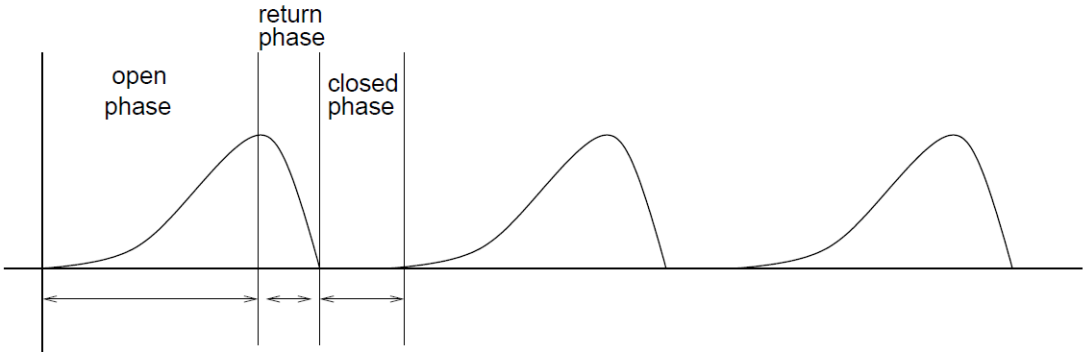

Figure 2.6: Plot of an idealised glottal source (from [Taylor, 2009])

Accurate modelling of the movement of the glottis is a difficult problem and has not been solved completely. Oscillation of the vocal folds is a non-uniform pattern which is too complex to be represented by simple mathematical equations. Many attempts have been made to improve the modelling of the glottal movement, such as a body-cover vocal-fold structure model introduced by Titze and Story. This consists of two "cover" masses coupled laterally to a "body" mass by non-linear springs and viscous damping elements to mechanically simulate the vocal fold vibration [Story and Titze, 1995; Titze and Story, 2002; Story, 2002]. In the absence of sufficiently adaptable physical models, parametric models for directly representing the glottal flow are often applied to fit the inverse filtered glottal waveform.

These models include the LF-model [Fant et al., 1985], the KLGLOTT88 model [Klatt and Klatt, 1990], the Flanagan model [Flanagan et al., 1975] and the R++ model [Veldhuis, 1998]. The most widely used is the LF-model [Fant et al., 1985], which describes the glottal flow derivative. The glottal open phase, return phase, and closed phase are given by equation (2.3), where $e(t)$ is the derivative signal.

$$e(t) = \begin{cases} \mathrm{E}_0 e^{\alpha t} \sin(\omega_g t), & 0 \leq t \leq t_e \quad \textit{open phase} \\ -\frac{E_e}{\varepsilon t_a} \left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)} \right], & t_e < t \leq t_c \quad \textit{return phase} \\ 0, & t_c < t \leq T_0 \quad \textit{closed phase} \end{cases} \tag{2.3}$$

A typical single pitch period LF-model and its undifferentiated equivalent

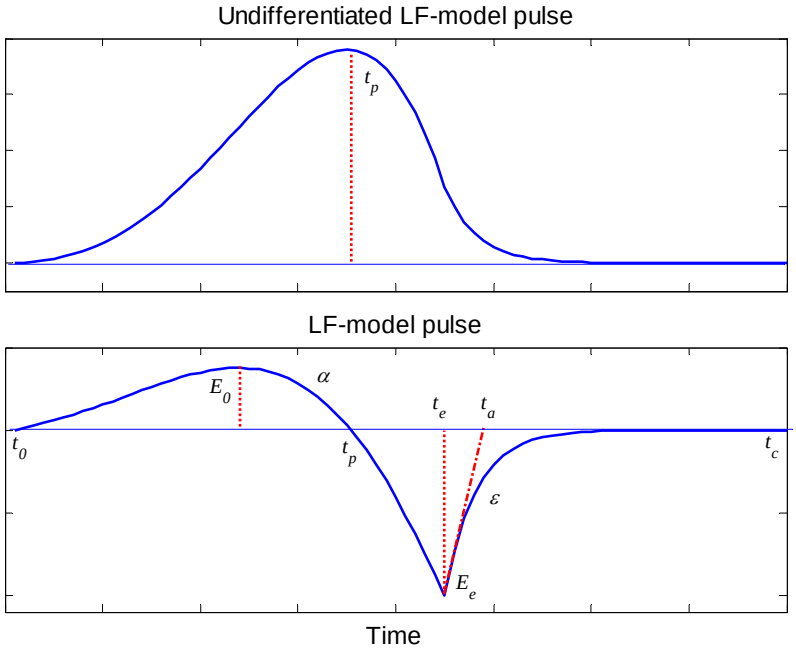

Figure 2.7: A typical LF-model pulse and its parameters in time domain

waveform are presented in Fig. 2.7, where t_0 is the instant of the glottal opening (here t_0 is equal to 0), t_p is the instant when the undifferentiated flow reaches its maximum, t_e is the time of the negative peak, which is also the glottal closing instant, t_c is the instant of glottal closure (t_c=T_0 in this example), t_a controls the return phase. The α, ω_g and ε parameters are the shape-controlling parameters. E_0 and E_e are the positive and negative peak values of the derivative function.

The transformed LF-model parameters were introduced by [Fant, 1995] and are presented in equation (2.4)

$$\begin{aligned} T_0 &= t_c - t_0 \\ R_g &= T_0/(2t_p) \\ R_k &= (t_e - t_p)/t_p \\ R_a &= t_a/T_0 \\ OQ &= t_e/T_0 = (1 + R_k)/(2R_g) \end{aligned} \tag{2.4}$$

where the R_g parameter increases with a decreasing t_p, R_k determines the dura-

tion of the falling interval from the glottal flow peak at time t_p to the glottal flow derivative negative peak at time t_e, R_g and R_k together determine the glottal open quotient OQ, and R_a is the return phase parameter t_a normalised by pitch period and it accounts for the degree of glottal spectral tilt. This set of transformed parameters describes the shape of the LF-model compared to the original timing parameters.

2.3.2 Vocal Tract Modelling

Generally, the vocal tract for most voiced sounds such as vowels can be modelled by an all-pole infinite impulse response (IIR) filter [Rabiner and Schafer, 1978; Deller et al., 1993]. This is because the vocal tract can be represented by the multiplication of transfer functions of a cascade of formant resonances. One single formant can be expressed by a second-order IIR filter. Therefore, the order for the all-pole model is double the number of formants. For some other types of sounds such as nasal sounds, the corresponding spectra contain not only poles but also zeros required to model the nasal cavity [Taylor, 2009]. Because of the complexity introduced by zeros, most researchers use the all-pole model for all types of voiced sounds. The all-pole vocal tract model is given by equation (2.5) in Z-domain

$$V(z) = \frac{G}{1 - \sum_{k=1}^{N} a_k z^{-k}} \tag{2.5}$$

where G is the gain and a_k are the coefficients related to the formant frequencies and bandwidth, and N is the order of the model. For a four-formant model for example, the order N is eight. Fig. 2.8 shows the spectrum (top) of the model and the poles calculated from a_k in Z-plane (bottom) for the vowel sound /a:/.

2.3.3 Lip Radiation Modelling

It is known that the speech signal waveform is influenced by the volume velocity at the lips through a radiation impedance, $R(z)$ in Z-domain. Unfortunately to accurately model the effect of lip radiation is a complicated and difficult problem [Taylor, 2009]. The overall effect of the lip radiation is to apply a +6 dB/octave

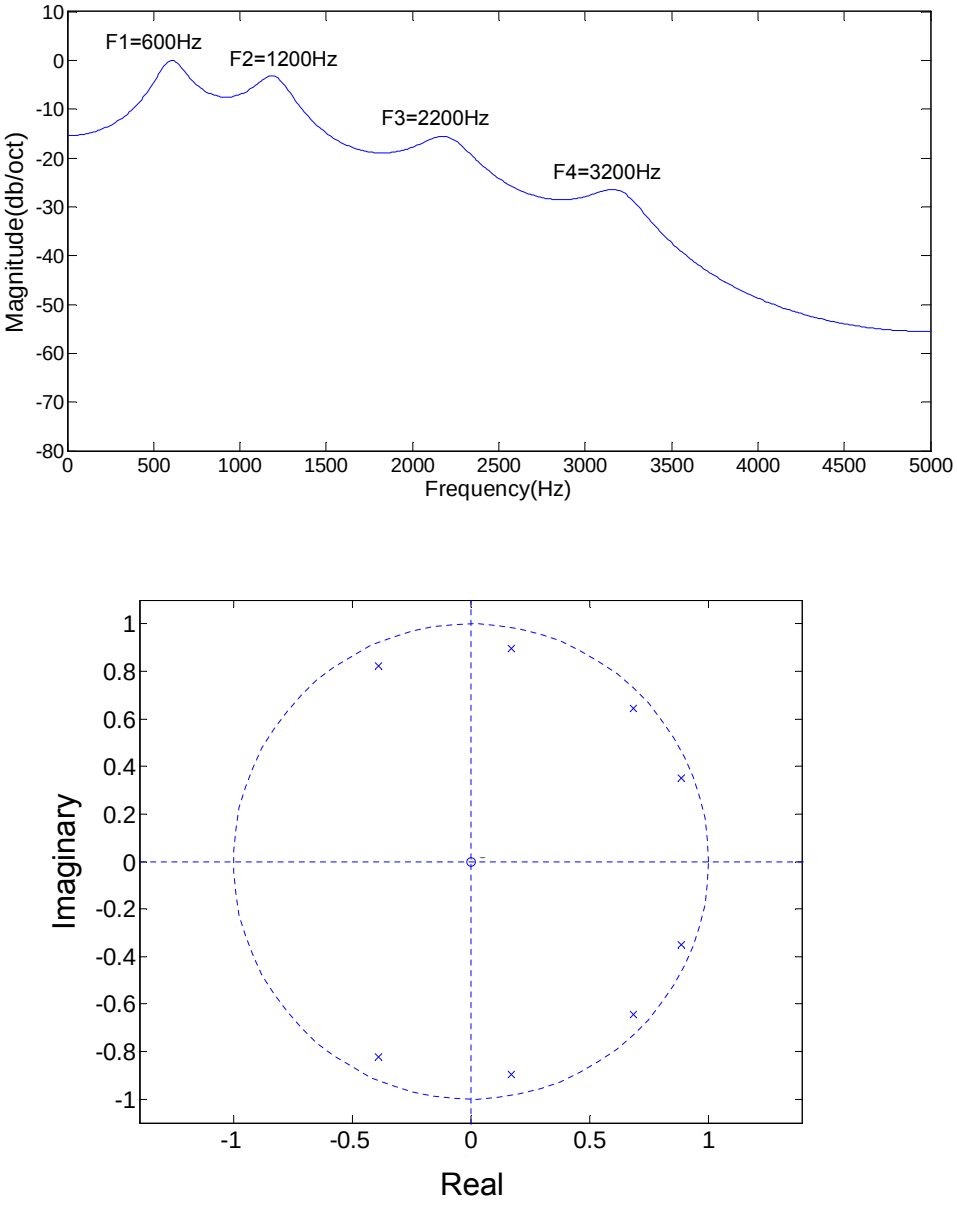

Figure 2.8: A four-formant, eight-pole model of vowel /a:/, top: spectrum plot and bottom: Z-plane plot.

emphasis to the air flow at the lips, thus researchers commonly use a first order differentiation function

$$R(z) = 1 - a_0 z^{-1} \tag{2.6}$$

where a_0 has a value less than but quite near to 1 (e.g. 0.95), to model the radiation function.

In most studies, the differentiation of the lip radiation function is applied directly to the glottal flow yielding the differentiated glottal flow modelled by the LF-model [Fant et al., 1985]. In this study, we also use the glottal flow derivative signal, or glottal waveform to refer to the glottal excitation.

2.4 Conclusion

In this chapter, the human speech production system was introduced. According to different excitation methods speech can be categorised as voiced, unvoiced and plosive. Important parameters of the speech signal, the fundamental frequency and vocal tract formants, were introduced. Subsequently, the source-filter model was presented. The speech signal can be effectively regarded as the multiplication of three components: glottal source, vocal tract filter and lip radiation in the frequency domain. We introduced the most widely used parametric model, the LF-model, which is used to represent the glottal excitation in voiced speech. We have shown that the vocal tract can be modelled by an all-pole IIR filter, and finally we noted that the lip radiation function can be effectively modelled by a first-order differentiator. In the next chapter we examine a range of algorithms whose aim is to recover the original glottal excitation from a given speech waveform.

Chapter 3

Glottal Waveform Extraction

3.1 Introduction

As mentioned in Chapter 2, speech production can be simply modelled by a source-filter model. Although this model has its limitations such as not taking into consideration the interaction between the glottal source and vocal tract [Fant, 1993], it has proved to be useful and works well for many speech processing applications.

The aim of glottal waveform extraction is to separate the glottal source from the vocal tract component by decomposing speech signals. To obtain a perfect 'clean' glottal waveform is a very difficult task for real speech. It is known that speech is a unique characteristic of individuals. Thus an approach which works well for one individual's speech may not work well for others. The fundamental frequency of female speech is higher than that of male speech, which means there is less data per individual pitch period to be used for analysis making it more difficult to analyse female speech [Walker and Murphy, 2007]. Efforts have been made attempting to obtain more reliable glottal and vocal tract estimates for speech with less data for e.g., female speech and transition sounds. A "multi-cycle covariance method" proposed by Yegnanarayana & Veldhuis utilises data from consecutive pitch cycles to average the estimates by the covariance LP analysis [Yegnanarayana and Veldhuis, 1998]. McKenna applied the Kalman filter to automatically detect the closed phase and utilise the non-independence of neigh-

bouring closed phases to combine estimates. This method overcomes the short duration of the closed phase, and more accurate LP estimates can be obtained [McKenna, 2001].

In addition, the environment (noise level) and the recording device may cause distortion in the recorded speech. Also, because of source-vocal tract interaction[Fant, 1993], it is difficult to perfectly and completely remove the formants from speech signals.

Many research efforts have been made to decompose speech into glottal source and vocal tract components. Although limitations exist, some of these approaches have proved useful and have been widely adopted. In Section 3.2 of this chapter, the most widely used speech decomposition approach - glottal inverse filtering (GIF) is introduced in detail, and three state-of-the-art fully automatic GIF algorithms are presented. Details of a performance study comparing the three are also presented in Section 3.3. In addition in section 3.4, other speech decomposition methods, such as mixed-phase speech deconvolution [Bozkurt et al., 2004b,a; Bozkurt, 2005] and the higher order statistics method [Nikias and Raghuveer, 1987; Mendel, 1991], are briefly described.

3.2 Glottal Inverse Filtering

Glottal inverse filtering is a technique which aims to remove the spectral effect caused by the vocal tract from speech signals and leave only the glottal source. Fig. 3.1 shows the process of inverse filtering in both the time and frequency domains.

In Chapter 2, the source-filter model was introduced, according to which most voiced speech can be regarded as the result of applying the glottal source signal through a particular vocal tract filter. Based on this, if the process is inverted, which means putting the speech signal through a filter which is the inverse of the vocal tract filter, the effect of the resonances will be cancelled, and ideally the voice source will be obtained. If we use $U_g(z)$ to represent the glottal flow, where $V(z)$ is the transfer function of the vocal tract filter, $R(z)$ is the lip radiation function and S(z) is the speech signal spectrum, the process of glottal inverse

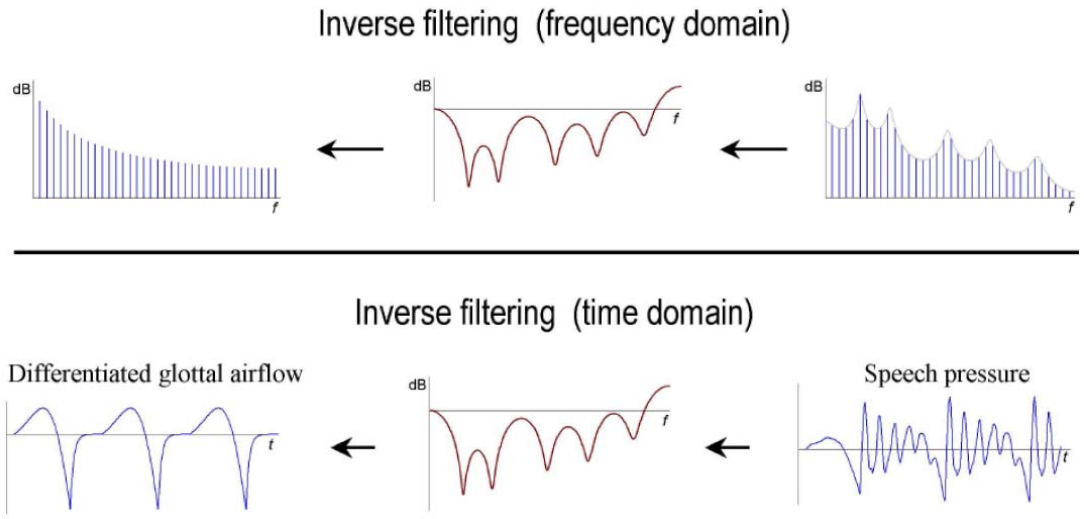

Figure 3.1: The process of glottal inverse filtering in the frequency (upper plot) and time domains (lower plot) (from [Gobl, 2003])

filtering can be expressed by equation (3.1).

$$U_g(z) = \frac{S(z)}{V(z)R(z)} \tag{3.1}$$

$R(z)$ can be considered as a first order FIR filter, thus $U_g(z)$ multiplied by $R(z)$ gives the glottal flow derivative $U_d(z)$.

$$U_d(z) = \frac{S(z)}{V(z)} \tag{3.2}$$

Currently there are many different glottal inverse filtering techniques available ranging from manual methods (see [Gobl et al., 1999]) to fully automatic algorithms (see [Wong et al., 1979; Alku and Laine, 1989; Alku, 1992; Alku and Vilkman, 1994; Vincent et al., 2007; McKenna, 2004; Moore and Clements, 2004; Akande and Murphy, 2005; Backstrom and Alku, 2006]). Most are based on the linear prediction techniques [Makhoul, 1975]. Using linear prediction, for an assumed linear and time invariant system, future values can be predicted by a difference equation which is derived from the previous input. For speech signals, it is known that the short-term speech signal can be regarded as linear and time

invariant; this means that linear prediction can be used to estimate the coefficients of the vocal tract all-pole filter by minimizing the error (covariance or autocorrelation methods) between the predicted sample values and the original inputs.

In this section we will introduce three glottal inverse filtering techniques which can be applied fully automatically to extract the glottal waveform: Closed Phase Inverse Filtering (CPIF), Iterative Adaptive Inverse Filtering (IAIF) and Weighted Recursive Least Squares analysis with Variable Forgetting Factor based inverse filtering (WRLS-VFF). These three techniques were successfully implemented by the author, their performance studied and the three methods will be used later in Chapter 6 for evaluating the proposed multi-estimate fusion framework.

3.2.1 Closed Phase Inverse Filtering (CPIF)

The assumption underlying closed phase inverse filtering is that during each pitch period of voiced speech there is an interval which is free of the influence of the glottal flow. In that interval, the glottis is closed and the interval is called the closed phase. During the closed phase, the speech signal consists only of the decaying vocal tract resonances. Thus, linear prediction analysis applied to this time interval will only model the vocal tract filter and the glottal excitation influence will be excluded [Wong et al., 1979]. The glottal waveform can then be extracted by performing inverse filtering with the vocal tract filter, with its coefficients estimated from the closed phase, on the entire pitch period of the original speech signal.

In the conventional CPIF method, the crucial step is to find the glottal closure instant (GCI) which is the start point of the closed phase. Inaccurate GCI estimates will introduce artefacts to the process of inverse filtering. Many epoch-detection algorithms exist for finding the GCIs, ranging from manual pitch-marking to fully automatic implementations [Cheng and O'Shaughnessy, 1989; Kounoudes et al., 2002; Ma et al., 1994; Naylor et al., 2007; Drugman and Dutoit, 2009].

One limitation of CPIF is that it is sensitive to inaccurate GCI estimates

[Alku et al., 2009]. In addition, for speech signals which have non-zero excitation during the closed phase interval, such as glottal leakage resulting from incomplete glottal closure [Gobl and N Chasaide, 1999], or have short or zero duration closed phases, CPIF may generate poor results [Walker and Murphy, 2005].

Many efforts have been made to improve the performance and accuracy of closed phase inverse filtering. In [Krishnamurthy, 1984], a second channel signal, called the electroglottograph (EGG), is used to better identify the closed phase in cases where the duration of the closed phase is short, e.g. in higher fundamental frequency speech (females, children), or breathy speech. In [Alku et al., 2009], Alku proposed a modified closed phase algorithm based on imposing certain predefined values on the gains of the vocal tract inverse filter at normalised angular frequencies of 0 and π in order to optimise filter coefficients, which makes the algorithm less sensitive to the location of the covariance frame position than the conventional closed phase technique. McKenna [McKenna, 2001, 2004] introduced an algorithm to determine the glottal closed phase locations and separate the source and filter from the speech signal using Kalman Filtering. The method overcomes some of the flaws in conventional linear prediction and closed phase analysis such as the non-stationary interval analysis issue, the requirement of a minimum number of closed phase data samples and high fundamental frequency speech issue.

In this work, an Iterative Closed Phase Inverse Filtering (ICPIF) method based on [Moore and Clements, 2004] was implemented. This method is less sensitive then standard CPIF to the identification of the glottal closing instants. A replotted block diagram of the approach is presented in Fig. 3.2.

$s_k[n]$ is the input speech signal frame covering three or more pitch periods. Firstly a Glottal Closure Instant detection technique is applied to extract the initial GCI locations. The estimated GCI points are then used as midpoints for an iterative procedure. The actual starting point (C) of the closed phase interval is determined by subtracting the order of the model (P) from the GCI locations. Afterwards, the vocal tract filter coefficients are obtained by applying covariance LP analysis (the poles outside of the unit circle are checked and reflected for stability assurance) to the time interval with length=$2P$. The corresponding glottal component G is obtained by filtering with the inverse vocal tract transfer

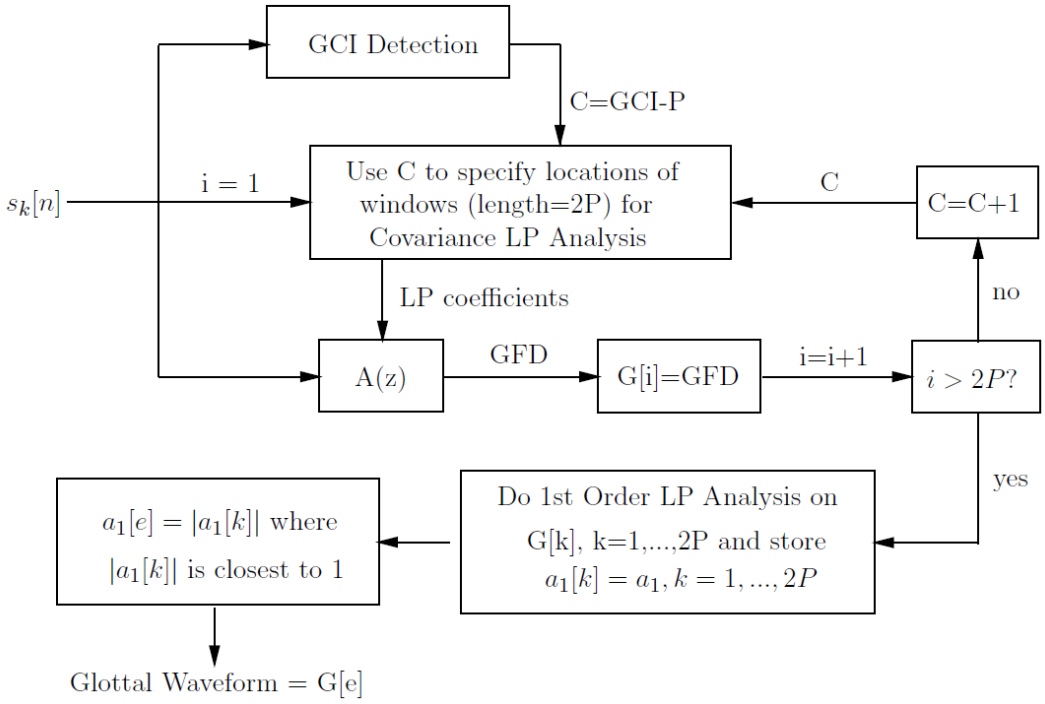

Figure 3.2: A block diagram of the ICPIF algorithm

function $A(z)$. For each pitch period, a total of 2P iterations are performed with the window centred at C being updated by one sample resulting in a series of sliding windows and the glottal estimates $G[k], k = 1, 2, ..., 2P$ are stored.

For each estimated glottal component G, the essential difference is that some exhibit noisy properties while others are relatively smooth. For each glottal estimate, a first-order LP autocorrelation analysis is applied and the coefficient $a_1[k], k = 1, 2, ..., 2P$ are stored. The reason for this is that the first term a_1 of the LP analysis describes the ratio of the autocorrelation function at lag 1 to the autocorrelation function at lag 0, given by equation (3.3).

$$a_1 = \frac{-r(1)}{r(0)} \tag{3.3}$$

In fact, a_1 represents the extent to which two consecutive samples are correlated with each other. For an ideal glottal flow derivative waveform (like an LF-model pulse), two consecutive samples are highly correlated, especially for the glottal

open phase (where in the LF-model it is described by a sinusoidal function). Although for real speech, incompletely cancelled formants and other noisy components will affect the smoothness of the glottal estimate, it is still reasonable to claim that a good glottal estimate should result in the value of a_1 closer to 1 than a more noisy estimate. Accordingly, the index (e) of the vector a_1 with value closest to 1 is picked from G as the best glottal waveform estimate. An example illustrating the glottal estimates and their corresponding a_1 values is presented in Fig. 3.3. It can be observed that values of a_1 which are closer to 1 are smoother (compared with, e.g., $|a_1| = 0.056981$ and $|a_1| = 0.16141$). (This 'a_1 criterion' was also used in [O'Cinneide et al., 2011b] for automatically selecting the order of the vocal tract filter.)

By running closed phase LP analysis iteratively with varying closed phase interval and choosing the best glottal waveform estimate from all iterations, ICPIF requires no precise glottal closure information. The limitation of this approach is the lack of robustness of the a_1 criterion. Sometimes inaccurate inverse filtered glottal flow derivatives (e.g. in the presence of incomplete formant cancellation) may show a_1 close to 1. Also, as for conventional CPIF, for speech signals with short or zero closed phase durations, ICPIF may produce poor estimates.

3.2.2 Iterative Adaptive Inverse Filtering (IAIF)

Introduced and developed by Alku [Alku and Laine, 1989; Alku, 1992], Iterative Adaptive Inverse Filtering (IAIF) operates based on the assumption that the overall spectral tilt of the speech signal can be attributed to the glottal source component, and the glottal flow waveform can be represented by a low order pole model. In the process of IAIF, the gross features of the glottal flow are repeatedly estimated by performing low order linear prediction analysis, and the effect of the glottal source is removed from the speech signal by means of inverse filtering. Subsequently, a more accurate estimation of the vocal tract filter can be obtained by using higher order LP analysis on the speech signal free of the estimated glottal effect. In the last step, the original speech signal is inverse filtered by the estimated vocal tract filter coefficients and the final glottal waveform estimate is derived. In his later work [Alku and Vilkman, 1994], Alku suggested using dis-

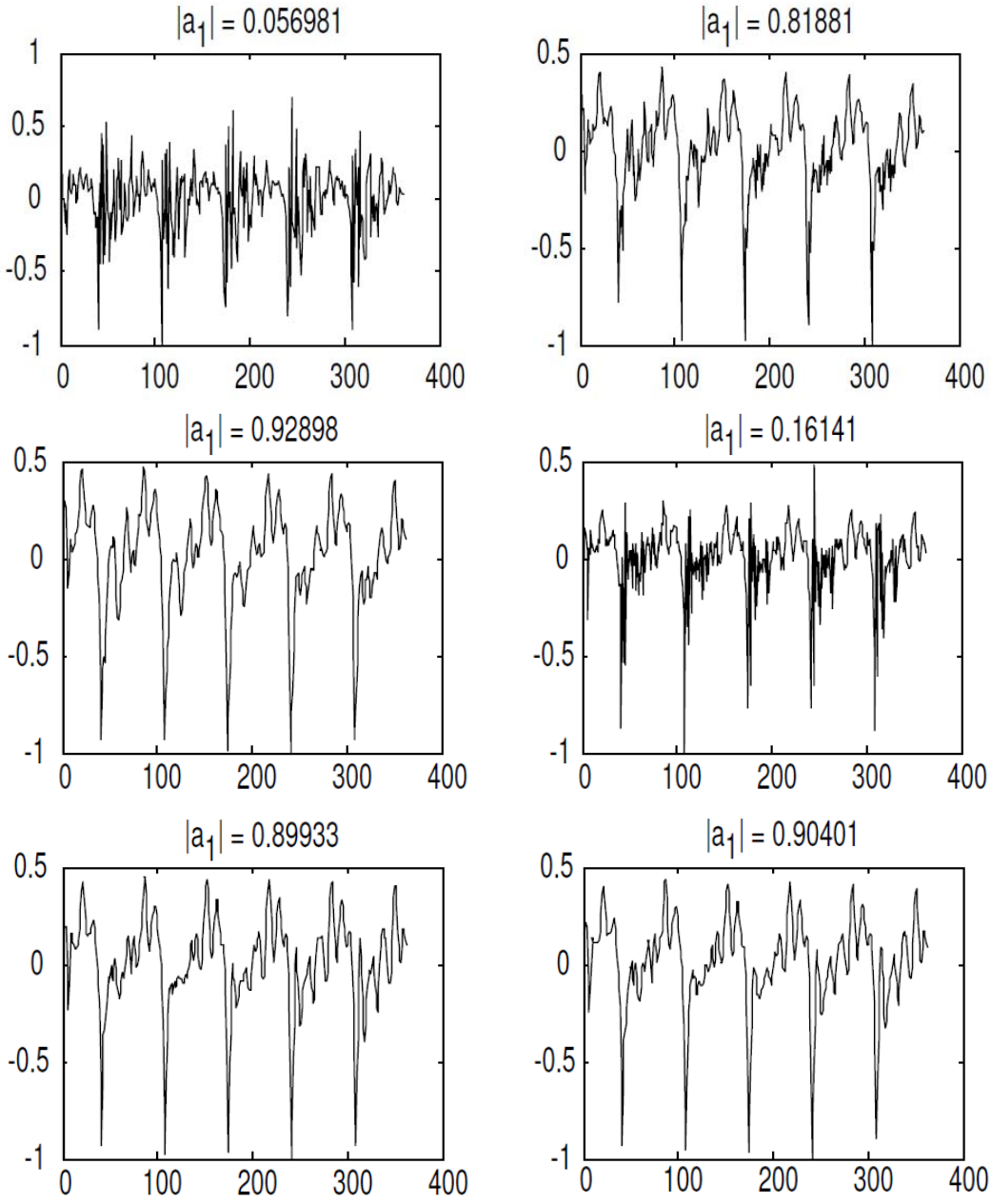

Figure 3.3: Glottal component estimates and the a_1 values by iteration (from [Moore and Clements, 2004])

crete all-pole (DAP) analysis to better estimate the vocal tract model coefficients instead of using the conventional linear prediction method. A schematic diagram of the IAIF algorithm is given in Fig. 3.4. Firstly, as shown in block 1 the input speech signal is high-pass filtered by a FIR filter to remove low frequency component under 50 Hz. The first estimation of the glottal flow is implemented from blocks 2 to 6. In block 2, a first-order DAP analysis is applied to model the effect of a combination of the glottal flow and lip radiation on the speech spectrum, after which the speech signal is inverse filtered to remove such effects in block 3. In block 4, an order p (double the number of formants so generally we choose $p = F_s/1000 + 2$) DAP analysis is applied to extract the first estimation of the vocal tract impulse response. The results are used in block 5 to inverse filter the speech signal to obtain a gross estimation of the glottal flow derivative (GFD). The GFD is integrated and high-pass filtered with a cut-off frequency 50 Hz to remove low frequency drift to obtain the first estimate of the glottal flow waveform in block 6, which will be used for further analysis. Blocks 7 to 12 make up the second phase of IAIF. In block 7, the gross estimated glottal flow signal is analysed by a g (2 or 4) order DAP analysis to obtain a new estimation of the glottal source contribution to the speech spectrum. For the remaining blocks 8 to 12, the process is the same as the first phase (generally the order r is set equal to p, but can be adjusted manually to improve the performance) and finally a refined estimation of the glottal flow waveform is obtained.

An example of applying IAIF to a male speech segment is shown in Fig. 3.5. The top plot presents the spectra of the speech signal, the final glottal flow estimate, the estimated glottis filter and the vocal tract filter, the bottom plot shows the glottal and vocal tract poles in Z-plane.

The limitation of IAIF lies with speech signals which have a low frequency formant, where the glottal formant (the peak of the glottal spectrum) overlaps the first formant of the vocal tract. In such situations, it is difficult to remove the glottal effect from the speech signal spectrum and the estimated vocal tract filter coefficients are not reliable.

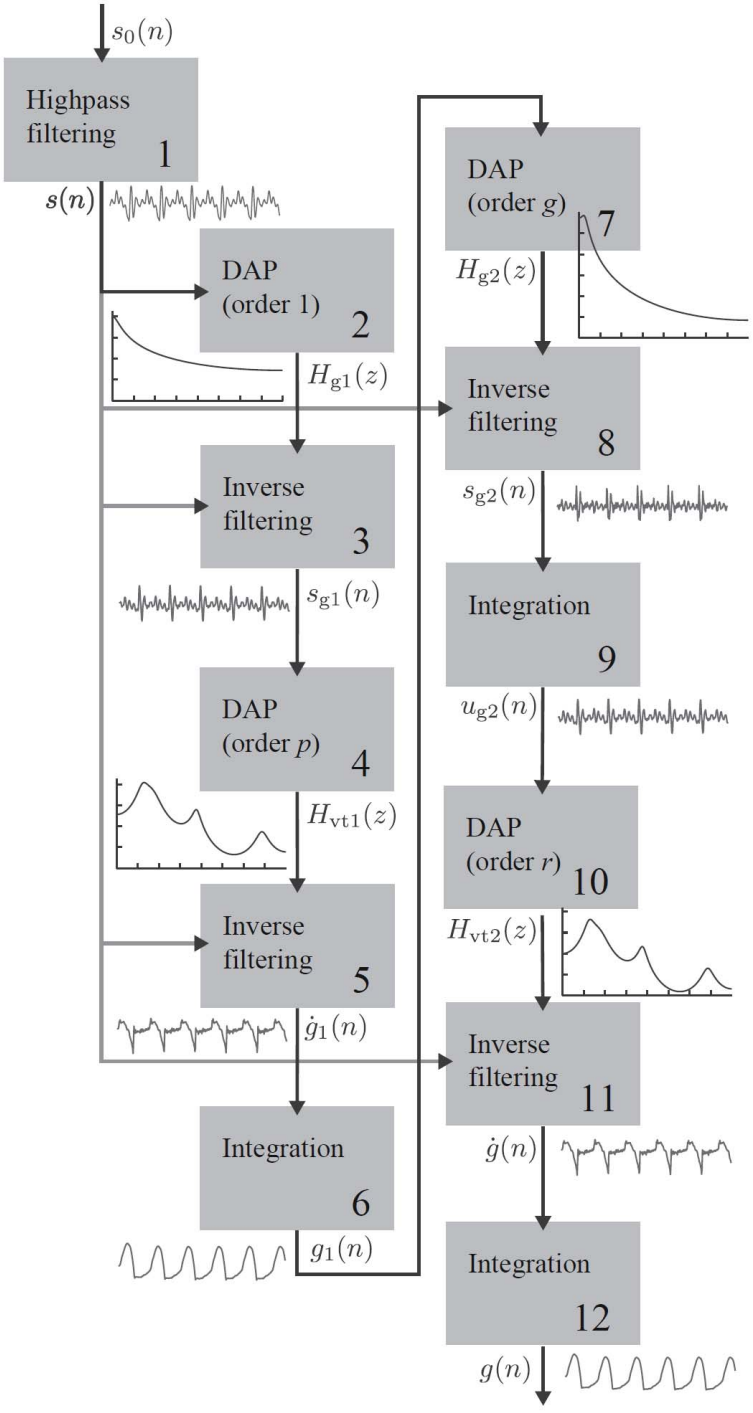

Figure 3.4: Block diagram of the Iterative Adaptive Inverse Filtering (IAIF) algorithm. (from [Airas, 2008])

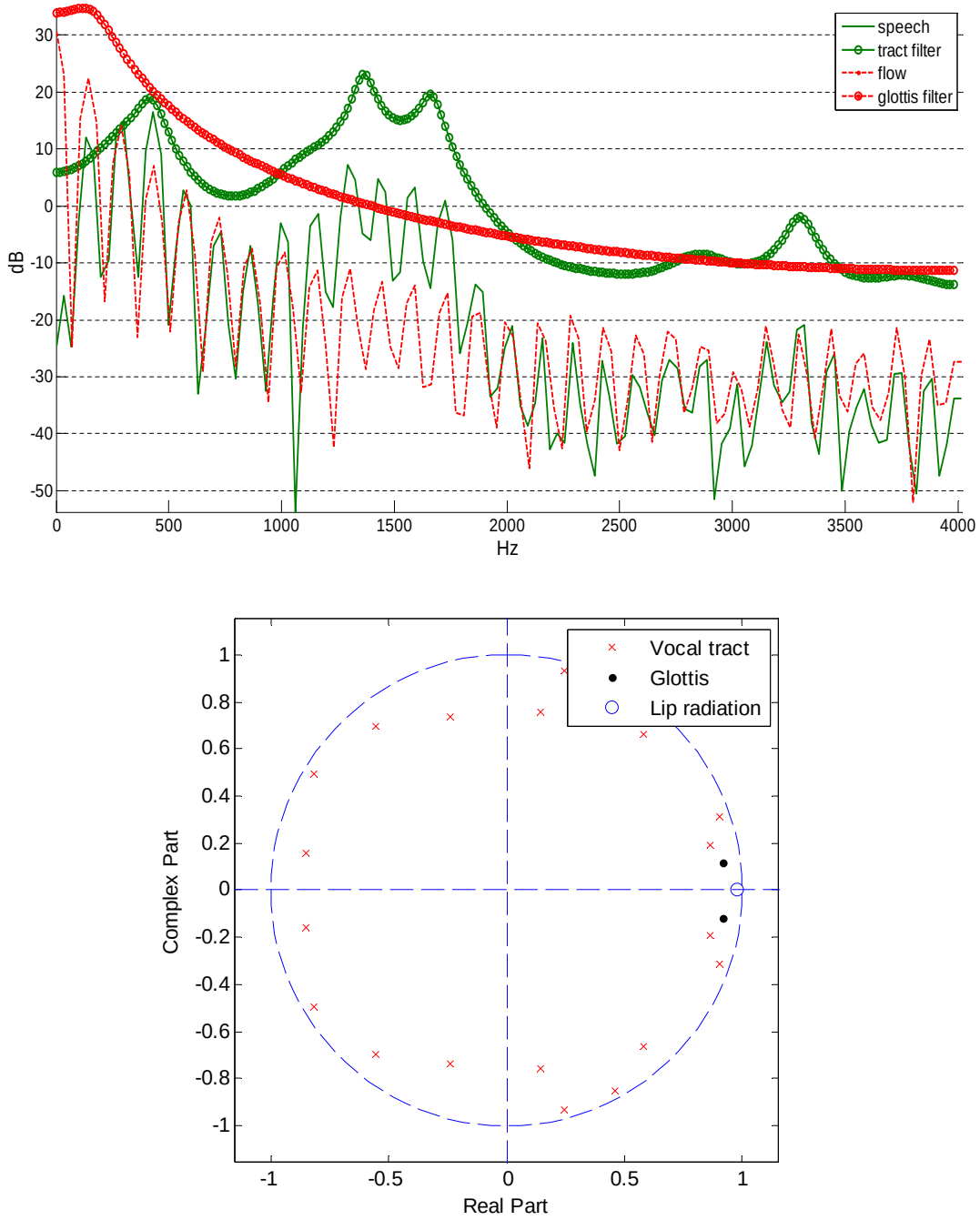

Figure 3.5: An example of applying IAIF to a male speech segment. Top: spectra plot and bottom: poles in Z-plane

3.2.3 Weighted Recursive Least Squares with Variable Forgetting Factor Analysis (WRLS-VFF)

Proposed and developed by Ting and Childers [Ting and Childers, 1990; Childers et al., 1995], the "Weighted Recursive Least Squares with a Variable Forgetting Factor" (WRLS-VFF) algorithm can be used to extract many features from speech signals. With the variable forgetting factor, which indicates the reliability of previous data for future prediction, WRLS-VFF can: 1) estimate the input excitation (white noise or periodic pulse train), 2) track vocal tract formants, 3) perform voiced/unvoiced speech segmentation, 4) detect glottal closure instant and 5) performn glottal inverse filtering [Lee and Park, 1999].

WRLS-VFF assumes that the speech signal is generated by an ARMA model given in equation (3.4):

$$y_k = -\sum_{i=1}^{p} a_i(k) y_{k-i} + \sum_{j=1}^{q} b_j(k) u_{k-j} + u_k \tag{3.4}$$

where y_k is the k^{th} speech sample, u_k is the input excitation, p and q are the number of the poles and zeros respectively of the ARMA model, a_i and b_j are the time-varying AR and MA parameters. It can be observed that to estimate the AR and MA coefficients, it is necessary to estimate the input excitation u_k.

Three vectors are defined as follows,

$$\begin{aligned} \theta_k^t &= [a_1(k), ..., a_p(k), b_1(k), ..., b_q(k)] \\ \hat{\theta}_k^t &= [\hat{a}_1(k), ..., \hat{a}_p(k), \hat{b}_1(k), ..., \hat{b}_q(k)] \\ \phi_k^t &= [-y_{k-1}, ..., -y_{k-p}, \hat{u}_{k-1}, ..., \hat{u}_{k-q}] \end{aligned} \tag{3.5}$$

where θ_k is the parameter vector, $\hat{\theta}_k$ is its estimate and ϕ_k is a data vector. Then the speech signal and its estimate can be calculated from

$$\begin{aligned} y_k &= \phi_k^t \theta_k + u_k \\ \hat{y}_k &= \phi_k^t \hat{\theta}_k + \hat{u}_k \end{aligned} \tag{3.6}$$

Accordingly, the error residual of the ARMA process is expressed as

$$r_k = y_k - \phi_k^t \hat{\theta}_k \tag{3.7}$$

Then a weighted least square error criterion or a cost function can be defined as the weighted sum of the residual error squares given in equation (3.8)

$$V_k(\theta) = \sum_{i=1}^{k} \lambda^{k-i} r_i^2 = \sum_{i=1}^{k} \lambda^{k-i} (y_i - \phi_i^t \hat{\theta}_i)^2 \tag{3.8}$$

where λ is the forgetting factor.

For an ARMA speech production process, the residual error r_k can be used to indicate the state of the estimator at each step k. If r_k is small, the forgetting factor λ should be close to unity and the current estimate is obtained by using most of the previous information in the speech signal. Accordingly, the estimated model parameters are sufficiently reliable. On the other hand, if r_k is large, a smaller λ value will decrease the weighting of the error and shorten the effective memory length of the estimation procedure. This results in estimating of the model parameters with the most recent data, and reducing the error. To obtain the appropriate weighting the algorithm should be able to choose the appropriate forgetting factor λ. one iteration of the WRLS-VFF algorithm is presented in Table 3.1 [Ting and Childers, 1990], and the results of applying the algorithm to a male vowel sound segment /aa/ are shown in Fig. 3.6.

$\hat{u}_k^P$ is the pulse input (for voiced sound) and its magnitude is estimated as the prediction error, and $\hat{u}_k^N$ is the white noise input (for plosive sound) determined by the residual error. It can be observed that when the prediction error is large, the forgetting factor decreases. A small value of λ_k implies an abrupt change of the current data, which occurs at the glottal closure instant (GCI). Accordingly, by using a suitable threshold value λ_0, the GCI positions can be identified and subsequently the glottal inverse filtering can be performed. A block diagram of the WRLS-VFF based glottal inverse filtering algorithm is presented in Fig. 3.7. The speech signal s_n is firstly pre-emphasised and then analysed by the WRLS-VFF method, where the variable forgetting factor is obtained sequentially (sample by sample). The smallest value of λ_k within a pitch period indicates the instant of

Table 3.1: Equations of the WRLS-VFF algorithm

Prediction error:	$\xi_k = y_k - \phi_k^t \hat{\theta}_{k-1}$
Gain update:	$K_k = P_{k-1}\phi_k[\lambda_{k-1} + \phi_k^t P_{k-1}\phi_k]^{-1}$
Forgetting Factor:	$\lambda_k = 1 - \xi_k^2(1 - \phi_k^t K_k)/V_1(\theta)$
Input Estimate:	a) Pulse train
	If $\lambda_k < \lambda_0$ then
	$\hat{u}_k^W = 0$
	$\hat{u}_k = \hat{u}_k^P = \xi_k = y_k - \phi_k^t \hat{\theta}_{k-1}$
	b) White noise
	If $\lambda_k > \lambda_0$ then
	$\hat{u}_k^P = 0$
	$\hat{u}_k = \hat{u}_k^W = r_k = y_k - \phi_k^t \hat{\theta}_k = \xi_k(1 - \phi_k^t K_k)$
Parameter update:	$\hat{\theta}_k = \hat{\theta}_{k-1} + K_k(y_k - \phi_k^t \hat{\theta}_{k-1} - \hat{u}_k^P)$
Covariance update:	$P_k = \lambda_k^{-1}[P_{k-1} + K_k \phi_k^t P_{k-1}]$

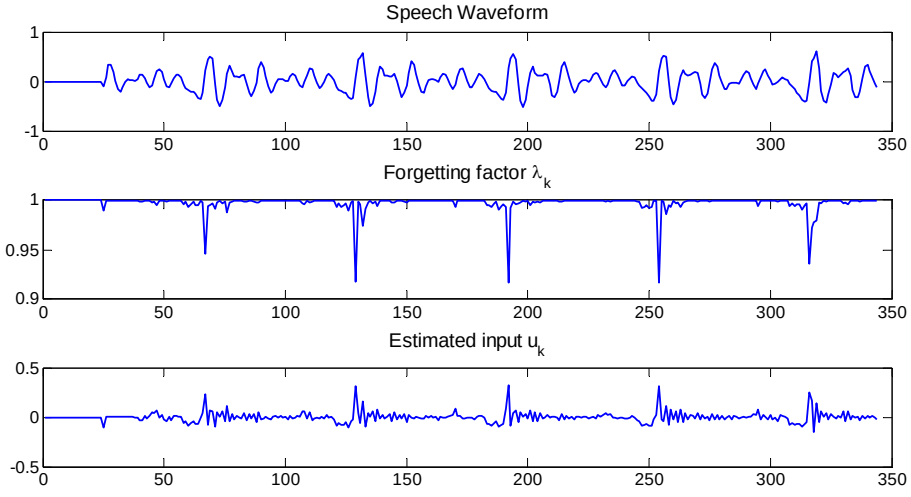

Figure 3.6: Applying WRLS-VFF analysis to a male speech segment /aa/

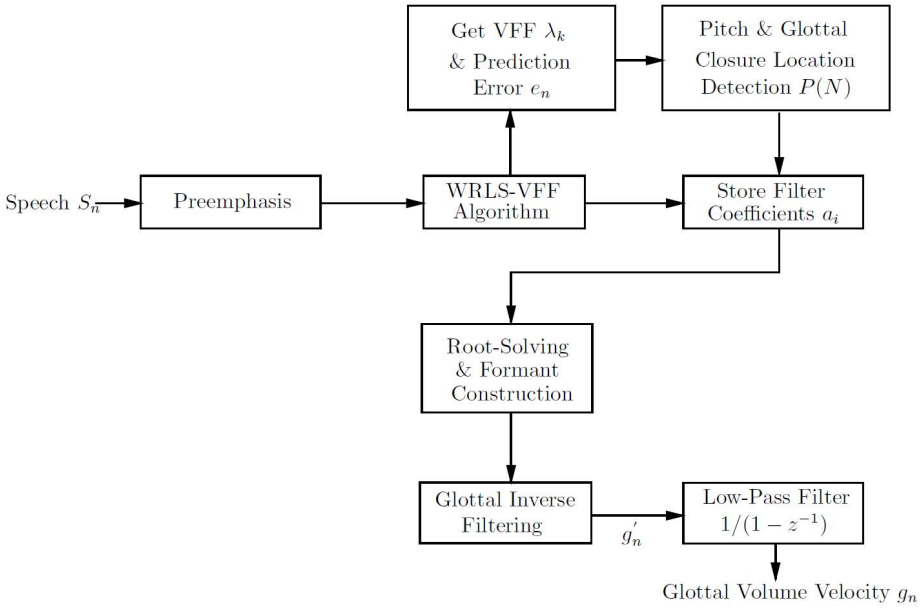

Figure 3.7: Glottal inverse filtering by WRLS-VFF algorithm (from [Ting and Childers, 1990])

the main excitation, which is actually the glottal closure point. All the estimated vocal tract filtering coefficients a_i during the adaptive process are stored and the set of coefficients for glottal inverse filtering is determined by the convergence of the adaptive process, where the minimal data length for convergence is equal to twice the filter order. In addition a root-solving and formant construction procedure is applied to check stability and ensure the poles outside the unit circle are reflected back into the unit circle. The refined filter coefficients are then used to inverse filtering the speech signal to achieve the glottal flow waveform.

WRLS-VFF and CPIF have similarities. The major difference of WRLS-VFF to standard CPIF is that CPIF calculates the vocal tract filter coefficients by applying the LP covariance method to the closed phase interval, while WRLS-VFF estimates multiple coefficients during the recursive least square procedure and chooses one set from the closed phase. According to Ting and Childers, there are two main limitations for the WRLS-VFF algorithm: 1) its computational complexity and 2) the lack of a priori information guiding the model type and model order selection for tracking the parameters. The first detracting factor can

be improved with more computational resources and the second disadvantage can be solved by ARMA model order selection techniques [Bhansali, 1993; Liang et al., 1993; Haseyama and Kitajima, 2001; Broersen and de Waele, 2004; Stoica and Selen, 2004].

3.3 Performance Study

To evaluate the quality of the estimated glottal waveform, it is necessary to use some numerical measure of quality. In this study, three glottal waveform quality measures (GQM) are used and described as follows:

• Phase-plane measures. [Backstrom et al., 2005] presented two glottal waveform quality measures based on phase plane analysis. As the vocal tract can be modelled by a cascade of second order resonators [Rabiner and Schafer, 1978], while the glottal waveform can be considered by a second order harmonic equation [Edwards and Angus, 1996]:

$$\frac{d^2x}{dt^2} + x = 0. \tag{3.9}$$

To analyse this system in the phase-plane (x=glottal flow,y=differentiated glottal flow), equation (3.10) can be obtained:

$$\frac{dx}{dt} = y \quad \text{and} \quad \frac{dy}{dt} = -x \tag{3.10}$$

Integrating the first order differential equation $dy/dx = -x/y$, we have $x^2 + y^2 = C$, where C is a constant. Accordingly, a glottal waveform which is resonance-free should be cyclic in the phase-plane corresponding to the pitch period. Vocal tract resonances yield different periodic solutions to form a subset of solutions. Therefore, formants which are not completely removed by glottal inverse filtering will appear as minor loops within the fundamental loop in the phase-plane. This results in two measures to quantify the quality of inverse filtering based on the phase-plane plot: the number of cycles per pitch period (pp_{cper}), and the the mean sub-cycle area (pp_{cyc}) which directly corresponds to the magnitude of formant ripple. The smaller the pp_{cper} and pp_{cyc}, the better the glottal estimates, and if pp_{cyc} is null it means that sub-cycles can hardly be detected. Examples to

illustrate the phase-plane measures are shown in Fig. 3.8 and Fig. 3.9. It can be observed for higher quality glottal estimates, where the formants are cancelled, the phase-plane plot is a closed loop and the corresponding pp_{cper} and pp_{cyc} are quite small. However for poor glottal estimates, where the formants are not fully removed, the phase-plane plot involves sub-cycles and the calculated pp_{cper} and pp_{cyc} values are higher.

• Kurtosis. Also proposed by [Backstrom et al., 2005], kurtosis can be used to measure the success of the deconvolution process of glottal inverse filtering. In probability theory and statistics, kurtosis is a measure of the 'peakedness' of the probability distribution of a real-valued random variable [Dodge et al., 2006]. Thus it describes the similarity of a distribution to the Gaussian distribution. For a discrete signal sequence, kurtosis is calculated by equation (3.11):

$$kurtosis = \frac{m_4}{m_2} = \frac{1/n \sum_{i=1}^{n} (x_i - \bar{x})^4}{(1/n \sum_{i=1}^{n} (x_i - \bar{x})^2)^2} \tag{3.11}$$

where m_4 is fourth sample moment about the mean and m_2 is the second sample moment about the mean (variance). For a normal Gaussian distribution, the kurtosis is 3. To make the kurtosis of the normal distribution equal to zero, the excess kurtosis is defined as $kurtosis - 3$, which has a range of [-3, $+\infty$], where positive and negative values correspond to supergaussian (sharper peak distribution) and subgaussian distributions (flatter peak distribution), respectively.

According to the central limit theorem, summation of equally distributed signals converges to the Gaussian distribution by increasing the number of samples. Convolution by the vocal tract transfer function involves summing copies of the glottal waveform at different time delays, thus the distribution of the convolution output should be closer to a normal Gaussian distribution than the glottal waveform. In addition, Backstrom observed that the idealised glottal flow, with a flatter distribution, is naturally subgaussian, and therefore should have a kurtosis value less than 3. Thus, the lower the value, the more accurate the glottal waveform estimate. Fig. 3.10 shows the histograms of a segment of voiced speech and the corresponding glottal estimate. It is visible that the speech signal has a distribution closer to normal distribution and thus a larger kurtosis value. Also, it can be observed in Fig. 3.8 and Fig. 3.9 that higher quality glottal estimate

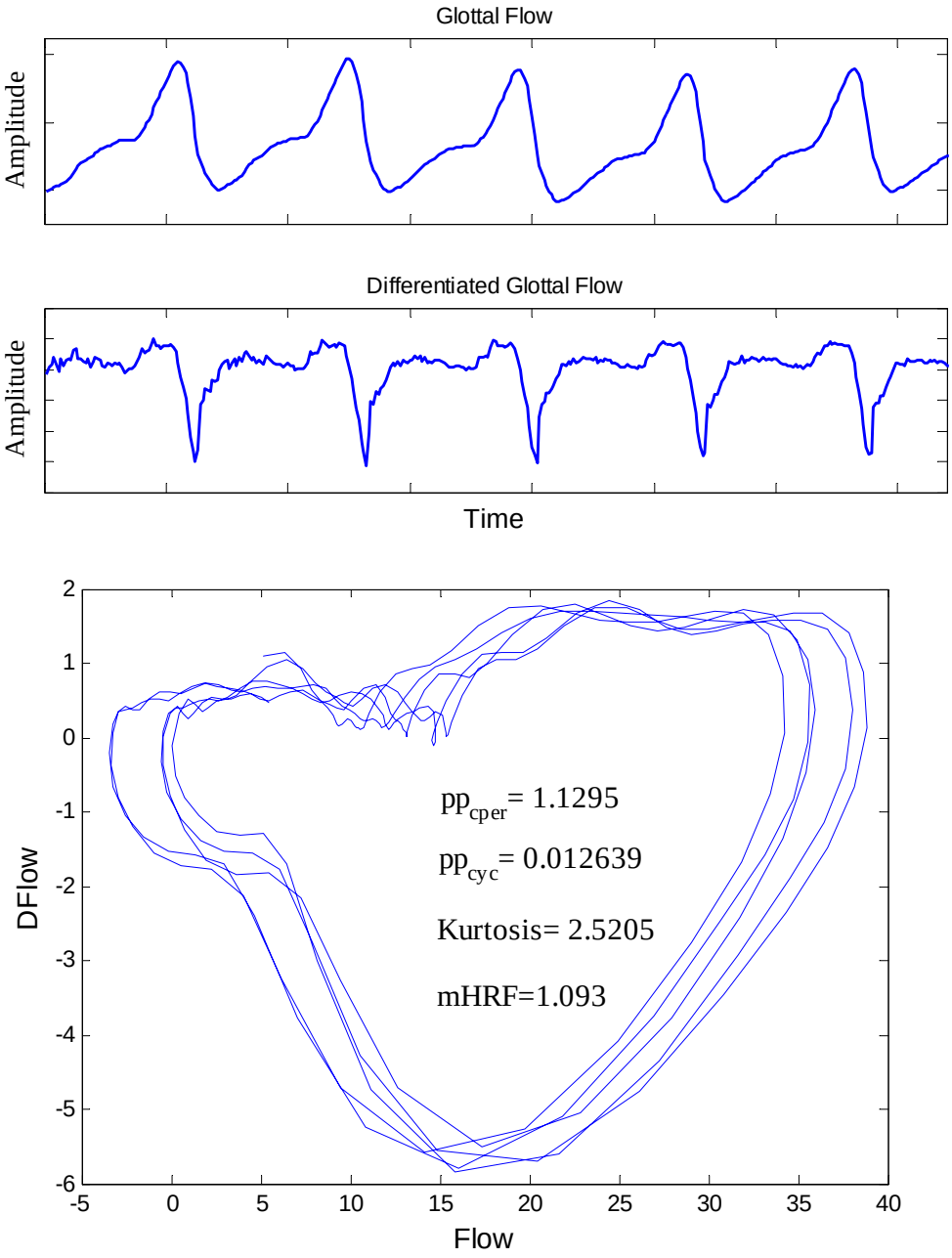

Figure 3.8: High quality glottal flow and differentiated glottal flow waveform estimate (above), Phase-plane plots and the four GQMs (below)

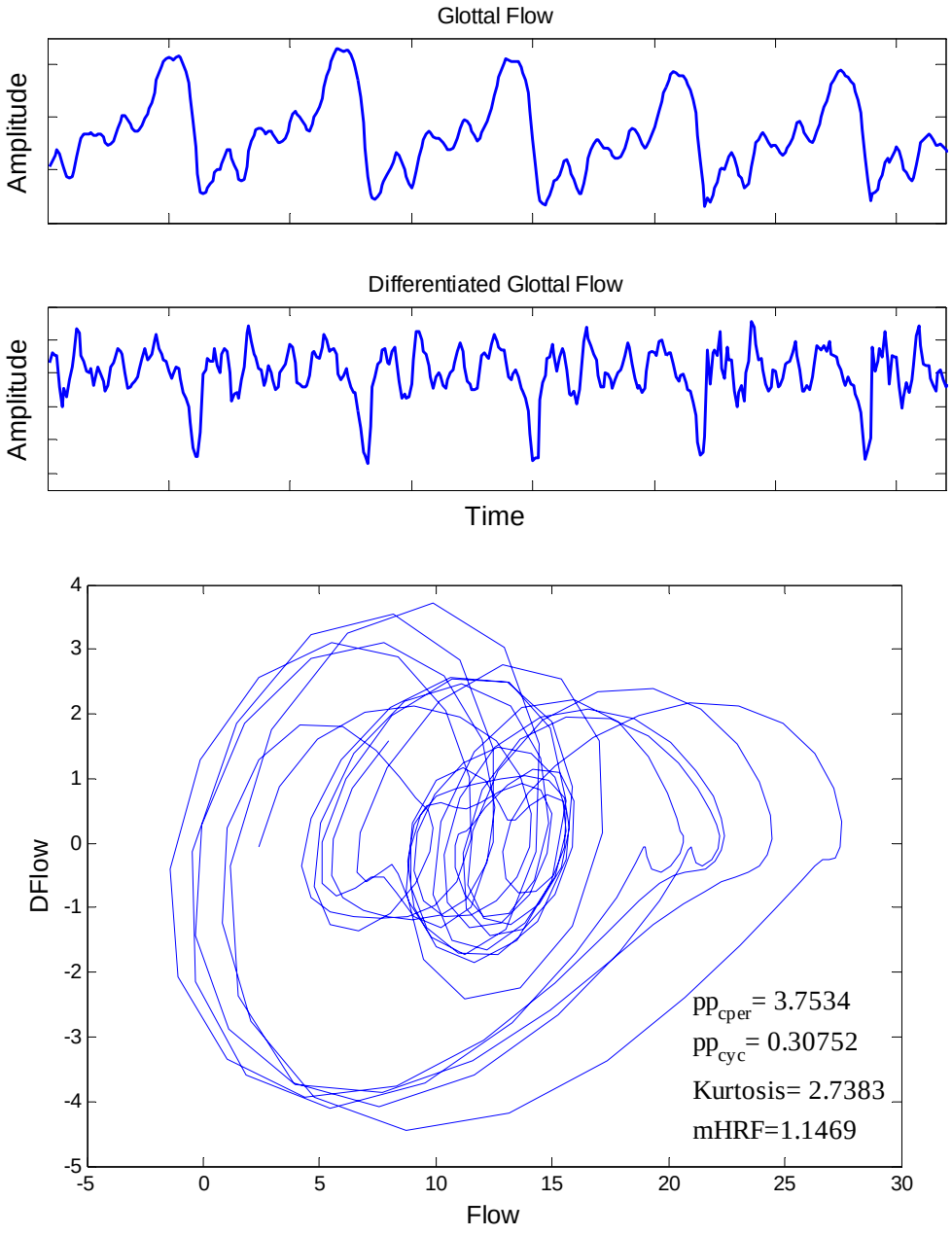

Figure 3.9: Poor quality glottal flow and differentiated glottal flow waveform estimate (above), Phase-plane plots and the four GQMs (below)

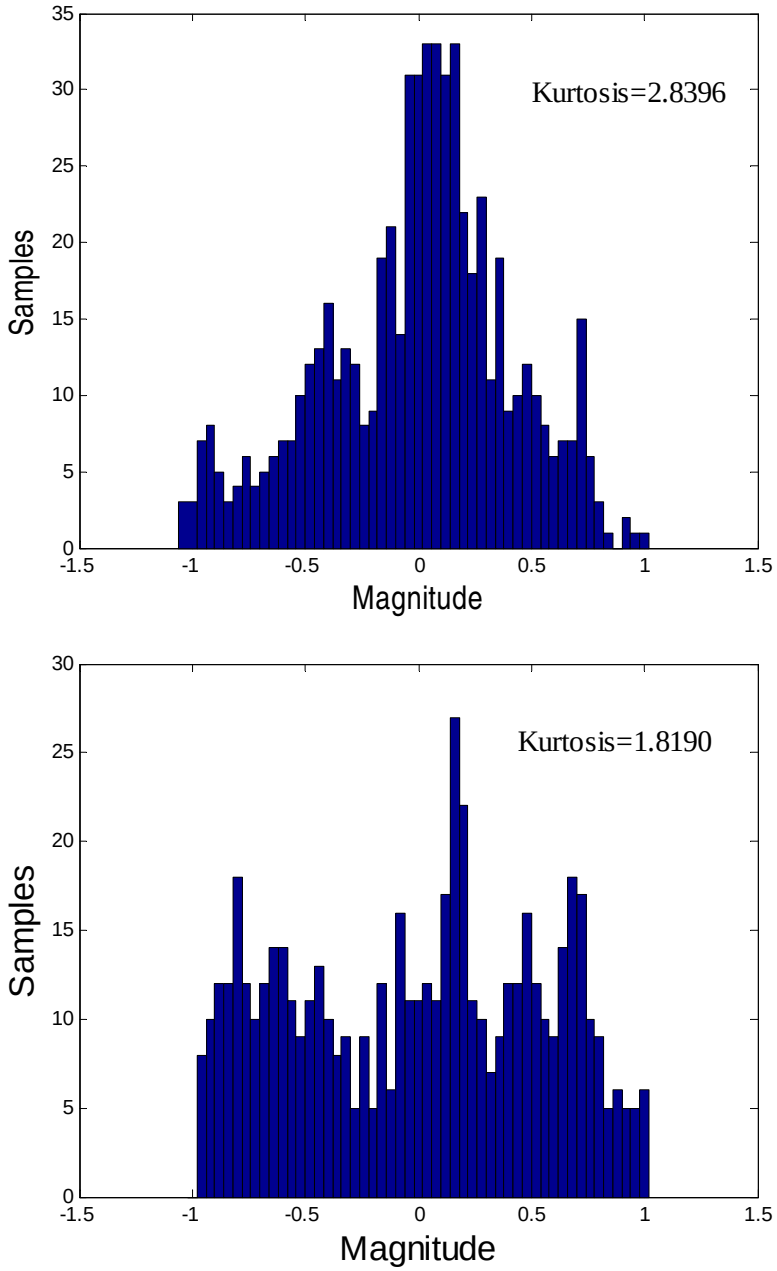

Figure 3.10: Histograms of a segment of voiced speech (top) and the corresponding glottal estimate (bottom)

has a smaller kurtosis value compared to the poor glottal estimate.

• Mean harmonic richness factor. The harmonic richness factor (HRF) is an objective frequency domain measure to characterise the inverse filtered glottal flow waveform [Childers and Lee, 1991; O'Cinneide et al., 2011b]. HRF is calculated as the ratio between the sum of the harmonic amplitudes above the fundamental frequency and the amplitude of the fundamental frequency. A large HRF value indicates the presence of energy in the higher harmonics of the glottal spectrum, and often corresponds to impulse-like signals. On the other hand, low HRF values imply that there is more energy around the fundamental frequency of the spectrum and the signal should be more like a sinusoidal wave. Since the idealised glottal waveform open phase is a sinusoidal function (the return phase affects just the spectral tilt), the corresponding HRF value should be low, as can be observed from Fig. 3.8 for a well estimated glottal waveform. For the poor glottal estimate in Fig. 3.9, the HRF is relatively higher.

To evaluate the performance of the GIF algorithms, six sustained vowel sounds were extracted from the CMU-Arctic database [Kominek and Black, 2004]. Examples of the vowels /æ/, /ə/and /ɔ/[1] were taken from recordings of the male speaker *bdl* and of the female speaker *slt*. The six vowel segments were applied to each of the GIF algorithms to extract the corresponding glottal flow derivative waveforms. The sampling frequency was 10kHz and a 256-point hamming window is applied for analysis. The results are presented in Figs. 3.11 - 3.16 and Tables. 3.2 - 3.7.

It can be observed from the waveforms that in most cases, reasonable glottal flow derivatives were successfully extracted by the three GIF methods. However, across different speech segments, performance varies. For the vowel sound /æ/ for both male and female speakers, it is visible that the ICPIF and WRLS-VFF outperform the IAIF method, by generating smoother glottal waveforms and relatively small GQM scores. Especially for the male /æ/, the glottal estimates by IAIF contain several noisy pitch periods and formant ripples that result in larger pp_{cper} and pp_{cyc} values. For male /ə/, ICPIF and WRLS-VFF generate similar results and overall the performances are similar, although IAIF has noisier estimates. For the female segment /ə/, noisy estimates and formant ripples

[1]phonetic transcriptions from the International Phonetic Alphabet (IPA)

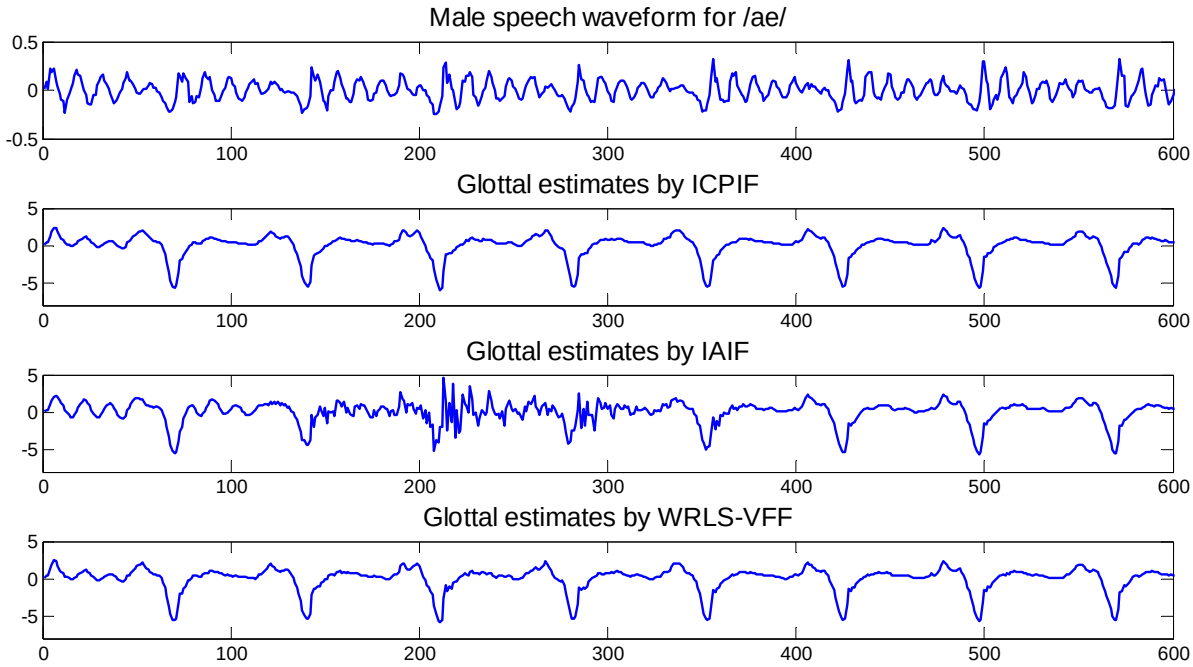

Figure 3.11: Waveform of male speech frame /æ/ and corresponding glottal estimates by the three inverse filtering algorithms.

Table 3.2: Glottal waveform quality measures by the three GIF methods for male speech frame /æ/

GIF	pp_{cper}	pp_{cyc}	Kurtosis	mHRF
ICPIF	1.4706	0.0228	2.2614	1.0976
IAIF	2.9408	0.1021	2.4507	1.1211
WRLS-VFF	1.5313	0.0163	2.2648	1.1206

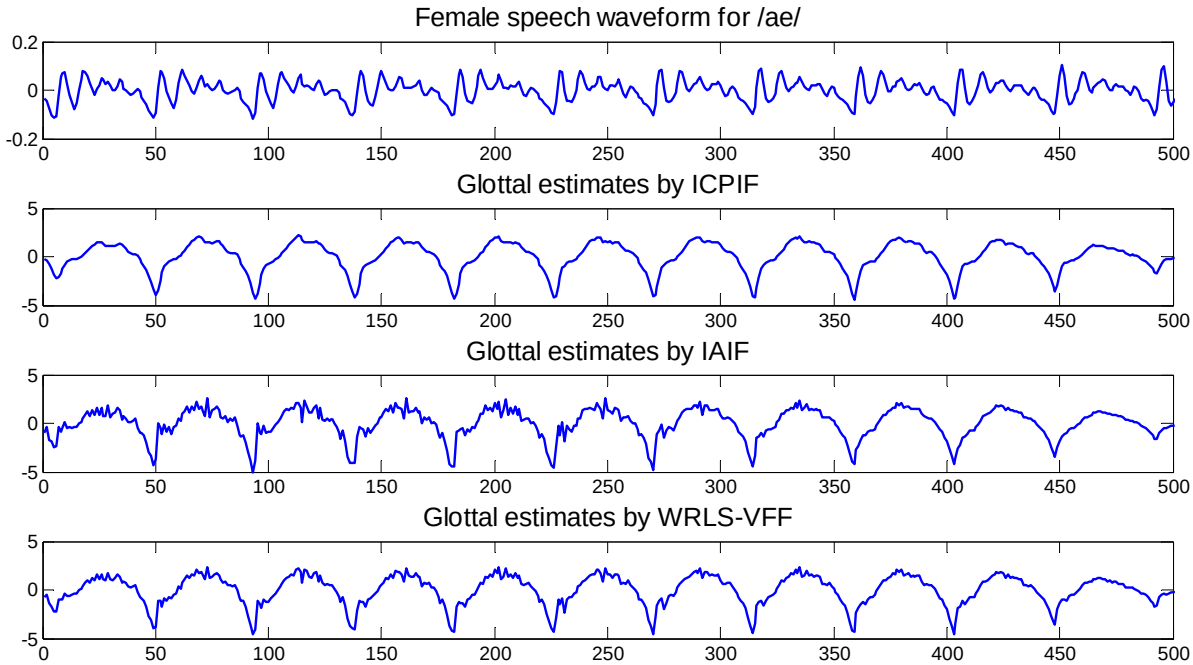

Figure 3.12: Waveform of female speech frame /æ/ and corresponding glottal estimates by the three inverse filtering algorithms.

Table 3.3: Glottal waveform quality measures by the three GIF methods for female speech frame /æ/

GIF	pp_{cper}	pp_{cyc}	Kurtosis	mHRF
ICPIF	0.9640	null	1.6498	0.3981
IAIF	0.9726	null	1.6746	0.8490
WRLS-VFF	0.9671	null	1.6851	0.8411

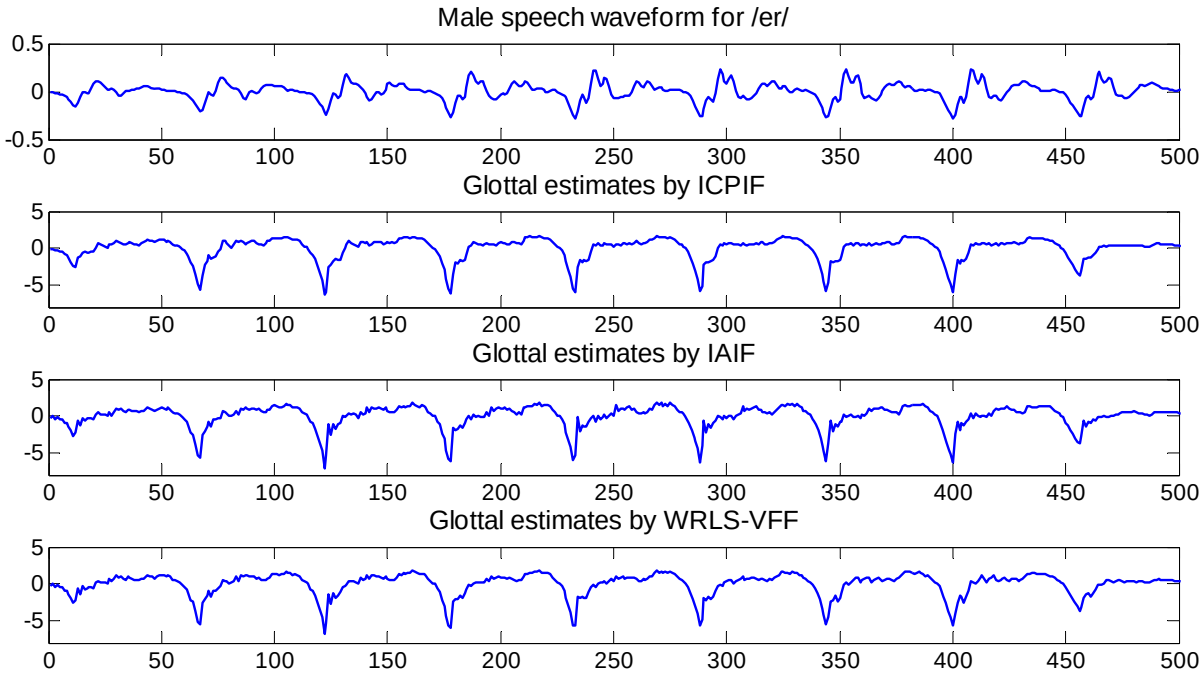

Figure 3.13: Waveform of male speech frame /ə/ and corresponding glottal estimates by the three inverse filtering algorithms.

Table 3.4: Glottal waveform quality measures by the three GIF methods for male speech frame /ə/

GIF	pp_{cper}	pp_{cyc}	Kurtosis	mHRF
ICPIF	0.9378	null	1.8627	0.9981
IAIF	1.1025	0.0147	1.8841	1.0334
WRLS-VFF	0.9365	null	1.8806	0.9966

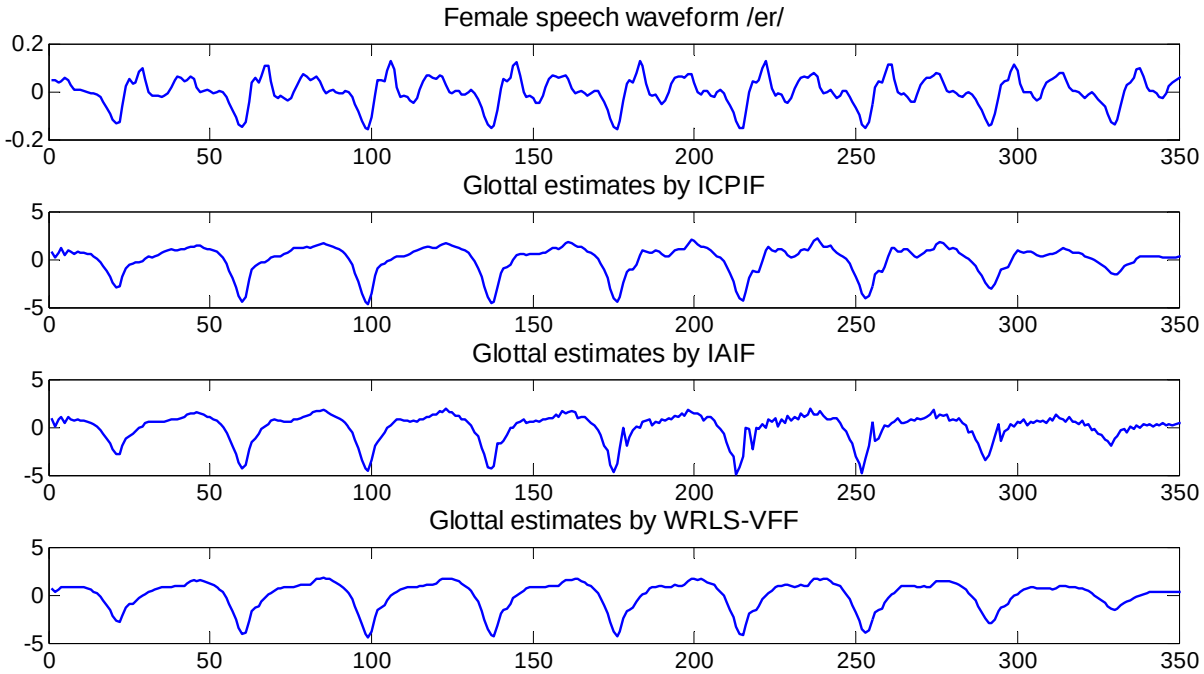

Figure 3.14: Waveform of female speech frame /ə/ and corresponding glottal estimates by the three inverse filtering algorithms.

Table 3.5: Glottal waveform quality measures by the three GIF methods for female speech frame /ə/

GIF	pp_{cper}	pp_{cyc}	Kurtosis	mHRF
ICPIF	0.9900	null	1.8476	0.8603
IAIF	0.9814	null	1.8307	1.1929
WRLS-VFF	0.9811	null	1.8625	0.8126

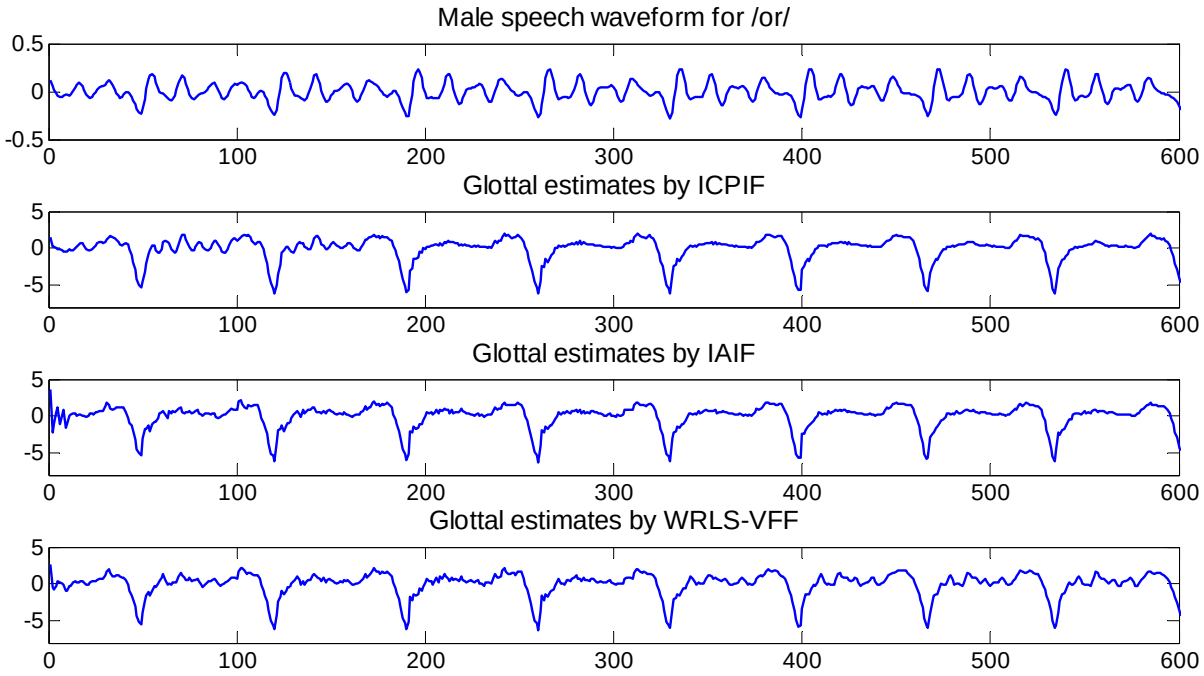

Figure 3.15: Waveform of male speech frame /ɔ/ and corresponding glottal estimates by the three inverse filtering algorithms.

Table 3.6: Glottal waveform quality measures by the three GIF methods for male speech frame /ɔ/

GIF	pp_{cper}	pp_{cyc}	Kurtosis	mHRF
ICPIF	2.0239	0.0393	2.2368	0.9195
IAIF	1.1674	0.0103	2.2582	0.9989
WRLS-VFF	2.6248	0.0192	2.2620	0.9167

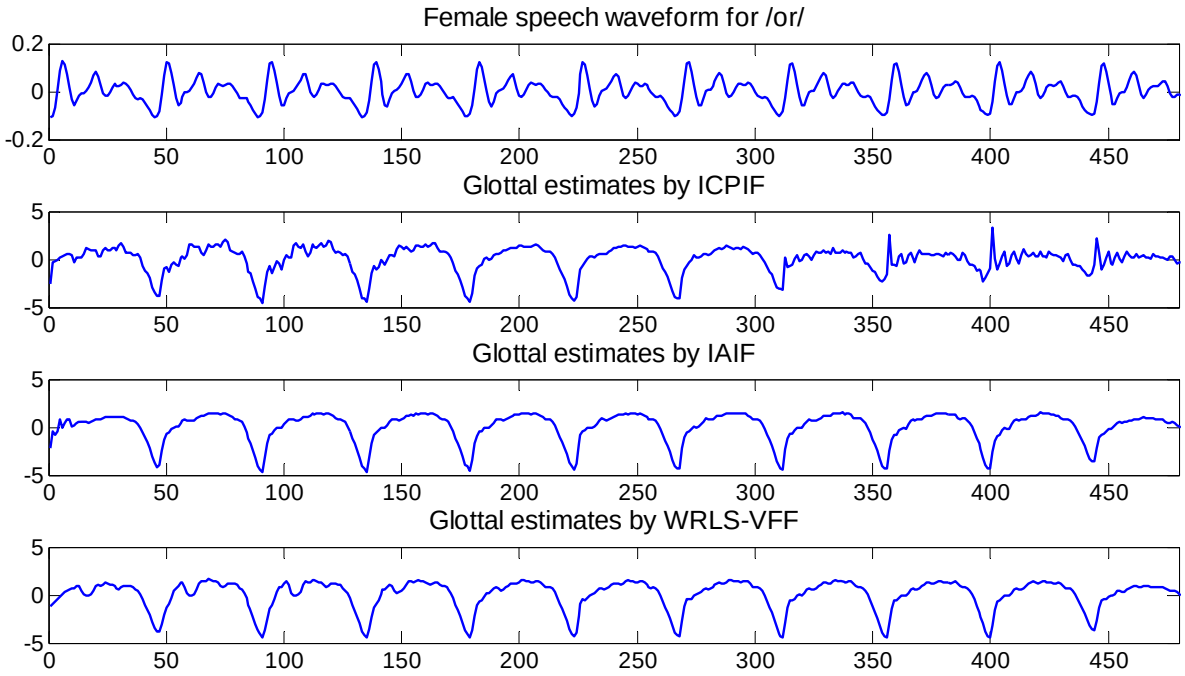

Figure 3.16: Waveform of female speech frame /ɔ/ and corresponding glottal estimates by the three inverse filtering algorithms.

Table 3.7: Glottal waveform quality measures by the three GIF methods for female speech frame /ɔ/

GIF	pp_{cper}	pp_{cyc}	Kurtosis	mHRF
ICPIF	1.1581	0.0628	2.0215	0.9150
IAIF	0.9281	null	1.5753	0.8393
WRLS-VFF	1.0592	0.0261	1.5924	0.7507

can be observed from the last several pitch periods of the IAIF and the ICPIF estimates, respectively. Consequently, ICPIF has a larger pp_{cper} score and IAIF generates a bigger mHRF value, where as WRLS-VFF results in reasonably small scores for all the three GQM and overall it outperforms the other two approaches. Finally, for both male and female /ɔ/, the IAIF method is superior, since it generates more consistent glottal estimates for most pitch periods, while noisy components and formant ripples can be observed from the estimates by the other two approaches. In addition, it is visible that the values of pp_{cper} and pp_{cyc} for IAIF are much smaller than ICPIF and WRLS-VFF's scores. The IAIF's mHRF score is slightly higher, which may be caused by several highly noisy samples appearing early in the estimated glottal waveform. The kurtosis score by IAIF for the male speaker is slightly higher than that of ICPIF, and for the female speaker is the smaller than the other two approaches.

From the experimental results and analysis above, we can conclude that no single algorithm performs best for all kinds of speech signal. Thus it is reasonable to suggest that improved estimate may be obtained by combining the algorithms in an informed fashion. In Chapter 5, we will present a general framework for multi-estimate fusion and study the performance of the framework when combining the glottal estimates from the above three glottal inverse filtering methods.

3.4 Other Speech Decomposition Methods

Despite the effectiveness of the glottal inverse filtering approach, many other methods have been proposed to decompose speech into its source and vocal tract components. Some of them are briefly described in this section. While not incorporated into our fusion framework, their addition is later suggested as a possible extension to the framework.

3.4.1 Mixed-phase Speech Decomposition

The mixed-phase speech decomposition method separates speech into its source and vocal tract components relying on the mixed-phase speech model [Bozkurt and Dutoit, 2003]. This model assumes that speech consists of both minimum-

phase (causal) and maximum-phase (anticausal) components. In [Doval et al., 2003], it has been shown that the voice source can be considered as a causal/anti-causal linear filter, where the glottal open phase is a maximum-phase component and the glottal return phase is a minimum-phase component. The vocal tract, which is generally modelled by an all-pole linear filter, is also minimum-phase. Accordingly, the key idea of the mixed-phased decomposition algorithm is to separate both maximum- and minimum-phase components from speech. There are primarily two mixed-phase decomposition methods: Zeros of Z-transform (ZZT) decomposition [Bozkurt et al., 2004b,a; Bozkurt, 2005] and Complex Cepstrum-based Decomposition (CCD) [Drugman et al., 2009b].

• **Zeros of Z-transform decomposition**. Unlike linear prediction that tries to predict future values, the ZZT [Bozkurt et al., 2004b,a; Bozkurt, 2005] decomposes the speech signal along two groups of zeroes which are the roots of the signal's Z-Transform. Using the distribution of zeros in the z-plane, the glottal flow contribution (zeros outside the unit circle) can be separated from vocal tract contributions (zeros inside the unit circle). For a discrete time sequence $x[n]$, the ZZT representation is defined as the set of roots (zeros) of its corresponding Z-Transform polynomial $X(z)$, which is shown in equation (3.12), where N is the length of the signal, and Z_m is the m^{th} root.

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} = x(0) z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \tag{3.12}$$

Accordingly, if $s(n)$ is the speech signal, the corresponding ZZT representation is given in (3.13)

$$S(z) = s(0) z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) = s(0) z^{-N+1} \prod_{g=1}^{K-1} (z - Z_g) \prod_{v=1}^{J-1} (z - Z_v) \tag{3.13}$$

where Z_g are the zeros of the open phase of the voice source and Z_v are the zeros for the vocal tract and the voice source return phase. Therefore, if we calculate the polynomial from Z_g, the obtained coefficients of the polynomial are the time domain signal samples contributed by the glottal open-phase component. An example showing the glottal waveform estimated by ZZT from a real speech

frame is given in Fig. 3.17, where there is no return phase information.

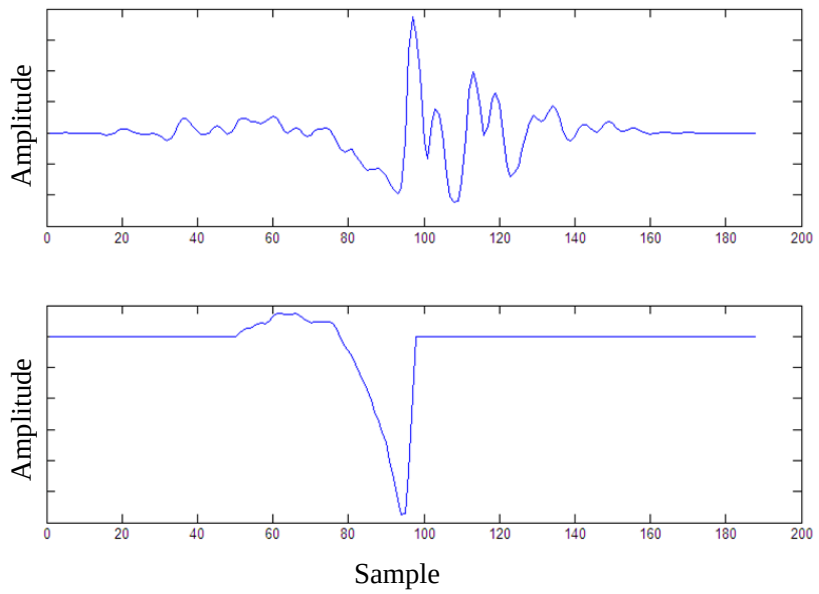

Figure 3.17: Waveforms of a voiced real speech frame and the corresponding glottal estimate by ZZT

• **Complex Cepstrum-based Decomposition**. Based on the same principles as ZZT, CCD [Drugman et al., 2009b] has been shown to be much faster to compute than ZZT. The complex cepstrum (CC) $\hat{s}(n)$ of a discrete speech signal $s(n)$ is calculated by the following equations:

$$S(\omega) = \sum_{n=-\infty}^{\infty} s(n)e^{-j\omega n} \tag{3.14}$$

$$\log[S(\omega)] = \log(|S(\omega)|) + j\angle S(\omega) \tag{3.15}$$

$$\hat{s}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[S(\omega)]e^{j\omega n}d\omega \tag{3.16}$$

where equations (3.14) - (3.16) are respectively the Discrete-time Fourier Transform (DTFT), the complex logarithm function and the inverse DTFD (IDTFT).

The CCD method works based on the assumption that the complex cepstrum $\hat{s}(n)$ of a causal component is zero for all negative n, and of an anticausal component is zero for all positive n. Thus, to estimate the glottal contribution which is anticausal, only the negative part of the CC should be retained.

The limitations of the mixed-phase speech decomposition approach lie in:

• Choice of windowing function is crucial to the algorithm performance. Bozkurt [Bozkurt et al., 2005] argued that the performance of ZZT decomposition depends highly on the windowing functions and the accuracy of the glottal closure instant (GCI) estimates. He suggested using a two-pitch-period long Blackman window centred on the GCI of the speech signal for better performance. Recently, Drugman proposed using chirp group delay processing to enhance the robustness of mixed-phase decomposition [Drugman et al., 2009a; Drugman and Dutoit, 2010].

• Noisy speech sensitivity. Bozkurt also tested the algorithm's sensitivity on synthetic speech with additive noise and concluded that the ZZT algorithm is sensitive to noisy speech, which distorts the maximum- and miminum-phase distribution of the speech signal [Bozkurt et al., 2005].

• No glottal return phase information. It should be pointed out that the glottal return phase is a minimum-phase system and is estimated together with the vocal tract component by the mixed-phase decomposition. It is difficult to separate the return phase from the vocal tract contribution.

3.4.2 Higher Order Statistics Approaches

Higher order statistics (HOS) refers to functions of the third or higher power of a sample as extensions to second order measures (such as the autocorrelation and power spectrum). For HOS techniques applied to speech analysis, additional properties of the input speech signals can be exploited. Taking the bispectrum (third-order spectrum) for example, it is theoretically immune to white Gaussian noise (in practice noise exists because of the fixed length of data) [Nikias and Raghuveer, 1987; Mendel, 1991]. Also, the phase information is retained in the bispectrum. Walker proposed an algorithm to analyse the glottal pulse based on bispectrum [Walker, 2003] with limited success. In the study, voiced speech is

modelled as a non-Gaussian coloured noise driven system, and linear bispectrum analysis can be applied to obtain glottal pulse and vocal tract estimates in a hybrid Iterative Adaptive Inverse Filtering (hIAIF) framework. In addition, the HOS approach has proved useful in recovering the transfer function of a speech production system, particularly for nasal sounds [Hinich and Shichor, 1991]. Such a system can be non-minimum phase (including both poles and zeros) and when it is inverse filtered, the residual is much closer to a pure pseudo-periodic pulse train than the outputs of pole-only inverse filtering. For example, in [Chen and Chi, 1993], a two-step method is applied to estimate the input pulse train and vocal tract filter. The first step is to estimate the input non-Gaussian pseudo-periodic positive pulse train by HOS based inverse filters; subsequently the ARMA coefficients are estimated by an input-output system identification method [Chi and Kung, 1992].

The main drawback with HOS methods is that they require a large amount of data to reduce the variance in the spectral estimates to ensure reliability [Hinich and Wolinsky, 1988]. Also, for ARMA system identification there is no a priori information about the number of poles and zeros. Finally, for HOS based deconvolution, the glottal return phase information is contained in the vocal tract effect and cannot be extracted easily.

3.5 Conclusion

Several approaches to glottal waveform extraction were introduced in this chapter. We described the most widely used method to estimate the glottal component: glottal inverse filtering (GIF). Subsequently, three effective GIF approaches (ICPIF, IAIF, WRLS-VFF) were described in detail. A performance study was carried out to apply the three methods to real speech segments. From the experimental results it can be observed that no single algorithm works best for all kinds of speech signals. Thus, it is reasonable to expect more reliable glottal estimates by combining the estimates from multiple algorithms. This is the motivation of the work in Chapters 5 and 6 of this study. In addition, some further speech decomposition methods, such as the mixed-phase (Zeros of Z-transform and Complex Cepstrum) decomposition and higher order statistic analysis-based

speech decomposition were introduced. The limitations of these approaches were discussed. With the multi-estimate fusion framework to be introduced in Chapter 5, it is possible to combine the estimates from different speech decomposition approaches.

Chapter 4

Automatic Glottal LF-model Fitting

4.1 Introduction

Once the glottal flow waveform has been obtained by the approaches described in Chapter 3, it is necessary to model the flow with a specific set of parameters. This leads us to fit a parametric model to the glottal flow waveform. Due to the popularity of the Liljencrants-Fant (LF) model [Fant et al., 1985] for modelling the voice source, much research had been carried out toward fitting this model to the extracted glottal flow derivative.

In Section 4.2, we introduce the basic concept of curve fitting for better understanding the theory of the LF-model fitting method. Related work is presented in Section 4.3, including both time-domain and the frequency-domain approaches for automatic LF-model fitting. Our own work, a new time-domain LF-model fitting algorithm based on the Extended Kalman Filter (EKF) is proposed in Section 4.4. To evaluate its effectiveness, in Section 4.5 we compare this method to both a standard time-domain LF fitting method and a spectral fitting approach. Experimental results are presented and discussed.

4.2 Curve Fitting

In real world applications, it is often difficult to interpret the observed output of a system. This is because the output of the system is affected by many factors such as the environment, the system operator, the stability of the system itself and other factors leading to noisy output data. To remove noise and provide insight, scientists and engineers often seek to represent the observed data with a mathematical model [Press et al., 1986; Motulsky and Christopoulos, 2004]. If the data is appropriately modelled, some important characteristics of the data can be determined conveniently, such as the rate of change anywhere on the curve (first derivative), the local minimum and maximum points of the function (zeros of the first derivative), and the area under the curve (integral) [Ledvij, 2003]. In general, the goal of curve fitting is to find the parameter values that best fit a series of data points.

One popular way for the curve fitting procedure is performed by minimising a pre-defined error function to optimise the model parameters. Initially, a set of parameters is given to bootstrap the model, and the corresponding fitting errors are calculated. Subsequently, the fitting algorithm runs in an iterative manner varying the parameters to minimise the error function. The algorithm stops when a certain condition is achieved, such as a maximum number of iterations or the value of the error function falls below a threshold. The estimated parameters are assumed to generate the best fit to the observed data. Two examples of the curve fitting procedure are shown in Fig. 4.1 and 4.2 by utilising functions from the Matlab curve fitting toolbox. It can be observed that for Fig. 4.1, a straight line is successfully fitted to the data samples and for Fig. 4.2, a fifth order polynomial model is fitted to the noisy measurements which are actually samples of a sinusoidal signal.

4.3 Automatic LF-model Fitting Related Work

In this section, the LF-model is firstly reviewed, then related work for automatic LF-model fitting is discussed along with the factors that affect the performance of the LF fitting methods.

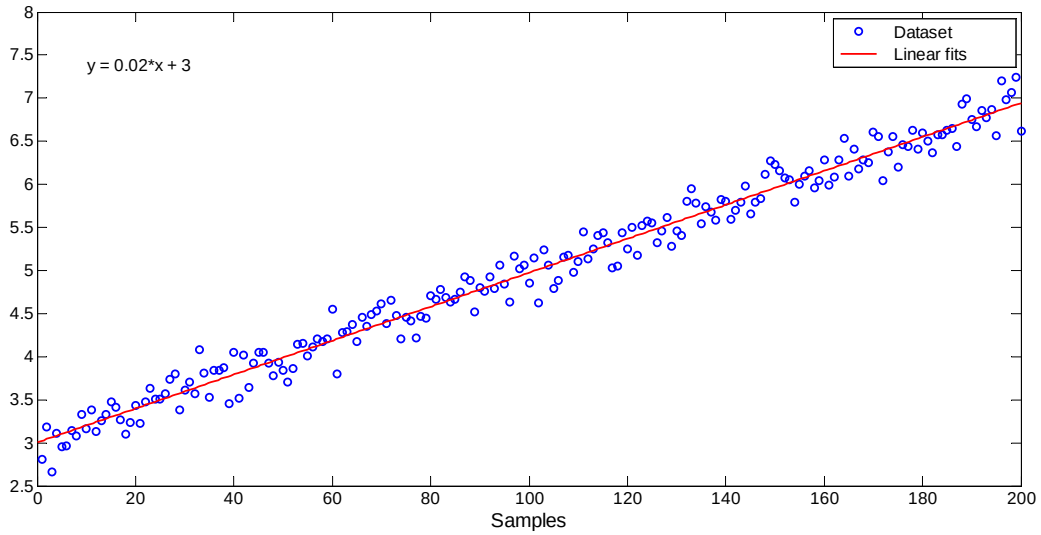

Figure 4.1: Data set and curve fitting results (linear fitting)

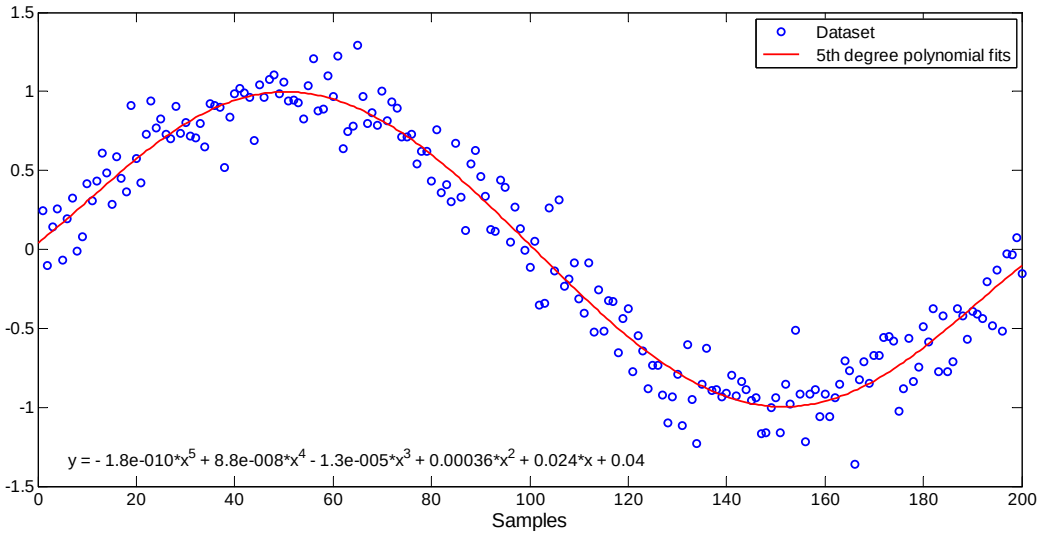

Figure 4.2: Data set and curve fitting results (fifth order polynomial fitting)

4.3.1 Review of the LF-model

The Liljencrants-Fant (LF) model [Fant et al., 1985] is a four-parameter model for describing the shape of the differentiated glottal flow. It is preferred by many researchers for a variety of reasons. Firstly, by choosing different sets parameters, the LF-model can accommodate a wide range of natural speech variations such as modal, vocal fried and breathy voices [Lu and Smith, 2002]. Also, Childers has shown that the LF-model is superior to other glottal source models for natural voice source modelling [Childers and Ahn, 1995; Childers, 1995] since the LF-model can not only represent the glottal open phase, but also the return phase which is important for describing voice quality. Strik pointed out that the LF-model is suitable for use in speech synthesis [Strik, 1998] and Cabral integrated the LF-model into an HMM-based speech synthesiser to improve the naturalness of synthetic speech [Cabral et al., 2008, 2011].

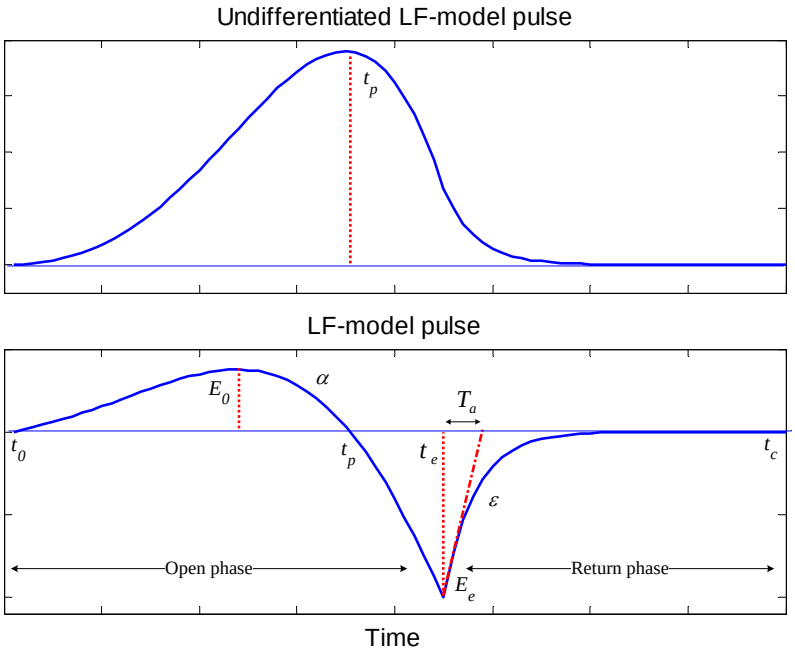

Figure 4.3: A typical LF-model pulse (bottom) and its undifferentiated waveform (top)

An illustration of a typical single pitch period LF-model pulse and its undifferentiated equivalent waveform are given in Fig. 4.3. The model consists of two

phases: the open phase and the return phase. The open phase starts from the glottal opening instant, t_0, until the closing instant t_e, the main glottal excitation point which has an amplitude value of $-E_e$. t_p is the positive peak of the undifferentiated flow. This segment is modelled by equation (4.1)

$$E(t) = E_0 e^{\alpha t} \sin \omega_g t \qquad \text{for } t_0 \leq t \leq t_e \tag{4.1}$$

where it can be observed that the LF-model open phase is a sinusoidal function which grows exponentially in amplitude. ω_g determines the frequency of the sine wave and α controls the amplitude increasing rate. E_0 is a scalar ensuring the required 'area-balance' of the open phase and the return phase (the positive area of the pulse should equal the negative area to ensure zero-flow across the whole pulse duration).

The LF-model return phase is given by equation (4.2)

$$E(t) = \frac{-E_e}{\varepsilon T_a}(e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}) \qquad \text{for } t_e < t \leq t_c \tag{4.2}$$

which is an exponential function from the maximum excitation to the end of the pitch cycle. The duration of this segment is $t_c - t_e$. The most important parameter for the return phase is T_a, which is a measurement of the duration of this segment and defined by a tangent fitted at t_e. Thus it has a duration from t_e to the point where the tangent intersects the horizontal axis. ε is the time-constant of the exponential function which is determined iteratively from T_a and $t_c - t_e$ by the following equation (4.3):

$$\varepsilon = \frac{1}{T_a}(1 - e^{-\varepsilon(t_c-t_p)}). \tag{4.3}$$

4.3.2 Time-domain LF-model Fitting

Many efficient approaches have been proposed to fit the LF-model to the inverse filtered time-domain glottal flow derivative waveform. In this section, an overview of the different methods proposed and developed by researchers in the past is given, and their effectiveness is discussed.

4.3.2.1 Strik's Method

Strik introduced an effective time-domain LF-model fitting approach in his works [Strik et al., 1993; Strik and Boves, 1994; Strik, 1998]. Firstly, the inverse filtered glottal flow derivative (GFD) signal is low-pass filtered to remove ripples and noise appearing in the waveform which might affect the initial parameter estimates. Strik suggested using a 7-point Blackman window as the low-pass filter because it does not cause significant alteration to the waveform. The initial estimates for the LF parameters t_0, t_p, t_e, E_e are obtained directly from the low-pass filtered GFD signal as follows:

- t_e: the instant of maximum negative peak.
- E_e: the amplitude of point t_e.
- t_p: the instant of the first zero-crossing to the left of t_e.
- t_o: the time-point to the left of t_p whose amplitude falls below a threshold value.

The return phase parameter T_a is initialised using a frequency-domain procedure: the value of the maximum amplitude (normalised by E_e) of the return phase spectrum is proved to be a good predictor of T_a, where probably this amplitude closely resembles the DC-component of the return phase and an increase of T_a would result in a larger DC-component [Strik et al., 1993]. This initial set of parameters is used to construct an initial LF-model pulse, before the following two optimisation procedures are applied.

The first optimisation technique used by Strik is the Nelder-Mead simplex search algorithm [Nelder and Mead, 1965]. It has been shown in [Strik and Boves, 1994] that simplex search approaches generally have a better performance at finding a global minimum than gradient descent methods when fitting the LF-model to the inverse filtered GFD signal.

When the estimate is close to a minimum, a gradient descent approach leads to faster convergence. Thus, during the second optimisation procedure Strik uses the Levenberg-Marquardt gradient descent algorithm [Marquardt, 1963] to refine the LF-model parameter estimates.

In further research [Strik, 1997], Strik studied the effect of low-pass filtering on fitting the LF-model to the inverse filtered GFD signal. He showed that

a low-pass filter might cause alteration of the shape of the GFD waveform. To obtain more accurate LF-model estimates, he suggested applying a 7-point Blackman windowing as the low-pass filter to the constructed LF-model pulse during the optimisation stage to ensure the consistency with the low-pass filtered GFD signal.

The basic idea of Strik's LF-model fitting method is adopted by many authors with various modifications for both the initialisation and optimisation procedures.

• Airas implemented an automatic LF fitting approach and integrated it into the TKK Aparat speech analysis toolkit [Airas, 2008]. The initial LF parameter estimates are obtained by searching the inverse filtered GFD waveform as in Strik's method. The difference is that the glottal openings t_0 are located by: firstly searching for the point t_{min} having the minimum amplitude of the glottal flow estimate after the glottal closing instant t_e, secondly a threshold is defined as 10% (relative to the maximum amplitude of the glottal flow) above the amplitude of t_{min} and the corresponding time instant is acquired, thirdly scanning backwards as long as the GFD is positive or the preceding 5% of the glottal flow shows limited variation. Also, instead of using a two-stage optimisation procedure to refine the estimates, Airas performs the optimisation in a single stage using an interior trust region least-squares non-linear optimisation algorithm [Coleman and Li, 1996].

• Lu proposed a modified LF-model fitting approach based on Strik's method [Lu and Smith, 2002]. Firstly, the inverse filtered GFD signal is low-pass filtered by a 7-point Blackman window to remove ripples and noisy components, and the LF-model timing parameters t_c, t_p, t_e and E_e are then directly obtained from the smoothed waveform. Different to Strik's method, the return phase parameter T_a is initialised as $T_a = 2/3 \cdot (t_c - t_e)$. The refinement of the LF estimates is based on a non-linear constrained optimisation by the sequential quadratic programming method [Gill et al., 1991]. At this stage, t_e, E_e and t_c are held constant, while t_p is allowed to vary within a range of 20% of its initial value. The value of T_a is also restricted to $0 < T_a < (t_c - t_e)$. The optimisation algorithm seeks to minimise the root-mean-square error between the reconstructed LF-model pulse and the GFD waveform.

4.3.2.2 Riegelsberger's Study

Riegelsberger studied the performance of three techniques applied to fit the LF-model to the inverse filtered GFD signal [Riegelsberger and Krishnamurthy, 1993].

First, the LF-model equations were modified and expressed in terms of complex exponentials for both the open phase and the return phase, which are given by equation (4.4):

$$E(n) = \begin{cases} C_o z_{go}^n + C_o^*(z_{go}^*)^n, n = 0, ..., N-1 \\ C_r z_{gr}^{n-N}, n = N, ..., M-1 \end{cases} \tag{4.4}$$

where N, M correspond to the glottal closing and closure instants, respectively and

$$\begin{aligned} C_o &= 0.5 A_{go} e^{j(\phi_{go} - \pi/2)}, \\ z_{go} &= e^{\alpha_{go} + j\omega_{go}}, \\ C_r &= -A_{gr}, \\ z_{gr} &= e^{-\alpha_{gr}} \end{aligned} \tag{4.5}$$

and A_{go}, A_{gr}, α_{go}, α_{gr}, ω_{go} are the corresponding LF-model scale factors and shape parameters, and ϕ_{go} is the phase term added to the open phase sinusoid. This modified LF-model is in the form of a sum of complex exponentials in both the open and return phases. Subsequently, Prony's method [de Prony, 1795] can be applied in each phase to fit these equations to the GFD signal. Experimental results show that the direct application of Prony's method may result in discontinuities when the glottal opening instants (GOIs) are poorly located.

To improve the discontinuity problem, Riegelsberger introduced an extended Prony's method. In this approach, the estimated glottal opening location is disregarded and the damped sinusoid of the open phase is simply extended backwards in time until it reaches zero, where this point is assumed to be the new GOI estimate. Consequently, the good fit over the open phase by Prony's method is retained and the fit at the opening point can be improved.

As there is still no guarantee that the GOI estimate by the extended Prony's method is accurate, Riegelsberger also implemented a gradient descent approach for the LF-model fitting. The major advantage of gradient descent is that since it is a search technique, the range of the parameters can be constrained. Ac-

cordingly, we can constrain the search to the valid ranges of these parameters regardless of the noise content of the GFD waveform. The disadvantage of this method is that manually adjusting the gradient descent solution is necessary to avoid convergence to poor local minima.

In this study, Riegelsberger concluded that the two Prony-based methods are inferior to the gradient descent approach, since although for synthetic waveforms both techniques can produce reasonable fits to clean glottal flow waveforms, in real speech a gradient descent method outperforms the Prony-based algorithms.

4.3.2.3 Childers and Ahn's Method

Childers and Ahn begin their LF fitting approach [Childers and Ahn, 1995] by locating the t_c parameter (the end of one pitch cycle) at the point when the GFD waveform falls to 1% of the maximum negative amplitude value. Subsequently, the glottal closing instant t_e and the corresponding amplitude E_e are approximated. t_p is estimated by searching for the first zero-crossing to the left of t_e. T_a is initialised by approximating the spectral tilt of the inverse filtered GFD signal. Afterwards, an initial set of LF estimates is obtained.

Subsequently, the GFD signal is divided into two segments: the open phase and the return phase. The open phase segment, from 0 to t_e, is used to optimise t_p, E_0, α and ω_g. T_a and ε are adjusted using the return phase signal, from t_e to t_c. All parameters are optimised by minimising the squared errors of the two phases between the reconstructed LF-model pulse and the inverse filtered data. The optimisation procedure acts iteratively and the LF-model estimates are varied until the total fitting error is minimised or reaches a certain threshold value.

4.3.3 Frequency-domain LF-model Fitting

Kane et al. introduced and developed a frequency-domain LF-model fitting approach for voice source parametrisation, since for a phase distorted speech signal, standard time-domain LF fitting algorithm cannot generate a valid estimate [Kane et al., 2010].

Firstly, a codebook is generated covering a wide range of LF-model parameter sets and the corresponding differences between the amplitudes (in dB) of the first two harmonics (H1*-H2*) are calculated. Fant proposed an empirical formulation relating the glottal open quotient (OQ) and H1*-H2* with respect to the LF-model: $H1^* - H2^* = -6 + 0.27exp(5.5OQ)$ [Fant, 1995]. Thus, the H1*-H2* measure can be utilised to find an initial set of LF estimates. For each pitch period of the GFD signal, a 256-point Hamming window (centred on the glottal closing instant) is applied. The GFD spectrum is calculated by the Fast Fourier Transform. Subsequently, the H1*-H2* of the glottal waveform spectrum is measured and a search procedure is applied to the codebook to find the closest H1*-H2* value and the corresponding set of LF-model parameters are selected as the initial estimates.

The refinement of the estimate uses a two-step optimisation procedure. The first step is to adjust the initial estimates of the LF-model parameters by minimising the difference between the first six harmonics (which as low frequency components characterise the glottal pulse) of the GFD spectrum H_{GFD} and the LF-model spectrum H_{LF}, by the Nelder-Mead multidimensional unconstrained non-linear algorithm [Nelder and Mead, 1965]. The difference for the first two harmonics is doubled to prioritise their matching (as they are more important to the glottal contribution) and the cost function is given by:

$$\begin{aligned} D_1 = 2 \cdot \sum_{n=1}^{2} (H_{GFD}(n) - H_{LF}(n))^2 \\ + \sum_{n=3}^{6} (H_{GFD}(n) - H_{LF}(n))^2. \end{aligned} \tag{4.6}$$

The second step is to adjust the estimate of the return phase parameter T_a, leaving T_p and T_e unchanged. As the LF-model return phase is an exponential function, T_a mainly contributes to higher frequency components. Thus, the fitting error, D_2 between the two spectra is minimised to obtain a more accurate T_a estimate. A flowchart of Kane's algorithm is given in Fig. 4.4.

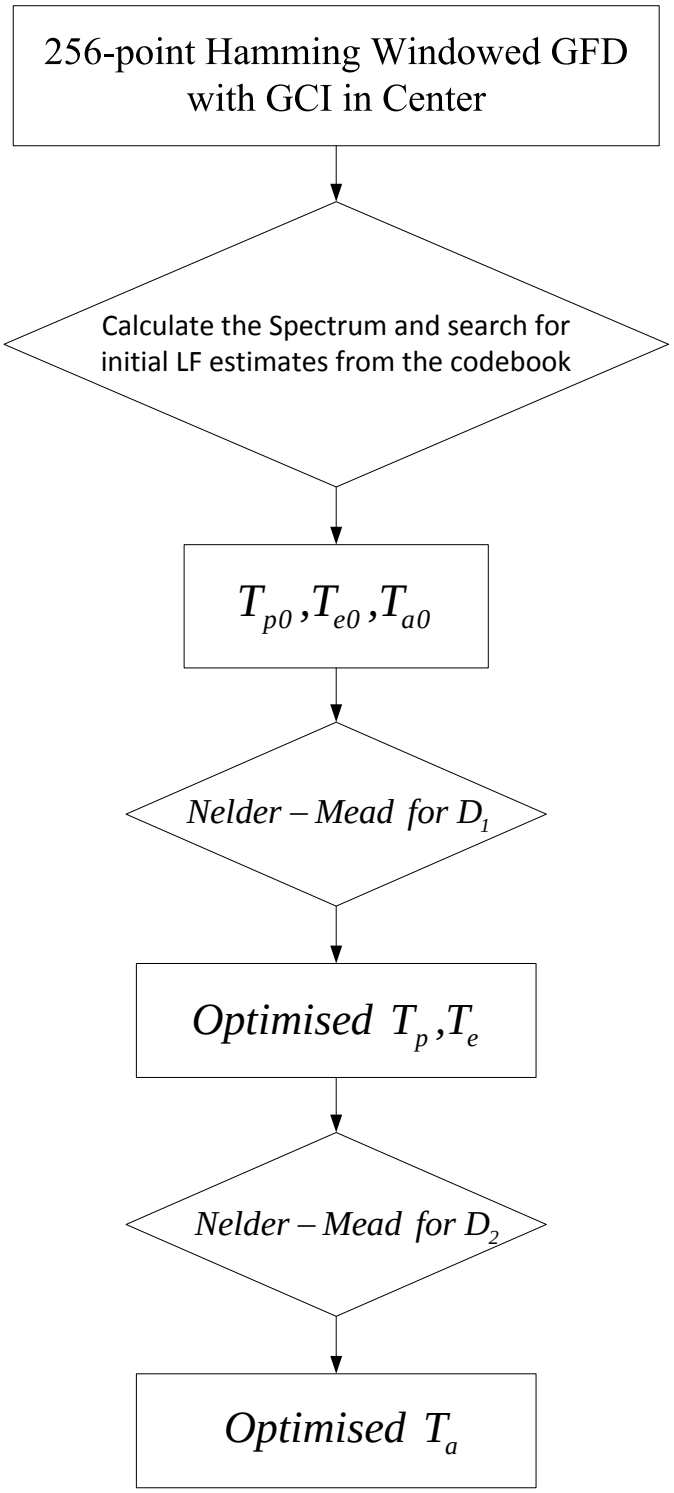

Figure 4.4: Flow chart of the spectral LF-fitting algorithm [Kane et al., 2010].

4.3.4 Factors Affecting LF-model Fitting

In order to obtain accurate estimates when fitting the LF-model to the glottal flow derivative, it is necessary to consider different factors that may affect the fitting procedure. The four main such factors are listed below:

• Quality of the extracted glottal waveform [Strik et al., 1993]. The goodness of inverse filtering is the most important factor that affects the fitting procedure. Clearly, the better the performance of the inverse filtering, the cleaner, the fewer ripples and noisy components in the extracted glottal waveform. This makes it easier for the fitting procedure to find a more accurate set of initial LF estimates. However, if the performance of the inverse filtering is poor, there will remain incompletely removed formants appearing in the glottal waveform. This can lead to an inaccurate estimate of the glottal openings for the time-domain method, and for the frequency-domain method, the amplitude values of the low frequency harmonics will be affected and poor initialisation ensues.

• Source-tract interaction [Fant, 1993]. The mechanism for real speech production is more complex than that assumed by the ideal source-filter model which assumes that the source and vocal tract components are independent of each other. In fact, because of the sudden closure and gradual opening of the vocal cords, a source-tract interaction effect exists, resulting in the true glottal flow always containing components that can not be easily modelled by an all-pole model. Thus, after inverse filtering, the derivative may exhibit some degree of such interaction in the glottal open phase as multiple peaks, which may affect the LF-model fitting by increasing the difficulty of finding the glottal opening instants.

• Optimisation techniques [Strik and Boves, 1994]. After initial estimation, the LF-estimates must be refined. Different optimisation techniques such as Nelder-Mead [Nelder and Mead, 1965] and Levenberg and Marquardt's gradient descent algorithm [Marquardt, 1963], can be applied to improve the accuracy of estimates. We should carefully choose among these approaches to avoid local minima.

• Error criterion. An error criterion is crucial for the optimisation procedure. Generally, the root-mean-square error measure is used to find a global minimum

[Childers and Ahn, 1995]. However, because of ripples resulting from poor inverse filtering and source-interaction, there is no guarantee of a global minimum error corresponding to the best fit of the LF-model. Consideration should be given to the fitting error especially the open phase fitting error which is directly related to poor glottal opening instant estimation.

4.4 A New Time-domain LF-model Fitting Algorithm by Extended Kalman Filtering

To overcome the difficulty of accurately estimating the LF-model return phase parameter in the time-domain [Strik et al., 1993], and improve the accuracy to locate the glottal opening instant (which is crucial of locating the estimation of the open phase parameters), a novel time-domain LF-model fitting approach is introduced in this section. Firstly, a brief introduction to extended Kalman filtering (EKF) is given, before the LF-model equation is re-written in a discrete time format to conveniently use the EKF equations. Subsequently, we show how to apply EKF to track the LF-model shape-controlling parameters. Finally, the full algorithm for LF-model parameter estimation is described.

4.4.1 Extended Kalman Filtering (EKF)

In estimation theory, the EKF is the nonlinear version of the classic Kalman filter [Kalman, 1960] which operates recursively on streams of noisy input data to produce a statistically optimal estimate of the underlying system state (a more detailed introduction to the Kalman filter is given in Chapter 5). The primary limitation of the Kalman Filter is that it offers the optimal estimate for linear system models with additive independent white noise in both the process and the measurement systems while most of the practical systems are nonlinear. The EKF was developed as a solution [Welch and Bishop, 1995; Ribeiro, 2004] and has been effectively utilised in many applications [Hoshiya and Saito, 1984; Dhaouadi et al., 1991; Lee and Ricker, 1994; Tuan Pham et al., 1998; Zeng et al., 2011; Kumari et al., 2011].

As with basic Kalman filtering, EKF makes use of past measurements to produce an a priori estimate. Subsequently, the current measurement is used to update and generate an a posteriori estimate. The process model and the measurement model are given by equation (4.7):

$$\begin{aligned} x_k &= f(x_{k-1}, u_{k-1}, k) + w_k \\ z_k &= h(x_k, k) + v_k \end{aligned} \tag{4.7}$$

where x_k is the state vector of the process model at step k, u_k is a control signal, z_k is the observed measurement, w_k and v_k are random variables representing the process and measurement noise with white Gaussian distribution $p(w) = N(0, Q)$ and $p(v) = N(0, R)$. f and h are the nonlinear functions controlling the transition and measure processes.

The EKF optimisation procedure has two steps: time update and measurement update. The time update equations are expressed by the following two equations:

$$\begin{aligned} \hat{x}_k^- &= f(\hat{x}_{k-1}, u_{k-1}, k) \\ P_k^- &= F_k P_{k-1} F_k^T + Q \end{aligned} \tag{4.8}$$

where $\hat{x}_k^-$ is the a priori estimate at step k, x_{k-1} is the a posteriori estimate at step $k-1$, P_k^- and P_{k-1} are their corresponding error covariances. F_k is the partial derivative function of with respect to x where

$$F_k = \frac{\partial f}{\partial x}(\hat{x}_{k-1}, u_{k-1}, k) \tag{4.9}$$

The EKF measurement update equations are given by equation (4.10):

$$\begin{aligned} K_k &= P_k^- H_k^T (H_k P_k^- H_k^T + R)^{-1} \\ \hat{x}_k &= \hat{x}_k^- + K_k(z_k - h(\hat{x}_k^-, k)) \\ P_k &= (I - K_k H_k) P_k^- \end{aligned} \tag{4.10}$$

where K_k is the Kalman gain, H_k is the partial derivative function of h with respect to x where

$$H_k = \frac{\partial h}{\partial x}(\hat{x}_{k-1}, k) \tag{4.11}$$

It can be seen that once a proper set of initial parameters $\hat{x}_0$, P_0, Q and R are given, the extended Kalman filter runs iteratively, and eventually an optimal estimate for the state vector of the process model can be obtained, although sufficient data is necessary for the convergence of the estimation process. For further information on EKF, see [Welch and Bishop, 1995].

4.4.2 Discrete Time LF-model representation

It is convenient to convert the LF-model timing parameters to ratio format for discrete time signals. For a single pitch period of the inverse filtered glottal flow derivative signal, if the length (or the number of samples) of the pitch cycle is N, the start point is the glottal opening t_0, the three LF-model timing parameters normalised by pitch period are expressed as:

$$\begin{aligned} T_e &= \tfrac{t_e - t_0}{N}, \\ T_p &= \tfrac{t_p - t_0}{N}, \\ T_a &= \tfrac{t_a}{N}, \end{aligned} \tag{4.12}$$

and because t_c is the end point of this pitch period (see Fig. 4.3), T_c is set to 1. The constraints of the LF-model are given by equation (4.13):

$$\begin{aligned} \int_0^{T_c} E(t)dt &= 0 \\ E_0 &= -\tfrac{E_e}{e^{\alpha T_e}\sin(\omega_g T_e)} \\ \omega_g &= \tfrac{\pi}{T_p} \\ \varepsilon N_a &= 1 - e^{-\varepsilon(1-T_e)} \end{aligned} \tag{4.13}$$

Accordingly, the LF-model equations (4.1) and (4.2) can be re-written as follows:

$$E(k) = \begin{cases} -\frac{E_e}{e^{\alpha T_e}\sin(\frac{\pi}{T_p}\cdot T_e)} e^{\frac{\alpha k}{N}} \sin(\frac{\pi}{T_p}\cdot\frac{k}{N}), 0 \le k \le T_e N \\ \\ -\frac{E_e}{\varepsilon T_a}[e^{-\varepsilon(\frac{k}{N}-T_e)} - e^{-\varepsilon(1-T_e)}], T_e N < k < N \end{cases} \tag{4.14}$$

where k is the k^{th} sample of the data.

4.4.3 LF-model Shape-controlling Parameter Tracking by EKF

For a single pitch period of the glottal flow derivative signal, the LF-model parameters are constant. Thus, the extended Kalman filtering can be applied to track the two LF-model shape controlling parameters, α for the open phase, and ε for the return phase, where the best fit of the LF-model can be obtained.

4.4.3.1 Tracking for α by EKF

The open phase signal E_o (normalised) extends from the glottal opening t_0 (which is set to 0 here) to the main excitation point t_e. Accordingly, a nonlinear open phase function h_o is defined by the following equation:

$$E_o(k) = h_o(\alpha_k, k) = -\frac{E_e}{e^{\alpha T_e}\sin(\frac{\pi}{T_p}\cdot T_e)} e^{\frac{\alpha k}{N}} \sin(\frac{\pi}{T_p}\cdot\frac{k}{N}), 0 \leq k \leq T_e N \qquad (4.15)$$

As α is the single constant for estimation, there is no input of the control signal u and the process noise vector w, and the corresponding covariance Q is zero. Subsequently, the process model and the measurement model can be expressed by equation (4.16):

$$\begin{aligned} \alpha_k &= \alpha_{k-1} \\ E_k &= h_o(\alpha_k, k) + v_k \end{aligned} \qquad (4.16)$$

where E_k is the k^{th} sample of E_o, h_o is a nonlinear function defined in equation (4.15), v_k is the measurement noise with a white Gaussian distribution $p(v) = N(0, R_o)$. The corresponding EKF time update equations are given by equation (4.17):

$$\begin{aligned} \hat{\alpha}_k^- &= \hat{\alpha}_{k-1} \\ P_k^- &= P_{k-1} \end{aligned} \qquad (4.17)$$

where the current a priori state estimate $\hat{\alpha}_k^-$ is simply a duplicate of the previous a posteriori estimate $\hat{\alpha}_{k-1}$, and the a priori covariance P_k^- is set to be equal to the previous calculated covariance P_{k-1}.

The EKF measurement update equations are presented as follows:

$$\begin{aligned} K_k &= P_k^- H_o(\hat{\alpha}_k^-)(H_o(\hat{\alpha}_k^-)P_k^- H_o(\hat{\alpha}_k^-) + R_o)^{-1} \\ \hat{\alpha}_k &= \hat{\alpha}_k^- + K_k(E_k - h_o(\hat{\alpha}_k^-, k)) \\ P_k &= (1 - K_k H_o(\hat{\alpha}_k^-))P_k^- \end{aligned} \tag{4.18}$$

K_k is the Kalman gain, which is updated in each iteration. $\hat{\alpha}_k$ is the current a posteriori state estimate and it is actually a linear weighted sum of the a priori estimate and the prediction error term $E_k - h_o(\hat{\alpha}_k^-, k)$. The covariance term P_k is also updated and will be used in the next iteration. H_o is a partial derivative function of h_o with respect to α:

$$H_o(\hat{\alpha}_k^-) = \frac{\partial h_o}{\partial \alpha}(\hat{\alpha}_k^-, k) \tag{4.19}$$

Unlike its linear counterpart, generally the EKF is not always an optimal estimator especially if the initial estimate of the state is not reasonable, or if the process is modelled incorrectly. Thus, to obtain an accurate estimate of the open phase shape-controlling parameter α, it is crucial to find appropriate initial values for the EKF parameters.

It was empirically found that $R_o = 0.01$ and $P_0 = 1$ is a reasonably good choice for the EKF tracking procedure. Selecting the value of α_0 is much more important, as it directly affects the accuracy of the state estimate $\hat{\alpha}$ by EKF. Without a priori information on the shape of the glottal flow derivative, it is difficult to set an appropriate value for α_0. Fortunately, the LF-model is a parametric model which has its constraints to construct a valid LF pulse. Thus, there is a limited range of variation of α values. Based on the discrete time LF-model representation given by equation (4.14), our experiments show that the open phase shape-controlling parameter α has a value in the range 0-100 for a valid LF-model. Thus it is reasonable to use multiple values for α_0 to perform the EKF tracking. Each α estimate according to different α_0 values is used to re-construct the LF-model open phase and the mean squared error (MSE) between open phase signals of the newly constructed LF-model and the glottal flow derivative is calculated. The best initial value of α_0 is the one resulting in the minimal MSE and the optimal estimate of α is obtained. Fig. 4.5 shows an example of the final

fitted LF-model open phase signals by different α_0 values. It can be observed that when $\alpha_0 = 15$, the EKF performs the best and the the optimal estimate of α is achieved.

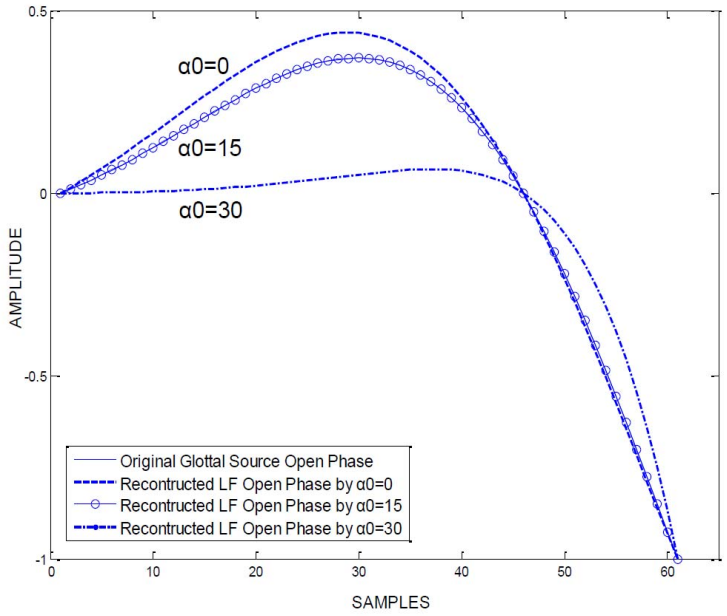

Figure 4.5: Fitted LF-model open phase signals according to different α_0 values

4.4.3.2 Tracking for ε by EKF

Tracking of the return phase shape-controlling parameter ε is similar to the procedure for estimating α. The return phase of the glottal flow derivative extends from the negative peak point t_e and ends at the last sample of this pitch period. The return phase derivative signal is normalised, and a nonlinear function for the return phase is defined by equation (4.20):

$$E_r(k) = h_r(\varepsilon_k, k) = -\frac{E_e}{\varepsilon T_a}[e^{-\varepsilon(\frac{k}{N}-T_e)} - e^{-\varepsilon(1-T_e)}], T_e N < k < N \tag{4.20}$$

Similar to the open phase parameter α, the process and measurement models for ε are:

$$\begin{aligned} \varepsilon_k &= \varepsilon_{k-1} \\ E_k &= h_r(\varepsilon_k, k) + v_k \end{aligned} \tag{4.21}$$

where E_k is the k^{th} sample of E_r, h_r is a nonlinear function defined by equation (4.20), v_k is the measurement noise with white Gaussian distribution $p(v) = N(0, R_r)$.

The EKF time update equations are expressed as:

$$\begin{aligned} \hat{\varepsilon}_k^- &= \hat{\varepsilon}_{k-1} \\ P_k^- &= P_{k-1} \end{aligned} \tag{4.22}$$

where the current a priori state estimate $\hat{\varepsilon}_k^-$ is equal to the previous a posteriori estimate $\hat{\varepsilon}_{k-1}$, and the a priori covariance P_k^- retains the value of the previous calculated covariance P_{k-1}.

The EKF measurement update equations are written as follows:

$$\begin{aligned} K_k &= P_k^- H_r(\hat{\varepsilon}_k^-)(H_r(\hat{\varepsilon}_k^-)P_k^- H_r(\hat{\varepsilon}_k^-) + R_r)^{-1} \\ \hat{\varepsilon}_k &= \hat{\varepsilon}_k^- + K_k(E_k - h_r(\hat{\varepsilon}_k^-, k)) \\ P_k &= (1 - K_k H_r(\hat{\varepsilon}_k^-))P_k^- \end{aligned} \tag{4.23}$$

K_k is the Kalman gain. $\hat{\varepsilon}_k$ is the current a posteriori state estimate updated by a linear weighted combination of the a priori estimate and the prediction error term $E_k - h_o(\hat{\alpha}_k^-, k)$ in terms of K_k. The error covariance P_k is also updated and will be used for the time update step in the next iteration. H_r is a partial derivative function of h_r with respect to ε:

$$H_r(\hat{\varepsilon}_k^-) = \frac{\partial h_r}{\partial \varepsilon}(\hat{\varepsilon}_k^-, k) \tag{4.24}$$

To initialise the return phase EKF tracking procedure, R_r is set to 0.01 and P_0 is set to 1, as good initial values for the normalised signal by our empirically observation. To obtain an accurate estimate of ε, it is necessary to find an appropriate value for ε_0 to avoid inaccurate estimate by deviated tracking. According to the LF-model [Fant et al., 1985], $\varepsilon \approx \frac{1}{T_a}$, and for different voice qualities T_a varies by 1% to 20% [Kane et al., 2010]; thus the value of a valid ε_0 has the range from 1 to 200. The optimal estimate of ε is obtained by fitting the re-built LF-model return phase signal to the original glottal flow derivative return phase using multiple ε_0 values to find the one giving the minimal mean squared error.

4.4.4 Algorithm Implementation

The implementation of the new time-domain LF-model fitting algorithm is described in this section. Above, we have shown that extended Kalman filtering can be used to track the LF-model shape-controlling parameters and the corresponding optimal fit to the glottal flow derivative can be obtained. According to the LF-model constraints (see equation 4.13), the full set of the LF parameters can be extracted by iteratively applying EKF tracking across a range of initial values.

For a segment of voiced speech signal, firstly a glottal waveform estimation approach (e.g., a glottal inverse filtering technique introduced in Chapter 3) is applied to extract the GFD signal. Next the GFD signal is segmented into individual pitch periods using initial estimates of the glottal opening instants derived from a threshold-based procedure [Airas, 2008]. Subsequently, for each pitch period of the GFD, the LF-model parameters are estimated by the new time-domain LF fitting approach presented in Table 4.1. The details of each step are explained as follows:

Step 1: The negative peak point t_e with its absolute amplitude E_e and the positive peak point t_m are found by searching the waveform, and the GFD signal is separated into the open phase and the return phase.

Step 2: The initial t_p estimate t_{p0} is obtained by searching for the first zero-crossing point to the left of t_e.

Step 3: The optimal fitting of the open phase mainly depends on the values of T_p (at this stage the initial estimate T_{p0} is used) and T_e, which are calculated by $T_p = (t_p - t_0)/N$, $T_e = (t_e - t_0)/N$. We set the dynamic range of t_0 from 1 to the point which is $0.1N$ samples before t_m. To locate the optimal t_0, a rectangular window from t_0 to t_e is used to extract the glottal open phase. Subsequently the windowed GFD open phase signal is used by the EKF to estimate the open phase shape-controlling parameter α with multiple values of α_0 as introduced in last section. The minimal mean squared fitting error ($MMSE_f$)is calculated and stored. For all of the t_0 points, the one which gives a global $MMSE_f$ is selected as the optimal glottal opening estimate t_{0opt}. Accordingly, T_{p0} and T_e is updated, and T_e and E_e are output.

Table 4.1: Implementation of the new time-domain LF-model fitting algorithm to the GFD train

For each pitch period of the signal E of length N **do**

Find the negative peak point t_e and its amplitude E_e, and the positive peak point t_m **Step 1**

Find t_{p0} which is the first zero-crossing point to the left of t_e **Step 2**

For $t_0 = 1 : t_m - 10\%N$ **do** **Step 3**

GFD open phase signal $E_o = E[k](k = t_0 : t_e)$

$T_{p0} = (t_{p0} - t_0)/N$

$T_e = (t_e - t_0)/N$

EKF for α with multiple α_0 (see Section 4.4.3.1), the optimal α results in the minimal mean squared fitting error

Store the $MMSE_f$ for the current t_0

End For

Find the optimal t_{0opt} with the smallest $MMSE_f$ across different t_0

Update $T_{p0} = (t_{p0} - t_{0opt})/N$

$T_e = (t_e - t_{0opt})/N$

Output T_e

For $T_p = T_{p0} - 5\% : T_{p0} + 5\%$ **do** **Step 4**

EKF for α as done in Step 3

End For

Find the optimal T_p with the smallest $MMSE_f$ across different T_p

Output T_p

GFD return phase signal $E_r = [E[k](k = t_e : N), zeros(1, t_0 - 1)]$ **Step 5**

EKF for ε with multiple ε_0 (see Section 4.4.3.2)

Find the optimal $\hat{\varepsilon}$ having the $MMSE_f$

Calculate $T_a = \frac{1-e^{(-\hat{\varepsilon}(1-T_e))}}{\hat{\varepsilon}}$

Output T_a

End For

Step 4: A dynamic searching procedure similar to Step 3 is used to refine the estimate of T_p: a reasonable range $T_{p0} \pm 5\%$ is applied to the EKF to find the optimal T_p as the output of this stage.

Step 5: To fit the return phase, $t_0 - 1$ zeros (associated with the closed phase of the previous pitch cycle and not used in the open phase fitting procedure) are appended to the current pitch period of the GFD signal to ensure there is a sufficient number of samples for the EKF. Subsequently, an optimal estimate $\hat{\varepsilon}$ of the return phase controlling parameter ε is obtained by using multiple initial values of ε_0 for initialising the EKF tracking procedure and searching for the global $MMSE_f$. Finally the return phase timing parameter T_a is calculated according to the LF-model constraints.

Applying the proposed new fitting algorithm, the whole set of LF-model parameters can be extracted. In the following section, the performance of the new LF fitting approach is studied by comparison with other LF fitting methods.

4.5 Performance Study

To test the effectiveness of the new fitting algorithm introduced above, we compared it with two other LF fitting methods: one a typical time-domain approach and the other a modified frequency-domain approach.

4.5.1 Comparison with a Standard Time-domain Method

In this section, the newly proposed Extended Kalman filtering LF-model fitting (EKFLF) algorithm was compared with a standard time-domain LF-model fitting algorithm (STDF) [Airas, 2008]. The latter and our new method were applied to both synthetic and real speech data. The evaluation results [Li et al., 2012b] are presented below.

4.5.1.1 Synthetic Speech

Three sets of LF-model parameters of different voice qualities including modal, vocal fry and breathy voice qualities (used in [Fu and Murphy, 2006] and presented in Table 4.2) were used to generate the glottal source signals, and the

corresponding glottal pulse trains were obtained by concatenating ten identical pitch periods. Afterwards, the three sets of LF-model pulse trains were passed through three all-pole vocal tract filters modeling three vowel sounds (formant frequencies and bandwidths were taken from [Akande and Murphy, 2005] and are presented in Table 4.3), and a total of nine sustained synthetic speech segments were created. In addition, for breathy voice, simulated noise of 30dB SNR was added to mimic breathy speech quality. All vowel segments were inverse filtered by an iterative closed phase inverse filtering approach, ICPIF (introduced in Chapter 3), to extract the GFD signal. The GFD signals were divided into individual pitch periods by the initial estimation of glottal opening points. Subsequently, the EKF fitting approach (EKFLF) and the standard time-domain LF-model fitting algorithm (STDF) were applied respectively to all pitch cycles of the GFD signals. The mean squared error (MSE) for the estimated LF-model timing parameters for both algorithms with respect to the true values were calculated, and the results are presented in Fig. 4.6. It is observed that for modal and vocal fry voice qualities the MSE scores are consistently lower for the proposed fitting algorithm compared to STDF. For breathy voice the results are less clear. The estimated T_p and T_e for breathy vowels /IH/, /UH/ by EKFLF are more accurate, however for T_a the standard fitting method performs better. This may be explained by additive noise to breathy voice and imperfect inverse filtering caused by short duration of the closed phase for breathy voices.

Overall these experimental results demonstrate the validity of the proposed LF-model fitting algorithm to estimate glottal LF-model parameters for a wide range of synthetic speech signals, and it is superior to the standard time-domain fitting method in most cases.

Table 4.2: LF-model parameters for three voice qualities

Voice Quality	$T_p(\%)$	$T_e(\%)$	$T_a(\%)$
Model	45.66	57.50	0.91
Vocal fry	18.99	25.14	0.83
Breathy	52.89	75.75	8.19

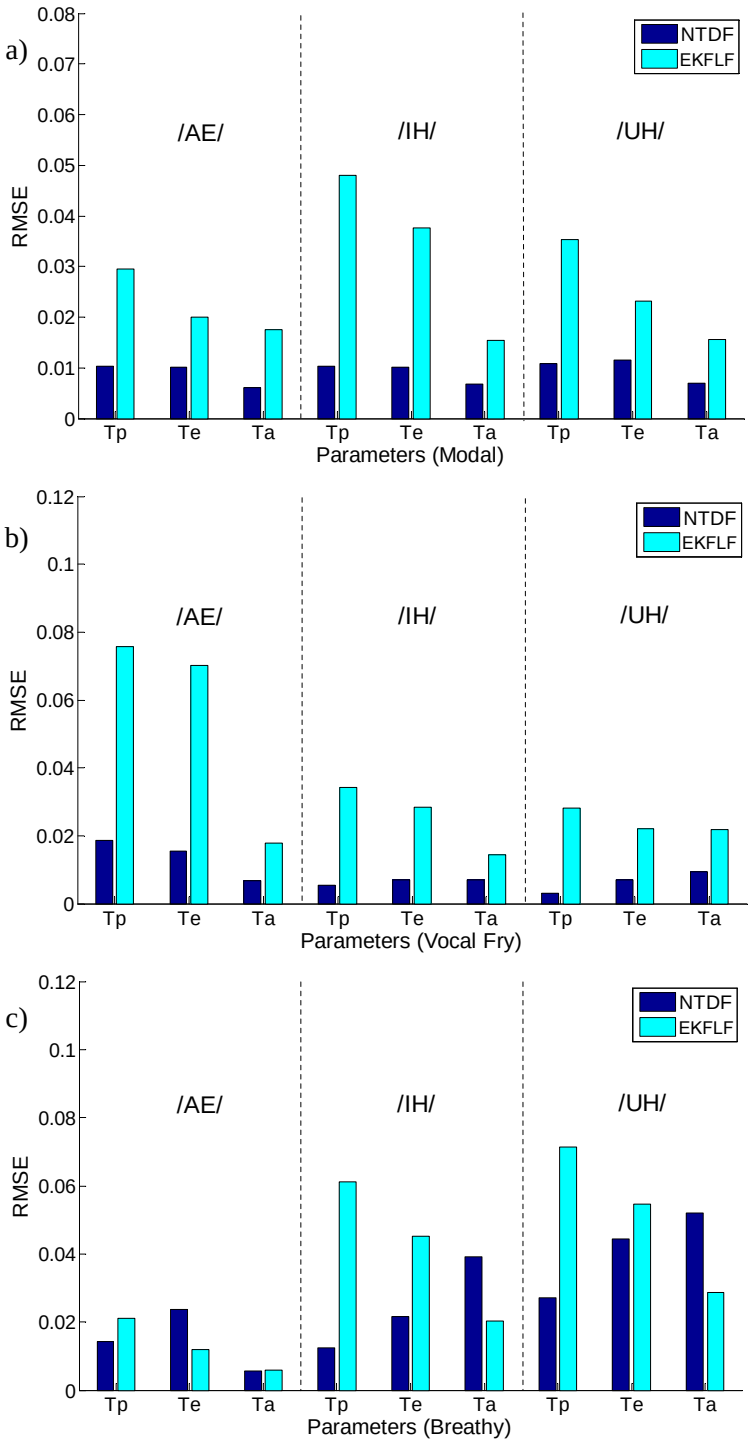

Figure 4.6: MSE scores of estimated LF-model parameters for a) modal voice, b) vocal fry voice and c) breathy voice

Table 4.3: Formant frequencies and bandwidths of three vowels

Vowel	Formant frequencies (Hz)				Formant bandwidths (Hz)			
	F1	F2	F3	F4	B1	B2	B3	B4
/AE/	826	1187	3081	4045	99	159	299	330
/IH/	319	2182	2890	3643	97	100	157	780
/UH/	273	753	2296	3185	90	105	91	105

4.5.1.2 Real Speech

Two segments of real speech were extracted from the CMU-ARCTIC database [Kominek and Black, 2004] for speakers bdl (a male voice) and slt (a female voice). Both segments were inverse filtered by ICPIF to extract the glottal flow derivative signals. Afterwards the two LF-model fitting algorithms were applied. The original speech waveforms, the GFD waveforms and the fitted LF-model pulses by the two methods (LFP_N and LFP_M) are presented in Figs. 4.7 and 4.8. A single pitch period of GFD and fitted LF-model waveforms are shown in Fig. 4.9. In the absence of a priori knowledge of the glottal source component for real speech, it is difficult to measure the accuracy of the estimated source parameters. Instead we compare the goodness of fit to the estimated GFD signals of the two algorithms. Therefore the mean squared error ($MSE = E[(r - r_{LF})^2]$) between the estimated GFD signals r and the reconstructed LF-model pulses r_{LF} across the full speech segments were calculated and the results are presented in Table 4.4. It can be observed from the waveforms and the MSE scores that for both male and female speech segments, the proposed algorithm outperforms the standard time-domain fitting approach by generating smaller MSE scores. For EKFLF, the male subject has a larger MSE than female because of the ripples appearing in closed phases.

From the results above it can be observed that the novel fitting algorithm outperforms the standard fitting approach. In the next section, the proposed fitting method is compared to a frequency-domain approach with both synthetic and real GFD signals.

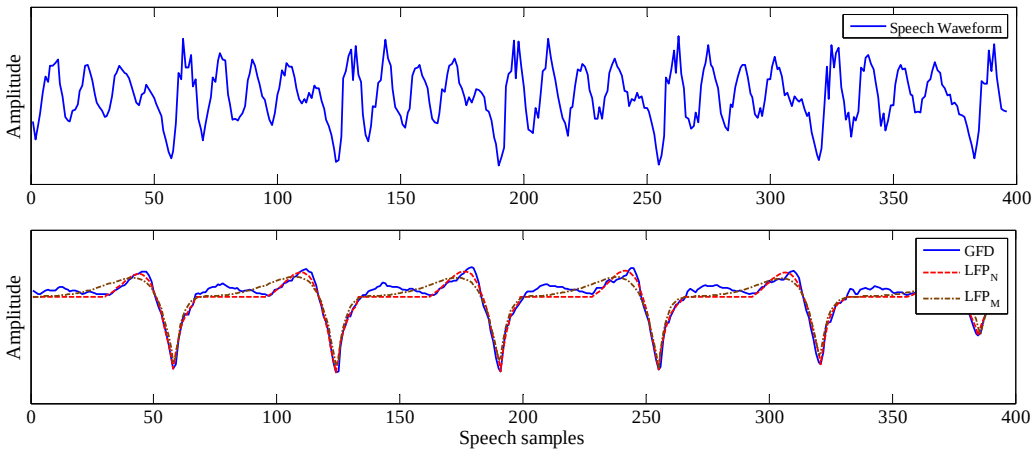

Figure 4.7: Top: male speech waveform, Bottom: GFD signals and fitted LF-model pulses (LFP) from the two algorithms

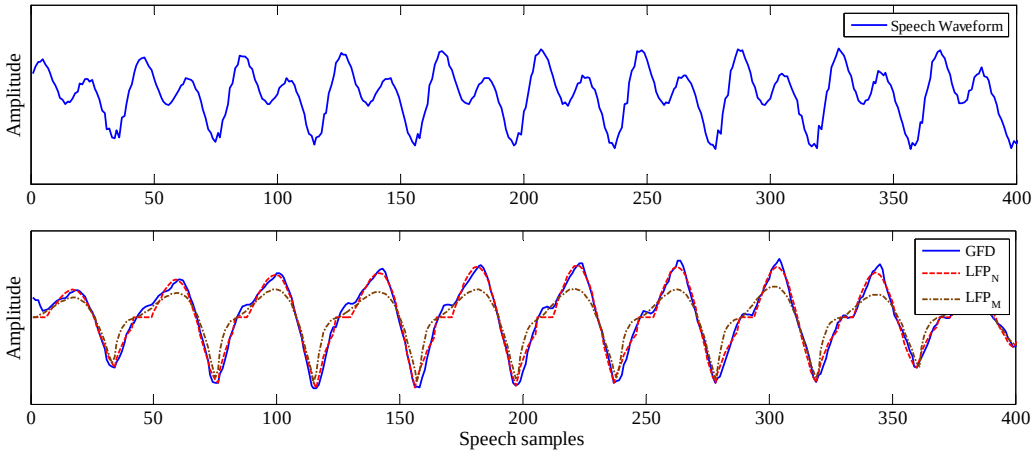

Figure 4.8: Top: female speech waveform, Bottom: GFD signals and fitted LF-model pulses (LFP) from the two algorithms

Table 4.4: MSE scores for real speech segments from two automatic time-domain LF-model fitting algorithms

Algorithm\ Subject	BDL	SLT
EKFLF	0.1851	0.0671
STDF	0.3448	0.7432

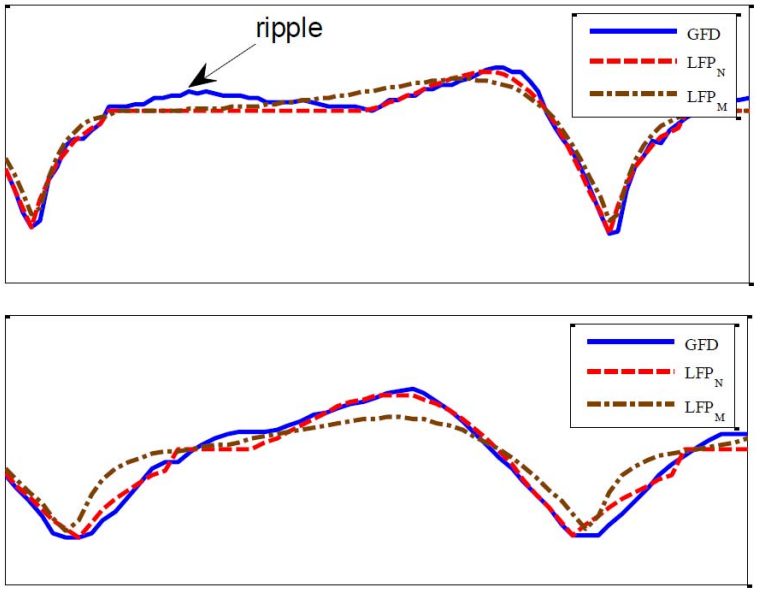

Figure 4.9: Single pitch period of GFD and fitted LF waveforms for top: male and bottom: female

4.5.2 Comparison with a Modified Frequency-domain Method

To compare our proposed fitting approach to a frequency-domain LF fitting method, a modified version of Kane's method [Kane et al., 2010] was implemented. Because of the correlation between the glottal open quotient and the difference between the amplitude of the first and second harmonics (H1*-H2*) of the glottal flow spectrum [Fant, 1995], Kane initialises the LF estimate by searching for the H1*-H2* value in the codebook closest to the GFD spectrum. However, it is shown in [Henrich et al., 2001] that multiple sets of LF-model parameters can generate very similar H1*-H2* values, therefore the spectral optimisation procedure may become stuck in a local minimum caused by poor initialisation, and correspondingly inaccurate estimates will be obtained.

We made some modifications to the approach. Firstly, a codebook is generated of over two thousand LF-model parameter sets and including the corresponding amplitudes of the first six harmonics. For each pitch period of the GFD, a 256-point Hamming window (with the glottal closing instant in the centre as with

Kane's method) is applied. The GFD spectrum is obtained by the Fast Fourier Transform. Subsequently, the mean squared error (MSE) between the first six harmonic amplitudes of the GFD spectrum and those in the codebook is calculated. The set of LF-model parameters generating the minimal MSE is selected as the initial estimate.

The second step of Kane's method to refine the estimate of T_a. How this is achieved is not clearly stated in his study. In our implementation the Itakura-Saito distance [Itakura and Saito, 1968] is minimised during this refinement procedure. This distance is given by equation (4.25):

$$D_{IS} = \frac{1}{N} \cdot \sum \left(\frac{P_{GFD}(\omega)}{P_{LF}(\omega)} - log\left(\frac{P_{GFD}(\omega)}{P_{LF}(\omega)} \right) - 1 \right). \tag{4.25}$$

where the frequency ω has a range from after the sixth harmonic to half the sampling frequency, P is the amplitude and N is the number of the frequency samples.

To compare these two LF-model fitting methods, both artificial and real glottal source signals were used. The experimental details [Li et al., 2012a] are presented below.

4.5.2.1 Artificial Glottal Source

50 sets of LF-model pulses were randomly generated from the range presented in Table 4.5 corresponding to a wide range of voice qualities. Subsequently, the time-domain EKF and the modified frequency-domain LF-model fitting (EKFLF and MFDF) algorithms were applied to extract estimates of the three LF parameters.

Results are presented in Fig. 4.10 (all values of T_e were increased by 0.3 for better illustration). In addition, the mean squared error (MSE) between the estimates and their true values were calculated and are presented in Table 4.6. It can be observed that for a clean, artificial glottal source signal, both EKFLF and MFDF can generate reasonably good estimates. It can also be seen that in most cases, EKFLF outperforms MFDF.

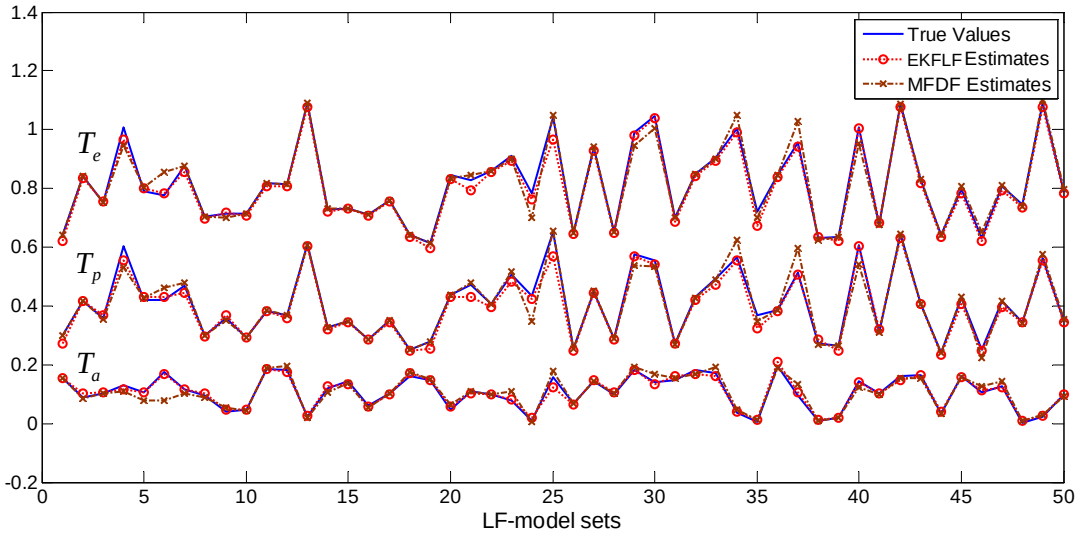

Figure 4.10: Artifical glottal source LF-model parameter true values and the estimates by EKFLF and MFDF

Table 4.5: Range of LF-model parameters

LF Parameter	Range
T_p	[0.2, 0.72]
T_e	[0.3, 0.8]
T_a	[0, 0.2]

Table 4.6: RMSE scores for the three LF-model parameters by EKFLF and MFDF applied to artificial speech data

LF Parameter	EKFLF	MFDF
T_p	0.0196	0.0257
T_e	0.0182	0.0258
T_a	0.0086	0.0191

4.5.2.2 Real Glottal Source

The real speech data used for the evaluation here was supplied by Dr. Irena Yanushevskaya of Trinity College Dublin[1]. The data are based on the all-voiced utterance "We were away a year ago". The inverse filtered glottal source waveform and corresponding hand-labelled LF-model parameters for one male speaker were selected, and 100 pitch periods of the source signal were extracted (the poorly inverse filtered glottal signal were excluded).

The EKFLF and MFDF algorithms were applied to the glottal source waveform. The hand-labelled LF-model parameters and the estimates by both algorithms are presented in Fig. 4.11 (again all values of T_e were increased by 0.3 for better illustration). The MSE scores between the hand-labelled data and estimates were calculated and are presented in Table 4.7. It can be observed that overall EKFLF has a better performance than MFDF. For pitch periods 2-12 and 18-27, MFDF generated inaccurate T_p and T_e estimates while the estimates obtained by EKFLF are very close to the true values. For the remaining 79 pitch periods, performance of the two approaches varies. Fig. 4.12 shows an example where EKFLF performs better. It can be observed that the LF-model waveform obtained by MFDF is poorly fitted to the glottal waveform open phase. An inaccurate fit of the third harmonic in the spectrum can also be seen. Fig. 4.13 illustrates a case where MFDF outperforms EKFLF. Due to the weak amplitude of the glottal waveform and a large number of ripples appearing in the open phase, EKFLF failed to locate the glottal opening instant. In addition, it can be observed in the spectrum that the first and third harmonic of the estimated

[1]Phonetics & Speech Laboratory, Centre for Language and Communication Studies, Trinity College Dublin

Table 4.7: RMSE scores for the three LF-model parameters by EKFLF and MFDF applied to real speech data

LF Parameter	EKFLF	MFDF
T_p	0.0543	0.0852
T_e	0.0527	0.1017
T_a	0.0111	0.0274

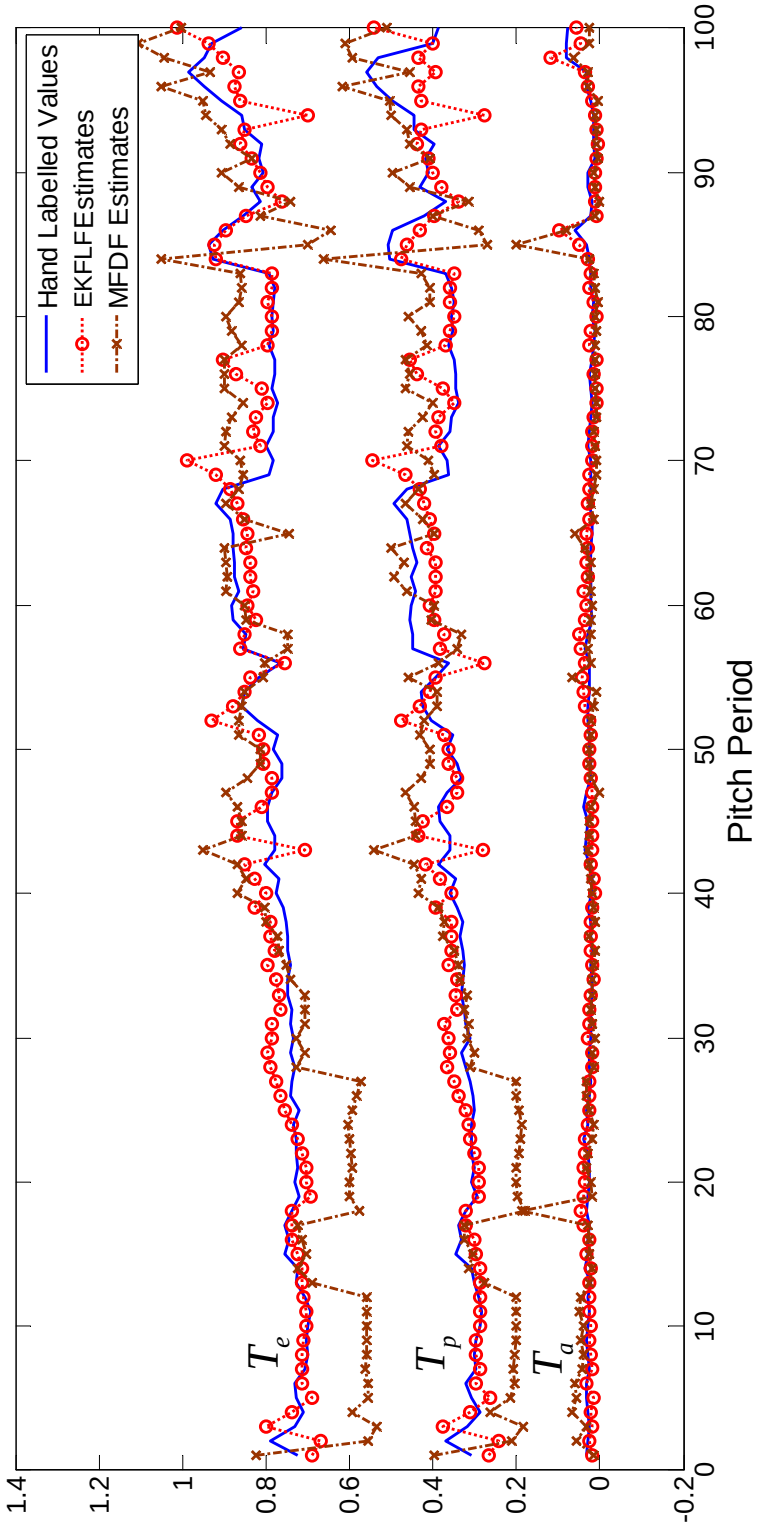

Figure 4.11: Real glottal source LF-model parameter hand-labelled values and the estimates by EKFLF and MFDF

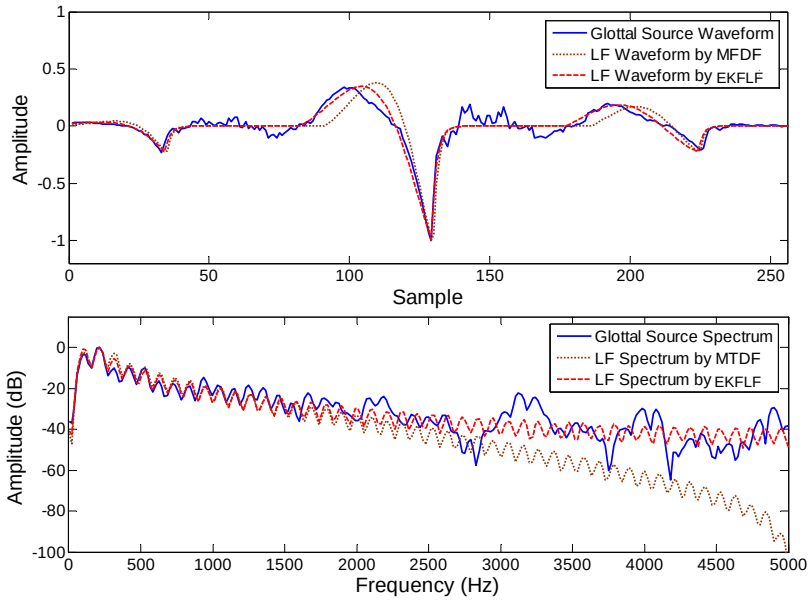

Figure 4.12: An example where the LF-model is better fitted by EKFLF to the real glottal source

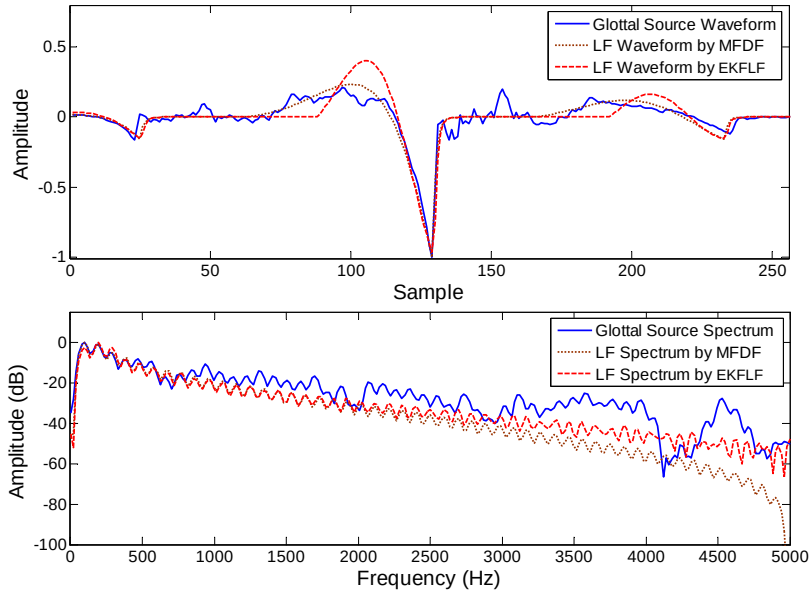

Figure 4.13: An example where the LF-model is better fitted by MFDF to the real glottal source

LF-model by EKFLF is not very well fitted to the glottal spectrum compared to MFDF.

It can be observed from the results above that, for an artificial glottal source signal, both methods can generate accurate LF-model estimates, where the time-domain approach performs better. For a real speech glottal source with hand-labelled data, the time-domain method shows generally more reliable estimates of the LF parameters compared to the spectral fitting method. It can be observed that in general the spectral fitting approach is more sensitive to the quality of the inverse filtered glottal source. Even small fitting errors to low frequency harmonics of the glottal spectrum may result in inaccurate estimates of the open quotient parameter T_e and the asymmetrical parameter T_p. In addition, Fant claimed that different sets of LF-model parameters may produce similar glottal spectra [Fant, 1995]. Therefore, the frequency-domain criterion is less robust compared to the time-domain criterion. Also, Kane put more weight on the time-domain information in his LF-model fitting algorithm [Kane, 2012].

To the above reasons, we decide to use the time-domain error criterion to evaluate the multi-estimate fusion framework incorporated with both the time- and frequency-domain LF-model fitting methods proposed in the following chapters. Further investigation is required by perception tests to measure which criterion is more significant when dealing with the LF-model fitting problem.

4.6 Conclusion

This chapter provided an introduction to glottal source parametrisation by LF-model fitting. General curve fitting was briefly discussed before an overview of widely used LF-model fitting algorithms (including both time-domain and frequency-domain approaches) was presented.

Subsequently, we proposed a new time-domain LF-model fitting algorithm based on extended Kalman filtering. The process of tracking the LF-model shape-controlling parameters was described in detail and a full implementation of the newly proposed algorithm was presented.

To test the effectiveness of this new fitting approach, it was firstly compared to a standard time-domain method and the experimental results showed that

in most cases the new algorithm is superior for both synthetic and real speech signals. In a further test, the new fitting approach was compared to a frequency-domain method. Results shown that for clean synthetic glottal source signal, both methods can result in reasonable LF-model estimates, while the time-domain approach performs slightly better. For a real speech glottal source with expertly hand-labelled data, the time-domain method is more robust than the frequency-domain approach for estimating the source parameters.

Improved performance of the LF-model fitting algorithms requires further developments not only of the fitting approach, but also in speech decomposition. It is obvious that the cleaner the glottal source waveform, the easier it is to fit the LF-model to it. It is interesting to note that (as with inverse filtering algorithms) no single fitting method consistently outperforms all others. This suggests that for accurate source parameter estimation, a hybrid approach that combines estimates from multiple algorithms is worthy of further investigation. In the next chapter, we will introduce a general framework to combine multiple sets of LF-model estimates (obtained from different glottal source extraction and LF fitting algorithms) to generate more consistent and reliable results.

Chapter 5

A Multi-estimate Fusion Framework for Glottal Source Parameter Estimation

5.1 Introduction

Previously, we observed that no single algorithm can give consistently reliable glottal source parameter estimates for different speech signals. Thus, it is reasonable to consider combining estimates from different approaches instead of proposing a 'flawless' algorithm. In this chapter, a general framework to generate and combine multiple sets of glottal LF-model parameter estimates by quantitative data fusion is introduced. We propose that an appropriate combination of estimates obtained from a range of speech decomposition and LF fitting algorithms should result in more reliable results than those from a single algorithm.

This chapter is scheduled as follows. In Section 5.2 an introduction is given to quantitative data fusion techniques. After a brief introduction to data fusion, two general data fusion structures, state vector fusion and measurement fusion, are described. Subsequently, a simple example illustrating Millman's fusion formula is presented. Also, Kalman filtering, which is one of the most widely used techniques for multi-sensor data fusion, is discussed. In Section 5.3, a general fusion framework is proposed for combining multiple glottal source estimates. The

hierarchical structure of the framework is described and the input and output of each level are explained in detail. The advantages and limitations of the fusion approach are discussed. In addition, the factors that may affect the performance of the fusion algorithm are considered.

5.2 Quantitative Data Fusion

With the continuing advancement of micro-electronics, integrated circuit and sensor technologies, multi-sensor fusion has received significant attention and is widely used in both military and commercial applications. A data fusion algorithm aims to combine data from multiple sensors and obtain improved accuracy of estimates of the target than obtainable from a single sensor alone. In the mid-1980s, the Joint Directors of Laboratories (JDL) data fusion group introduced a model for data fusion [Hall and Llinas, 1997]. The JDL fusion model is a universal model covering all levels of data fusion and is widely utilised. In this section we will give a general outline of data fusion and introduce some commonly used techniques for combining quantitative data from different sensors, which can be applied to construct the glottal LF-model multi-estimate fusion framework.

5.2.1 An Introduction to Data Fusion

Data fusion, or information fusion, while a useful concept, is one for which no unified definition exists. Generally, data fusion is regarded as a framework for, or the process of combining data from multiple sources to provide a robust representation of a target [Klein, 2004]. Data fusion is an information processing procedure used in information technology to automatically analyse and integrate the observations from several sensors according to certain criteria, and consequently to complete the anticipated tasks such as making decisions and target estimations. Because most of the information sources in data fusion systems are sensors, data fusion is also widely known as multi-sensor data fusion (MSDF) [Mitchell, 2007]. There is a wide variety of sensors, e.g. according to different measurands sensors may be needed as acoustic, biological, chemical, optical, magnetic etc [Richard, 1987]. Thus different sensors and fusion targets lead to different data fusion algo-

rithms. Data fusion is an inter-disciplinary research area, although at its core are mathematical techniques [Hall and McMullen, 2004]. All data fusion applications can be considered as finding optimal solutions for uncertain problems.

Generally multi-sensor data fusion systems have the following advantages [Mandic et al., 2005]:

1) System reliability and robustness can be enhanced when more than one sensor is used.

2) Time and space measurement range can be extended if sensors are located in different positions and activated in different periods.

3) Estimation accuracy improvement. Various noise components are unavoidable during the process of sensor measuring; utilising multiple pieces of information describing identical features may reduce the uncertainty caused by inaccurate measurements.

4) Target inspection and identification. Different sensors can represent different features of the target; these complementary pieces of information result in the reduction of distortion for comprehending the target.

With the above benefits, multi-sensor data fusion found applications originally mainly in the military field [Tong et al., 1987; Linn et al., 1991; Bass, 2000]. For example, in the Second World War an optical ranging system was incorporated into anti-aircraft guns in order to utilise information from both the radar and optical sensors and consequently to improve the distance measurement accuracy and defense interference. Nowadays the significance of multi-sensor data fusion is widely accepted; together with the appearance of novel sensor devices, enhancement and improvement of processing techniques and software, data fusion systems are widely used not only for military purposes, e.g. territorial waters monitoring, air-to-air and ground-to-air defence and strategic early-warning, but also for civilian convenience, e.g. medical diagnosis, air traffic management, automatic industrial process control, non-destructive testing and remote sensing [Luo et al., 1988; Hall and Linn, 1991; Franklin and Blodgett, 1993; Filippidis et al., 2000; Majumder et al., 2001].

Data fusion algorithms in general can be classified into three groups [Sasiadek, 2002] : probabilistic model based data fusion, least-squares techniques based data fusion, and intelligent fusion. The probabilistic model fusion method includes

Bayesian reasoning, evidence theory, robust statistics and recursive operators; the least-squares techniques are Kalman filtering, optimal theory, regularization and uncertainty ellipsoids; the intelligent information fusion algorithms are fuzzy logic, genetic algorithms and neural networks. For practical applications, there is no absolute best choice and different algorithms may result in equivalent answers. One should choose one or some of the data fusion algorithms according to the requirements of the problem.

For glottal source estimation applied to real speech signal, there is no a priori information of the true glottal component. Also, there is no a priori information can be used for combining data from different source estimation algorithms. Kalman filtering is shown to be effective [Chang et al., 1997; Gao and Harris, 2002] to track the state vectors of a system lacking of useful information. Thus, for our multi-estimate fusion framework we choose to use a measurement fusion method based on Kalman filtering to fuse the estimates obtained by individual algorithms, where the corresponding multiple sets of estimates are combined by a Millman's fusion formula and then smoothed by a Kalman filter. Details of these techniques are presented below.

5.2.2 Two General Fusion Structure: State-Vector Fusion and Measurement Fusion

State vector fusion and measurement fusion are two broad approaches for data fusion based on Kalman filtering. Measurement fusion directly combines the observations without pre-processing and there is no loss of information from the observations, thus the optimal state vector estimate is obtained [Raol, 2009]. In practical applications, this may not always be feasible because the amount of data to be transmitted to the fusion centre would be quite large, and problems might occur due to the channel's limited transmission capacity. For such situations, state-vector fusion is preferable, since each sensor uses a Kalman filter to track the state vector and its associated covariance matrices from the measurements at the corresponding sensor. These state vector estimates are transmitted to the fusion centre and combined to obtain the fused estimate. Transmitting state vector rather than raw data reduces the overload of the channels. The problem

with state-vector fusion arises since the input noise associated with the target and different target sensors could be correlated, and the corresponding cross-correlation is difficult to calculate without a priori information thus often ignored by real applications, which might result in inaccurate estimation. The two fusion structures are described in more detail below.

5.2.2.1 State-Vector Fusion

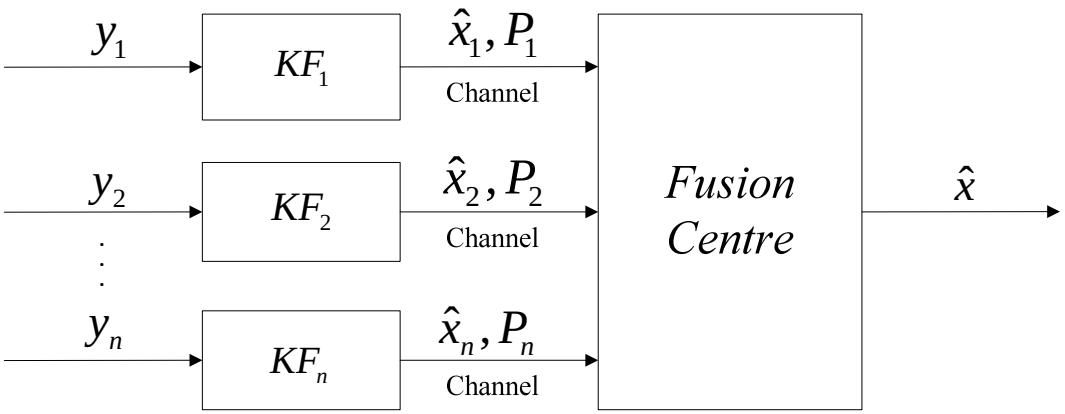

Figure 5.1: State-vector fusion structure

The basic structure of state-vector fusion is shown in Fig. 5.1, where $y_n (n = 1, 2, ..., n)$ is the measurement of the n^{th} sensor, KF_n is the n_{th} local Kalman filter, $\hat{x_n}$ and P_n are the state vector estimate and corresponding error covariance obtained by Kalman filtering, and $\hat{x}$ is the optimal state vector estimate produced by combining all the local estimates. It can be observed that in state-vector fusion, all the measurements y_n of individual sensors are put through a group of Kalman filters to obtain individual sensor-based state estimates, before these estimated state vectors are combined in the fusion centre according to the corresponding estimated state error covariances from n Kalman filters, to achieve an optimal fused state estimate.

Further details on state-vector fusion and some of its applications can be found in [Roecker and McGillem, 2002; Saha, 1996; Chang et al., 1997; Saha and Chang, 1998; Wang et al., 2003; Mitchell, 2007].

5.2.2.2 Measurement Fusion

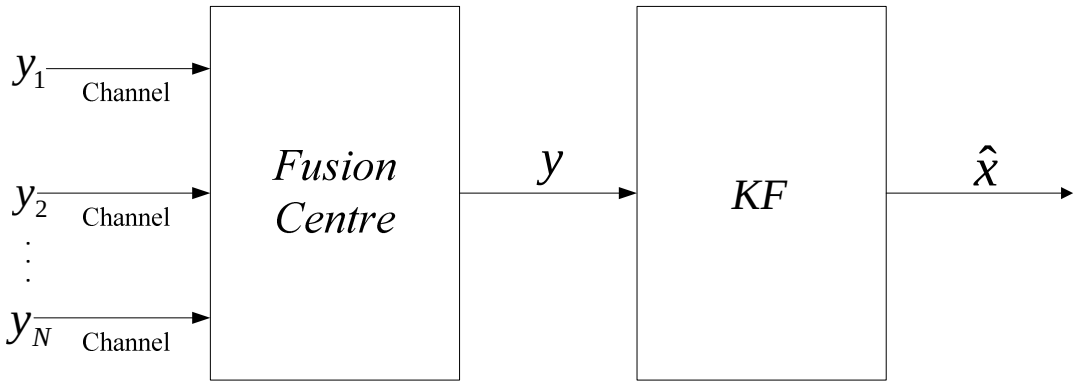

Figure 5.2: Measurement fusion structure

Fig. 5.2 presents the basic structure of measurement fusion, where $y_n (n = 1, 2, ..., n)$ is the measurement of the n^{th} sensor, y is the fused measurement and $\hat{x}$ is the optimal state vector estimate obtained by Kalman filtering. It is clear that in contrast to state-vector fusion, measurement fusion directly fuses the measurements from individual sensors to obtain a composite measurement and the final state estimate is achieved by applying a single Kalman filter to the fused observations.

Extended descriptions of measurement fusion and corresponding applications can be found in [Palmieri et al., 2001; Gan and Harris, 2002; Lee, 2003; Deng et al., 2006; Ran et al., 2008; Gao et al., 2009].

For speech signals, since the number of pitch periods and corresponding LF-model parameter estimates for a voiced frame is relatively small, the analysis is done off-line and requires no transmission. Also, there is no a priori information that can be used to measure the cross-correlation between the "noise" by individual algorithms, thus the fused estimate by state-vector fusion might be inaccurate because of the weight assigned to poor estimates. In such a scenario, a measurement fusion structure is more appropriate to combine the estimates from different algorithms and thus is applied in this study.

5.2.3 A Basic Fusion Formula: Millman's Formula

Millman's formula is a useful tool used in many data fusion applications for directly combining data from different sources [Ajgl et al., 2009]. To illustrate how Millman's fusion formula works, we provide a simple example.

Suppose we have some apples of weight x that we seek to estimate. We have two scales, and they generate two measurements z_1 and z_2, with random, independent and unbiased measurement errors (zero mean), v_1 and v_2. Because we do not know which one of the two scales is more reliable, it is reasonable to combine the two measurements and obtain an optimal estimate of x. Accordingly, the measurements are described by equation (5.1):

$$\begin{aligned} z_1 &= x + v_1 \\ z_2 &= x + v_2 \end{aligned} \tag{5.1}$$

Without a priori information, we could assume that an estimate of x (denoted by $\hat{x}$) is a linear combination of the measurement as followings:

$$\hat{x} = a_1 z_1 + a_2 z_2 \tag{5.2}$$

where a_1 and a_2 are the weights we need to derive, where $a_1 + a_2 = 1$. If the estimation error $\tilde{x}$ is defined as:

$$\tilde{x} = \hat{x} - x \tag{5.3}$$

we should minimise the mean squared value of $\tilde{x}$ as the criterion of optimality. Given v_1 and v_2 are unbiased and independent of x, from equations (5.1)-(5.3) we can derive the following:

$$E[\tilde{x}] = E[\hat{x} - x] = E[a_1(x + v_1) + a_2(x + v_2) - x] = 0 \tag{5.4}$$

$$E[v_1] = E[v_2] = 0, \quad E[x] = x \tag{5.5}$$

$$E[v_1^2] = \sigma_1^2, \quad E[v_2^2] = \sigma_2^2, \quad E[v_1 x] = E[v_2 x] = 0 \tag{5.6}$$

where E is the mean, and σ_1^2 and σ_2^2 are the covariances of v_1 and v_2. Then the mean squared error of $\hat{x}$ is given by equation (5.7):

$$MSE[\hat{x}] = E[\tilde{x}^2] = a_1^2\sigma_1^2 + (1 - a_1)^2\sigma_2^2 \tag{5.7}$$

If we differentiate this quantity with respect to a_1 and set the result to zero, the following expressions can be obtained:

$$2a_1\sigma_1^2 - 2(1 - a_1)\sigma_2^2 = 0 \tag{5.8}$$

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{5.9}$$

Accordingly the minimum mean squared estimation error is:

$$E[\tilde{x}^2] = (\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2})^{-1} \tag{5.10}$$

which it can be observed is smaller than either of σ_1^2 and σ_2^2. Finally we have an optimal estimate of x:

$$\hat{x} = (\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2})z_1 + (\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2})z_2 \tag{5.11}$$

For this example, if $\sigma_1^2 = \sigma_2^2$, the estimate is simply the average value of the two measurements; and if one measurement is perfect, which means σ_1 or σ_2 equals to zero, the other has in fact a zero weight and makes no contribution to the estimate.

This example illustrates the application of Millman's fusion formula to combine the measurements from two sensors. A more general form for combining an arbitrary number of local estimates, called the generalised Millman's formula [Choi et al., 2004; Shin et al., 2006], is given by the following equations:

$$\hat{x} = a_1\hat{x}_1 + a_2\hat{x}_2 + a_3\hat{x}_3 + \cdots + a_n\hat{x}_n \tag{5.12}$$

$$a_1 + a_2 + a_3 + \cdots + a_n = 1 \tag{5.13}$$

$$a_i = \frac{1}{P^{ii}}(\frac{1}{P^{11}} + \frac{1}{P^{22}} + \cdots + \frac{1}{P^{nn}})^{-1} \tag{5.14}$$

$$P^{ii} = E[(x - \hat{x}_i)^2] \tag{5.15}$$

where $\hat{x}_1$,..., $\hat{x}_n$ are local unbiased estimates of an unknown random variable x obtained by independent sensors and P^{11},..., P^{nn} are the associated error covariances. It can be observed that given a set of estimates from different sensors and a priori information of the corresponding covariances, the weights for each estimate can be calculated and an optimal estimate is obtained from linear combination of all sensor estimates.

Millman's formula assumes that the measurement noises from different sensors are uncorrelated, which simplifies the fusion procedure and is very efficient for many problems. In fact, for practical applications the measurement noises of multiple sensors are correlated and the corresponding cross-covariance is not zero. Bar-Shalom and Campo [Bar-Shalom and Campo, 1986] proposed their fusion formula while taking into account the correlated noise from different sensors. However, a priori information of such correlation is required for appropriate initialisation.

5.2.4 A Data Fusion Tool: Kalman Filter

A Kalman filter is an optimal recursive data processing state estimator first introduced by Kalman [Kalman, 1960]. It has been widely used for over thirty years in the areas of control, multi-sensor data fusion, navigation, military, computer image processing, etc.

The basic idea behind the Kalman filter is: update the current state vector by the previous estimate and the current measurement according to the state space model of signal and noise, and subsequently calculate the current estimate. The Kalman filter is in fact a data fusion technique since the current estimate is obtained from combining previous estimate and the current measurement.

The basic Kalman filter equations are given below. An under-test system is described by the following linear stochastic difference equation

$$x_k = Ax_{k-1} + Bu_k + w_k \tag{5.16}$$

and the measurement equation

$$z_k = Hx_k + v_k \tag{5.17}$$

where x_k is the system state at time step k, u_k is the controlling variable, A and B are the system parameters which are matrices for multi-model (linearly parameterised) systems. z_k is the measurement at the k_{th} step, H is the parameter of the measurement system which is a matrix for multi-model systems. w_k and v_k are the process noise and measurement noise which are assumed to be white Gaussian noise with zero mean and covariance Q and R respectively.

Kalman filter can be applied to obtain an optimal state vector estimate. The Kalman filter can be simply expressed by two sets of equations: two time update equations and three measurement update equations [Welch and Bishop, 1995] which are presented in Fig. 5.3. $\hat{x}_k$ is the a posteriori state estimate at time

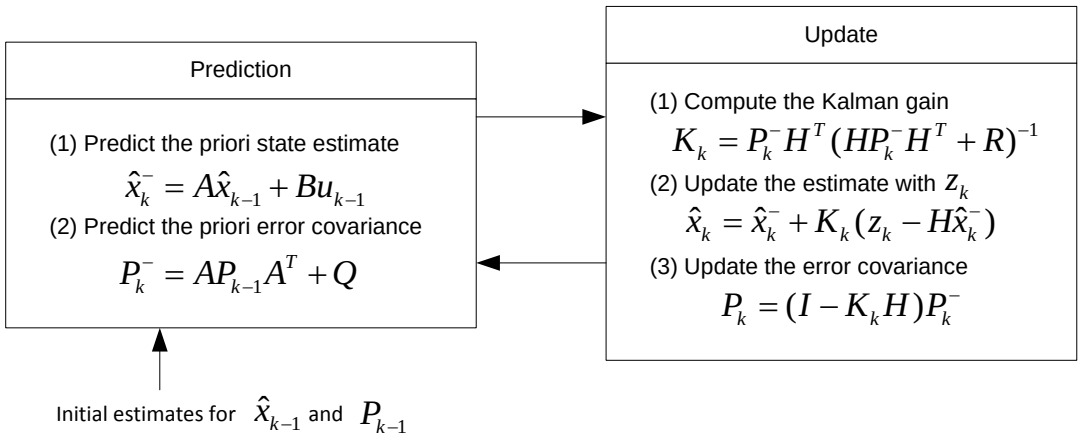

Figure 5.3: Kalman filter prediction and update equations

step k, P_k is the corresponding error covariance, $\hat{x}_k^-$ is the a priori state estimate with an error covariance P_k^-, and K_k is the Kalman gain. The two time update equations predict the estimate at step k according to the estimate at previous

step $k-1$. Afterwards the measurement update equations correct and update the current state estimate using the measurement at the current time step. The obtained a posteriori state estimate and the error covariance then will be used for the next iteration. This filtering procedure runs recursively and stops when there is no more data.

The Kalman filter is an optimal estimator which can be used for prediction, filtering and smoothing. The standard Kalman filter can deal only with linear systems, but it has been extended by researchers to provide a solution to nonlinear problems, e.g. the extended Kalman filter and the Unscented Kalman filter [Julier and Uhlmann, 1997]. It is often difficult to get the best implementation of Kalman filtering for a real system due to the difficulty of obtaining good estimates of the system and measurement noise covariances. To enhance performance, a machine learning technique such as the Expectation Maximisation (EM) algorithm [Dempster et al., 1977] can be applied to optimise the initial parameters for Kalman filtering.

5.3 Glottal LF-model Parameter Multi-estimate Fusion

It has been shown in our previous experimental results that different algorithms (glottal source extraction and LF-model fitting) result in different estimates. As there is no a priori information about the true glottal source parameter values, it is reasonable to regard different algorithms as individual sensors and a more reliable estimate can be obtained by combining the measurements from these sensors. In this section, a general framework for glottal LF-model parameter multi-estimate fusion is introduced and the main factors affecting the performance of the fusion algorithm are discussed.

5.3.1 Multi-estimate Fusion Framework

The proposed fusion framework is depicted in Fig. 5.4 and the details are described in the following sections.

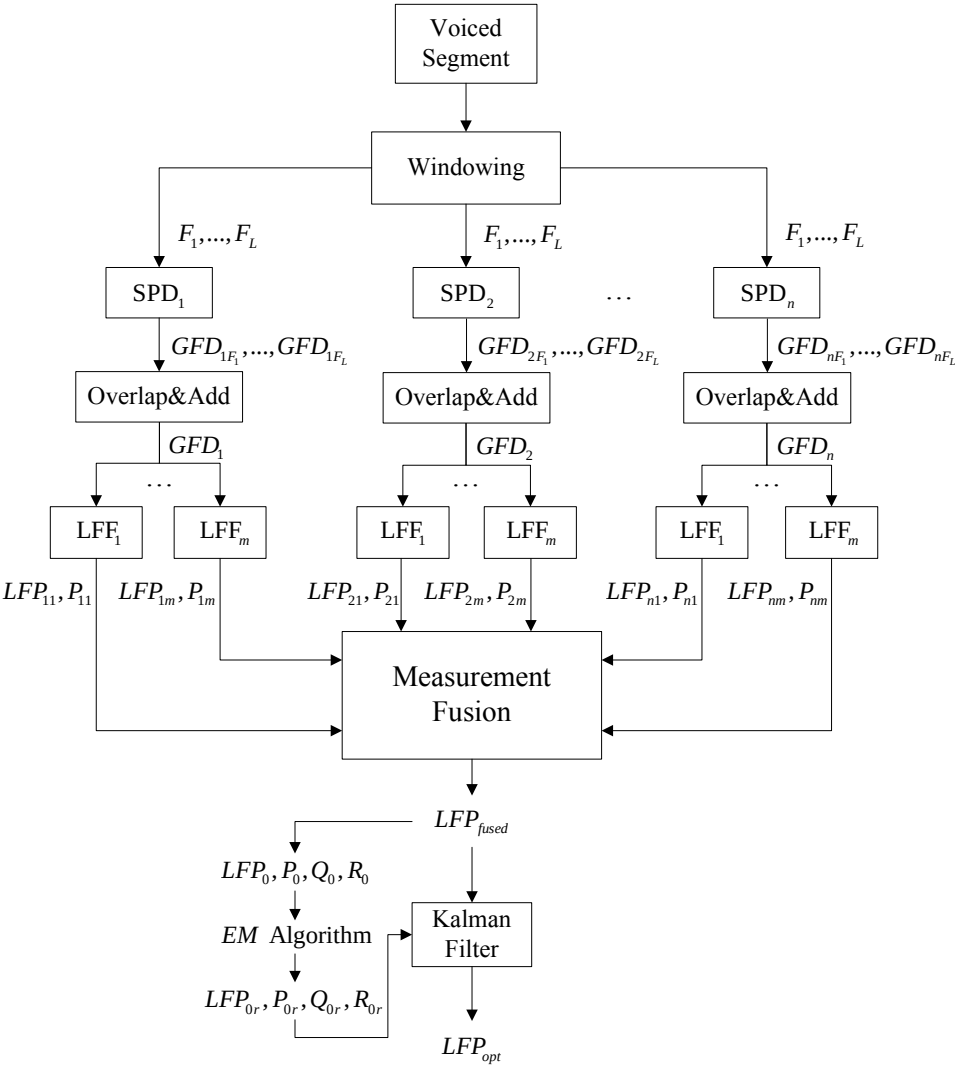

Figure 5.4: A general framework of the multi-estimate fusion algorithm

5.3.1.1 Multiple Glottal Source Extractions

A voiced speech segment is firstly divided into overlapping frames F_1,...,F_L. Next, two or more speech decomposition (SPD) algorithms are applied to each frame in parallel to extract multiple sets of glottal estimates. These algorithms may include Linear Prediction-based glottal inverse filtering techniques [Alku, 1992; Childers et al., 1995; Moore and Clements, 2004] or any other source-vocal tract separation methods [Bozkurt et al., 2004a; Walker and Murphy, 2005, 2007; Drugman et al., 2009b]. Each SPD algorithm separates the speech signal into glottal source and vocal tract components. With the extracted vocal tract coefficients the speech signal can be inverse filtered to cancel the vocal tract effect. Once the vocal tract component is removed the glottal flow derivative is obtained. For each of the overlapping frames the corresponding inverse filtered GFD signals obtained from the same SPD algorithm are concatenated by an overlap-and-add procedure and the outputs of this stage are n GFD (where n is the number of speech decomposition algorithm applied) signals for the original voiced segment.

5.3.1.2 Multiple LF-model Fitting algorithms

Next the glottal source parameters are estimated. Because of its effectiveness for approximately 83% of natural phonations [Strik and Boves, 1992], the LF-model [Fant et al., 1985] is currently utilised in the proposed framework for representing the glottal source. Each GFD signal is divided into consecutive pitch periods according to the initial estimated glottal opening instants [Airas, 2008]. Subsequently, one or more LF-model fitting (LFF) algorithms are applied to each pitch period of the GFD signal. An LFF algorithm is used to estimate the glottal LF-model parameters (LFP) by fitting the LF-model to the GFD estimate. Given n speech decomposition algorithms and m LF-model fitting algorithms, a total of $n \times m$ sets of LF-model parameter estimates are obtained for each pitch period. In addition, for each set of estimated LF parameters, an error covariance P is calculated from the fitting error between the reconstructed LF pulse and the GFD signal.

5.3.1.3 Multiple Estimates Combination

At this stage, the fusion procedure is applied. For a single pitch period the $n \times m$ sets of estimated LF-model parameters are combined by the generalised Millman's fusion formula (introduced by equations (5.12) - (5.15)) given as follows:

$$LFP_{fused} = a_{11}LFP_{11} + \cdots + a_{ij}LFP_{ij} + \cdots + a_{nm}LFP_{nm} \tag{5.18}$$

$$a_{11} + a_{12} + \cdots + a_{ij} + \cdots + a_{nm} = 1 \tag{5.19}$$

$$a_{ij} = \frac{1}{P_{ij}}(\frac{1}{P_{11}} + \cdots + \frac{1}{P_{ij}} + \cdots + \frac{1}{P_{nm}})^{-1} \tag{5.20}$$

where $i = 1, ..., n$, $j = 1, ..., m$, and a_{ij} is the weighting factor of the corresponding set of LFP, which is calculated from the error covariances P_{ij}. It can be observed that the smaller the covariance of a particular estimate, the more weight is given to that estimate. The measurement fusion procedure is applied across all pitch periods in the speech signal to produce a single, fused set of LF-model parameter estimates.

5.3.1.4 Fused Estimates Smoothing

In order to obtain reliable parameter trajectories, it is necessary to smooth the fused LF-model parameters across all pitch periods. It is reasonable to assume limited variation in glottal source parameters across adjacent pitch periods especially for sustained vowel sounds. It has been shown that the variation of LF-model parameters can be regarded as a linear process [Tooher and McKenna, 2003]. Thus, assuming that the true glottal source parameters are the system state to be tracked, and the fused LF-model parameter estimates are the measurement, the corresponding state-space process and measurement equations can be described by equations (5.21) and (5.22):

$$rLFP_k = \Phi rLFP_{k-1} + w_k \tag{5.21}$$

$$estLFP_k = rLFP_k + v_k \tag{5.22}$$

where $rLFP$ is the vector of true glottal source parameters for estimation, $estLFP$ is the vector of fused voice source parameter estimates, Φ is a parameter controlling the linear variation of the glottal source parameters, w and v are the process and measurement noise respectively, with white Gaussian distributions $p(w) = N(0, Q)$ and $p(v) = N(0, R)$.

The above equations are based on the following two premises: 1) that the variation of the voice source parameters across adjacent pitch periods is small, so that the true glottal source parameters of the pitch cycle can be represented by source parameters of the pitch cycle plus process noise; 2) that the estimated source parameters can be considered as a summation of true parameter values and measurement noise. With such correlation, it is reasonable to use a Kalman filter (KF) to track the glottal source parameters. The KF performs best when the process and measurement noise covariances are exactly known. However, for real speech signals there is no such a priori information available. The expectation-maximisation (EM) algorithm [Dempster et al., 1977] is a machine learning technique for optimisation by iteratively adjusting the estimates to maximise the corresponding log-likelihood. Here in our fusion framework, implementation of the EM technique by Shumway [Shumway and Stoffer, 1982] is utilised to refine the KF parameters. The details and equations of the EM algorithm are discussed and listed below.

Firstly, the EM algorithm involves a backward smoothing procedure [Rauch et al., 1965], the corresponding equations are given below

$$\begin{aligned}
J_{n-1} &= P_{n-1}\Phi_n(P_n^-)^{-1} \\
x_{n-1}^N &= x_{n-1} + J_{n-1}(x_n^N - \Phi_n x_{n-1}) \\
P_{n-1}^N &= P_{n-1} + J_{n-1}(P_n^N - P_n^-)J_{n-1}^T \\
P_{n,n-1}^N &= P_n^N J_{n-1}^T + J_n(P_{n+1,n}^N - \Phi_{n+1}P_n^N)J_{n-1}^T
\end{aligned} \tag{5.23}$$

where J_{n-1} is a smoothing gain, x_{n-1}^N and P_{n-1}^N are the smoothed backward estimate and error covariance, respectively. $P_{n,n-1}^N$ is the lag-1 estimate error covari-

ance for $n = N-1, N-2, ..., 1$. For $n = N$,

$$P_{N,N-1}^{N} = (I - K_N H_N)\Phi_N P_{N-1}^{N-1} \tag{5.24}$$

Subsequently, three terms are calculated as

$$\begin{aligned} A &= \sum_{n=1}^{N} [P_{n-1}^{N} + (x_{n-1}^{N})(x_{n-1}^{N})^T] \\ B &= \sum_{n=1}^{N} [P_{n,n-1}^{N} + (x_{n}^{N})(x_{n-1}^{N})^T] \\ C &= \sum_{n=1}^{N} [P_{n}^{N} + (x_{n}^{N})(x_{n}^{N})^T] \end{aligned} \tag{5.25}$$

which are used to reestimate the Kalman filters for the next iteration after the rth iteration:

$$\begin{aligned} &\Phi(r+1) = BA^{-1} \\ &Q(r+1) = \tfrac{1}{N}(C - BA^{-1}B^T) \\ &R(r+1) = \tfrac{1}{N}\sum_{n=1}^{N} [(z_n - H_n x_n^N)(z_n - H_n x_n^N)^T + H_n P_n^N H_n^T] \\ &x_0(r+1) = x_0^N \end{aligned} \tag{5.26}$$

The log-likelihood function of each iteration's forward pass to the end of the data is calculated by

$$\begin{aligned} \log L = &-\tfrac{1}{2}\sum_{n=1}^{N} \log \left| H_n P_n^{n-1} H_n^T + R_n \right| \\ &-\tfrac{1}{2}\sum_{n=1}^{N} (z_n - H_n x_n^{n-1})^T (H_n P_n^{n-1} H_n^T + R_n)^{-1} (z_n - H_n x_n^{n-1}) \end{aligned} \tag{5.27}$$

To integrate the EM algorithm into the fusion framework, x is the LF-model parameter to be optimised, z is the fused estimate, H is a constant of value 1. The initial KF parameters are empirically selected as: $\Phi = 1, Q = 1e{-}5, R = 0.01$ and $x_0 = mean(x)$. The EM algorithm stops at the convergence of the log-likelihood function, or the maximum number of iterations, which is set to 300 in this study, is reached. Afterwards, with these re-estimated parameters the Kalman filtering is applied to estimate the optimal glottal source parameters across pitch periods

of the full voiced segment.

5.3.2 Advantages and Limitations

Compared to a single glottal source parameter estimation algorithm, the fusion algorithm has the following **advantages**:

1) **More reliable**. Based on quantitative data fusion technology, the proposed fusion algorithm combines estimates obtained from multiple glottal source extraction and LF-model fitting methods and outputs an overall optimal set of estimates. Thus the fusion approach is more reliable than applying a single algorithm to estimate the glottal source parameters.

2) **Convenient for extension**. Instead of inventing another new algorithm, the fusion algorithm is basically a framework that intelligently integrates different voice source estimation algorithms. Thus, it is convenient to add new algorithms to the framework keeping other components unaffected.

3) **Fusion centre updating**. The fusion algorithm has a clear structure across all functional levels. Thus, it is not difficult to update the fusion centre if different fusion formulas and fusion rules are proved to be more effective.

4) **High flexibility**. The local glottal source extraction and LF-model fitting algorithms can be regarded as individual sensors running in parallel. It is possible to open or close certain "sensors" to satisfy the requirements of different applications.

5) **Optimal estimation for an individual algorithm**. Although the fusion algorithm is described as combining estimates from different glottal source estimation methods, in fact it also works for an individual algorithm with different configurations. Taking the iterative adaptive inverse filtering for example, if different filter orders or pre-emphasis values are utilised, the extracted glottal waveform can be quite different. Generally, higher filter orders result in a lower least squares error. However, it is possible that a lower analysis order is in fact more appropriate. To deal with this situation, we can utilise multiple error criteria in the fusion framework, such as the fitting errors for both the waveform and the spectrum, and some measures to evaluate the quality of the glottal estimate, to decide which set of configuration should have more weight.

The proposed multi-estimate fusion algorithm has three main limitations:

1) **High computational load**. By applying multiple algorithms to extract the glottal source component and fit the LF-model to it in parallel, the fusion approach has a higher computational load compared to utilising a single method.

2) **Contribution from poor estimates**. The fusion approach tries to combine the estimates obtained from a series of local algorithms; it guarantees that the fused estimates are globally optimal. However, estimates from a local algorithm although having the best performance for a specific segment of speech are weighted and combined with estimates from all other algorithms make contributions. This is a limitation for all data fusion approaches: all estimates, even poor ones, make some contribution.

3) **Inaccurate glottal source modelling**. In the proposed fusion algorithm, we utilise the LF-model to track the shape of the glottal source waveform because of its effectiveness for most voiced sounds. However, there are some characteristics of the glottal source that the LF-model cannot model, such as source and vocal tract interaction. Such information cannot be captured by the fusion method. To solve this problem, a more complex and accurate glottal source model is required.

5.3.3 Factors Affecting the Performance of the Proposed Fusion Algorithm

Before implementing and evaluating the glottal LF-model multi-estimate fusion approach, it is necessary to discuss the possible factors that might affect the algorithm's performance. The main factors are listed below:

• **Quality of the recorded speech**. Obviously, the cleaner the speech recordings, the easier to extract the glottal source parameters accurately. On the contrary, if the speech signal has a low SNR level, the characteristics of the speech and its corresponding glottal waveform and spectra will be distorted, resulting in an inaccurate source parameter estimation.

• **Glottal source extraction quality**. The glottal source extraction methods are the 'sensors' in the fusion framework to generate the measurements of the glottal source components. Thus, outputs of these 'sensors' may affect the performance of the fusion algorithm. It is reasonable to consider that if all the

selected algorithms perform well the fused estimate will be accurate. On the contrary if all approaches generate poor estimates the combined estimate will be less accurate, although one poorly performing local algorithm will not significantly affect the performance if other approaches are reliable.

• **LF-model fitting quality**. A LF-model fitting approach is responsible for estimating the shape parameters of the glottal flow derivative. For a similar reason to that discussed above, outputs of such methods are also important in determining the effectiveness of the fusion approach, since they directly generate the observations to be combined in the fusion centre.

• **Error criterion**. The multi-estimate fusion approach is a fully automatic method to estimate the voice source parameters. Based on no a priori information of the real source parameters and the goodness and poorness of the algorithms combined at the fusion centre, it is critical to select an appropriate error criterion as the numerical measure of goodness of estimate. A reasonable error measure is the fitting error covariance, which can be calculated from the fitting error between the re-constructed LF-model pulses from a particular set of estimates and the original inverse filtered glottal flow derivative signals. Such covariance values can be utilised by the generalised Millman's formula as the error covariance.

• **Duration of the voiced speech segment for analysis**. The length of the voiced speech frame to be analysed affects the effectiveness of the smoothing procedure. Each pitch period of the glottal flow derivative corresponds to a fused set of LF-model estimates. If the speech segment contains a sufficient number of pitch periods, the smoothing procedure can generate optimal trajectories of the LF parameters; however, if the number of pitch periods of the GFD is insufficient for the smoothing procedure, the estimated parameter trajectories may be of poor quality.

5.4 Conclusion

The main contribution of this chapter is the introduction of a general glottal LF-model multi-estimate fusion framework. To combine the multiple estimates from different source estimation and fitting algorithms, it is necessary to understand the basic idea and techniques of quantitative data fusion. State-vector fusion,

measurement fusion, generalised Millman's fusion formula and Kalman filtering were introduced. The framework for the multi-estimate fusion was presented and each of the functional levels in the framework was fully detailed. The advantages of the fusion algorithm over a single approach were presented and potential limitations were discussed. In the next chapter we describe the implementation and evaluation of the proposed fusion framework.

Chapter 6

Multi-estimate Fusion Evaluation

6.1 Introduction

The aim of this chapter is to evaluate the multi-estimate (ME) fusion algorithm proposed in Chapter 5. Firstly, in Section 6.2 the ME-fusion approach is implemented with three glottal inverse filtering methods and one time-domain LF-model fitting algorithm, all of which were introduced previously. To combine the estimates from different approaches, it is necessary to synchronise the estimates according to pitch cycles. To solve the problem, a synchronisation by glottal closing instants procedure is then described in Section 6.3.

To test the effectiveness, in Section 6.4 the implemented fusion algorithm is firstly applied to a synthetic voiced segment, for which the true values of the glottal source parameters are known and comparison can be made conveniently. Subsequently, in Section 6.5 the ME-fusion algorithm is applied to utterances selected from the CMU-ARCTIC database. Each sentence is segmented into voiced/unvoiced frames and thus we can observe the estimated LF-model trajectories across a complete utterance by local individual algorithms and the fusion approach.

In addition, in Section 6.6 several voiced frames are extracted from the utterances used in the test above to study the performance variation across individual algorithms, which is useful for further investigation of glottal source parametrisation by different approaches and improvement of the multi-estimate fusion al-

gorithm.

In a further evaluation in Section 6.7, the ME-fusion method is applied to an all voiced utterance spoken by different speakers for which hand-labelled "gold standard" LF-model estimates exist in order to provide numerical evidence for the validity of the fusion approach.

In Section 6.8, a final evaluation is carried out, to measure the effect of adding another algorithm to the framework, where the ME-fusion algorithm is extended by adding an additional LF-model fitting methods. In one experiment a poorly performing frequency-domain LF-model fitting approach is integrated in the fusion framework to test its effectiveness to cope with poor estimates. In a second experiment a well performing hybrid TD/FD fitting approach is incorporated. The experimental results from the two new versions of the ME-fusion approach are presented and discussed.Finally, a critical assessment of the the fusion approach is made in Section 6.9.

6.2 Implementation of the Fusion Algorithm

The flow chart of our implementation of the multi-estimate fusion algorithm is shown in Fig. 6.1. The details of this implementation are described below with a brief review of the relevant source parameter estimation approaches.

The input voiced speech segment is divided into individual frames of length 40ms with 50% overlap. Each frame is processed by three glottal inverse filtering speech decomposition approaches which were introduced and described in Chapter 3. Here we give a review of these algorithms as follows:

• The first algorithm is the iterative adaptive inverse filtering (IAIF) [Alku and Laine, 1989; Alku and Vilkman, 1994], which is based on the assumption that the glottal flow waveform can be represented by a low order all-pole model. The IAIF algorithm operates by repeatedly estimating and removing the glottal and radiation effects using low order Linear Prediction analysis and inverse filtering. This removes the overall spectral tilt of the speech and allows estimation of the vocal tract filter using higher order linear prediction analysis. The estimated vocal tract filter is used to inverse filter the original speech signal to extract the glottal flow derivative (GFD).

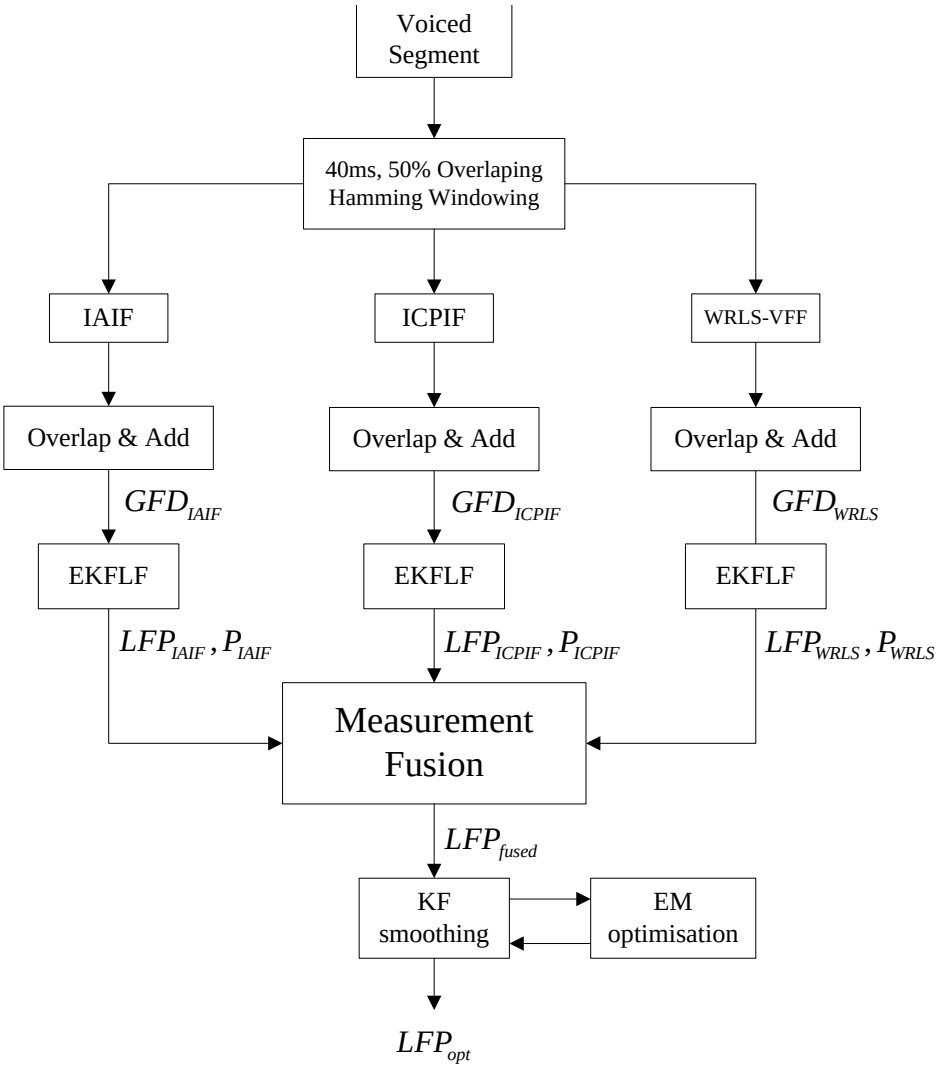

Figure 6.1: Implementation of multi-estimate fusion algorithm

• Iterative closed phase inverse filtering (ICPIF) [Moore and Clements, 2004] is also used in this implementation. Typically, closed phase inverse filtering [Walker and Murphy, 2007] operates on the assumption that for several milliseconds after the glottal closing instant the glottis remains closed and during this time the speech signal is due solely to the decaying vocal tract response. Thus, linear predictive analysis performed across this time interval models only the vocal tract filter and excludes any components due to the glottal source. The glottal waveform can be determined by inverse filtering the entire pitch period with the coefficients obtained from the closed phase. The first-order autocorrelation parameter can be used to measure the smoothness of the estimated glottal source waveform [Moore and Clements, 2004; O'Cinneide et al., 2011b] and an iterative analysis procedure is applied to select the smoothest GFD signal across several estimated closed phase intervals. In our implementation of ICPIF, the initial glottal closing instants are estimated by analysing the variable forgetting factor obtained from the "weighted recursive least squares with variable forgetting factor" (WRLS-VFF) method [Childers et al., 1995].

• WRLS-VFF analysis is the third speech decomposition method. The approach assumes that the speech signal is generated by an ARMA model. WRLS-VFF analysis operates by recursively minimising the prediction error for speech samples and allowing variation of the forgetting factor. During the analysis process, the forgetting factor and the ARMA coefficients are obtained. Generally the maximum prediction error occurs at the glottal closure instant. Accordingly, the model coefficients at the instant of glottal closure can be used to do the inverse filtering.

For each inverse filtering algorithm applied to the speech segment frame by frame, the corresponding consecutive and overlapped GFD frames are obtained. These GFD frames are concatenated to generate the entire glottal flow derivative signal containing all pitch periods corresponding to the original input voiced speech segment. Thus, three inverse filtering approaches result in three sets of GFD signals.

Subsequently, the Extended Kalman filtering LF-model fitting (EKFLF) algorithm introduced in Chapter 5 is applied to the three sets of GFD signals to

estimate the glottal LF-model parameters period by period. For one particular set of LF estimates, the corresponding error covariance P is calculated from the fitting error between the inverse filtered GFD signal and the reconstructed LF-model pulse.

Thus, for each pitch period of the speech signal, three sets of LF-model parameter estimates are obtained. The fusion centre applies the generalised Millman's formula to combine these estimates with the corresponding covariance P and generates a fused set of estimates across all pitch periods. Subsequently, a Kalman filtering with its parameters re-estimated by the EM algorithm described in Section 5.3.1 is applied to smooth the fused parameter trajectories.

6.3 Synchronisation by Glottal Closing Instants

The fusion algorithm applies different speech decomposition approaches to extract multiple sets of glottal estimates. If the estimated glottal components have the same glottal closing instants, representing individual pitch periods, then the corresponding LF-model parameter estimates are ready to be combined. However, glottal estimates obtained by different algorithms may have inconsistent glottal closing instants. This is because the multiple sets of GFD estimates may contain unanalysable frames. A simple example is presented in Table 6.1 for illustration.

Table 6.1: Asynchronous glottal estimates

Pitch cycle	1	2	3	4	5	6	7	8	9	10
$Detected_{IAIF}$	√	×	√	√	√	√	√	√	√	×
$Detected_{ICPIF}$	√	√	√	×	√	√	√	√	×	√
$Detected_{WRLS-VFF}$	√	√	√	√	×	×	×	√	√	√

The first row means that for the true GFD signal there are ten pitch periods numbered from 1:10. The remaining three rows stand for the GFD estimates by three different inverse filtering methods. '√' represents successful glottal pitch

cycle detection and '×' means this pitch period of the glottal component is not extracted by the related algorithm. In this example it can be observed that IAIF failed to generate a GFD estimate at pitch cycles 2 and 10, ICPIF yielded unanalysable GFD estimates for pitch cycles 4 and 9 and WRLS-VFF generated three continuous problematic GFD estimates from pitch periods 5 through 7. A problematic GFD estimate means the glottal closing instant cannot be located.

Subsequently, if we apply the LF-model fitting approach to the three sets of glottal estimates, the corresponding LF parameter (LFP) estimates are not pitch synchronous, which is illustrated in Table 6.2. It is obvious that these estimates cannot be directly combined because they are not in sync with the same glottal pitch cycle.

Table 6.2: Asynchronous glottal estimates

LFP_{True}	1	2	3	4	5	6	7	8	9	10
LFP_{IAIF}	1	3	4	5	6	7	8	9		
LFP_{ICPIF}	1	2	3	5	6	7	8	10		
$LFP_{WRLS-VFF}$	1	2	3	4	8	9	10			

Thus, it is necessary to develop a procedure to deal with the inconsistent multiple glottal estimates problem. The procedure applied is described below:

Step 1 Three sets of GCI estimates (GCI_{IAIF}, GCI_{ICPIF}, and $GCI_{WRLS-VFF}$) are estimated from the three GFD estimates respectively.

Step 2 Calculate the neighbourhood differences for the three sets of GCI estimates, e.g., $diff[i] = GCI_{IAIF}[i+1] - GCI_{IAIF}[i]$ is the difference between the i^{th} and the $(i+1)^{th}$ GCI estimates by IAIF. If the difference is smaller than a threshold (for e.g., $0.3 * T0$), this GCI estimate is considered as invalid and removed from the full set. This step validates the GCI estimates.

Step 3 Synchronise the three sets of GCI and the corresponding LF-model parameter estimates. Here an iterative procedure is applied. The GCI estimates set 1 and 2, set 2 and 3, set 3 and 1, are compared by finding the optimal alignment of the GCIs time markers (sample numbers), respectively. If a certain

set has a smaller number of estimates it means that there are one or more missing estimates due to unanalysable pitch cycles from this set compared to the other set. In such a case, the GCI estimates from the compared set are assigned to this set and the LF parameters are obtained by linear interpolating the previous and succeeding estimates. The checking procedure runs iteratively until all sets have the same number of estimates.

Step 4 T0 values for each pitch period are updated by calculating the differences between adjacent GCI time markers. This step is necessary to ensure the final LF-model parameter trajectories are in sync with the original input speech signal.

Afterwards, the GCI estimates and the corresponding LF parameter estimates are synchronised period by period. The measurement fusion procedure can then be applied to combine the multiple LF-model estimates.

6.4 Evaluation on Synthetic Voiced Segments

This evaluation was carried out as a preliminary performance study. An artificial voiced segment was used for this test and the experimental details are presented below.

To test the validity of the fusion algorithm, a segment of synthetic speech was generated as follows:

1. 50 LF-model pulses were created from a set of LF parameters $t_0 = 0$, $t_p = 0.48$, $t_e = 0.65$, $T_a = 0.035$, $T_0 = 1$.

2. The first 20 were passed through a formant synthesizer for the vowel /AH/ and the last 20 pulses for the vowel /IH/ (thus two sustained vowel segments were obtained).

3. A "coarticulatory" segment was generated by synthesizing the middle 10 pitch periods with line spectral frequencies calculated by linear interpolation from /AH/ to /IH/.

4. The three segments were concatenated.

The multi-estimate fusion algorithm was applied to this synthetic speech segment and the LF-model shape parameters were calculated, which are the open quotient $O_q = (t_p - t_0)/T_0$, the asymmetry coefficient $\alpha_m = (t_p - t_0)/(t_e - t_0)$ and

the return phase parameter T_a, where

$$\begin{aligned} O_q &= (t_p - t_0)/T_0 \\ \alpha_m &= (t_p - t_0)/(t_e - t_0). \end{aligned} \tag{6.1}$$

The root mean square error (calculated by $RMSE(\hat{x}) = \sqrt{1/n \sum((\hat{x} - x)^2)}$, where x is the true value and $\hat{x}$ is the corresponding estimate) of the estimated LF-model shape parameters by each algorithm and by the fusion method are presented in Table 6.3, with the corresponding mean error covariances P_m in the last column. It can be observed that the fusion algorithm shows consistently smaller RMSE scores compared to other methods. For all three LF-model parameters, both IAIF and ICPIF performed well, and a relatively bigger RMSE was generated by WRLS-VFF. It is worth mentioning that more weight was given to the IAIF estimates by the fusion procedure, due to its producing the smallest mean error covariance.

Table 6.3: RMSE scores of LF-model parameters estimated by different algorithms for clean synthetic speech

	O_q	α_m	T_a	P_m
IAIF	0.0298	0.0174	0.0247	0.0317
ICPIF	0.0311	0.0194	0.0289	0.0626
WRLS-VFF	0.0422	0.0205	0.0383	0.0691
ME-FUSION	0.0272	0.0142	0.0239	

In this preliminary performance study, it is observed that the fusion approach can accurately estimate the LF-model parameters for the coarticulatory synthetic speech segment. Based on limited knowledge to produce more realistic synthetic speech, where the source-tract interaction, variation of the glottal source parameters especially for transition sounds and other speaker correlated characteristics should be considered, it is reasonable to test the fusion algorithm on real speech recordings. When generation of synthetic speech could be better controlled, a more comprehensive evaluation should be made to facilitate the evaluation of the

fusion method. In the following sections, the evaluations will made on natural speech signals.

6.5 Evaluation on Utterance Recordings

In this section, several recorded utterances were chosen from the CMU-ARCTIC database [Kominek and Black, 2004] to test the performance of the fusion algorithm. These utterance signals were firstly divided into voiced and unvoiced segments and then analysed by the multi-estimate fusion approach.

The proposed fusion algorithm works only on voiced speech signals having glottal contributions, thus it is necessary to extract the voiced segments from an utterance. Many existing approaches can be utilised for this task including the wavelet transform [Janer et al., 1996; Erelebi, 2003], bispectrum analysis [Wells, 1985], LPC distance measure [Rabiner and Sambur, 1977b,a], pattern classification [Siegel and Steiglitz, 1976; Siegel, 1979] and zero-crossing rate and energy [Bachu et al., 2010]. In our work, a robust pitch tracking approach (RAPT) [Talkin, 1995] implemented as a function in the VOICEBOX toolbox [1] is applied to identify and extract individual voiced segments, since it is open source and convenient for integration.

Three sentences spoken from each of *bdl* and *slt* speakers were selected giving six utterances for analysis. The sentences were:

1. Robbery, bribery, fraud.
2. Shall I carry you.
3. He did not rush in.

For each utterance, all voiced segments were identified by the RAPT algorithm and for each segment the multi-estimate fusion algorithm was utilised to estimate the LF-model parameters. To make a fair comparison with the fusion algorithm, estimates from individual algorithms were smoothed by the same Kalman smoothing procedure as used in the fusion method. Estimates by all approaches were concatenated and padded with zeros for unvoiced frames to generate the LF parameter trajectories for the full utterance. The experimental results are presented

[1] available from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox. html

in Fig. 6.2 - Fig. 6.7, where the six utterances are marked as M1, F1, M2, F2, M3, F3, and the voiced segments of each utterance are named by V1, V2,...,Vn. Detailed analysis of the results is presented below.

M1 Analysis For the utterance M1, there are five voiced segments V1-V5. For V1-V4, all approaches generated similar trajectories of T_p and T_e. The duration of the four segments are relatively short and consequently the performance of the smoothing procedure can be affected by outliers, which is observed from the rapidly increasing trajectory for T_a of V2 by IAIF. For V5, IAIF resulted in smoother trajectories for the whole segment, although an outlier of the Tp estimate can be observed. However, the last several pitch cycles contain the transitions and generally there should be variations of the glottal source parameters and thus estimates by ICPIF and WRLS-VFF might be more accurate. Compared to ICPIF and IAIF, T_p and T_e estimated by WRLS-VFF are smaller. The fused trajectories are most similar to the ICPIF ones.

F1 Analysis For F1, there are three voiced segments (V1-V3) according to the automatic segmentation procedure. V1 covers the entire word "robbery" and among the three glottal inverse filtering approaches, WRLS-VFF achieves the most consistent LF-model parameter estimates. By combining the three sets of estimates, the fusion algorithm gives smoother T_p and T_e trajectories than IAIF and ICPIF while exhibiting decrease of T_p at the transition sound. In addition, it can be observed an increase of T_a at the transition from /rɔ/ to /b/ for all approaches. V2 consists of the phonemes in the word "bribery" and the estimated LF-model parameter trajectories are less consistent for ICPIF. The fused estimates show consistent trajectories since more weight is given to the IAIF and WRLS-VFF estimates. The segment V3 contains the /rɔ/ segment and the corresponding source parameter trajectories by all methods are relatively smooth except the T_e estimates by WRLS-VFF which are noisier across the segment.

M2 Analysis The utterance M2 has two voiced segments. V1 contains the phonetic /əl/ from "shall" and "I". From the extracted parameter trajectories of all algorithms we can see that estimated parameters are consistent for the main part of this segment, except for the final pitch periods where there is an increase for T_e and T_a. ICPIF and WRLS-VFF generated similar results although it is evident that estimates by the fusion method are more similar to ICPIF. V2 covers

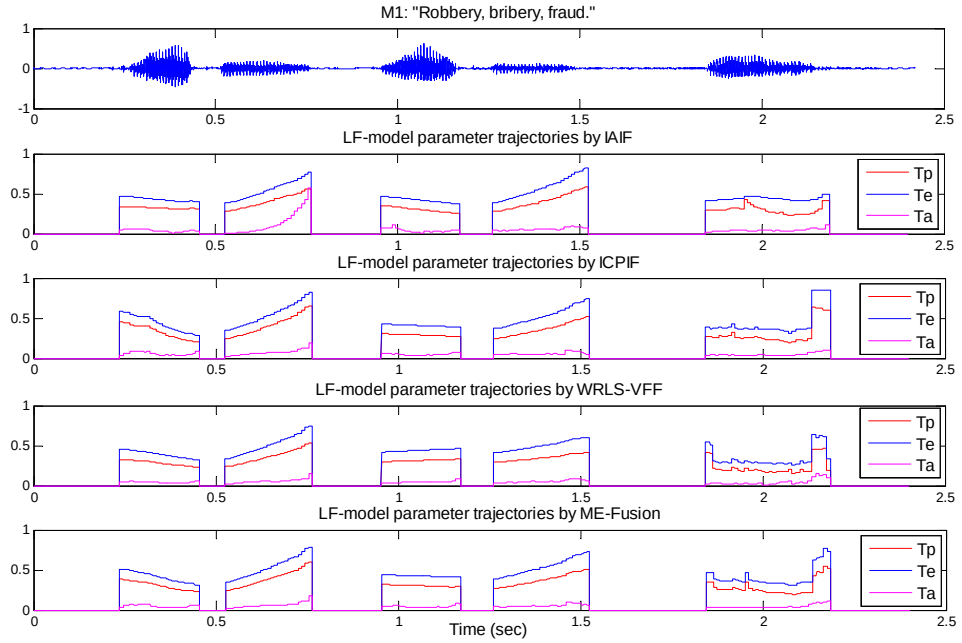

Figure 6.2: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the male utterance “Robbery, bribery, fraud.”

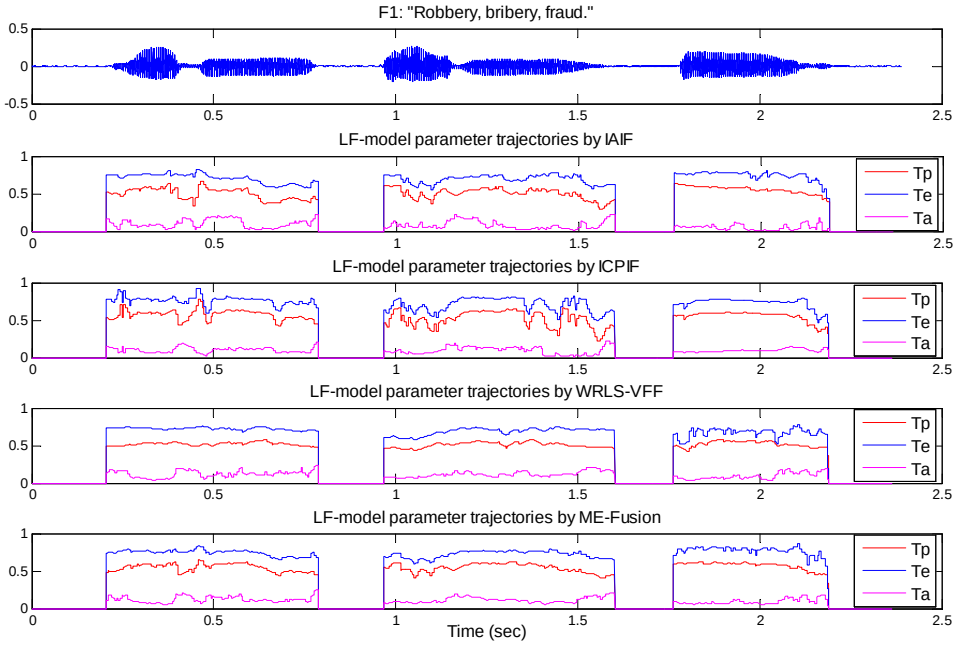

Figure 6.3: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the female utterance “Robbery, bribery, fraud.”

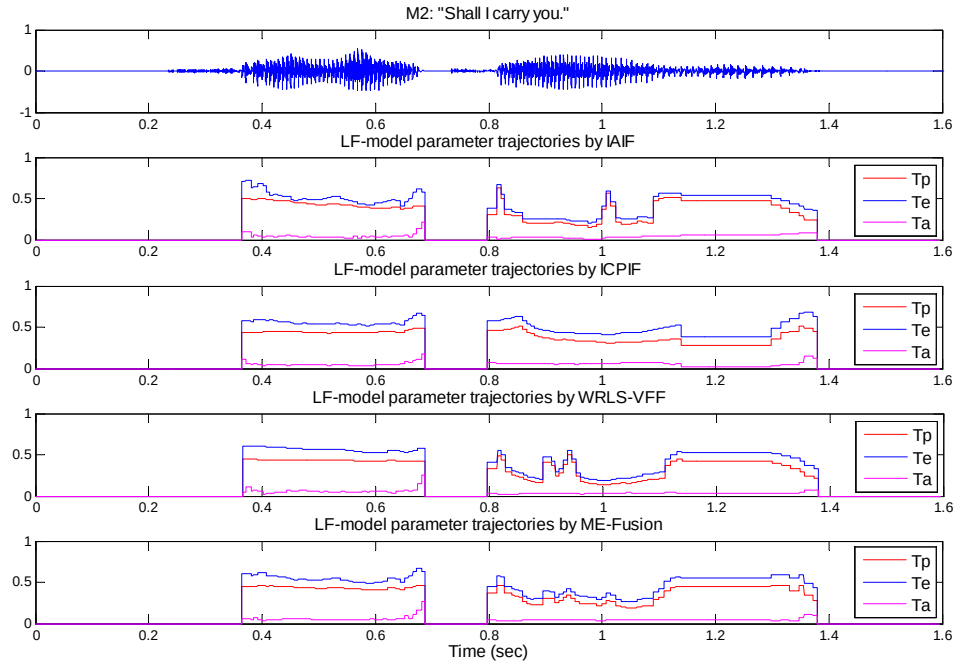

Figure 6.4: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the male utterance “Shall I carry you.”

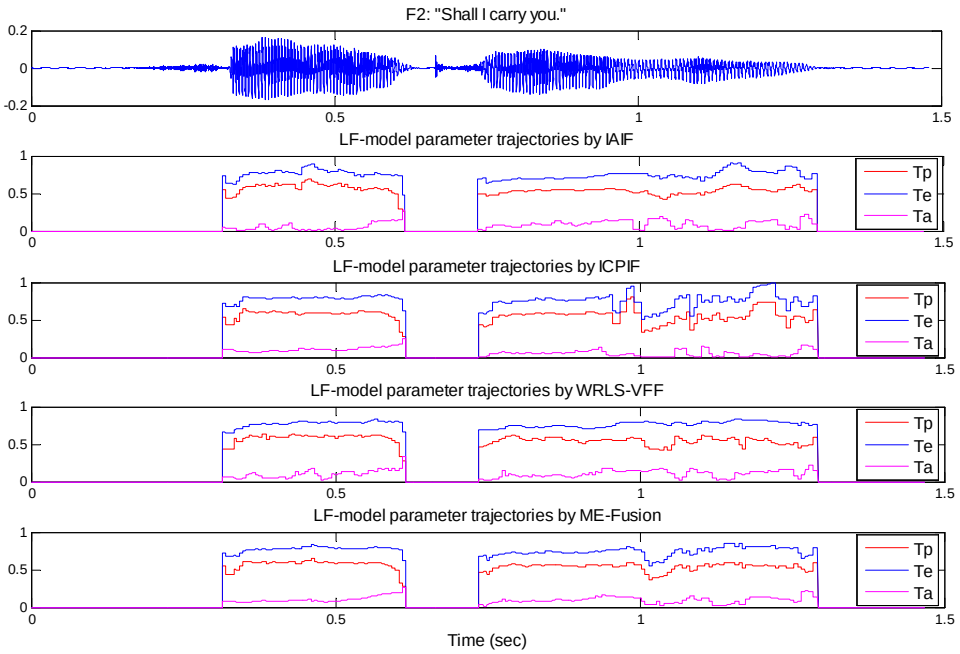

Figure 6.5: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the female utterance “Shall I carry you.”

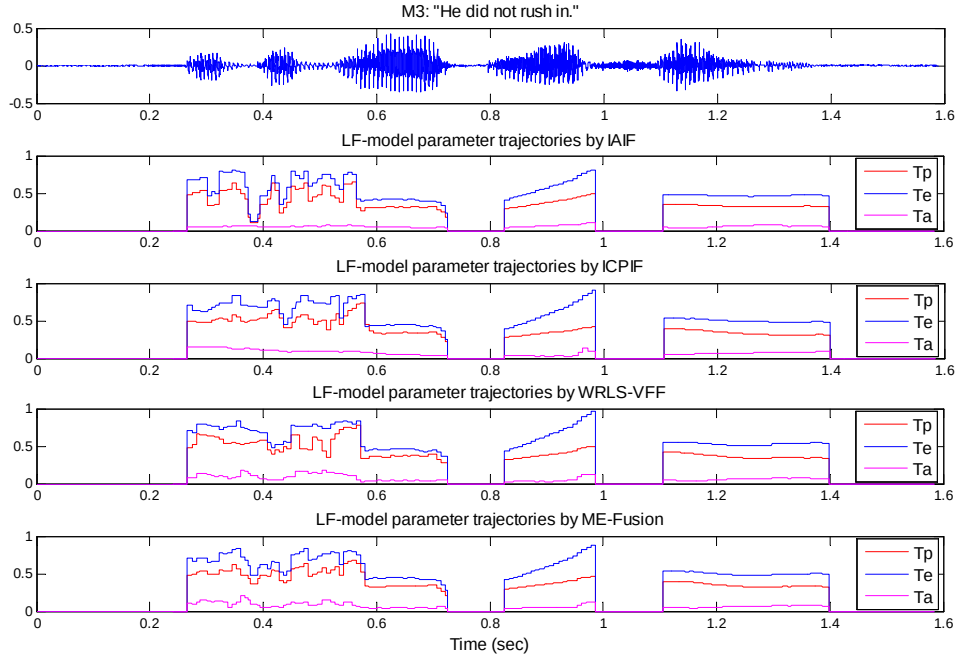

Figure 6.6: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the male utterance "He did not rush in."

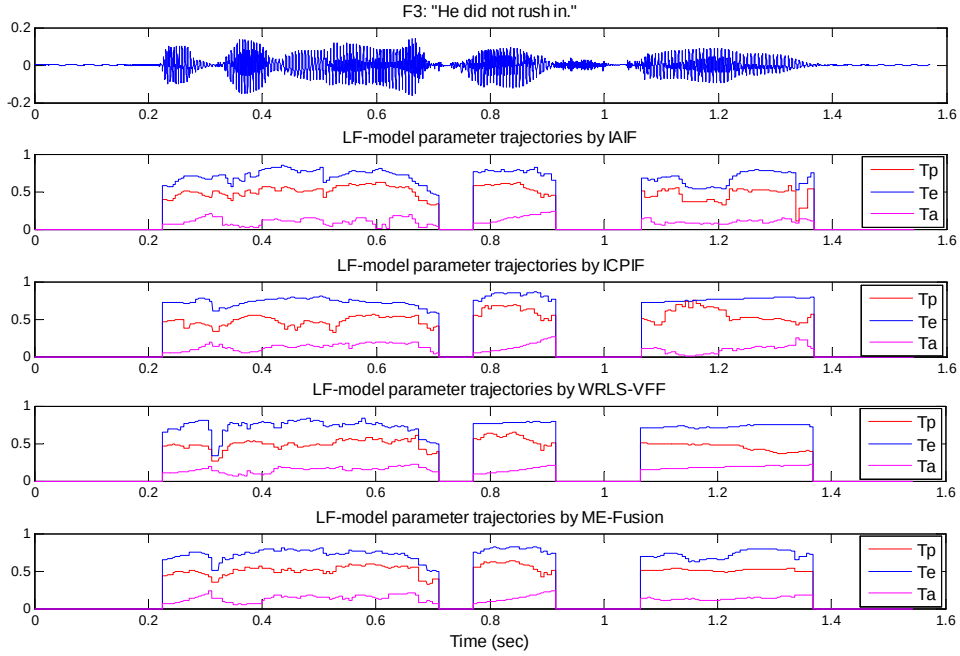

Figure 6.7: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for the female utterance "He did not rush in."

the sound /æri/ from "carry" and the entire word "you". There are visible outliers of the T_p and T_e estimates by IAIF and WRLS-VFF. Estimated trajectories by the fusion algorithm are similar to the IAIF and WRLS-VFF ones but more consistent.

F2 Analysis Similar to M2, the female utterance F2 also has two voiced segments. For V1, the estimated LF-model parameter trajectories by all the four approaches are quite smooth except that at the end there is a decrease of T_p and an increase of T_a. For V2, IAIF and WRLS-VFF generate more consistent estimates compared to ICPIF. Also, it is visible that the fused estimates draw more weight from the estimates by IAIF and WRLS-VFF for the similar trajectories.

M3 Analysis Three voiced segments (V1-V3) are extracted from utterance M3. V1 is a long segment containing the phonetic /i/ in "he", and the words "did" and "not". For this segment, it is visible that ICPIF and WRLS-VFF perform better than IAIF by generating relatively smoother trajectories. Consequently, estimates from these two algorithms make more contribution to the estimates by the fusion method. For V2, which is the sound /rʌ/ from "rush", all approaches show an increasing trend for all the three LF-model parameters especially for T_e, which is probably caused by the large T_e estimates by the three algorithms at the end of this segment. For the voiced segment V3, estimated trajectories by all methods are similar, although the fused estimates were mainly contributed by ICPIF and WRLS-VFF.

F3 Analysis The last utterance F3 has three voiced segments, which is the same as M3. For V1, WRLS-VFF generates relatively smooth parameter trajectories except for the first transition and it is visible that these estimates contribute more to the fused estimates. The estimated trajectories for V2 by all algorithms are quite similar, although the fusion algorithm resulted in estimates closer to ICPIF. For V3, the fused trajectories are comparably consistent to WRLS-VFF and retained some properties of IAIF estimates.

From the experimental results above, we can see the performance of the tested algorithms vary across different speech signals. It cannot be concluded that one algorithm always outperforms the other two methods. In fact, variation of performance can be observed across different segments and pitch periods by different approaches. The performance of IAIF, ICPIF and WRLS-VFF is highly related

to the type of the sustained voiced sound, the length of the transition sound, the consistency of continuous pitch cycles.

The proposed multi-estimate fusion algorithm makes combinations of the estimates obtained from the three individual algorithms and can generate reasonably consistent LF-model parameter trajectories. Because there is no a priori information about the real glottal source, it is difficult to measure the accuracy of the fused estimates numerically. The convergence of the error covariance by Kalman filter can show some degree of confidence of the estimates since the prediction error is getting smaller and smaller. One reasonable way to evaluate the accuracy of the estimates is to re-synthesise the speech signals with the estimates and make perception tests. Also, by applying the fusion approach to a large number of data with respect to individual speakers, some properties of the glottal source parameters could be revealed.

However, according to the assumption of limited variation for the glottal source parameters across continuous pitch periods, it is reasonable to conclude that the corresponding parameter trajectories should exhibit consistency. In addition, it can be observed that the performance of the fusion algorithm depends on various factors such as, the length of the segment, the reliability of the estimates and the corresponding error covariances, the number and the position of the outliers.

6.6 Detailed Analysis of Performance Variation across Individual Algorithms

It is still unclear which algorithm performs the best for arbitrary speech signals because of the complexity and variety of human speech. However, it is worthwhile to extract some speech frames and observe the results obtained by different algorithms. Such analysis may lead to further investigation of glottal source extraction by different approaches and improvement of the multi-estimate fusion algorithm.

Three voiced speech frames from the utterances tested in the last section were extracted and the corresponding waveform, spectrum, estimated GFD signals and

the pole positions by the three inverse filtering approaches are presented in Fig. 6.8 - Fig. 6.10. The details of the analysis are as follows.

In the first example, a voiced frame of sound /ɔ/ in "robbery" was extracted from utterance M1, and the corresponding spectrum, GFD estimates and poles by the three approaches are presented in Fig. 6.8. It can be observed that as a typical vowel sound of this frame, the spectrum shows clear formants at frequencies around 700Hz, 1200Hz, 2100Hz and 3000Hz, where the corresponding poles can be accurately predicted by linear prediction. This can be proved by the clean GFD estimates and pole locations by all the three approaches. GFD estimates by IAIF and ICPIF are quite similar, which is in accordance with the fact that poles for the first four formants by the two algorithms fall into almost the same locations. Although poles for formants of higher frequencies are different, they make relatively small contributions to the inverse filtering procedure. The GFD estimated by WRLS-VFF shows similarity to the estimates by the other two approaches but is more noisy. This is caused by a poorly estimated pole around 1300Hz which increases the bandwidth of the second formant.

A frame of sound /i/ in "robbery" was chosen from utterance M1 as the second example. The waveform, spectrum and the estimates are shown in Fig. 6.9. IAIF works firstly by removing the glottal effect and subsequently applying the linear prediction technique to estimate the vocal tract filter coefficients. For this frame of speech signal, the formants are less clear in the spectrum plot, and the first formant has a large bandwidth from the spectrum, which may lead to a poor estimate of the glottal pole. In such a case, the estimated vocal tract parameters might be inaccurate. This can be seen from the poles and zeros plot by IAIF, where the first two poles have large bandwidth values judging by their distances to the unit circle, thus they were incorrectly estimated. For ICPIF and WRLS-VFF, the estimated GFD signals are reasonably good. A possible reason is that ICPIF extracts the vocal tract filter coefficients by utilising only data of the closed phase. If the closed phase interval is sufficiently long and is free from noise, accurate estimates can be obtained; and WRLS-VFF tries to minimise the prediction error recursively and optimal estimates are obtained after convergence of the algorithm. The GFD estimate by ICPIF is smoother than that of WRLS-VFF, which is because the second and third formants were relatively inaccurately

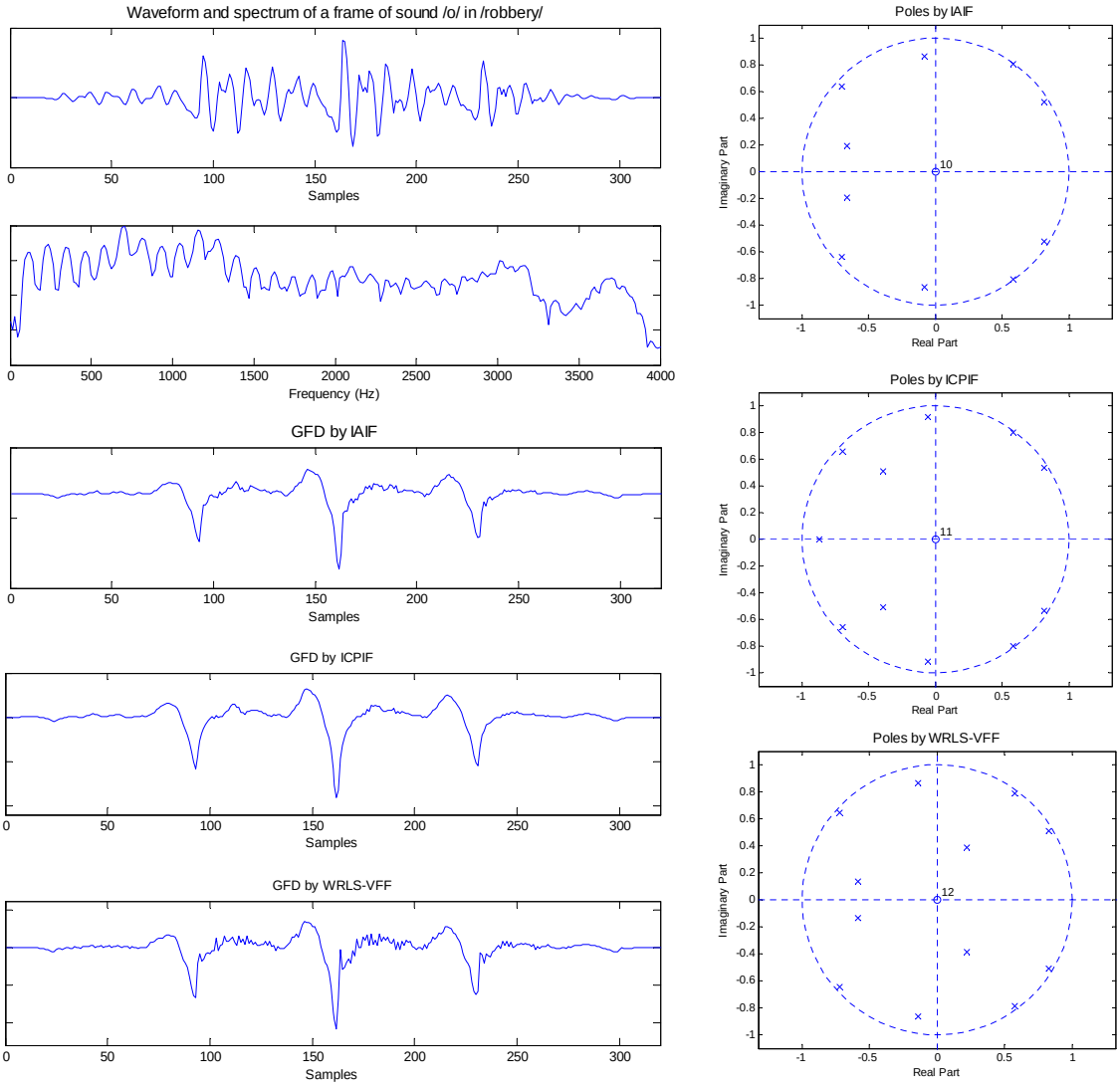

Figure 6.8: Example 1: Waveform, spectrum, estimated GFD and poles by multiple algorithms of sound /ɔ/ in "robbery" by male speaker

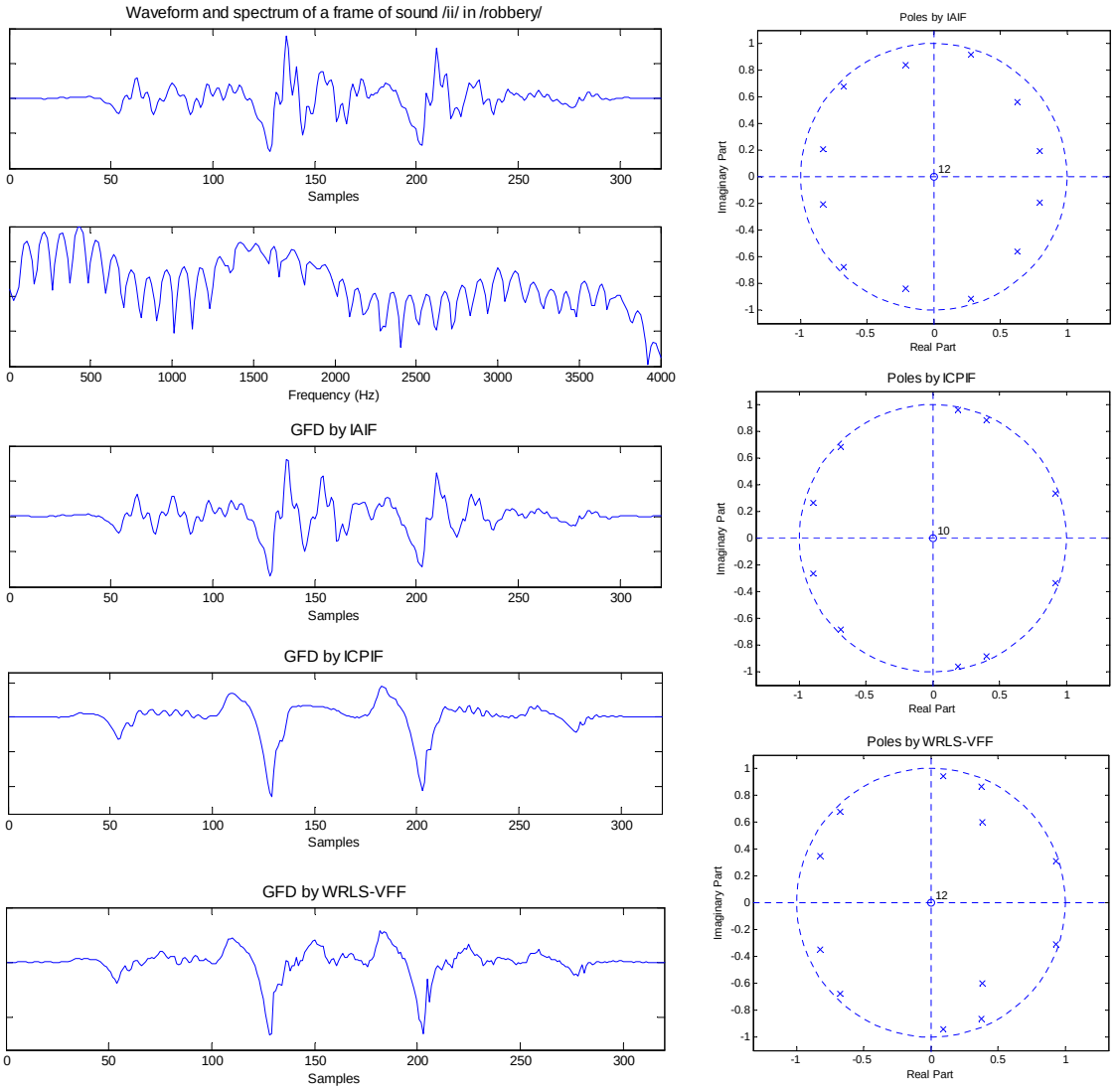

Figure 6.9: Example 2: Waveform, spectrum, estimated GFD and poles by multiple algorithms of sound /i/ in "robbery" by male speaker

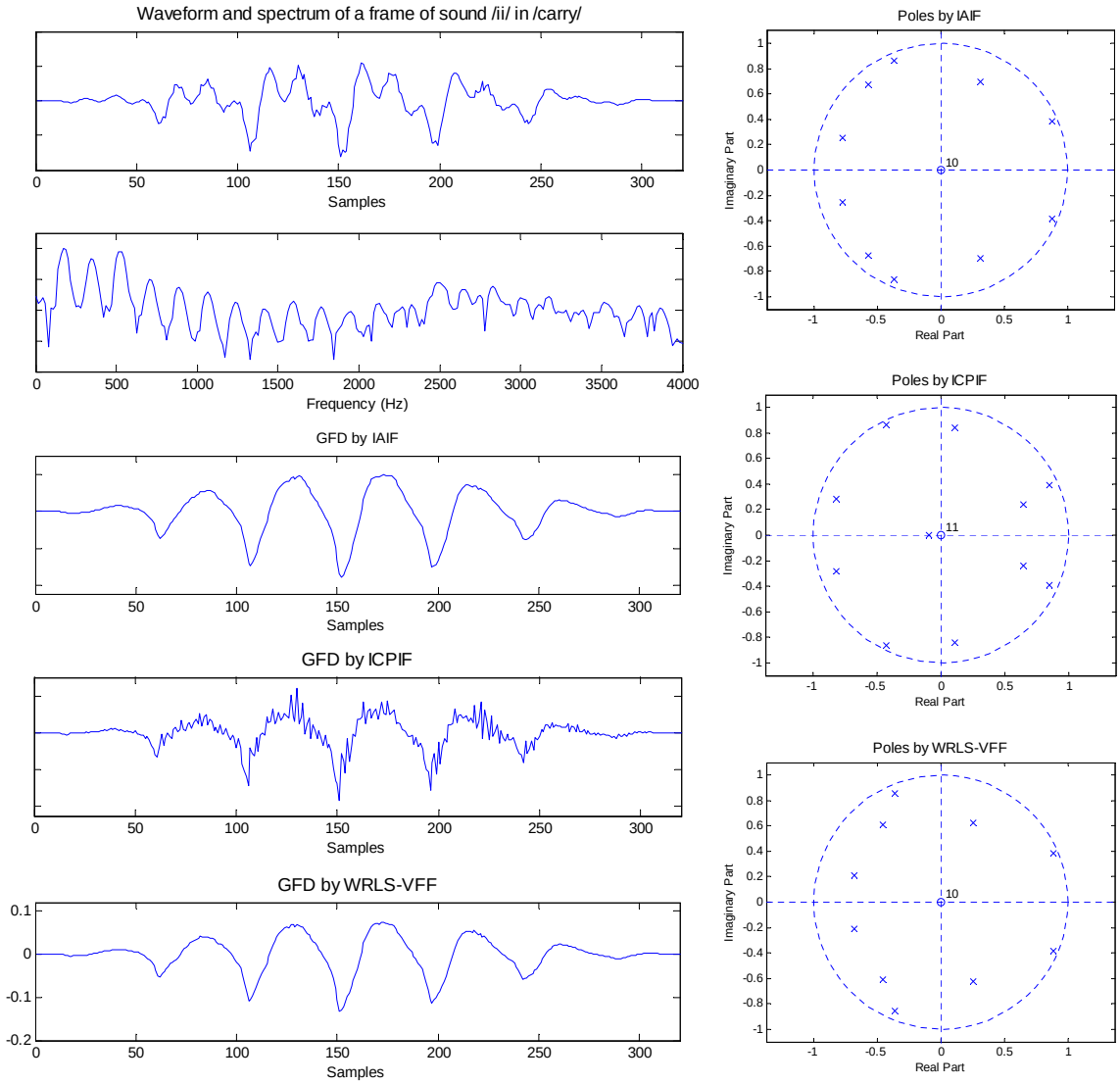

Figure 6.10: Example 3: Waveform, spectrum, estimated GFD and poles by multiple algorithms of sound /i/ in “carry” by female speaker

estimated according to the second, third, and fourth poles locations.

The third frame was obtained from the utterance F2, the speech waveform and the corresponding estimates are presented in Fig. 6.10. It can be observed from the spectrum of the speech signal that the amplitude of the first harmonic is relatively large which has the effect of increasing the bandwidth of the first formant, where the frequency of the first formant is about 500 Hz. For this frame, IAIF and WRLS-VFF provided similar GFD estimates which are clean and consistent for contiguous pitch periods, since it can be seen that the poles, especially those correlated to the first three formants estimated by the two approaches, have a similar distribution. For IAIF, the removal of the glottal effect might help decrease the amplitude of the first harmonic and subsequently the vocal tract parameters can be estimated accurately. For WRLS-VFF it is probably the recursive optimisation procedure sample by sample which enhances the accuracy of the vocal tract parameter estimation. For ICPIF, the corresponding GFD estimate is quite noisy, which is caused by the wrongly estimated poles. The first two poles contribute to a wide bandwidth first formant, the second formant corresponding to the third pole has a frequency about 1900 Hz which is inaccurate and the fourth pole is close to the unit circle which resulted in a narrow bandwidth third formant. The possible reasons for poor performance of ICPIF are the short duration and inappropriate selection of the closed phase of the female speech.

The three examples above show that for different speech by the same speaker, or the same sound by different speakers, the performance varies across multiple approaches. More research is required to fully explore the strengths and weaknesses of each algorithm. We concentrate here on their intelligent combination.

6.7 Evaluation on Hand-labelled Data

In the absence of a priori information of the glottal component, it is difficult to measure the accuracy of the glottal source estimate for a real speech signal. Although it is useful to observe the variation of the estimated parameters, to further demonstrate the effectiveness of the multi-estimate fusion algorithm, it is necessary to test the algorithm against reliable LF estimates. In this section,

the fusion approach is applied to an all-voiced utterance "we were away a year ago" spoken by three speakers and the estimated LF-model parameters by both individual algorithms and the fusion approach are compared with expertly hand-labelled LF parameter data[1]. These estimates were obtained by semi-automatic inverse filtering and LF-model fitting procedures [Gobl and N Chasaide, 1999]. For the inverse filtering procedure, a sliding window is used to automatically and continuously select the voiced frames, the corresponding speech waveform, inverse filtered glottal flow and the spectrum plots are presented. The optimal vocal tract parameters can be obtained by tuning the formant frequencies and bandwidths by the experimenter while observing the resulted glottal derivative waveform and spectrum plots. For the LF-model fitting, each pitch period of the inverse filtered glottal pulse and the corresponding spectrum was plotted, then by marking the timing points and the negative amplitude point in the waveform manually, the LF-model pulse is constructed and fitted to both the waveform and the spectrum. The optimal set of the LF-model parameters is obtained by observing the goodness-of-fit and fine-tunes of the markers. Although time-consuming, these manually estimated LF-model parameters are more reliable then other automatic approaches. Therefore, it is worthy of making comparisons of different algorithms by using the hand-labelled data as a 'gold-standard', where good estimates should be closer to these estimates.

In this test, the ME-fusion method was utilised to estimate the LF-model parameter trajectories for the utterance by three speakers (JK, JD and LP). The speech waveforms and the smoothed LF parameter trajectories obtained by different algorithms and the hand-labelled TCD data are presented in Fig. 6.11, 6.13 and 6.15. Weights applied to each algorithm are plotted in Fig. 6.12, 6.14 and 6.16 and from where the contribution to the fused estimates by different algorithms can be observed. The corresponding LF-model estimates were compared to the hand-labelled data and their root mean squared error (RMSE) scores were calculated and given in Tables. 6.4, 6.6 and 6.8. Also, the mean error covariances and weights are shown in Tables. 6.5, 6.7 and 6.9.

[1]Kindly supplied by Dr. Irena Yanushevskaya of Phonetics & Speech Laboratory, Centre for Language and Communication Studies, Trinity College Dublin, Ireland

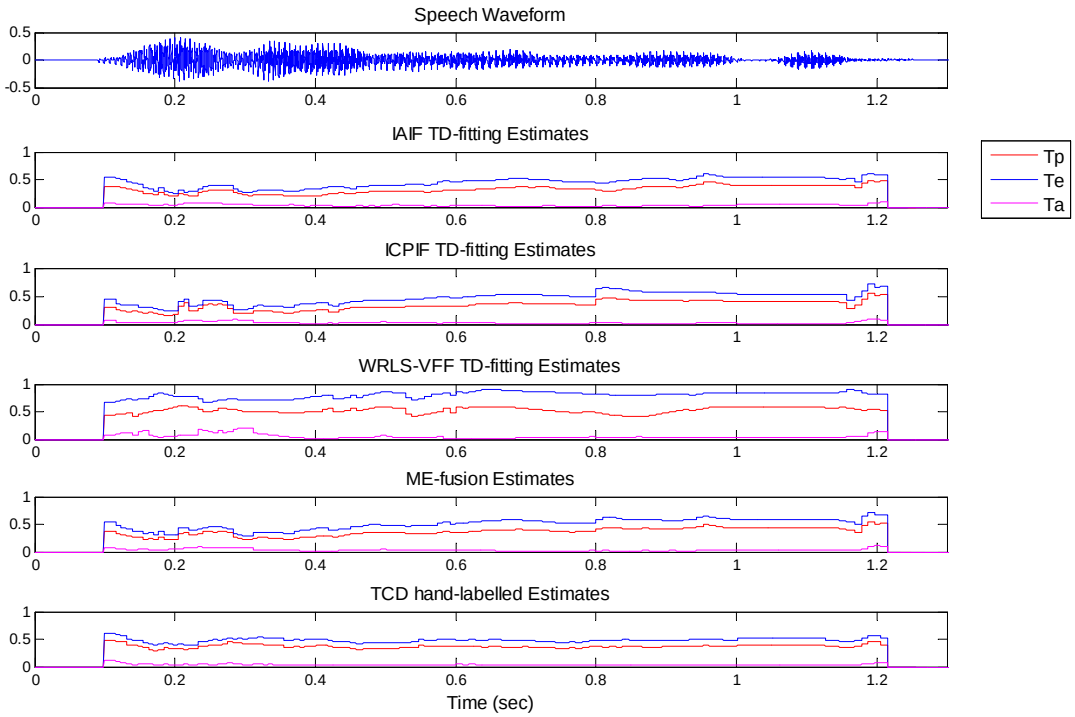

Figure 6.11: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for JK

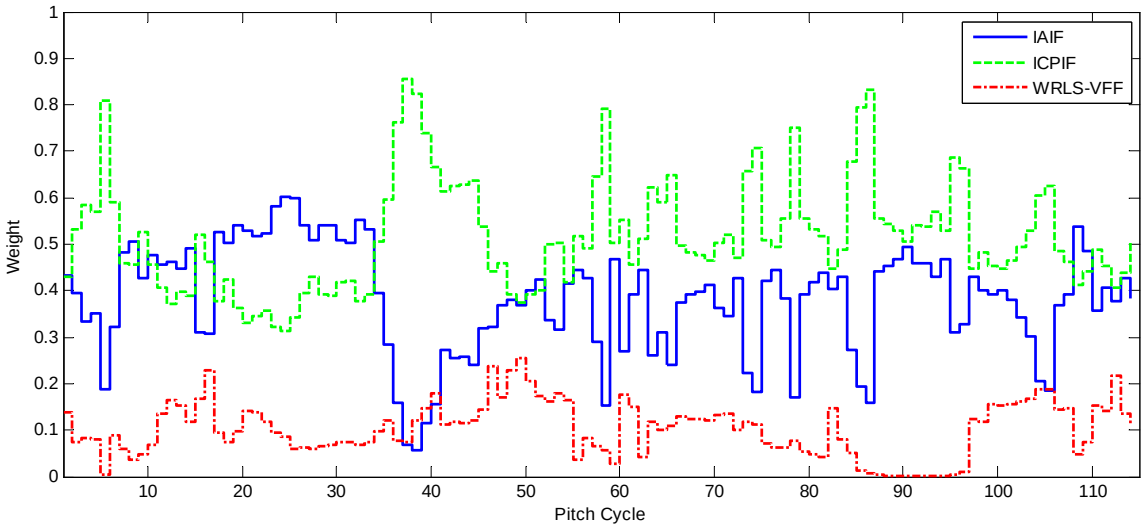

Figure 6.12: Weights by different approaches across JK

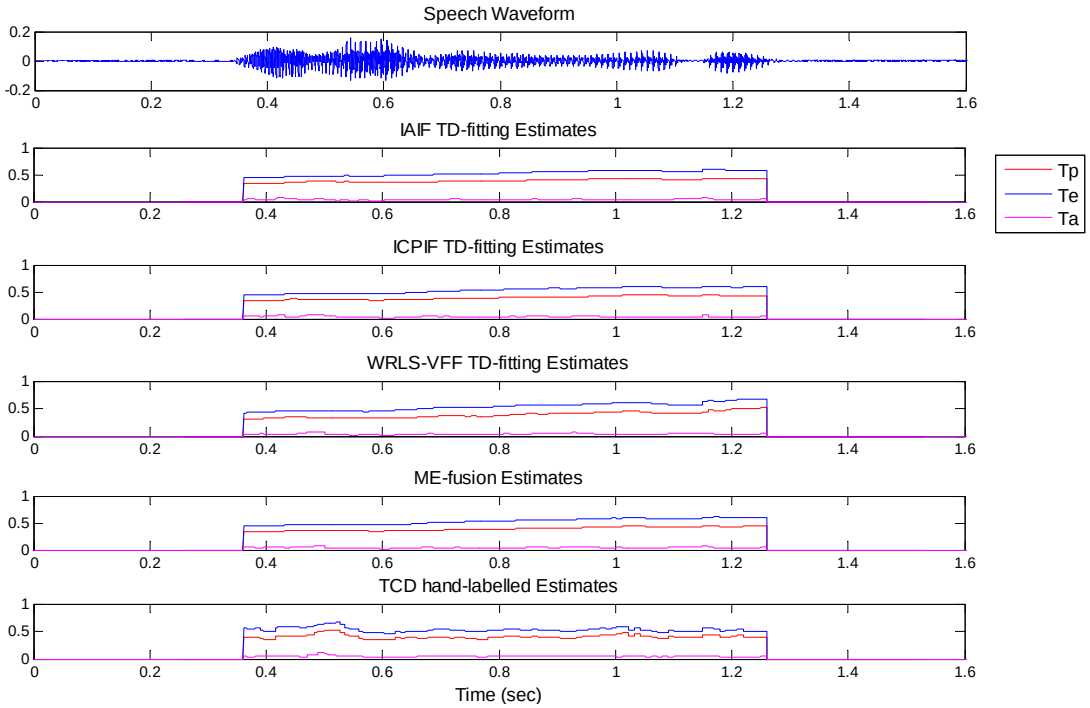

Figure 6.13: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for JD

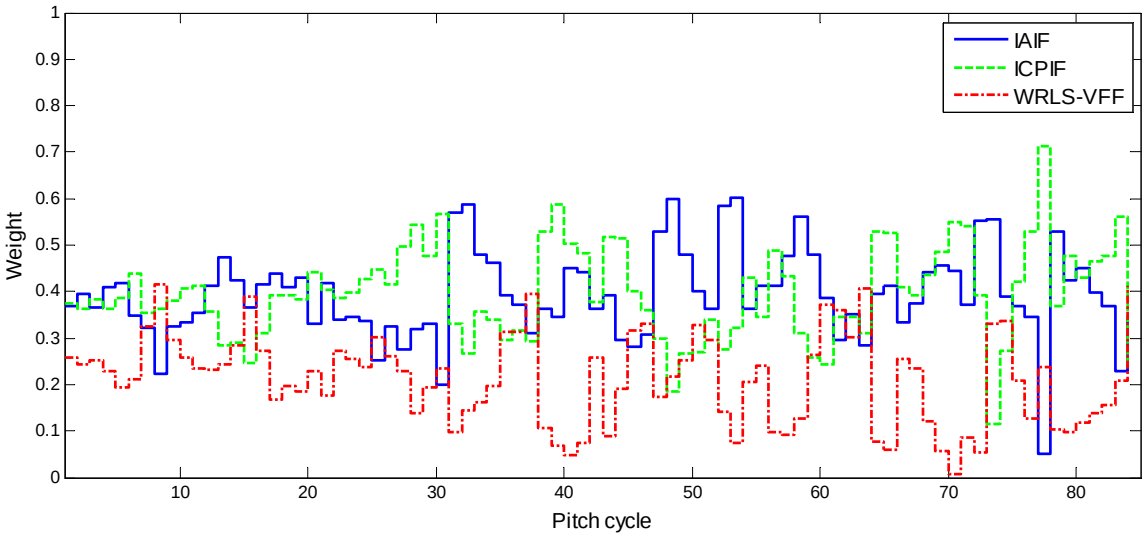

Figure 6.14: Weights by different approaches across JD

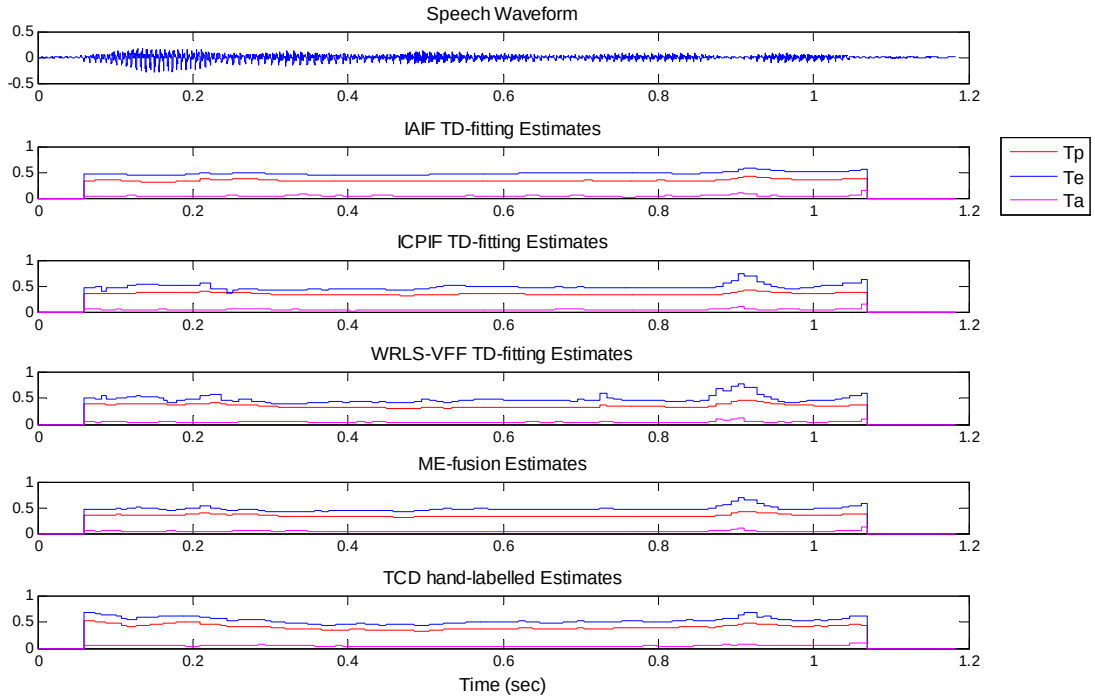

Figure 6.15: Speech waveform and the corresponding LF-mode parameter trajectories by different approaches for LP

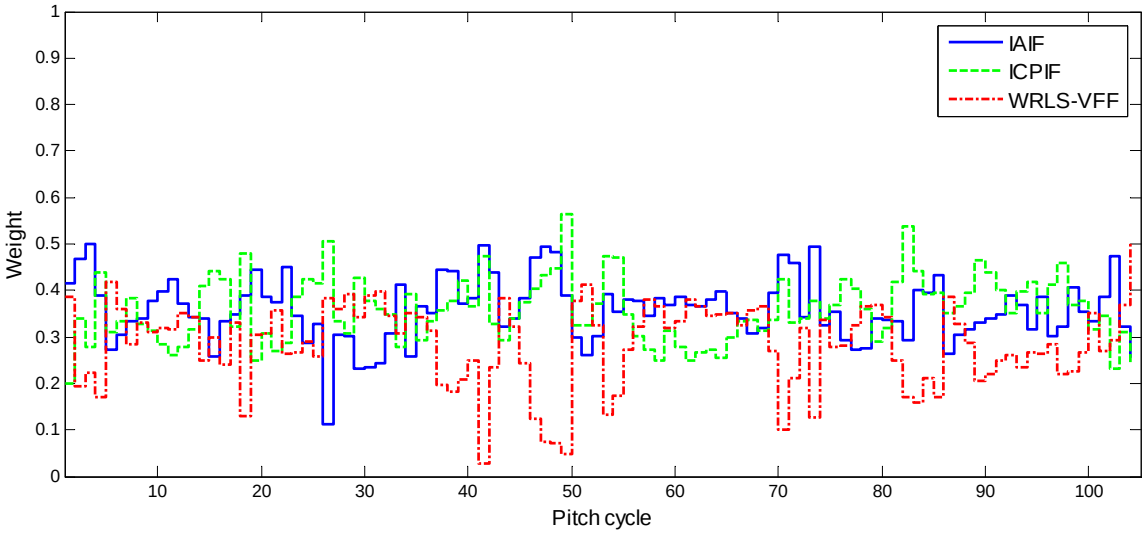

Figure 6.16: Weights by different approaches across LP

Table 6.4: RMSE scores to hand-labelled data for JK

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	ME^{TD}_{fusion}
T_p	0.0932	0.1014	0.1623	**0.0813**
T_e	0.0997	0.1133	0.3256	**0.0982**
T_a	**0.0144**	0.0174	0.0464	0.0172

Table 6.5: Means of fitting error covariance and weight for JK

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$
Covariance	0.0325	0.0207	0.5990
Weight	0.3815	0.5158	0.1026

• Speaker JK. It can be observed from Table. 6.4 that the fusion method has the smallest RMSE scores for T_p and T_e with respect to the hand-labelled data. For T_a, the fusion algorithm has the second smallest RMSE. From the smoothed trajectories in Fig. 6.11 we can see that IAIF, ICPIF and the fusion approach resulted in similar trajectories to the hand-labelled estimates, while WRLS-VFF obviously generated larger, inaccurate estimates. Accordingly, the mean error covariance and weight in Table 6.5 by WRLS-VFF are larger compared to the other two algorithms. Also, it can be observed in Fig. 6.12 that for most pitch periods, it is the estimates by ICPIF mainly contributing to the fused estimates. Estimates by WRLS-VFF only have small contributions when combined with the other two sets of estimates. Especially for pitch cycles 85 through 97, the weights by WRLS-VFF are nearly zero. The discrepancy in larger RMSE scores but lower mean error covariance and weight by ICPIF is probably because only the waveform fitting error criterion is used for the current implementation of the fusion algorithm, where it is possible that a poor GFD estimate can be well fitted and result in a low error covariance.

• Speaker JD. It is clear that for this speaker the glottal source parameters are more consistent across the full utterance, by observing the estimated trajectories by different approaches in Fig. 6.13. In this case, the fusion algorithm generated

Table 6.6: RMSE scores to hand-labelled data for JD

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	ME^{TD}_{fusion}
T_p	**0.0449**	0.0478	0.0644	0.0493
T_e	**0.0692**	0.0743	0.0885	0.0747
T_a	0.0174	**0.0133**	0.0175	0.0137

Table 6.7: Means of fitting error covariance and weight for JD

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$
Covariance	0.0160	0.0155	0.0330
Weight	0.3921	0.3949	0.2131

the RMSE scores for T_p and T_e close to but slightly higher than IAIF and ICPIF. For T_a, the fusion approach has an RMSE value near to the smallest one by ICPIF. WRLS-VFF resulted in the highest RMSE scores for all three parameters. It can be observed from Table 6.7 that IAIF and ICPIF have similar mean error covariances, and approximately 39% weights are assigned to the estimates by these two methods. WRLS-VFF has a higher mean error covariance score and approximately 21% weight in average is given to its estimates. Contributions to the fused estimates by different algorithms can also be seen from the weight plot in Fig. 6.14.

• Speaker LP. It can be seen from the trajectory plot in Fig. 6.15 that the three algorithms and the fusion method generated similar LF-model trajectories. From Table 6.8, WRLS-VFF has the lowest RMSE score for T_p, while T_e and T_a are most accurately tracked by the fusion approach. For this utterance, WRLS-VFF resulted in the largest error covariances and thus less weight was given to its estimates, by observing the mean error covariance and weight shown in Table 6.9. Also it can be seen from Fig. 6.16 that for several pitch cycles, WRLS-VFF resulted in weights less than 0.2. Accordingly, contributions by WRLS-VFF to the fused estimate are less than the other two approaches. This explains the slightly higher RMSE for T_p by the fusion method than WRLS-VFF.

Table 6.8: RMSE scores to hand-labelled data for LP

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	ME^{TD}_{fusion}
T_p	0.0719	0.0613	**0.0589**	0.0627
T_e	0.0704	0.0692	0.0793	**0.0690**
T_a	0.0147	0.0120	0.0129	**0.0111**

Table 6.9: Means of fitting error covariance and weight for LP

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$
Covariance	0.0188	0.0198	0.0263
Weight	0.3579	0.3574	0.2847

From the experimental results presented above, it can be seen that the ME-fusion algorithm can be applied to accurately estimate the glottal source parameters. In most cases, the estimated LF-model parameters by the fusion approach are close to the hand-labelled data. In other cases, the fusion method resulted in less accurate estimates compared to the estimates by certain local algorithm, which is caused by various reasons such as the weakness of quantitative data fusion [Raol, 2009], the imperfect selection of the error criterion, the complexity of speech signals and limitations of local algorithms. In the future, the most effective way to improve the fusion algorithm is to investigate multiple error criteria and constraint based LF-model fitting, avoiding the problem of a good fit to a poor GFD estimate. To sum up, compared to single algorithms, the ME-fusion algorithm is more reliable at tracking the glottal source parameters across a wide range of speech signals.

6.8 Extending the Fusion Framework

To measure the benefit of integrating another method with an existing framework, a further evaluation is carried out. The three inverse filtering methods (IAIF, ICPIF and WRLS-VFF) are kept the same, but another LF-model fitting

approach is added to the framework with two alternatives: one is a spectral fitting method similar to [Kane et al., 2010] which was introduced in Chapter 4, the other is a new LF-model fitting method proposed by Kane in a recent paper[1] [Kane, 2012].

The spectral fitting approach was shown to be less robust than our time-domain method in Chapter 4. Kane's new method utilises both time and frequency features for the fitting. In this evaluation, we will examine the effect of adding both a poorly and a well performing LF fitting algorithm to the fusion framework. The experimental details and results are presented in the following contents.

6.8.1 Adding FD-LF Fitting to the Fusion Framework

In this test, the frequency domain fitting method is added to the fusion framework. Three inverse filtering approaches multiplied by two LF-model fitting algorithms results in six sets of LF estimates. As discussed in Section 4.5.2.2, the frequency-domain error criterion is less robust than the time-domain one. Thus, we decide to use the time-domain fitting error as the error criterion for both the TD and FD fitting methods. The estimated LF-model parameters by the FD fitting approach are used to re-construct the time-domain waveforms corresponding to the inverse filtered GFD signals. Afterwards, the error covariances for both the TD and FD approaches can be obtained from the waveform fitting errors and then are used to calculate the weights in the fusion centre. At the measurement fusion stage, the fusion formula is also expanded to combine all sets of estimates. Subsequently, this modified multi-estimate fusion algorithm was applied to the utterance "we were away a year ago" by three speakers (JK, JD and LP) with hand-labelled data. The smoothed estimated LF-model parameters by individual algorithms and estimates by the fusion approach were compared to the TCD hand-labelled data, and the RMSE scores were calculated and presented in Tables. 6.10, 6.12, and 6.14 for the three speakers respectively, where in the last columns are the results from the original framework. Also, the mean error

[1]John Kane, Phonetics & Speech Laboratory, Centre for Language and Communication Studies, Trinity College Dublin, Ireland, provided the matlab code for the implementation used in this study

Table 6.10: RMSE to TCD-labelled data by different algorithms for JK

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	FD_{iaif}	FD_{icpif}	$FD_{wrlsvff}$	ME^{TD-FD}_{fusion}	ME^{TD}_{fusion}
T_p	0.0932	0.1014	0.1623	0.1303	0.1422	0.1536	**0.0810**	0.0813
T_e	0.0997	0.1133	0.3256	0.1217	0.1578	0.1715	0.0995	**0.0982**
T_a	**0.0144**	0.0174	0.0464	0.0210	0.0352	0.0320	0.0181	0.0172

Table 6.11: Means of fitting error covariance and weight for JK

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	FD_{iaif}	FD_{icpif}	$FD_{wrlsvff}$
Covariance	0.0325	0.0207	0.5990	0.0420	0.0355	0.1028
Weight	0.2117	0.2804	0.0574	0.1736	0.1968	0.0801

covariances and weights are shown in Tables. 6.11, 6.13, and 6.15.

It is observable that estimates by the FD fitting method are less accurate than the TD approach in most cases for all the three speakers for their higher RMSE scores. For speaker JK, WRLS-VFF resulted in poor glottal estimates and thus the two fitting methods generated higher RMSE and larger error covariances in Tables. 6.10, 6.11, and small weights were assigned to their estimates. For IAIF and ICPIF, the FD fitting method performed poorly and resulted in higher RMSE scores for all the three LF-model parameters. Also it can be seen that TD_{icpif} has the highest mean weight and thus the corresponding estimates contributed most to the fused estimates. Compared to the framework using only TD fitting in the last section, the fused estimates were not significantly affected by adding another poorly performed fitting approach, since their RMSE scores for the three parameters are similar.

Similarly for speaker JD, the FD fitted estimates are consistently mere inaccurate compared to the three sets of TD estimates for their higher RMSE in Table 6.12. The $FD_{wrlsvff}$ has the poorest performance and on average contributed approximately 7% to the fused estimates from Table 6.13. TD_{iaif} and TD_{icpif} performed best and generated lower RMSE scores for all three LF parameters, and greater weights were given to their estimates in the fusion procedure. Accord-

Table 6.12: RMSE scores to TCD-labelled data by different algorithms for JD

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	FD_{iaif}	FD_{icpif}	$FD_{wrlsvff}$	ME_{fusion}^{TD-FD}	ME_{fusion}^{TD}
T_p	**0.0449**	0.0478	0.0644	0.0937	0.1022	0.1273	0.0496	0.0493
T_e	0.0692	0.0743	0.0885	0.1006	0.0898	0.1288	**0.0620**	0.0747
T_a	0.0174	**0.0133**	0.0175	0.0359	0.0259	0.0332	0.0157	0.0137

Table 6.13: Means of fitting error covariance and weight for JD

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	FD_{iaif}	FD_{icpif}	$FD_{wrlsvff}$
Covariance	0.0160	0.0155	0.0330	0.0414	0.0403	0.0763
Weight	0.2698	0.2736	0.1450	0.1253	0.1177	0.0688

ingly, the combined estimates by the fusion method show similar RMSE scores to the two approaches, and are only slightly different compared to the results from the original framework.

For LP's utterance, RMSE scores by the spectral fitting methods are much higher than the TD approaches observed from Table 6.14. Since in total about 30% weight was given to the estimates by the three FD algorithms, the fused estimates exhibit slightly higher RMSE scores for T_p and T_e than the TD based estimates and the fused estimates by original framework. The RMSE for T_a by the fusion approach is the smallest one of all the approaches, which is probably because the fused T_a values place more weight on the TD estimates because of their smaller fitting error covariances where these estimates are close to the hand-labelled data, and according to the smoothing procedure the T_a trajectory was well captured because of its slower variation compared to the other two parameters.

In this test, we can see that the frequency-domain LF-model fitting algorithm is not as robust as the time-domain approach, since the FD methods consistently generated higher RMSE scores and the fitting error covariances are large. This is because initialisation is crucial for a FD fitting method. However, the H1-H2 value or the mean squared error of the first six harmonic amplitudes is not sufficiently

Table 6.14: RMSE scores to TCD-labelled data by different algorithms for LP

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	FD_{iaif}	FD_{icpif}	$FD_{wrlsvff}$	ME_{fusion}^{TD-FD}	ME_{fusion}^{TD}
T_p	0.0719	0.0613	**0.0589**	0.1889	0.1343	0.1601	0.0916	0.0627
T_e	0.0704	0.0692	0.0793	0.2064	0.1313	0.1282	0.0911	**0.0690**
T_a	0.0147	0.0120	0.0129	0.0346	0.0219	0.0472	**0.0107**	0.0111

Table 6.15: Means of fitting error covariance and weight for LP

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	FD_{iaif}	FD_{icpif}	$FD_{wrlsvff}$
Covariance	0.0188	0.0198	0.0263	0.0446	0.0316	0.0576
Weight	0.2262	0.2266	0.1830	0.1252	0.1455	0.0936

robust for choosing the initial estimate values. This conclusion was also reached in Chapter 4 by comparing the TD and FD fitting methods for artificial and real glottal source signals.

Also, it can be observed that by combining all six sets of estimates, the fusion algorithm can be affected by the FD fitted estimates (especially for speaker LP). However, although the FD fitted estimates are not as accurate as the estimates by TD approaches, the multi-estimate fusion algorithm can still generate good estimated results for different utterances. This is because the combination of the estimates is based on the fitting error covariances of all estimates: the smaller the covariance the more weight is given to that set. Obviously in this test, more weight was given to the TD-based estimates because of their smaller error covariances in most cases. Also, the Kalman filter smoothing procedure ensures the optimal estimate trajectory can be obtained.

In the next section, we will replace the poorly performing frequency-domain LF fitting method with a more robust approach, and the results will be presented following with analysis.

Table 6.16: RMSE to TCD-labelled data by different algorithms for JK

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	JKN_{iaif}	JKN_{icpif}	$JKN_{wrlsvff}$	ME^{TD-JKN}_{fusion}	ME^{TD}_{fusion}
T_p	0.0932	0.1014	0.1623	**0.0470**	0.0611	0.0820	0.0629	0.0813
T_e	0.0997	0.1133	0.3256	0.0705	**0.0666**	0.1182	0.0820	0.0982
T_a	**0.0144**	0.0174	0.0464	0.0312	0.0308	0.0318	0.0168	0.0172

Table 6.17: Means of fitting error covariance and weight for JK

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	JKN_{iaif}	JKN_{icpif}	$JKN_{wrlsvff}$
Covariance	0.0325	0.0207	0.5990	0.0363	0.0301	0.0814
Weight	0.2121	0.2849	0.0579	0.1708	0.1986	0.0756

6.8.2 Adding JKN-LF Fitting to the Fusion Framework

In this test, the frequency domain fitting method is replaced by John Kane's new (JKN) LF-model fitting algorithm [Kane, 2012]. Instead of considering only the time-domain waveform or the spectral information, this algorithm takes both into account and accordingly tries to minimise a weighted (more weight is given to the waveform fitting error) sum error criterion to obtain the optimal R_d estimates which control the main shape of the LF-model [Fant, 1995]. Subsequently, T_p, T_e and T_a can be derived from R_d and subsequently refined by the "simplex" optimisation procedure [Nelder and Mead, 1965].

This new version of the multi-estimate fusion algorithm was also applied to the utterance "we were away a year ago", spoken by three speakers as in the previous section. The RMSE scores for the three LF-model parameters by different methods are shown in Tables 6.16, 6.18, and 6.20 and the corresponding mean values of the fitting error covariances and weights were calculated and presented in Tables 6.17, 6.19, and 6.21.

Obviously, the new JKN estimates are more accurate compared to the FD fitted estimates presented in the last section, since the RMSE scores for all the three parameters are consistently lower than the FD methods across different speakers. For JK's utterance, the JKN estimates are more accurate than estimated LF pa-

Table 6.18: RMSE to TCD-labelled data by different algorithms for JD

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	JKN_{iaif}	JKN_{icpif}	$JKN_{wrlsvff}$	ME_{fusion}^{TD-JKN}	ME_{fusion}^{TD}
T_p	**0.0449**	0.0478	0.0644	0.0662	0.0910	0.1005	0.0596	0.0493
T_e	**0.0692**	0.0743	0.0885	0.0719	0.1132	0.1264	0.0780	0.0747
T_a	0.0174	0.0133	0.0175	0.0176	0.0193	0.0220	**0.0122**	0.0137

Table 6.19: Means of fitting error covariance and weight for JD

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	JKN_{iaif}	JKN_{icpif}	$JKN_{wrlsvff}$
Covariance	0.0160	0.0166	0.0330	0.0216	0.0206	0.0340
Weight	0.2380	0.2377	0.1291	0.1535	0.1506	0.0911

rameters by the TD-based methods for T_p and T_e in Table 6.16. RMSE for T_a by JKN is higher compared to TD estimates, which is mainly because the JKN method derives T_a under the assumption that T_a and R_d are linearly correlated, although this relationship may not be robust across different sounds and speakers. The discrepancy for lower RMSE scores but larger mean error covariance in Table 6.17 by JKN methods is mainly caused by the error criterion: although T_p and T_e were accurately estimated and the open phase was fitted well, T_a is less accurate which might yield a large fitting error for the return phase and result in an increase of the overall error. Also it can be observed that TD_{icpif} has the largest weight in average, and different from the FD-fitting-added framework, there are similar weights given to TD_{iaif}, JKN_{iaif} and JKN_{icpif}. WRLS-VFF generated an inaccurate glottal estimate and thus the two fitting methods contributed less to the fusion procedure. Accordingly, RMSE scores of the fused estimates are close to the JKN estimates and lower than the TD estimates, and the results from the previous two tests, thus the performance of the fusion method is improved.

For speaker JD, RMSE scores in Table 6.18 for T_p and T_e by JKN_{icpif} and $JKN_{wrlsvff}$ are higher than the corresponding TD estimates. However, compared to the mean covariances by FD methods in Table 6.13, the error covariances by JKN are relatively smaller compared to the FD estimates in the last section and

Table 6.20: RMSE to TCD-labelled data by different algorithms for LP

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	JKN_{iaif}	JKN_{icpif}	$JKN_{wrlsvff}$	ME_{fusion}^{TD-JKN}	ME_{fusion}^{TD}
T_p	0.0719	0.0613	0.0589	0.0582	**0.0486**	0.0583	0.0575	0.0627
T_e	0.0704	0.0692	0.0793	0.0583	**0.0475**	0.0575	0.0571	0.0690
T_a	0.0147	0.0120	0.0129	0.0205	0.0206	0.0197	0.0135	**0.0111**

Table 6.21: Means of fitting error covariance and weight for LP

	TD_{iaif}	TD_{icpif}	$TD_{wrlsvff}$	JKN_{iaif}	JKN_{icpif}	$JKN_{wrlsvff}$
Covariance	0.0188	0.0198	0.0263	0.0228	0.0254	0.0277
Weight	0.1976	0.1982	0.1570	0.1627	0.1499	0.1346

accordingly there was more weight given to their estimates during fusion. Consequently, the fused estimates have lower RMSE values than the JKN estimates but higher scores than the TD estimates and the FD-fitting-added framework. In this case, the overall performance of the fusion approach is affected.

For the third utterance by LP, all the six approaches performed well for their relatively low RMSE scores for the three LF parameters in Table 6.20. Also, the mean error covariances by these algorithms do not have significant differences. Consequently, the weights are almost averagely assigned to the corresponding estimates as can be observed in Table 6.21. For T_p and T_e, the fused estimates are less accurate compared to the estimates by JKN_{icpif} but more reliable than the other five sets of estimates. The fused T_a has a slightly higher RMSE score than that of TD_{icpif} and $TD_{wrlsvff}$. Compared to the results by the original framework and the FD-fitting-added, the overall performance of this implementation is enhanced.

In this section, two new versions of the multi-estimate fusion algorithm were applied to an all voiced utterance spoken by three speakers and the results were compared to the corresponding hand-labelled data. From the experimental results, it can be observed that with an added frequency-domain fitting method lacking of robustness, the ME-fusion approach can still generate accurate LF-

model estimates due to more weight being assigned to the time-domain fitted estimates.

By integrating Kane's new fitting algorithm, the ME-fusion approach can generate estimates closer to the hand-labelled data compared to the FD-fitting-added framework for speakers JD and LP. This is consistent with the fact that in the fusion algorithm the better the performance of an individual algorithm, the more accurate the fused estimates obtained.

Some discrepancies can be observed in the results, such as lower RMSE scores resulting from larger error covariances. This is mainly caused by the imperfect error criterion currently used in the framework, where only the time-domain waveform fitting errors are considered. As mentioned in the last section, to improve the performance of the fusion algorithm, further investigation on criteria to calculate the error covariance much be carried out.

6.9 Assessment of the Fusion Framework

After the evaluations and discussions presented in previous sections, the fusion framework can be assessed as follows.

Firstly, multiple estimates extracted by individual algorithms are combined in the fusion centre to produce an optimal set of glottal source parameters. Thus, for a wide range of speech signals, the fusion approach can more accurately estimate the voice source parameters compared to individual algorithms. From the experimental results, it can be observed that overall the fusion algorithm is superior than other approaches, although in some cases certain local algorithm performed the best. This is because the lack of a perfect error criterion for combining data. The fitting error criterion used currently in the framework is a significant indicator for the performance of individual algorithms, however, it is sensitive to formant ripples and noise. A further investigation should be made to use multiple error criteria, such as the GQMs used in Chapter 3, to improve the performance of the fusion method.

Secondly, by utilising appropriate algorithms, the most important glottal source parameters can be extracted. In our implementation of the fusion framework, different glottal inverse filtering and LF-model fitting approaches are ap-

plied and the three timing parameters of the LF-model describing the shape of the full cycle of the glottal waveform are estimated. Other algorithms such as the mixed-phase decomposition method which cannot estimate the glottal return phase parameter, are not incorporated. As a general framework for glottal source parametrisation, procedures for estimating other glottal source features such as the glottal formant, amplitudes of the positive and negative peaks of the glottal waveforms can be conveniently incorporated.

Thirdly, an automatic speech segmentation procedure is incorporated to separate arbitrary input speech signal to individual voiced/unvoiced frames, thus the corresponding full glottal source parameter trajectories can be extracted. In addition, the Kalman smoothing procedure ensures the consistency of the estimates. Further investigation can be made to analyse these trajectories for the purpose of improving the naturalness of synthetic speech, or enhancing the accuracy for speaker identification system. It is always a problem to give quantitative measures for the quality of the glottal parameter estimates based on no a priori information about the true glottal component, however it is worthy of looking at the 'by-product' of the fusion approach, such as the covariances by Kalman filtering and the log-likelihood functions obtained from the EM algorithm, to present some pictures.

Finally, the fusion framework offers the measure to compare the performance of different algorithms. Glottal source parameter estimates by multiple approaches are calculated corresponding to the same input speech segment, thus we can make comparisons of these methods based on the information obtained by the fusion approach, such as the fitting error covariance and weights. Also, because of the flexibility, it is convenient to incorporate a 'gold' algorithm for specific speech signals into the framework and evaluate the performance of other or new proposed algorithms. Therefore, the multi-estimate fusion framework is a useful scientific tool for researchers in this area.

6.10 Conclusion

In this chapter, the effectiveness of the multi-estimate fusion algorithm introduced in Chapter 5 was evaluated. The ME-fusion method was firstly implemented

according to the fusion framework by utilising three inverse filtering techniques (IAIF, ICPIF and WRLS-VFF) and a time-domain LF-model fitting algorithm. A procedure to synchronise the multiple sets of estimates by glottal closing instants was introduced. By applying this implementation to an artificial "coarticulatory" voiced speech segment, it can be observed that the fusion algorithm can generate more accurate estimates than local individual methods.

In the second evaluation, the implemented ME-fusion algorithm was applied to three utterances spoken by a male and a female speaker taken from the CMU-ARCTIC database. The estimated LF-model parameter trajectories from different algorithms were presented. It is demonstrated that the estimated trajectories by the fusion method are reasonably consistent, the outliers obtained by local algorithms were smoothed to be more correlated to neighbouring estimates and some properties of estimates by individual algorithms were retained.

Several voiced segments were extracted and three inverse filtering methods were applied. Analysis was made of the performance variation across individual algorithms. It is clear that no single algorithm outperforms others and more comprehensive research is required to fully explore the strengths and weaknesses of each algorithm.

A further evaluation was carried out by applying the ME-fusion algorithm to an all voiced utterance for which expertly hand-labelled data exist. Results showed that the fusion algorithm can automatically lock to estimates with small error covariances and the fused estimates are comparable to the hand-labelled data. Overall, the fusion algorithm is more reliable than individual algorithms.

In a final evaluation, to measure the effect of adding another algorithm to the current fusion framework, two alternative LF fitting methods were integrated: one is a spectral fitting approach which lacks robustness, and the other a new method taking into consideration both waveform and spectral fitting errors which is more robust. It is observed that the fusion framework can intelligently avoid to poor estimates since integration of the poorly performing FD method does not significantly affect the performance. In addition, further improvement can be obtained when a well performing approach is added.

Chapter 7

Summary & Conclusions

7.1 Introduction

The aim of this chapter is to firstly summarise the research described in this thesis, secondly, to list the contributions of this study to the field of speech processing and thirdly to suggest possible directions for future work.

7.2 Summary of the thesis

In this thesis, we aimed to investigate approaches to estimate the glottal source parameters. For this purpose, four studies were undertaken:

- speech signal source-filter model representation,
- glottal waveform extraction,
- automatic LF-model fitting and finally,
- proposal and evaluation of the multi-estimate (ME) fusion algorithm.

Chapter 2 As background to the thesis, we introduced a basic model of human speech production. The classic source-filter model assumes that speech signals can be considered as the response of an IIR filter to a source signal. For voiced speech the source signal is a simple pulse train, and for unvoiced speech it is white noise. By changing the positions of the poles of the IIR filter the generated speech signals will vary. To more realistically model the voice source, the Liljencrants-Fant (LF) model [Fant et al., 1985] was introduced, which is

currently the most widely used parametric model to describe the shape of the glottal source. It not only describes the open phase shape of the glottal flow derivative, but also captures the return phase of the GFD signal which can affect the speech quality. The LF-model was applied throughout the thesis for the purpose of glottal source parametrisation.

Chapter 3 Based on the assumption that the speech signal can be decomposed into its source and vocal tract filter components, various approaches were introduced to extract the glottal waveform. Many speech decomposition methods are based on Linear Prediction (LP) analysis [Makhoul, 1975]. For voiced speech, successive samples are highly correlated and by applying the LP technique the obtained prediction coefficients are in fact the coefficients of the vocal tract filter. If the original speech signal is put through the inverse of the vocal tract filter, theoretically the vocal tract effect can be removed and the glottal source component is recovered. In this work, three LP analysis based algorithms were described in detail. A performance study was carried out to show the performance variations across different approaches for different speech signals. In addition, two additional glottal waveform extraction methods, mixed-phase speech decomposition [Bozkurt et al., 2004b,a; Drugman et al., 2009b] and higher order statistics analysis [Chen and Chi, 1993; Walker, 2003], were briefly introduced and their limitations were discussed.

Chapter 4 Once the glottal waveform estimate is obtained, we need to fit a parametric model to it to capture the parameters of the model. As mentioned previously, the LF-model is utilised as the parametric glottal source model and generally there are two approaches to fitting the LF-model to the GFD signal: time-domain and frequency-domain methods. For a time-domain fitting method, it is necessary to find suitable initial values a priori to optimisation because this is a non-linear multi-dimensional optimisation problem. In general, the initial values are directly obtained by searching the GFD waveform. One or more optimisation algorithms are then applied to refine the estimates by minimising the fitting error between the LF-model pulse and the GFD signal, period by period. In Chapter 4, we reviewed three different time-domain LF fitting algorithms. For a frequency-domain fitting method, the LF-model parameter estimates are obtained by minimising the spectral distance between the LF-model and the in-

verse filtered GFD spectra. A typical spectral fitting approach was introduced in the thesis, which initialises the estimates by searching a codebook. Subsequently, optimisation procedures are applied to adjust the estimates. As the extended Kalman filter (EKF) can be used to track the state vector of a non-linear process, we proposed a new time-domain fitting algorithm based on EKF. This method separates the GFD signal into its open phase and return phase and the corresponding LF-model shape controlling parameters can be estimated by EKF. By comparing the proposed fitting algorithm to both standard time-domain and modified frequency-domain fitting approaches, it was observed that the new method is superior.

Chapter 5 Several techniques for quantitative data fusion were introduced such as Millman's fusion formula, the Kalman filter for data fusion and the state-vector fusion and measurement fusion structures. The main purpose of this thesis is to propose an approach that can accurately track the glottal source parameters. To this end, we proposed a multi-estimate fusion framework which is general and flexible enough to support varying requirements. As we saw from our experimental results and other references, there is no single algorithm that performs best for all kinds of speech signals. To obtain more reliable glottal source estimates, the multi-estimate fusion algorithm firstly utilises several different approaches in parallel to generate multiple sets of estimates. Subsequently, all sets of estimates are combined by a measurement fusion procedure where the corresponding weights are calculated from their respective fitting error covariances. In addition, a Kalman smoothing procedure is applied to reduce the variation of the estimates and make them more correlated to their neighbouring estimates along the time axis, which is consistent with the assumption that the glottal source parameters should have limited variation across adjacent pitch periods especially for sustained vowel sounds. Afterwards, the advantages, limitations and factors which may affect the performance of the fusion framework were discussed.

Chapter 6 Several evaluations were carried out to test the effectiveness of the algorithm. In a preliminary evaluation, in which the fusion approach was applied to an artificial voiced speech segment and compared to the estimates obtained by individual algorithms, the validity of the fusion method was established. In addition, the fusion method was applied to different male and female utterances

to extract LF-model trajectories. Results were compared to the smoothed trajectories by individual algorithms and it was observed that the fused estimates are more consistent than those of a poorly performing single algorithm across different utterances. A detailed analysis was made of performance variation across individual algorithms over several extracted voiced segments. In a further evaluation, an all-voiced utterance by different speakers with hand-labelled data was used to test the effectiveness of the fusion method. The RMSE scores of different approaches were calculated and compared, and the mean error covariances and weights were presented. It can be observed that the fusion algorithm can automatically give greater weight to more accurate estimates. Consequently, the fused estimates were more reliable than estimates by single algorithms across different speakers. To measure the effect of adding another fitting algorithm to the existing framework, two tests were carried out. The first one integrated the frequency-domain fitting method into the framework and the augmented framework was applied to the same set of utterances with hand-labelled data. Results showed that the poorly performing FD fitting method did not significantly affect the performance of the fusion algorithm. In the second test, a more robust LF-model fitting method was added to the original framework and it can be observed that its performance was further improved.

7.3 Contribution of the thesis

The thesis has made a number of contributions to voice source parametrisation including a review and evaluation of existing techniques and the proposal of new algorithms. These contributions are reviewed below.

1. **Review and performance evaluation of existing glottal inverse filtering algorithms**. Three Linear Prediction based glottal inverse filtering techniques were reviewed in detail, which were closed phase inverse filtering (CPIF), iterative adaptive inverse filtering (IAIF) and weighted recursive least square with variable forgetting (WRLS-VFF) inverse filtering. The three approaches were evaluated by applying them to different speech signals and observing the glottal source estimates and calculating corresponding quality measures. It was

observed that no single algorithm consistently outperformed other methods across different speech signals.

2. **Review of LF-model fitting algorithms**. Several glottal source LF-model fitting algorithms were reviewed in this study, including three time-domain methods [Strik et al., 1993; Riegelsberger and Krishnamurthy, 1993; Childers and Ahn, 1995] and a frequency-domain fitting method [Kane et al., 2010]. Time-domain approaches estimate the LF parameters by fitting the LF-model pulses to the inverse filtered glottal source signals. The main challenge is obtaining accurate initial estimates from the glottal waveform to facilitate multi-dimensional optimisation. The spectral fitting approach attempts to find a set of LF parameters generating the spectrum best matching the glottal source spectrum. To achieve accurate estimates by this method, high quality inverse filtering of the glottal source signal is required.

3. **A new time-domain LF-model fitting algorithm was proposed and evaluated**. This approach aims to improve the accuracy of the estimated LF-model parameters. The basic theory of the proposed approach is that the LF-model has two phases: the open phase which is an exponentially growing sinusoidal waveform and the return phase which is a decaying exponential, and the shape-controlling parameters for both phases can be tracked by an extended Kalman filter (EKF) [Welch and Bishop, 1995]. Because of the non-linear property of the model, multiple initial values are utilised by the EKF to estimate the two LF-model shape-controlling parameters and the one giving the minimum mean squared fitting error between the LF-model pulse and the glottal flow derivative signal is chosen as optimal. The proposed time-domain LF-model fitting method was compared to both standard time-domain method and spectral fitting approach. Results showed that the new algorithm can more accurately estimate the glottal source parameters.

4. **Proposal of a multi-estimate fusion framework for glottal source parametrisation**. Aiming to obtain more reliable estimates of the glottal source parameters, a multi-estimate (ME) fusion framework was introduced. Different from previous approaches which may generate poor estimates under certain circumstances, the ME fusion method utilises multiple algorithms to extract multiple sets of glottal LF-model estimates and combines them appropriately to en-

sure one overall optimal set of estimates can be obtained. A measurement fusion structure [Raol, 2009] is applied to the fusion framework by the generalised Millman's formula [Shin et al., 2006]. Afterwards, the fused set of estimates are smoothed by a Kalman filtering [Kalman, 1960] procedure, where the parameters are re-estimated by the EM algorithm [Shumway and Stoffer, 1982], with the assumption that the variation of the glottal source parameters is roughly a linear random process across continuous pitch periods.

5. **Evaluation of the ME-fusion algorithm**. The ME fusion algorithm was implemented and evaluated to test its effectiveness. Firstly, the fusion algorithm was implemented with three inverse filtering methods and one LF-model fitting approach. This implementation was tested with a synthetic voiced speech segment as a preliminary evaluation. In the second evaluation, the ME-fusion algorithm was applied to both male and female utterances to extract the LF-model parameter trajectories across complete utterances. In a further test, the ME-fusion approach was applied to an all-voiced utterance spoken by three speakers with hand-labelled LF-model parameter data, the corresponding RMSE scores, mean error covariances and weights of different algorithms were calculated and compared. A final evaluation was carried out by integrating an additional fitting algorithm, with two alternatives, to the current framework to test its performance. Results from all evaluations showed the effectiveness of the ME-fusion algorithm.

7.4 Further work suggestions

Several potential opportunities for future research have been inspired by the work described in this thesis. They are discussed below.

• *More appropriate error criteria.* One crucial factor that affects the performance of the fusion algorithm is the error criterion used to combine multiple sets of estimates. Currently, the fitting error covariance calculated from the fitting residuals between the reconstructed LF-model pulses and the extracted glottal pulses is the only error criterion considered by the fusion method. However, because of the complexity of real speech production and the limitation of glottal

source extraction techniques, vocal tract filtering cannot be completely removed from speech signals. Consequently, the fitting error criterion may result in poor decisions for the fusion procedure, e.g., ripples appearing in the glottal flow estimate may increase the fitting error even if the estimated LF-model parameters are accurate. Thus additional criteria should be considered to enhance the fusion approach, such as spectral distance and glottal estimate quality measures. In addition, how to allocate the weights among these measures should be investigated.

• *Investigation of more robust candidate algorithms.* To obtain better estimates by the fusion algorithm, it is necessary to improve the performance of individual algorithms including both the glottal source extraction method and the LF-model fitting approach. The inverse filtering methods utilised in this thesis have been demonstrated to generate reasonably accurate glottal flow estimate under most circumstances. However, it is still a problem to obtain a robust glottal estimate when the first formant of the vocal tract is close to the fundamental frequency. Also, source-tract interaction cannot be fully accounted for by state-of-the-art inverse filtering techniques. Therefore, it is difficult to obtain accurate glottal source parameter estimates by a curve fitting when a poor glottal flow estimate is obtained. Further investigation is required into these problems, and a more robust LF-model fitting algorithm which is less sensitive to glottal open phase ripple is necessary for better locating the glottal opening instant.

• *Integration of joint source-filter estimation methods.* Besides the approaches introduced and utilised in this thesis for glottal source extraction and LF-model parameter estimation, it is also possible to add joint source-filter estimation algorithms into the multi-estimate fusion framework. Generally, a joint estimation method firstly tries to find an initial set of glottal source parameter estimates and then applies one or more optimisation techniques to minimise an error criterion between the original speech signal and the source-filter model. In [Fu and Murphy, 2006], Fu proposed a joint estimation algorithm for glottal source estimation, where the LF-model and the vocal tract parameters are optimised by minimising the predictive errors between the speech samples and the ARX model representation. O'Cinneide [O'Cinneide et al., 2011a] introduced a frequency-domain method to jointly estimate the glottal LF-model and vocal tract parameters. This method uses a codebook to find a set of parameters giving the minimal spectral

distance and refines the estimates in later stages.

• *Improvement of automatic speech segmentation.* Another factor that may affect the performance of the fusion algorithm is the automatic speech segmentation. If a voiced segment is too short, there may not be sufficient data to be used by the Kalman smoothing procedure. In this thesis a pitch tracking [Talkin, 1995] based speech segmentation procedure is utilised. To achieve further improvement, it is necessary to investigate some other approaches such as the method proposed in [Rabiner and Sambur, 1977b] that uses the Itakura distance to identify voiced and unvoiced frames; in [Janer et al., 1996; Erelebi, 2003] the author applied the wavelet transform to segment speech.

• *Investigation of fuzzy Kalman filtering of more complex speech signals.* Currently the ME fusion algorithm works under the assumption that the variation of glottal source parameters across continuous pitch periods is roughly a linear process. However, there are situations where this assumption is invalid such as a voiced speech segment containing voice quality variation. In such a case, the variation of voice source parameters is roughly a piecewise linear process and a standard Kalman filter may generate poor results. For this problem, it is necessary to utilise a more complex tracking technique to achieve robust estimates and fuzzy logic adaptive Kalman filtering [Remus, 1992; Chen et al., 1998; Han, 2004] may be a solution. A fuzzy logic controller allows adjusting of the KF parameters according to the changes of mean and covariance of the state estimate and uses a weighting factor to deal with variation of process and measurement noise covariance, thus helping KF perform better than standard KF for non-linear tracking tasks.

• *Integration of state-vector fusion structure.* Measurement fusion is applied to the current fusion framework, where it has the advantage that estimates from individual algorithms are directly combined and thus there is no information lost. However, it is worth studying the performance of the fusion approach when state-vector fusion is applied. A rational implementation of the modified fusion framework is shown in Fig. 7.1, where the output LFP of all algorithms are firstly smoothed by individual Kalman filters and subsequently the corresponding smoothed estimates are combined at the fusion centre with the covariances P obtained from LF fitting. Comparison of the two fusion structures should be

made. It would be also possible to investigate a mixed structure.

Finally, it will be interesting to investigate integrating the ME fusion algorithm output into glottal source parametrisation related applications such as speaker characterisation, voice transformation, speech synthesis and pathological voice diagnosis.

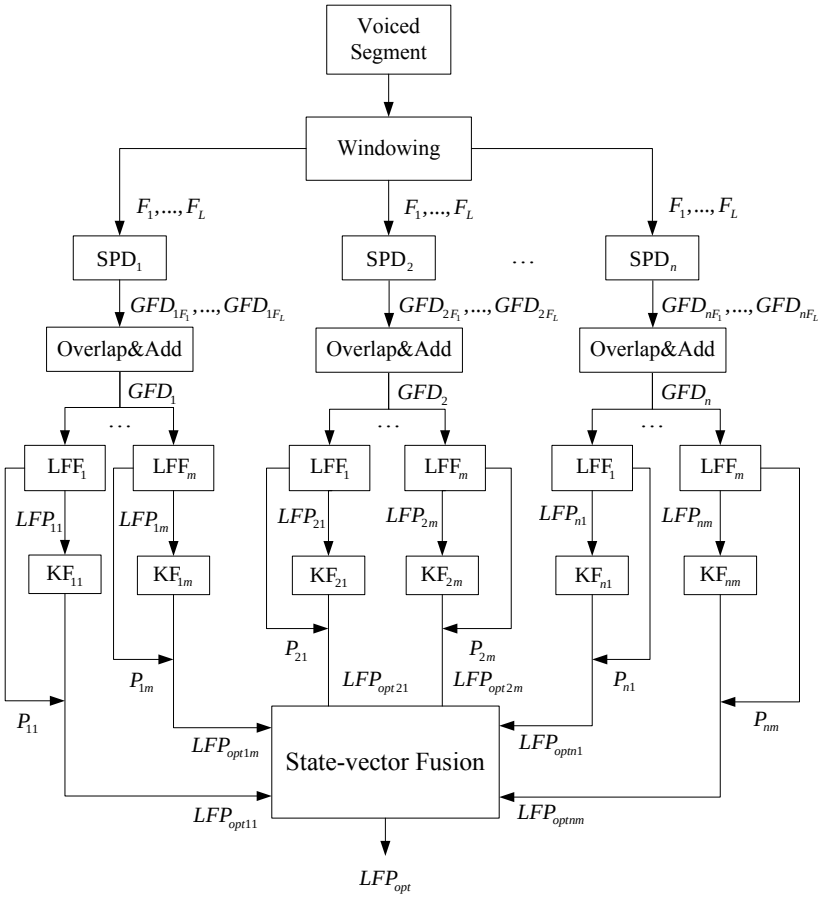

Figure 7.1: Implementation of the state-vector multi-estimate fusion algorithm

References

M. Airas. TKK aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33:49–64, 2008.

J. Ajgl, M. Simandl, and J. Dunik. Millman's formula in data fusion. In *Proceedings of the 10th International PhD Workshop on Systems and Control*, pages 1–6, 2009.

O. O. Akande and P. J. Murphy. Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication*, 46(1):15–36, May 2005.

P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

P. Alku and U.K. Laine. A new glottal LPC method for voice coding and inverse filtering. In *Proc. IEEE International Symposium on Circuits and Systems*, volume 3, pages 1831–1834, 1989.

P. Alku and E. Vilkman. Estimation of the glottal pulseform based on discrete all-pole modeling. In *Proc. Third International Conference on Spoken Language Processing*, 1994.

P. Alku, C. Magi, and T. Backstrom. Glottal inverse filtering with the closed-phase covariance analysis utilizing mathematical constraints in modelling of the vocal tract. *Logopedics Phoniatrics Vocology*, 34(4):200–209, 2009.

R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. *Advanced*

Techniques in Computing Sciences and Software Engineering, pages 279–282, 2010.

T. Backstrom and P. Alku. Harmonic all-pole modelling for glottal inverse filtering. In *Proceedings of the 7th Nordic Signal Processing Symposium*, pages 182–185, 2006.

T. Backstrom, M. Airas, L. Lehto, and P. Alku. Objective quality measures for glottal inverse filtering of speech pressure signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 897–900, 2005.

Y. Bar-Shalom and L. Campo. The effect of the common process noise on the two-sensor fused-track covariance. *IEEE Transactions on Aerospace and Electronic Systems*, (6):803–805, 1986.

T. Bass. Intrusion detection systems and multisensor data fusion. *Communications of the ACM*, 43(4):99–105, 2000.

R. J. Bhansali. Order selection for linear time series models: a review. *Developments in Time Series Analysis*, pages 50–56, 1993.

B. Bozkurt. *Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals*. PhD thesis, Faculte Polytechnique de Mons, Belgium, 2005.

B. Bozkurt and T. Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *Proc. ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

B. Bozkurt, B. Doval, C. DAlessandro, and T. Dutoit. Zeros of z-transform (ZZT) decomposition of speech for source-tract separation. In *Proc. International Conf. Speech, Language Processing*, 2004a.

B. Bozkurt, T. Dutoit, B. Doval, and C. D'Alessandro. A method for glottal formant frequency estimation. In *Proc. Eighth International Conference on Spoken Language Processing*, 2004b.

B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12(4):344–347, 2005.

P. M. T. Broersen and S. de Waele. Finite sample properties of ARMA order selection. *IEEE Transactions on Instrumentation and Measurement*, 53(3): 645–651, 2004.

J. P Cabral, S. Renals, K. Richmond, and J. Yamagishi. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. *ISCA SSW6*, 2007.

J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi. Glottal spectral separation for parametric speech synthesis. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond. HMM-based speech synthesiser using the LF-model of the glottal source. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4704–4707, May 2011.

K. C. Chang, R. K. Saha, and Y. Bar-Shalom. On optimal track-to-track fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 33(4):1271–1276, 1997.

G. Chen, Q. Xie, and L. S. Shieh. Fuzzy kalman filtering. *Information Sciences*, 109(1):197–209, 1998.

W. T Chen and C. Y Chi. Deconvolution and vocal-tract parameter estimation of speech signals by higher-order statistics based inverse filters. In *Proc. IEEE Signal Processing Workshop on Higher-Order Statistics*, pages 51–55, 1993.

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

C. Y Chi and J. Y Kung. A new cumulant based inverse filtering algorithm for identification and deconvolution of nonminimum-phase systems. In *Proc. IEEE Sixth SP Workshop on Statistical Signal and Array Processing*, pages 144–147, 1992.

D. G. Childers. Glottal source modeling for voice conversion. *Speech Communication*, 16(2):127–138, 1995.

D. G. Childers. *Speech Processing and Synthesis Toolboxes*. Wiley, 1999.

D. G. Childers and C. Ahn. Modeling the glottal volume-velocity waveform for three voice types. *The Journal of the Acoustical Society of America*, 97, 1995.

D. G. Childers and CK Lee. Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991.

D. G. Childers, J. C. Principe, and Y. T. Ting. Adaptive WRLS-VFF for speech analysis. *IEEE Transactions on Speech and Audio Processing*, 3(3):209–213, 1995.

D. Choi, V. Shin, B. H Ahn, and J. I Ahn. Fusion of local filters. In *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, pages 22–27, 2004.

T. F. Coleman and Y. Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6(2):418–445, May 1996.

B. de Prony. Essai experimental et analytique: sur les lois de la dilatabilite de fluides elastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, a differentes temperatures. *Journal de l'cole Polytechnique*, 1:24–76, 1795.

J. R. Deller, J. Proakis, and J. Hansen. *Discrete-Time Processing of Speech Signals*. 1993.

A. P Dempster, N. M Laird, and D. B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Z. L. Deng, Z. Kong, and L. Li. On functional equivalence of two measurement fusion methods. *Control Theory $0 Applications*, 23(2):319–323, 2006.

R. Dhaouadi, N. Mohan, and L. Norum. Design and implementation of an extended kalman filter for the state estimation of a permanent magnet synchronous motor. *IEEE Transactions on Power Electronics*, 6(3):491–497, 1991.

Y. Dodge, D. Cox, and D. Commenges. *The Oxford dictionary of statistical terms*. Oxford University Press, USA, 2006.

B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *Proc. ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. *Proc. Interspeech*, 2009.

T. Drugman and T. Dutoit. Chirp complex cepstrum-based decomposition for asynchronous glottal analysis. In *Proc. Eleventh Annual Conference of the International Speech Communication Association*, 2010.

T. Drugman, B. Bozkurt, and T. Dutoit. Chirp decomposition of speech signals for glottal source estimation. In *Proc. ISCA Workshop on Non-Linear Speech Processing*, 2009a.

T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. Interspeech*, pages 116–119, 2009b.

T. Drugman, T. Dubuisson, and T. Dutoit. On the mutual information between source and filter contributions for voice pathology detection. In *Proc. Interspeech*, 2009c.

T. Dubuisson, T. Drugman, and T. Dutoit. On the mutual information of glottal source estimation techniques for the automatic detection of speech pathologies. In *Sixth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2009.

J.A. Edwards and J.A.S. Angus. Using phase-plane plots to assess glottal inverse filtering. *Electronics Letters*, 32(3):192 –193, 1996.

E. Erelebi. Second generation wavelet transform-based pitch period estimation and voiced/unvoiced decision for speech signals. *Applied acoustics*, 64(1):25–41, 2003.

G. Fant. *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*. Mouton De Gruyter, 1970.

G. Fant. The source filter concept in voice production. In *IV FASE Symposium on Acoustics and Speech, Venezta*, 1981.

G. Fant. Some problems in voice source analysis. *Speech Communication*, 13 (1-2):7–22, 1993.

G. Fant. The LF-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2:3, 1995.

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

A. Filippidis, L. C Jain, and N. Martin. Multisensor data fusion for surface land-mine detection. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(1): 145–150, 2000.

J. L. Flanagan. *Speech analysis: Synthesis and perception*. 1972.

J. L. Flanagan, K. Ishizaka, and K. L. Shipley. Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *Bell Syst. Tech. J*, 54(3):485–506, 1975.

S. E. Franklin and C. F. Blodgett. An example of satellite multisensor data fusion. *Computers $0 Geosciences*, 19(4):577–583, 1993.

Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):492–501, 2006.

Q. Gan and C. J. Harris. Comparison of two measurement fusion methods for kalman-filter-based multisensor data fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 37(1):273–279, 2002.

J. B. Gao and C. J. Harris. Some remarks on kalman filters for the multisensor fusion. *Information Fusion*, 3(3):191–201, 2002.

Y. Gao, W. J. Jia, X. J. Sun, and Z. L. Deng. Self-tuning multisensor weighted measurement fusion kalman filter. *IEEE Transactions on Aerospace and Electronic Systems*, 45(1):179–191, 2009.

P. E. Gill, W. Murray, M. H. Wright, et al. *Numerical linear algebra and optimization*, volume 5. Addison-Wesley Redwood City, CA, 1991.

C. Gobl. *The voice source in speech communication-production and perception experiments involving inverse filtering and synthesis*. KTH, Department of Speech, Music and Hearing, 2003.

C. Gobl and A. N Chasaide. Techniques for analysing the voice source. *Coarticulation: Theory, Data and Techniques*, pages 300–320, 1999.

C. Gobl, A. Chasaide, and P. Hoole. Techniques for investigating laryngeal articulation and the voice-source. *Coarticulation: Theory, Data and Techniques*, pages 105–143, 1999.

D. L. Hall and R. J. Linn. Survey of commercial software for multisensor data fusion. In *Proceedings of SPIE*, volume 98, 1991.

D. L. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

D. L. Hall and S. A. H. McMullen. *Mathematical techniques in multisensor data fusion*. Artech House Publishers, 2004.

L. R. Han. *A Fuzzy-Kalman filtering strategy for state estimation*. PhD thesis, University of Saskatchewan, 2004.

M. Haseyama and H. Kitajima. An ARMA order selection method with fuzzy reasoning. *Signal processing*, 81(6):1331–1335, 2001.

N. Henrich, C. d'Alessandro, and B. Doval. Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data. In *Proc. Seventh European Conference on Speech Communication and Technology*, 2001.

M. J. Hinich and E. Shichor. Bispectral analysis of speech. In *Proc. 17th Convention of Electrical and Electronics Engineers in Israel*, pages 357–360, 1991.

M. J. Hinich and M. A. Wolinsky. A test for aliasing using bispectral analysis. *Journal of the American Statistical Association*, pages 499–502, 1988.

M. Hoshiya and E. Saito. Structural identification by extended kalman filter. *Journal of Engineering Mechanics*, 110:17–57, 1984.

F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the International Congress on Acoustics*, pages 17–20, 1968.

L. Janer, J. J. Bonet, and E. Lleida-Solano. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. In *Proc. Fourth International Conference on Spoken Language*, volume 2, pages 1209–1212, 1996.

S. J. Julier and J. K. Uhlmann. A new extension of the kalman filter to nonlinear systems. In *Int. symp. aerospace/defense sensing, simul. and controls*, volume 3, page 26, 1997.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

J. Kane. Exploiting time and frequency domain measures for precise voice source parameterisation. In *Proce. Speech Prosody*, 2012.

J. Kane, M. Kane, and C. Gobl. A spectral LF model based approach to voice source parameterisation. In *Proc. Eleventh Annual Conference of the International Speech Communication Association*, 2010.

D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, 87:820, 1990.

L. A. Klein. *Sensor and data fusion: a tool for information assessment and decision making*, volume 138. Society of Photo Optical, 2004.

J. Kominek and A. W. Black. The CMU arctic speech databases. In *Proc. Fifth ISCA Workshop on Speech Synthesis*, 2004.

A. Kounoudes, P. A. Naylor, and M. Brookes. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 349–352, 2002.

A. Krishnamurthy. Two channel (speech and egg) analysis for formant and glottal inverse filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 60–63, 1984.

L. Kumari, P. Raju, and C. V. Y. V. Application of extended kalman filter for a free falling body towards earth. *International Journal of Advanced Computer Sciences and Applications*, 2(4), 2011.

M. Ledvij. Curve fitting made easy. *Industrial Physicist*, 9(2):24–27, 2003.

J. H. Lee and N. L. Ricker. Extended kalman filter based nonlinear model predictive control. *Industrial $0 engineering chemistry research*, 33(6):1530–1541, 1994.

K. Lee and K. Park. Glottal inverse filtering (GIF) using closed phase WRLS-VFF-VT algorithm. In *Proc. IEEE Region 10 Conference (TENCON 99)*, volume 1, pages 646–649, 1999.

T. G. Lee. Centralized kalman filter with adaptive measurement fusion: its application to a GPS/SDINS integration system with an additional sensor. *International Journal Of Control Automation And Systems*, 1:444–452, 2003.

H. Li, D. O'Brien, and R. Scaife. Comparison of time- and frequency-domain based LF-model fitting methods for voice source parametrisation. In *Proc. 23nd IET Irish Signals and Systems Conference*, 2012a.

H. Li, R. Scaife, and D. O'Brien. Automatic LF-model fitting to the glottal source waveform by extended kalman filtering. In *Proc. 20th European Signal Processing Conference*, 2012b.

G. Liang, D. M. Wilkes, and J. A. Cadzow. ARMA model order estimation based on the eigenvalues of the covariance matrix. *IEEE Transactions on Signal Processing*, 41(10):3003–3009, 1993.

R. J. Linn, D. L. Hall, and J. Llinas. Survey of multisensor data fusion systems. In *Proc. Sensor Fusion Conference*, volume 1470, 1991.

H. L. Lu and J. O. Smith. *Toward a high-quality singing synthesizer with vocal texture control*. Stanford University, 2002.

R. C. Luo, M. H. Lin, and R. S. Scherp. Dynamic multi-sensor data fusion system for intelligent robots. *IEEE Journal of Robotics and Automation*, 4(4):386–396, 1988.

C. X. Ma, Y. Kamp, and L. F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, 1994.

S. Majumder, S. Scheding, and H. F. Durrant-Whyte. Multisensor data fusion for underwater navigation. *Robotics and Autonomous Systems*, 35(2):97–108, 2001.

J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63 (4):561–580, 1975.

D. Mandic, D. Obradovic, A. Kuh, T. Adali, U. Trutschell, M. Golz, P. De Wilde, J. Barria, A. Constantinides, and J. Chambers. Data fusion for modern engineering applications: An overview. *Artificial Neural Networks: Formal Models and Their Applications*, pages 752–752, 2005.

R. Mannell. Overview of vocal tract. 2009. URL `http://clas.mq.edu.au/phonetics/phonetics/introduction/vocaltract.html`.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2): 431–441, 1963.

J. G. McKenna. Automatic glottal closed-phase location and analysis by {k}alman filtering. In *Proc. 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

J. G McKenna. *Kalman Filtering Towards Automatic Speaker Characterisation.* PhD thesis, University of Edinburgh, 2004.

J. M. Mendel. Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proceedings of the IEEE*, 79(3):278–305, 1991.

H. B Mitchell. *Multi-sensor data fusion: an introduction.* Springer, 2007.

E. Moore and M. Clements. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 2004.

H. Motulsky and A. Christopoulos. *Fitting models to biological data using linear and nonlinear regression: a practical guide to curve fitting.* Oxford University Press, USA, 2004.

P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):34–43, 2007.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

C. L. Nikias and M. R. Raghuveer. Bispectrum estimation: A digital signal processing framework. *Proceedings of the IEEE*, 75(7):869–891, 1987.

A. O'Cinneide, D. Dorran, M. Gainza, and E. Coyle. A frequency domain approach to ARX-LF voiced speech parameterization and synthesis. In *Proc. Twelfth Annual Conference of the International Speech Communication Association*, 2011a.

A. O'Cinneide, D. Dorran, M. Gainza, and E. Coyle. Glottal inverse filtering with automatic filter order selection. In *Proc. Irish Signal Processing Conference*, 2011b.

F. Palmieri, S. Marano, and P. Willett. Measurement fusion for target tracking under bandwidth constraints. In *Proc. IEEE Aerospace Conference*, volume 5, pages 2179–2190, 2001.

M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *Speech and Audio Processing, IEEE Transactions on*, 7(5):569–586, 1999. ISSN 1063-6676.

W. H. Press, B. P. Flannery, Teukolsky, W. T. Vetterling, et al. *Numerical recipes*, volume 547. Cambridge Univ Press, 1986.

L. Rabiner and M. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(4):338–343, 1977a.

L. Rabiner and M. Sambur. Voiced-unvoiced-silence detection using the itakura LPC distance measure. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 323–326, 1977b.

L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*, volume 100. Prentice-hall Englewood Cliffs, NJ, 1978.

C. J. Ran, Y. S. Hui, L. Gu, and Z. L. Deng. Correlated measurement fusion steady-state kalman filtering algorithms and their optimality. *Acta Automatica Sinica*, 34(3):233–239, 2008.

J. R. Raol. *Multi-sensor data fusion with MATLAB*. CRC, 2009.

H. E. Rauch, C. T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

M. R. Remus. *Fuzzy logic applied to adaptive Kalman filtering*. PhD thesis, University of Nebraska - Lincoln, 1992.

M. I. Ribeiro. Kalman and extended kalman filters: Concept, derivation and properties. *Institute for Systems and Robotics*, 2004.

M. Richard. A sensor classification scheme. 1987.

E.L. Riegelsberger and A.K. Krishnamurthy. Glottal source estimation: Methods of applying the LF-model to inverse filtering. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 542–545, 1993.

J. A. Roecker and C. D. McGillem. Comparison of two-sensor tracking methods based on state vector fusion and measurement fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 24(4):447–449, 2002.

R. K. Saha. Track-to-track fusion with dissimilar sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 32(3):1021–1029, 1996.

R. K. Saha and K. C. Chang. An efficient algorithm for multisensor track fusion. *IEEE Transactions on Aerospace and Electronic Systems*, 34(1):200–210, 1998.

J. Z. Sasiadek. Sensor fusion. *Annual Reviews in Control*, 26(2):203–228, 2002.

V. Shin, Y. Lee, and T. S Choi. Generalized millman's formula and its application for estimation problems. *Signal Processing*, 86(2):257–266, 2006.

R. H Shumway and D. S Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, 3(4): 253–264, 1982.

L. Siegel. A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(1):83–89, 1979.

L. Siegel and K. Steiglitz. A pattern classification algorithm for the voiced/unvoiced decision. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 326–329, 1976.

P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

Brad H. Story. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206, 2002.

Brad H. Story and Ingo R. Titze. Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, 97(2): 1249–1260, 1995.

H. Strik. The effect of low-pass filtering on estimated voice source parameters. In *Proc. Fifth European Conference on Speech Communication and Technology*, 1997.

H. Strik. Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *The Journal of the Acoustical Society of America*, 103:26–59, 1998.

H. Strik and L. Boves. On the relationship between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11:167–174, 1992.

H. Strik and L. Boves. Automatic estimation of voice source parameters. In *Proc. International Conference on Spoken Language Processing*, pages 155–158, 1994.

H. Strik, B. Cranen, and L. Boves. Fitting a LF-model to inverse filter signals. In *ESCA 3rd European Conference on Speech Communication and Technology: EUROSPEECH 93, Berlin*, pages 103–106, 1993.

D. Talkin. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495, 1995.

Paul Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

Y. T. Ting and D. G. Childers. Speech analysis using the weighted recursive least squares algorithm with a variable forgetting factor. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 389–392, 1990.

Ingo R. Titze and Brad H. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112(3):1064–1076, 2002.

C. W. Tong, S. K. Rogers, J. P. Mills, and M. K. Kabrisky. Multisensor data fusion of laser radar and forward looking infrared (FLIR) for target segmentation and enhancement. In *Proce. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 782, pages 10–19, 1987.

M. Tooher and J. G McKenna. Variation of glottal LF parameters across f0, vowels, and phonetic environment. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

D. Tuan Pham, J. Verron, and M. Christine Roubaud. A singular evolutive extended kalman filter for data assimilation in oceanography. *Journal of Marine systems*, 16(3):323–340, 1998.

R. Veldhuis. A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation. *The Journal of the Acoustical Society of America*, 103, 1998.

D. Vincent, O. Rosec, and T. Chonavel. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM

modeling. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, 2007.

J. Walker. Application of the bispectrum to glottal pulse analysis. In *ISCA Tutorial and Research Workshop on Non-Linear Speech Processing*, 2003.

J. Walker and P. Murphy. Advanced methods for glottal wave extraction. *Nonlinear Analyses and Algorithms for Speech Processing*, pages 139–149, 2005.

J. Walker and P. Murphy. A review of glottal waveform analysis. *Progress in nonlinear speech processing*, pages 1–21, 2007.

J. Wang, R. Achanta, M. Kankanhalli, and P. Mulhem. A hierarchical framework for face tracking using state vector fusion for compressed video. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, 2003.

G. Welch and G. Bishop. An introduction to the kalman filter. *University of North Carolina at Chapel Hill, Chapel Hill, NC*, 7(1), 1995.

B. Wells. Voiced/unvoiced decision based on the bispectrum. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 1589–1592, 1985.

D. Wong, J. Markel, and A. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):350–355, 1979.

B. Yegnanarayana and R. N. J. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *Speech and Audio Processing, IEEE Transactions on*, 6(4):313327, 1998.

Y. Zeng, J. Kirkland, J. F. Anderson, L.J. Leftin, and R.W. Briske. *Methods and systems for implementing an iterated extended Kalman filter within a navigation system.* Google Patents, 2011.