Multi-parametric source-filter separation of speech and prosodic voice restoration

THÈSE N^{O} 5474 (2012)

PRÉSENTÉE LE 26 OCTOBRE 2012
À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE TRAITEMENT DES SIGNAUX 1
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Olaf SCHLEUSING

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Dr J.-M. Vesin, directeur de thèse
Dr A. Drygajlo, rapporteur
Prof. E. Grivel, rapporteur
Dr T. Kinnunen, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2012

The question of whether computers can think is like the
question of whether submarines can swim.
— Edsger W. Dijkstra

The greatest enemy of knowledge is not ignorance,
it is the illusion of knowledge.
— Stephen Hawking

The more you know,
the more you realize you know nothing.
— Socrates

To my wife . . .

Acknowledgements

First of all I would like to address a huge thank-you to my thesis director, Dr. Jean-Marc Vesin, for having given me the possibility to do this thesis under his supervision. His countless technical inputs, remarks and advices gave the general technical direction for this thesis. Even more important, his great personality for me was a true inspiration and often an anchor point in emotionally difficult phases of this journey. Although our interactions were less frequent due to geographical distances than I would have wished, I could not imagine to have come so far without his help. Thank you, Jean-Marc.

A huge appreciation also goes to my supervisors at CSEM, Dr. Philippe Renevey and, former CSEM employee Dr. Rolf Vetter. They have set up the financial and administrative framework for this project and guided me through the project. Their numerous technical ideas and our countless joint discussions provided immeasurable inspirations and constitute a big portion of the outcome of this thesis.

Furthermore, I would like to address a great thank-you to Dr. Tomi Kinnunen of the University of Eastern Finland in Joensuu. His great technical knowledge, his admirable work attitude, our countless technical, personal and general discussions concerning scientific speech processing in general will always be remembered and are a profound building stone of this thesis. It is people like him that made this journey truly memorable and make this research community the fantastic and exciting place it is.

An aspect of profound importance for a successful conclusion of a work like this one is a supportive and encouraging work environment and atmosphere. For the past four years I was lucky enough to have been in exactly such a place and it is therefore that I would like to thank the respective people at CSEM for having created such a great infrastructure and working climate. It is the late Thomas Hinderling, Mario El Khoury, Jens Krauss, Dr. Olivier Chételat and in particular Dr. Mattia Bertschi that are responsible for having created such a truly inspirational environment.

I would also like to thank the Systems Engineering Division of CSEM for partly funding my PhD position at CSEM. Furthermore I would like to thank the members of the Signal Processing Group and the Control Group of the Systems Engineering Division of CSEM for many memorable moments. It was also always a pleasure to work with and to have fruitful discussions with my colleagues Stephan Dasen, Josep Sola, Olivier Grossenbacher, Martin Proença, Enric Muntane Calvo, Mathieu Lemay, Andrea Ridolfi, Damien Ferrario, Patrick Theurillat, Christophe Verjus, Yves Stauffer, Victor Neuman, Leopoldo Rossini, Emmanuel Onillon and Patrick Çelka. Special thanks also go to Michele Siggen, Jana Ackermann and

Mayssa Cherif for their special care-taking of all administrative matters. Their overwhelming efficiency is unmatched. Thanks to everybody of you for all technical help you gave me.

Furthermore, I would also like to thank some people that were not directly involved in this work, but who nevertheless played an important role on my path. I had the pleasure to meet and to work with some truly admirable and inspirational people at the National University of Singapore. In particular, I would like to mention Prof. Ye Wang and Dr. Bingjun Zhang, who provided me the opportunity to look beyond the horizon. Besides that I would also like to express my thanks to Pierre-François Laquerre for the time spent together at EPFL working on student projects.

I also would like thank various people who helped throughout the project with helpful technical and also non-technical discussions: Prof. Paavo Alku, Prof. Brad Story, Prof. Bart Verkerke, Joanna Bitterli, Roland Bitterli, Stefan Weber, Luca Dassa, Matteo Contaldo, Fabio Jutzi and many others. Dr. Ward van der Houwen I would like to thank for his honesty, his inspirational character, for his humour and especially for his friendship.

A very special thank-you goes to my parents and to my sister. It is their believe in me and their support in so many aspects that has enabled me to become who I am and to reach this far. Thank you for always looking after me, pushing me into the right direction (which was sometimes very necessary) and for the countless discussions, practical tips and the priceless guidance. You are truly a gift. Above all this, I would like to thank my wife, Shamla, for her patience, understanding and unconditional support, in particular during these last, very exhausting and intensive months. I am sure, without you I would not have come so far. You are the most supportive person one can imagine. Thank you.

A very distinguished thank-you also goes to the *Recherche Suisse Contre le Cancer* for their help in enabling this research through their financial support (KFS 02681-08-2010 and OCS - 01777-08-205). The same gratitudes are expressed to the *Gebert-Rüf-Stiftung* for their financial support (GRS-045/05).

It is impossible to name everyone who has contributed to such a large undertaking. If ever I forgot anyone who helped me or contributed to this work I would like to apologize and express my thanks to him/her as well!

Neuchâtel, 13 June 2012 O. S.

Abstract

In this thesis, methods and models are developed and presented aiming at the estimation, restoration and transformation of the characteristics of human speech. During a first period of the thesis, a concept was developed that allows restoring prosodic voice features and reconstruct more natural sounding speech from pathological voices using a multi-resolution approach. Inspired from observations with respect to this approach, the necessity of a novel method for the separation of speech into voice source and articulation components emerged in order to improve the perceptive quality of the restored speech signal. This work subsequently represents the main part of this work and therefore is presented first in this thesis. The proposed method is evaluated on synthetic, physically modelled, healthy and pathological speech. A robust, separate representation of source and filter characteristics has applications in areas that go far beyond the reconstruction of alaryngeal speech. It is potentially useful for efficient speech coding, voice biometrics, emotional speech synthesis, remote and/or non-invasive voice disorder diagnosis, etc.

A key aspect of the voice restoration method is the reliable separation of the speech signal into voice source and articulation for it is mostly the voice source that requires replacement or enhancement in alaryngeal speech. Observations during the evaluation of above method highlighted that this separation is insufficient with currently known methods. Therefore, the main part of this thesis is concerned with the modelling of voice and vocal tract and the estimation of the respective model parameters.

Most methods for joint source filter estimation known today represent a compromise between model complexity, estimation feasibility and estimation efficiency. Typically, single-parametric models are used to represent the source for the sake of tractable optimization or multi-parametric models are estimated using inefficient grid searches over the entire parameter space. The novel method presented in this work proposes advances in the direction of efficiently estimating and fitting multi-parametric source and filter models to healthy and pathological speech signals, resulting in a more reliable estimation of voice source and especially vocal tract coefficients. In particular, the proposed method is exhibits a largely reduced bias in the estimated formant frequencies and bandwidths over a large variety of experimental conditions such as environmental noise, glottal jitter, fundamental frequency, voice types and glottal noise. The methods appears to be especially robust to environmental noise and improves the separation of deterministic voice source components from the articulation.

Alaryngeal speakers often have great difficulty at producing intelligible, not to mention prosodic, speech. Despite great efforts and advances in surgical and rehabilitative techniques,

currently known methods, devices and modes of speech rehabilitation leave pathological speakers with a lack in the ability to control key aspects of their voice. The proposed multiresolution approach presented at the end of this thesis provides alaryngeal speakers an intuitive manner to increase prosodic features in their speech by reconstructing a more intelligible, more natural and more prosodic voice. The proposed method is entirely non-invasive. Key prosodic cues are reconstructed and enhanced at different temporal scales by inducing additional volatility estimated from other, still intact, speech features. The restored voice source is thus controllable in an intuitive way by the alaryngeal speaker.

Despite the above mentioned advantages there is also a weak point of the proposed joint source-filter estimation method to be mentioned. The proposed method exhibits a susceptibility to modelling errors of the glottal source. On the other hand, the proposed estimation framework appears to be well suited for future research on exactly this topic. A logical continuation of this work is the leverage the efficiency and reliability of the proposed method for the development of new, more accurate glottal source models.

Keywords: Global optimization, differential evolution, joint source-filter estimation and separation, glottal inverse filtering, time-varying vocal tract estimation, alaryngeal voice restoration, prosodic voice restoration, expressive speech synthesis

Zusammenfassung

In dieser Arbeit werden Methoden und Modelle entwickelt und vorgestellt, welche die Schätzung, Restaurierung und Umgestaltung von Merkmalen der menschlichen Sprache zum Ziel haben. Zu Beginn der Arbeit wurde ein Konzept entwickelt und vorgestellt, welches es ermöglicht, prosodische Merkmale in pathologischen Stimmen mit Hilfe eines Multi-Resolution-Ansatzes wiederherzustellen. Des weiteren kann eine natürlichere Sprachwiedergabe von pathologischen Stimmen erreicht werden. Andererseits zeigten die Ergebnisse welche mit diesem Ansatz erreicht wurden auch, dass ein neues, verbessertes Verfahrens für die Trennung von Sprache in Stimm- und Artikulations-Komponenten notwendig ist um eine höhere Qualität der rekonstruierten Stimme zu erreichen. Dieser Teil der Arbeit bildete im Folgenden den Schwerpunkt dieser Arbeit aus und wird aus diesem Grund im ersten Teil dieser Thesis präsentiert. Eine solche separate Modellierung der Stimm-und Filtereigenschaften hat Anwendungen in vielen Bereichen welche weit über die Rekonstruktion von pathologischen Stimmen hinausgehen. Es ist potentiell nützlich für eine effiziente Sprachkodierung-, Sprecherkennung und -identifizierung, emotionale Sprachsynthese, Fern- und/oder nicht-invasive Diagnose von Stimmerkrankungen, etc.

Ein wesentlicher Aspekt dieser Methode ist die zuverlässige Trennung des Sprachsignals in Stimm- und Artikulationsanteile. In der oben beschriebenen Anwendung, zum Beispiel, ist nur die Stimme zu ersetzen, während die Artikulation erhalten bleiben soll. Die Beobachtungen während der Auswertung des obigen Verfahrens haben hervorgehoben, dass derzeit bekannte Verfahren diese Trennung nur unzureichend gewährleisten. Aus diesem Grund widmet sich der Hauptteil dieser Arbeit der Modellierung von Stimmbildung und Vokaltrakt sowie der Schätzung der jeweiligen Modellparameter.

Die meisten heute bekannten Methoden für die gemeinsame Schätzung von Stimme und Artikulation stellen einen Kompromiss zwischen Komplexität des Modells, dessen Optimierbarkeit und der Effizienz der Optimierung dar. Typischerweise werden sehr einfache Stimm-Modelle benutzt, um die Optimierung zu vereinfachen oder multi-parametrische Modelle werden mittels recht ineffizienter Methoden wie der systematischen Suche über den Raum aller möglicher Parameterkombinationen geschätzt. Das in dieser Arbeit neu vorgestellte Verfahren setzt genau an dieser Stelle an und ermöglicht die Schätzung und die effiziente Anpassung eines multi-parametrischen Stimm- und Vokaltrakt-Modells an gesunde und pathologische Sprachsignale. Die vorgestellte Methode ermöglicht eine zuverlässigere Schätzung von Modell-Parametern, insbesondere der Vokaltrakt-Koeffizienten. Vor allem zeigen die Experimente, dass das vorgeschlagene Verfahren eine stark verbesserte Genauigkeit der geschätzten

Formant-Frequenzen und Bandbreiten ermöglicht. Dies wurde unter einer Vielzahl von experimentellen Bedingungen wie Umgebungsrauschen, Jitter, variierender Grundfrequenz, Stimmtypen und Stimm-Rauschen untersucht. Die vorgestellte Methode ist sehr verlässlich in Gegenwart von Umgebungsrauschen und stellt eine klare Verbesserung dar wenn es darum geht Sprache in seine Stimm- und Artikulationsanteile zu zerlegen.

Pathologische Sprecher haben oft grosse Schwierigkeiten bei der Formulierung verständlicher Sprache, oft ganz zu schweigen von prosodischen Sprachmerkmalen. Trotz grosser Anstrengungen und Fortschritte in chirurgischen und rehabilitativen Techniken sind derzeit bekannte Verfahren und Geräte zur Sprachrehabilitation nicht in der Lage einen Mangel in der Kontrollierbarkeit von wichtigen Stimmeigenschaften auszugleichen. Der im letzten Teil dieser Arbeit vorgestellte Multi-Resolution-Ansatz bietet pathologischen Sprechern die Möglichkeit auf intuitive Weise prosodische Merkmale zu einer natürlich klingenden Stimme zu rekonstruieren. Die vorgeschlagene Methode ist komplett nicht-invasiv. Prosodie wird dabei auf verschiedenen zeitlichen Skalen rekonstruiert und durch Erhöhung zusätzlicher Volatilität verstärkt. Dazu werden andere, intakte Sprachmerkmale benutzt, um relevante Aspekte für die Prosodie-Rekonstruktion zu schätzen. Die restaurierte Stimme ist somit durch den pathologischen Sprecher in einer intuitiven Weise steuerbar.

Trotz der oben genannten Vorteile hat die vorgestellte Methode auch Schwachpunkte in Form von einer Anfälligkeit für grobe Fehler der Modellierung der Stimmquelle. Nichtsdestotrotz erscheint die vorgeschlagene Methode sehr geeignet für die zukünftige Forschung auf genau diesem Thema. Eine logische Fortsetzung dieser Arbeit ist daher die Ausnutzung der Effizienz und Zuverlässigkeit des vorgeschlagenen Verfahrens für die Entwicklung von neuen, genaueren Stimm-Modellen, um genau diese Schwachstelle zu adressieren.

Stichwörter: Globale Optimierung, differential evolution, gemeinsame Stimm-Vokaltrakt-Schätzung und -Trennung, inverse Stimmfilterung, zeitlich variierende Vokaltrakt-Schätzung, pathologische Stimmrestaurierung, prosodische Stimmrestaurierung, ausdrucksvolle Sprachsynthese

Contents

Acknowledgements **iii**

Abstract (English/Français/Deutsch) **v**

Lists of figures, tables and symbols **xi**

Introduction **1**

1 Introduction **1**

1.1 Problematics 2

1.1.1 Source-Filter Separation of Speech 2

1.1.2 Restoration and Transformation of Voice 3

1.1.3 The Joint Source-Filter Separation Methods and their Evaluation 4

1.2 Outline 4

I Speech Processing Models **7**

2 Speech Production **9**

2.1 Speech Production Anatomy 9

2.1.1 Voice Source 12

2.1.2 Voice Source in Pathological Speakers 16

2.1.3 Vocal Tract 17

2.2 Mechanical-Acoustic Model 19

2.2.1 Mechanical-Acoustic Source Model 19

2.2.2 Acoustic Vocal Tract Model 21

2.3 Source-Filter Model 22

2.3.1 Glottal Source Models 24

2.3.2 Non-deterministic Voice Components 32

2.3.3 The Vocal Tract Filter 33

2.3.4 Speech Production Process 38

2.4 Conclusion 39

3 Theory and Tools for Source-Filter Separation 41
3.1 General Considerations for Source-Filter Separation . . . 41
3.2 Autoregressive methods . . . 43
  3.2.1 The Normal Equations . . . 43
  3.2.2 Solution of Normal Equations . . . 44
  3.2.3 Stability of the Vocal Tract Filter . . . 46
  3.2.4 Limitations of Linear Prediction . . . 47
3.3 Source-Filter Separation Methods . . . 51
  3.3.1 Pre-emphasis filtering . . . 52
  3.3.2 Analysis-by-Synthesis Methods . . . 52
  3.3.3 Iterative Adaptive Inverse Filtering (IAIF) . . . 53
  3.3.4 Adaptive Estimation of the Vocal Tract (AEVT) . . . 53
  3.3.5 Closed-Phase Linear Prediction . . . 53
  3.3.6 Cepstral Analysis . . . 54
  3.3.7 Joint Estimation Approaches using Parametric Glottal Models . . . 56
3.4 Global Optimization using Differential Evolution . . . 57
3.5 Conclusion . . . 65

II Proposed Methods 67

4 Source-Filter Separation using Differential Evolution 69
4.1 Formulation of Optimization Problem . . . 69
4.2 Conditions for Convergence . . . 71
4.3 Necessary Constraints on Voice and VT model . . . 72
4.4 Implementation Details . . . 75
  4.4.1 Optimization Parameters . . . 75
  4.4.2 Compensation of Overlapping Resonances . . . 77
4.5 Experimental Validation . . . 78
  4.5.1 General Considerations and Problematics . . . 78
  4.5.2 Reference Methods . . . 80
  4.5.3 Generation of Synthetic Speech . . . 81
  4.5.4 Performance Comparison . . . 82
4.6 Conclusion . . . 91

5 Application of Source-Filter Separation 95
5.1 Source-Filter Separation on Physically Modelled Speech . . . 95
5.2 Source-Filter Separation of Real Speech Signals . . . 99
  5.2.1 Test material description . . . 99
  5.2.2 Discussion of Formant Estimation . . . 101
5.3 Conclusion . . . 111

6 Prosody Restoration **115**
6.1 Introduction 115
6.2 Speech Rehabilitation for Alaryngeal Speakers 117
6.3 Characteristics of Pathological Speech 117
6.4 Multi-Resolution Voice Restoration 118
6.4.1 Method 118
6.4.2 Long-Term f_0-Estimation 119
6.4.3 Middle-Term f_0-Restoration 124
6.4.4 Short-Term f_0-Restoration 124
6.5 Results 125
6.6 Conclusion 128

7 General Conclusions **129**

Bibliography **132**

List of Publications **149**

Curriculum Vitae **151**

List of Figures

2.1 Speech production models 10
2.2 Schematic midsagittal head section 11
2.3 Three-layer model of the vocal folds 13
2.4 Illustration of the mucosal wave 14
2.5 Comparison of different vocal tract vowel configurations 18
2.6 Three-mass model of the vocal folds 20
2.7 Vocal tract area function 22
2.8 Glottal flow model descriptions 25
2.9 Influence of E_e 30
2.10 Influence of Oq 30
2.11 Influence of α_m 31
2.12 Influence of Q_a 31
2.13 First order all-pole filter 36
2.14 Second order all-pole filter 37
2.15 All-pole filter modelling four resonance regions 38
2.16 Speech Production in the source-filter model 39

3.1 Associativity of prediction filter 46
3.2 Influence of f_0 on the VT envelope estimated using LP 49
3.3 Spectrogram of Vowel Transition with overlaid Formant Estimates 51
3.4 Cepstral smoothing 55
3.5 Peaks Function 62
3.6 Differential Evolution Example: Stage 1 62
3.7 Differential Evolution Example: Stage 2 63
3.8 Differential Evolution Example: Stage 3 63
3.9 Differential Evolution Example: Stage 4 64
3.10 Differential Evolution Example: Stage 5 64

4.1 Z-transform of LF Model 74
4.2 Overlapping of Successive Glottal Cycle Resonances 76
4.3 Example of a Glottal Cycle Optimization Process 79
4.4 Formant estimation error with respect to environmental noise 84
4.5 Formant estimation error with respect to fundamental frequency 85

4.6 Formant estimation error with respect to glottal jitter 86
4.7 Glottal source error with respect to background noise, fundamental frequency and pitch jitter 88
4.8 Formant estimation error with respect to glottal noise 92
4.9 Glottal source error with respect to glottal source distortion 93

5.1 Glottal Source Signal estimated using DE 96
5.2 Real speech: example A 102
5.3 Formant estimation on real speech: example A 103
5.4 Real speech: example B 104
5.5 Formant estimation on real speech: example B 105
5.6 Real speech: example C 106
5.7 Formant estimation on real speech: example C 107
5.8 Glottal source parameters estimated from real speech: example A 108
5.9 Glottal source parameters estimated from real speech: example B 110
5.10 Glottal source parameters estimated from real speech: example C 113

6.1 Comparison of a laryngeal and an alaryngeal voice source signal 119
6.2 Block diagram of the multi-resolution f_0-restoration method 120
6.3 Performance comparison of f_0-estimation methods 123
6.4 Illustration of the correlation between the speech signal intensity and f_0 125
6.5 Spectrograms of alaryngeal and reconstructed speech signals 127

List of Tables

4.1 Formant Boundary Constraints 77
4.2 Boundary constraints of Source Model 77
4.3 Formant frequencies and bandwidths (Bw) in Hz used for synthesizing the test material for the first two experiments. 82
4.4 Voice model parameters used for synthesis of DE validation experiments 83

5.1 Formant frequency estimation results using speech with a pressed voice synthesized by the physical model. 98
5.2 Formant frequency estimation results using speech with a modal voice synthesized by the physical model. 99
5.3 Formant frequency estimation results using speech with a breathy voice synthesized by the physical model. 100

6.1 Results of f_0-estimation method evaluation. 124
6.2 Results of Voice Restoration 126

List of Symbols

AR	Autoregressive Process	33
ARMA	Autoregressive Moving Average Process	43
ARX	Autoregressive Model with exogenous input	69
AWGN	Additive White Gaussian Noise	124
AWO	Adaptive Wavetable Oscillator	115
CPLP	Closed-Phase Linear Prediction	48
DAP	Discrete All-Pole Modeling	48
DE	Differential Evolution	57
DFT	Discrete Fourier Transform	29
EA	Evolutionary Algorithm	57
EGG	Electroglottography	52
ES	Evolution Strategies	57
GA	Genetic Algorithms	57
GCI	Glottal Closure Instant	29
GIF	Glottal Inverse Filtering	41
GOI	Glottal Opening instant	29
IAIF	Iterative Adaptive Inverse Filtering	53
IS	Itakura-Saito distance	82
LF	Liljencrants-Fant Voice Source Model	26
LP	Linear Prediction	43
MA	Moving Average Process	43
MOS	Mean Opinion Score	123
MRAE	Mean Relative Absolute Error	123
PDF	Probability Distribution Function	58
PE	Preemphasis Filter	43
PSLP	Pitch-Synchronous Linear Prediction	48
PSO	Particle Swarm Optimization	61
ROC	Region of Convergence	34
SFO	Source-Filter Optimization	56

SNR	Signal-To-Noise Ratio	82
SPL	Sound Pressure Level	29
VT	Vocal Tract	10
VTF	Vocal Tract Filter	17
WSS	Wide Sense Stationary	44
ZZT	Zeros of Z-Transform	56

1 Introduction

Speech is an incredibly versatile and tremendously important evolutionary gift passed to humans. It enables us to formulate and communicate abstract concepts that convey information far more complex than simple emotions or desires. Since the days that scientific methodologies earned their first merits it has been of interest to researchers to understand and reproduce the mechanisms underlying the production of speech.

Despite decades of research and countless brilliant ideas, many questions surrounding speech remain unanswered. This is partly due to the aforementioned complexity of the process itself, but also due to the impossibility to directly observe the speech production process without actually distorting it. No method is known that allows to measure parameters relevant for voice production at the vocal folds without actually altering the ultimately produced speech signal.

In order to gain tractability over such complexity, researchers build models that simplify reality to a degree that is acceptable for the application at hand. Globally, two types of models have emerged. On one hand side are models with a relatively large degree of freedom that attempt to depict physical, physiological and aero-dynamic phenomena as exactly as possible, either in the form of mechanical or numerical models. Typically, these types of models have tens up to hundreds of parameters (Story and Titze, 1995; Yanga et al., 2011; Inwald et al., 2011). The objective pursued with these models is to obtain a thorough understanding of the processes involved in speech production. This is done by posing hypotheses followed by their validation using simulation with the goal of obtaining simulation results that match real, observed voices as close as possible.

On the other hand, models used for speech processing applications are generally rather simple. They typically have a lower degree of freedom and are not necessarily inspired by the anatomy and physiology of speech production. The main aspect of these models is to provide a parametrical representation of those aspects of the speech that are relevant for the particular application of interest. This correlation between parameters and application-relevant cues is not necessarily intuitive. Its complete understandung is often hindered by the lack of understanding of the perception of speech and sound in general. For instance, prosody is a perceptual concept describing the semantics of language well beyond what is literally being said. It is understood that the fundamental frequency of the voice source and the perceived pitch are a fundamental cue related to prosody, but it is also known that it is not the only cue leading to the perception of prosody. It is not well understood, which cues in particular are missing and how they exactly would correlate with the perceived prosody. Due to this lack of understanding, amusingly in speech processing it is often found that certain phenomena are merely defined as what they are *not*. For instance, the timbre of a sound is often described as what differentiates one sound from another despite their equality in terms of fundamental frequency and loudness (Klapuri and Davy, 2006), but it is left open what exactly makes this difference.

This lack of understanding of the relationship between physcially measurable and perceived cues has contributed to the success of machine learning techniques in speech processing applications. New models were developed that do not necessarily provide an intuitive link between the model parameters and speech production anatomy. Cepstral coefficients are a well-known example of such a model. They have proved very successful in applications such as speech and speaker recognition, yet it is not very well understood why they work so well.

In this work we are concerned with models used for the alteration of voice source characteristics. For this purpose, the voice source signal needs to be separated first from other components of an observed speech signal. In a first period of this project, concerned with the restoration of pathological speech, this separation was carried out using the universal standard tool for this task, linear prediction. An altered voice source was reconstructed from healthy voice source patterns and using modified prosody cues. While the main objective of this first project was achieved by improving prosodic cues, it also became clear that this approach imposes several drawbacks. The estimation of voice-related parameters from the speech signal proved too inaccurate for a successful, unobstructed restoration of the voice source signal. Based on these observations, the main objective of this work then concentrated on the improvement of the accuracy of the separation of speech into its components in order to facilitate their modification and alteration. Since this latter part constitutes the majority of the contribution of this work it is presented in the first part of this thesis, followed by the description of the work on alaryngeal voice restoration.

1.1 Problematics

1.1.1 Source-Filter Separation of Speech

The most general description of speech production is provided by the well-known source-filter model of speech (Fant, 1960). It is inspired by the anatomical presence of a voice source, most prominently the vocal folds, and the modulation of the resulting signal by resonances occurring in the vocal tract. Many speech signal processing applications employ a more or less complex parametric representation of this source-filter model. In the simplest case, the source is modelled as a series of Dirac pulses, representing the consecutive, high-energy closing instants of the vocal folds. Our observations during the first phase of the project have led us to believe that this model is insufficient for a successful restoration of pathological speech signals or for the alteration of voice source characteristics in general. In such a simple model, the source and the filter are jointly represented in a single model, which is commonly dedicated to the vocal tract alone. The source characteristics are assumed to be neutralized before this estimation takes place. This model proved too simple. Instead, we decided to employ a multi-parametric model of the voice source in order to capture the source characteristics in a dedicated model, separate from the filter characteristics. A joint estimation process simultaneously finds optimal parameters for both models and ideally captures source and filter characteristics in the dedicated models. Assuming that real speech indeed may be separated in this way, this separation would allow to modify the observed voice source in a parametric way and to reconstruct speech by subsequently joining the modified source parameters with the previously estimated filter parameters in an appropriate manner.

As our survey of the state-of-the-art in Chapter 3 shows, a significant amount of research has addressed this issue of source-filter separation during the past twenty years. Many methods approach the topic by utilizing the commonly used frame-based analysis, in which an averaged, spectral representation of the voice source and respective models are estimated. More advanced methods employ a dedicated, time-domain description of the glottal signal waveform to capture source characteristics. The methods for jointly optimizing such models usually represent a trade-off between numerical efficiency and model complexity. This thesis proposes efficient estimation methods utilizing multi-parametric models.

1.1.2 Restoration and Transformation of Voice

As mentioned above, the first part of this project focused on a method and a device addressing the restoration of healthy-like and prosodic features in pathological speech. The method is intended for use in a real-time scenario in a device for the restoration of authentic, f_0-related characteristics in pathological speech uttered by subjects with laryngeal disorders. The original speech signal is acquired and analyzed by the device and a speech signal with improved, healthy-like features is reconstructed. For the reconstruction, different cues of the original, acquired signal are used.

In order to obtain a perceptionally superior voice in the reconstructed speech signal, the pathological excitation is replaced by a concatenation of healthy, glottal waveform patterns, which are randomly chosen from a reference database. Furthermore, to increase the naturalness of the f_0-variability in the reconstructed voice source, a multi-resolution approach is used to determine the instantaneous intervals between subsequent reference patterns. In particular, f_0-variability is reconstructed using different cues for its reconstruction at different time scales. The long-term f_0-trend is estimated from the remaining pitch variability found in pathological voices. Furthermore, the middle-term f_0-variability is restored through its correlation with speech intensity or loudness. For the reconstruction of short-term f_0-variability, a statistical noise model is used to induce jitter based on the instantaneous loudness of the speech signal. Two authentic features are used to assess the method's performance, namely breathiness and prosody. Preliminary results indicate that breathiness of the restored signal is reduced and prosody related features are improved. On the other hand, it also became apparent that better methods for source-filter separation were required to obtain more reliable VTF estimates, which was the main motivation for the methods introduced below.

1.1.3 The Joint Source-Filter Separation Methods and their Evaluation

We propose a novel joint SFO approach, in which the voice source is modeled using the multi-parametric Liljencrants-Fant model described in Section 2.3.1. The proposed method is based on a pitch-synchronous analysis-by-synthesis approach. In contrast to traditional analysis-by-synthesis methods, we do not use a codebook to generate reference speech signal patterns. Instead, a time-varying auto-regressive VT model with exogenous input is used to generate candidate solutions. *Differential evolution* serves as a computational tool to optimize the source and the filter parameters. The objective function is constructed so as to reduce the effect of inter-glottal-cycle resonances to increase the effective duration of the analysis window. The efficiency of the DE method allows us to carry out extensive experiments on different speech signals. The proposed optimization method converges reliably under a variety of conditions such as environmental and glottal noise, varying fundamental frequencies, jitter and vowel transitions.

In this thesis we aim to address the following objectives:

- Develop a better understanding of the particular problems occuring in pathological speech, in particular problems related to the perception and production of pathological speech as well as the inherent challenges from a signal processing point of view.
- Propose, develop and evluate a pathological speech restoration method addressing and taking into account the identified problems.
- Understand and highlight the problems inherent to currently used source-filter separation methods.
- Investigate and propose an approach that aims for a more reliable and more distinctive separation of speech components.
- Validate the proposed approach in a series of experiments using controlled environment variables.
- Explore and evaluate the proposed method's performance in a series of experiments using physically modelled and real speech signals.
-

1.2 Outline

This document is structured as follows:

Chapter 2 provides a detailed description of anatomical and physical processes involved in the production of speech. We will further detail how various models represent the particularities of speech production with different levels of detail.

Chapter 3 presents an overview of the current state-of-the-art of source filter separation. We present different approaches and highlight some of the details of the most prominent methods. We describe some of their advantages and disadvantages, which provide the motivation for the methods presented in this work. The chapter is then concluded with a description of a computational tool used as a cornerstone for the proposed methods called differential evolution.

Chapter 4 presents a detailed description of the proposed joint source-filter separation approach, a formulation of the optimization problem as well as an illustration of implementation details. In the second part of that chapter, the proposed approach is validated by a pre-liminary evaluation using synthetic speech signals.

Chapter 5 is concerned with the application and evaluation of the proposed method to more realistic speech signals. Since a quantitative evaluation using real speech signals is nearly impossible, this evaluation is split into two parts. In a first part, physically modelled speech is used for an objective analysis and in a second part a qualitative analysis using real speech signals is presented.

Chapter 6 presents a multi-resolution approach addressed at prosody restoration in pathological speech that was developed in an early phase of this work. The observations and lessons learned from this approach sparked the development of the methods proposed in the earlier chapters.

Chapter 7 summarizes the presented work and discusses the obtained results. Furthermore, application scenarios are discussed and directions for future research based on the obtained results are pointed out.

Speech Processing Models Part I

2 Speech Production

The production of speech in humans is an incredibly complex process. Naturally, complexity is best handled by building models that simplify reality. A model makes processes and systems tractable while conveying sufficient of the original complexity to fulfil an application's needs. Various models of the process of speech production exist, many of them *complete* but *approximate*. Globally it is known how people use their voice organs and articulators to produce the various sounds of speech. Nevertheless, our knowledge of this process is very approximate. No model as yet can considerably accurately predict how a speech waveform from a particular speaker would look like assuming that his or her intended linguistic message is known. Given the incredible complexity of the production process, this is not very surprising. For instance, we can make general assumptions about how a sound is pronounced, but it is much more difficult to describe how speech production varies from speaker to speaker, how it is affected by prosody or surrounding sound. This lack of knowledge has in some research areas led to the abandoning of trying to mimic the human production process entirely.

Instead, models of various degrees of complexity have been adopted for different areas of speech signal processing. In the following we provide a brief overview of the anatomy involved in speech production and relate it to two common speech production models providing different degrees of abstraction. The first model is a highly detailed description of the physical and aerodynamic properties leading to the production of speech. The second model is called the source-filter model. It is popular in many fields of speech processing due to its simplicity and low computational complexity. For a first comparison, the left part of Fig. 2.1 schematically illustrates various anatomical component involved in speech production, whereas the middle and right parts depict the two models just mentioned.

2.1 Speech Production Anatomy

Fig. 2.2 depicts a schematic midsagittal section of the head to illustrate the physiological components involved in speech production. Speech is generated by the coordinated use of various anatomical articulators known collectively as the *vocal organs*. These mainly comprise the lungs, the bronchi (both not depicted), the larynx, the *pharynx*, the *oral* and the *nasal*

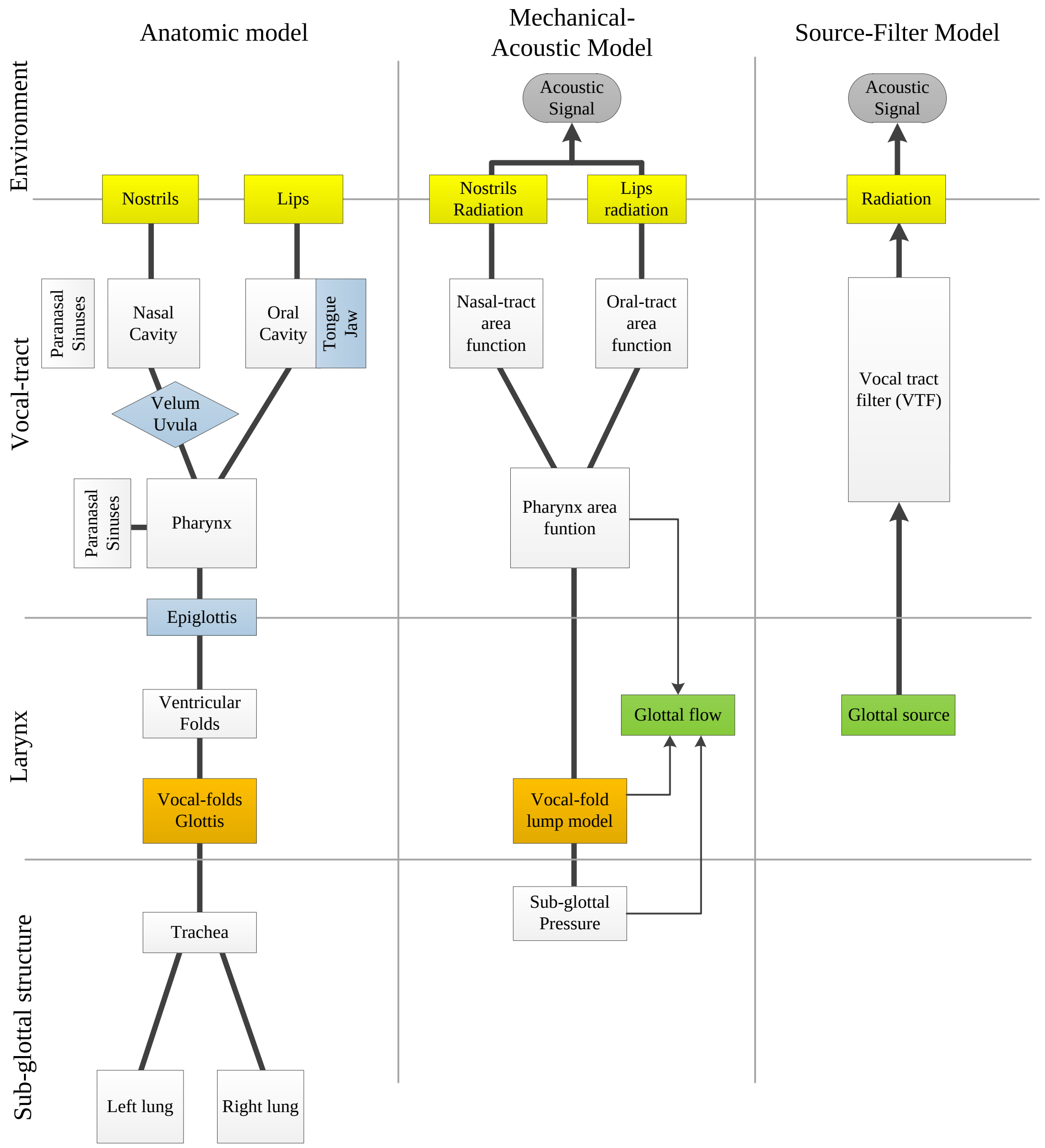


Figure 2.1: Comparison of different models of speech production.

cavities, where the latter three commonly are referred to as the *vocal tract* (VT).

In general, an air flow is induced by exhaling air from the lungs through the vocal organs to the lips or nostrils, from where it is radiated to the environment. During the passage of the air flow, one or more *constrictions* is applied, as a result of which a sound is generated. In particular, the opening between the vocal folds is called the *glottis* and sounds generated while air is passing through the oscillating vocal folds are referred to as *glottal source*. Besides, sound

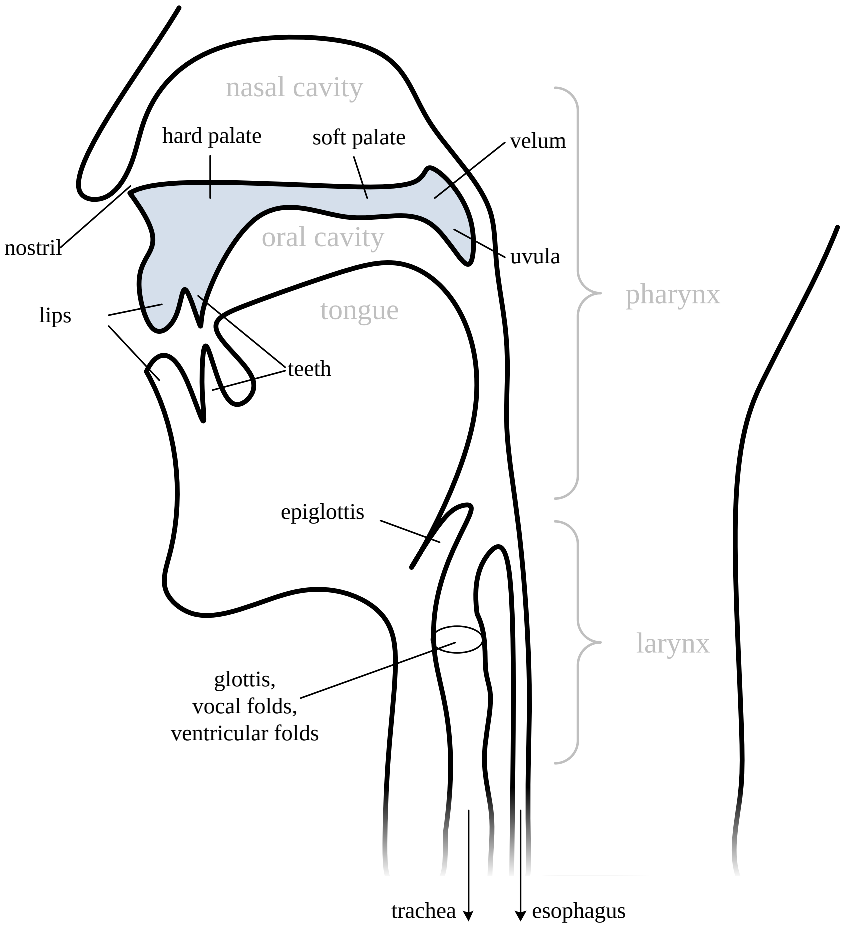


Figure 2.2: A schematic midsagittal section of the head highlighting anatomical parts relevant for speech production. In voiced speech, air passes from the lungs through the vocal folds acting as glottal source and eventually through the vocal tract, where articulation takes place.

can be generated by any other constriction other than the glottis along the VT, such as at the tongue, the lips or the teeth. Progressing further, the air passes through the oral cavity (*oral* sounds) or in a few cases through the nasal cavity (*nasal* sounds), depending on the position of the velum. For instance, in English language, the first sound of *Mother* is a nasal sounds. In other languages such as French, also *nasalized* sounds exist, where the air passes through the oral and nasal cavity at the same time (*e.g.* words ending on -*ont*). Generally, the modification of the constellation of the VT organs and the resulting resonances are what allows a speaker to articulate.

2.1.1 Voice Source

Anything in the respiratory airway that can vibrate or pose a resistance to the exhaled flow of air represents a potential sound source. That includes mainly the vocal folds, the tongue, the teeth, but also the velum or even the vestibular folds. However, in this work we are mainly concerned with the estimation and alteration of the *glottal* source, which originates from the periodic lateral and medial motion of the vocal folds in healthy speakers. In addition, we are concerned with the voice source in *pathological* voices — those produced by speakers with malfunctioning or even completely excised vocal folds as a result of laryngeal surgery — and how the proposed method is applicable to them.

The vocal folds are located inside a cartilaginous structure called *larynx*. As a primary function, the larynx provides a carefully guarded passageway between the pharynx (throat) and the trachea, leading to the lungs. Above the vocal folds sits the *epiglottis*. It is normally pointed upward during breathing or phonation, but while swallowing the epiglottis is drawn downward to a more horizontal position. Thereby it prevents food from descending into the trachea and instead directs it to the esophagus positioned posterior. The larynx is innervated and surrounded by muscles and its position may vary significantly during swallowing, breathing and phonation. For this need of mobility, the larynx is mainly in the form of cartilage. The only bone directly related to the larynx is the hyoid bone, and it happens to be the only bone in the human body not directly articulated to the remaining skeleton (Titze, 2000). The hyoid bone is anchored by muscles from the anterior, posterior and inferior directions. It provides attachment to the muscles of the floor of the mouth and the tongue above, the larynx below, and the epiglottis and pharynx behind. It acts as an anchor point for many movements of the larynx.

The muscles of the larynx may be divided into an extrinsic and intrinsic group of muscles. The former connect the larynx to its surroundings such as the hyoid bone or the sternum while the latter interconnect the cartilages of the larynx itself. The group of intrinsic muscles may be further divided into thyroarytenoid, cricothyroid, lateral cricoarytenoid, posterior cricoarytenoid and interarytenoid muscles, where each of these muscles is considered to be consisting of two parts. Jointly, these muscles allow for a very fine-grained control over the vocal folds' position, thickness, length and shape. For instance, viewed from the top, the vocal folds are able to open partially, where half of the glottis is closed and the other half is open. The cricothyroid muscles lengthen the vocal folds and thereby act as a primary means for pitch control. The vocal folds themselves represent a natural point of division between the subglottal and supraglottal airways due to their location and their ability to abduct (move apart) during respiration and to adduct (move together) during phonation.

The vocal folds are made of several layers of soft tissue (see also Fig 2.3). The outermost layer is a thin skin named *epithelium* and is between 0.05 and 0.1 mm thick. It wraps several layers of a softer, fluid-like tissue called *lamina propria*, similar to a balloon filled with water. The innermost and thickest layer is made of muscle fibers, belonging to the aforementioned thyroarytenoid muscle. Most vocal fold models usually combined these layers of skin and nonmuscular/muscular tissue into two or three groups, depending on the physiology that

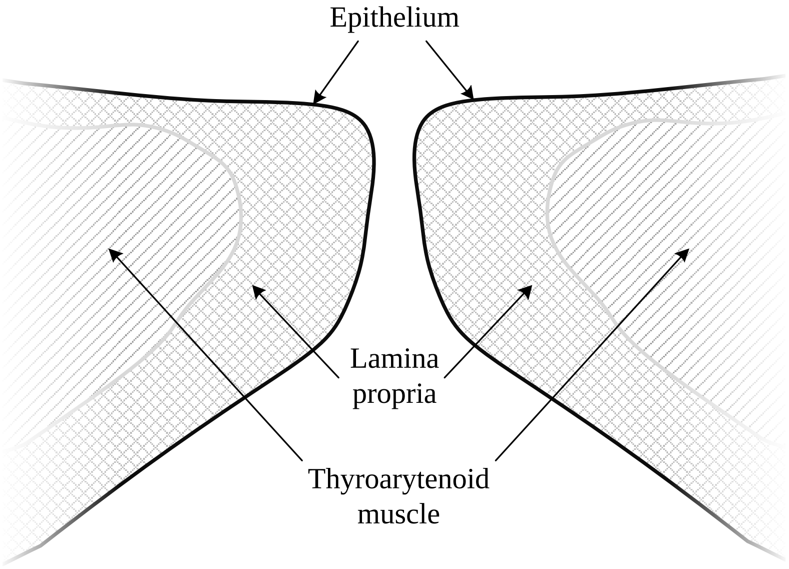


Figure 2.3: Coronal view of a schematic section of the vocal folds (redrawn from (Story and Titze, 1995)). The very thin, skin-like *epithelium* wraps the thicker, fluid-like *lamina propria* to form a structure similar to a ballon filled with water. The innermost layer is part of the *thyroarytenoid* muscle being partly responsible for the fine-grained vocal control.

is to be described. It has been argued in the past that these soft-tissue layers of the vocal folds have adapted for phonation in an evolutionary sense (Pressman, 1942; Hast, 1983), although phonation is not the primary biological function of the larynx. This can be observed in certain reinforcements of the ligament tissues in locations that are exposed to physical stress during vocal fold motion, which is the primary function of the vocal folds during phonation. Comparisons of the human vocal folds with other mammals have revealed that this layered structure is unique (Berke et al., 1987; Slavit and McCaffrey, 1991) and it has been argued that this combination of thin and highly flexible structures is of high importance to the ability of humans to produce and sustain vocal fold oscillation at a large range of frequencies (Titze, 2000).

It is this self-sustained vocal fold oscillation that intrigues some researchers since decades. Classical descriptions of vocal fold vibration usually attribute the periodic movement to a negative Bernoulli pressure in the glottis (van den Berg, 1958; Ladefoged, 1963; Lieberman, 1977). According to this *myoelastic-aerodynamic* theory, the vocal folds are sucked together if the glottis is sufficiently narrow, the airflow is sufficiently high and the medial surface of the vocal folds is soft enough to yield. After collapsing, the glottis is closed and subglottal pressure restarts to build up until the vocal folds yield and start moving lateral (outward). Lateral movement continues until elastic forces in the tissue retard the motion and reverse it. The

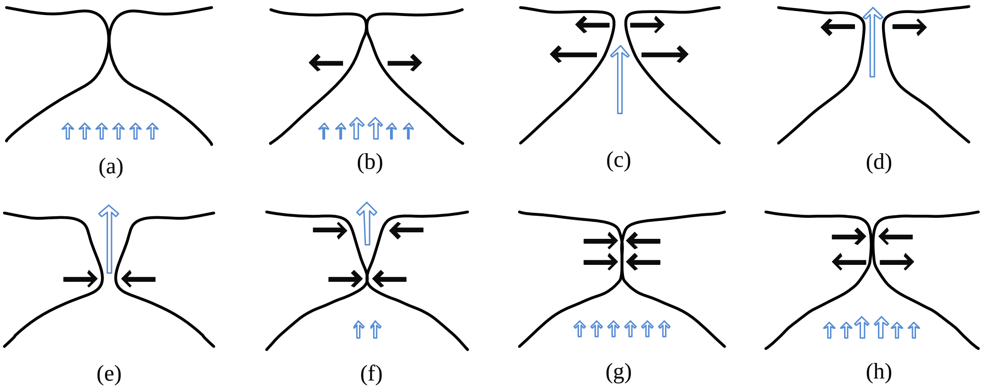


Figure 2.4: The asymmetric opening and closing of the vocal folds is referred to as mucosal wave. It is this mucosal wave in interaction with the aerodynamic forces produced within the glottis that are viewed to being the main driving forces behind vocal fold oscillation (redrawn from (Titze, 2000) and augmented).

tissue will restart moving medially and the Bernoulli effect reinforces the closing of the glottis again. However, recently developed techniques such as laryngeal video stroboscopy (Hirano, 1981) or high-speed videoendoscopy (Deliyski, 2007) combined with subsequent theoretical studies attributed only a secondary role to the Bernoulli effect. This effect alone can *not* be the driving force behind sustained vocal fold oscillation. Instead, an asymmetry between the aerodynamic driving forces and the actual opening and closing of the glottis appears to be more important for a self-sustained oscillation (Titze, 1976; Ishizaka and Matsudaira, 1972; Stevens, 1977; Titze, 1988). Two key factors contribute to this asymmetry, the *mucosal wave* and the *inertial acoustic loading* presented by the vocal tract.

The mucosal wave is best explained with the idealized representation of the vocal folds in the coronal plane (front view) in Fig. 2.4. One cycle of the vocal fold motion can be considered to consist of an *open phase* and a *closed phase.* In the first frame (Fig. 2.4a) the airway is closed while the vocal folds are initially in contact. The next frame (Fig. 2.4b) indicates the beginning of a lateral movement which separates the left and right vocal folds. This lateral movement is led by the inferior portion of cover surface, the superior portions follows as shown in Fig. 2.4c. Upon reaching the maximum lateral displacement (Fig. 2.4d), it is again the lower portion of the vocal folds that begins to move medially with the upper portion following (Fig. 2.4e). Eventually, the lower portions on the left and right sides make contact and close the airway (Fig. 2.4f). Medial displacement stops as the upper portions collide (Fig. 2.4g). The mucosal wave is effectively leading the inner-glottal pressure wave and thereby actively supporting its opening and closing.

The second phenomenon aggravating self-sustained oscillation is the inertial acoustic loading due to the vocal tract. This term generally refers to the pressure difference arriving at the glottis due to back propagating airwaves from the vocal tract. The delayed response of the air column in the vocal tract actively supports the opening and closing of the glottis by adding to the aforementioned asymmetry.

Following the discussion above it becomes clear that the human voice is an incredibly versatile organ that allows the speaker to produce a large variety of different sounds. From a phonetic point of view voices can be characterized under different aspects. The rate of the vocal fold oscillation is referred to as the *fundamental frequency*, $f_0 = 1/T_0$, where T_0 is the *fundamental period* between subsequent vocal fold oscillation cycles. Often the term *pitch* is used interchangably instead of f_0, although pitch refers to the rate of vibration perceived by a listener. In many cases these are the same however. Typical values for f_0 found in male speakers range from 80 Hz to 250 Hz, whereas female speakers mostly exhibit values between 120 Hz and 400 Hz.

Besides a peak at the fundamental frequency, energy is also present at multiples of f_0. These signal components are called the *harmonics*. Harmonics are very important for pitch perception. Nevertheless, the currently no general theory exists that explains all phenomena of perception. Instead, the prevalent theories of acoustic perception, place theory and temporal theory, may each explain only some observations regarding pitch perception (Moore, 2012). For instance, the perception of the *missing fundamental* cannot be explained by the place theory. Besides pitch perception, the harmonics also contribute to the perception of the *timbre* of a voice. In general, the higher harmonics of f_0 tend to have less energy and their ratio contributes to the unique sound of a speaker's voice and also to the *voice quality*. An analogy from music might help to illustrate the effect of the timbre. Two different musical instruments creating sounds of exactly the same fundamental frequency are still perceived to sound different. It is the difference in energy between the harmonics that eventually leads to very different perceptions.

Another important aspect of the voice is the occurrence of *aspiration noise*. Aspiration noise is produced when the glottis remains fully or partially open or does not close completely during a vocal fold oscillation cycle. The result is a non-periodic turbulent air flow. Unvoiced sounds like the *H* in the English word *house* are created using only aspiration noise and also whispering is produced with abducted vocal folds.

In this work we are mainly interested in voiced speech since this kind of speech sounds are most affected in alaryngeal speech. Furthermore, the scope of this work may be enlarged easiliy to address also problems in laryngeal (non-pathological) speech processing because voiced sounds comprise the largest category of sounds in human speech. We conclude the introduction of the voice source at this point with a brief description of different *phonation types* that are commonly differentiated (Roubeau et al., 2009; Childers, 2000; Degottex, 2010):

Modal or Chest voice is the normal mode of phonation. Both, body and cover of the vocal folds are vibrating and the duration after closing of the vocal folds until the beginning of the next opening is comparable to the open phase of the vocal folds. A common average value of f_0 for male speakers is 120 Hz and 180 Hz for female speakers.

Falsetto, Soft or Head voice The laryngeal muscles stretch the vocal folds, limiting the degree of freedom to move and vibrate. Consequently, in this phonation type it is mainly the lamina propria that vibrates while the muscles are assumed to be fixed. The fundamental frequency can be easily twice as in modal voice. This mechanism is mainly used by children and often used by female speakers.

Whistle or whispery voice The glottis is open, and there is no periodic vibration of the vocal folds. The harmonic richness is very low, and the voice is rather soft. This type of phonation is often observed in alaryngeal speakers before an alternative phonation such as tracheo-esophageal speech is mastered (Max et al., 1996).

Vocal fry or creaky voice There is much structural aperiodicity in the source (jitter, shimmer), often jumps of f_0 are observed. This phonation is often perveived at the end of sentences, when the lung pressure slackens. The voice can be soft or loud, pressed or not pressed.

Breathy voice The glottal closure is incomplete, there is a much additive aperiodicity in the source. In terms of spectral parameters, the amount of aspiration noise is relatively high, and the voice can be soft or loud, pressed or not pressed.

Pressed voice The vocal effort is high, but the signal is not necessarily efficient, and thus energy can be rather low. The main correlate of pressed voice is a relatively short open phase, resulting in a high glottal formant frequency.

Soft voice the vocal fold are vibrating, the vocal effort is weak. In terms of spectral parameters, the ratio of periodic and aperiodic content is low, the glottal formant is low and f_0 is generally lower.

Loud voice both the vocal effort and the signal energy are high. In terms of spectral parameters, the ratio of periodic and aperiodic content is high, the glottal formant is low and f_0 is generally high

2.1.2 Voice Source in Pathological Speakers

Injury to our vocal folds can stem from a variety of causes occurring during every-day-life, *e.g.* talking too much, screaming, constantly clearing your throat or smoking can make you hoarse. Other problems can occur such as the development of nodules, polyps and sores on the vocal folds, complete or partial vocal fold paralysis, paradoxical vocal fold movement or spasmodic dysphonia. More serious causes of voice disorders include infections or growths due to a virus or cancer. For a successful treatment it is vital to diagnose these voice problems as early as possible. Unfortunately, in cases of advanced disease patterns, a partial or total excision of the vocal folds is the only residing treatment.

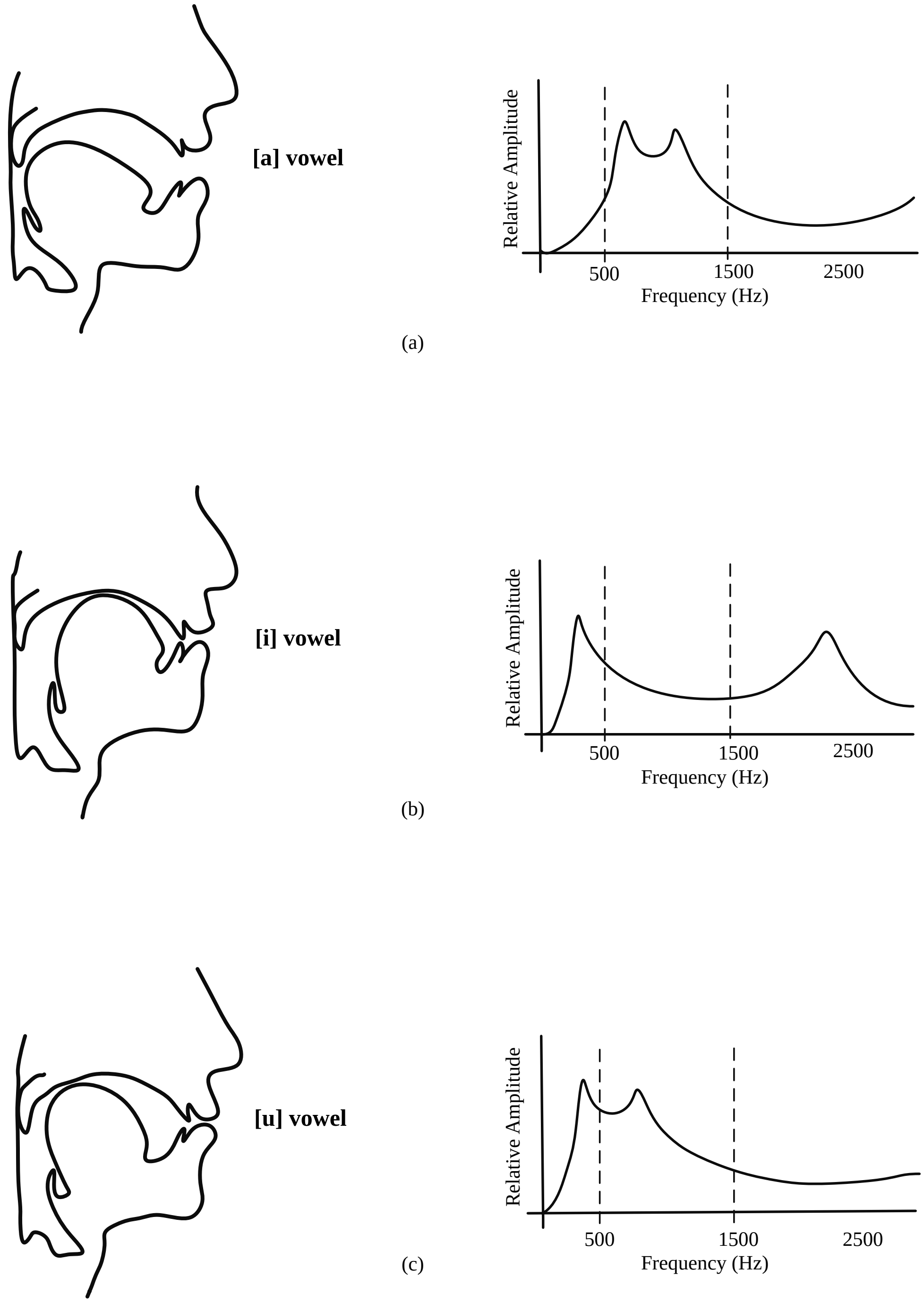
The degree of degradation in disordered voices depends on the acute problem and naturally engenders a decrease in a patient's speech intelligibility and thereby a severe limitation in his social life and oral interaction (Weinberg, 1986). For example, subjects who have undergone laryngectomy suffer from degradation of their natural vocal excitation (Williams and

Barber Watson, 1987; Most et al., 2000; Moerman et al., 2004). Laryngectomy is the common treatment after diagnosis of larynx cancer in an advanced stage and constitutes the partial or total removal of the larynx. During the surgery, commonly a neoglottis is created to permit phonation after laryngectomy. Therefore, the pharyngeal mucosa is sutured over the superior end of the transected trachea, thereby making a permanent stoma in the mucosa. Effectively, the trachea secludes directly into a hole at the location where the larynx used to reside. Consequently, the aspiratory airway ends below the vocal tract. By using a one-way valve allowing the exhalation of air through the vocal tract and with a considerable amount of rehabilitation effort, the subject may acquire the ability to use the neoglottis for phonation (tracheo-esophageal speech). Nevertheless, the patient's ability to produce voiced sounds due to the reduced or missing vocal fold functionality is significantly reduced (van As, 2001; Pindzola and Cain, 1988). The resulting voice sounds often unpleasant and unnatural and exhibits a fluctuating and often intermitted periodicity (Kasuya et al., 1986). In addition, the speaker loses most of its control over pitch variability. Furthermore, the position and shape of the neoglottis vary significantly (Qi et al., 1995), altering the formant locations. Often incomplete glottal closure can be observed. Furthermore, the flexibility and controllability of the neoglottis lacks greatly when compared with a healthy glottis, especially due to the absence of the laryngeal muscles. The high mass of the neoglottis and low resistance to mucus aggregation influence the absolute value and stability of the fundamental frequency in a disadvantageous manner. Eventually, the resulting voice source has an unnaturally low and instable pitch and often is found to have a hoarse, croaky and breathy voice quality (Verma and Kumar, 2005). The methods for separation of voice source and articulation proposed in this work aim at providing a fundamental building block for signal processing systems targeting the restoration of such pathological voices.

2.1.3 Vocal Tract

The pharynx, the oral cavity and the nasal cavity are collectively named vocal tract. To pronounce phonemes, the speaker may articulate the sounds produced by the voice source by varying the constellation of the surrounding vocal organs. Thereby the sound coming from the source is further enriched by the spectral shaping due to the acoustical characteristics of the vocal tract. A wider variety of sounds is created by modifying the basic sound source. Recall that all voiced sounds from the glottis comprise a fundamental frequency and its harmonics. The vocal tract functions by weighing these harmonics, which has the effect of changing the timbre of the sound. In effect, the vocal tract filter does not alter the harmonic structure of the signal, but it does alter the relative strengths of the harmonics.

Whereas the pharynx and nasal tract are relatively static, it is mainly the vocal cavity that assumes the role of an articulation filter by further enriching and modifying the source sounds. This filter is called *vocal tract filter* (VTF). The vocal organs surrounding the oral cavity allow the speaker to considerably vary the shape and size of the vocal tract and thereby alter the resulting speech. The pharynx and nasal cavity are relatively fixed, but the tongue, lips and jaw can all be used to change the shape of the oral cavity and hence modify the sound. If one



(a)

(b)

(c)

Figure 2.5: Comparison of different vocal tract vowel configurations (Titze, 2000).

looks for instance at the tongue in isolation, it becomes clear that it may freely move in all three dimensions and thereby may create a complex variety of movements and trajectories, leading to aforementioned complexity of the voice production process.

Acoustically, during the articulation of vowels, the vocal tract represents a quarter-wave resonance tube of approximately 15 – 17 cm in length. It is open at one end (the lips) and quasi-closed at the other (the glottis). Although technically the glottis is not closed, most models assume it to be so at the cost of a minor modelling error. The diameter of this tube varies across its length, depending on the position of the vocal organs, in particular the tongue, but also the jaws and the lips. Effectively, over time the speaker varies the diameter by modulating the position of the respective vocal organs. The pressure waves emanating from the glottal source propagate through the vocal tract and are reflected at the transition into open space at the lips with a negative reflection coefficient. Resonances occur in the vocal tract as a result of the interference of acoustic waves travelling in opposite directions. Depending on the constellation of the vocal organs, those resonances occur at different frequencies. Thereby the speaker effectively modulates the spectral characteristics of the VTF to pronounce different vowels. Fig. 2.5 conceptually illustrates examples of different vocal tract configurations and the resulting vowel spectra.

2.2 Mechanical-Acoustic Model

2.2.1 Mechanical-Acoustic Source Model

As detailed in (Titze, 2002), the understanding and accurate modelling of the physical phenomena leading to self-sustaining vocal fold oscillation has been a topic of research for several decades. As already introduced in 2.1.1, the first formulation of a model was presented in (van den Berg, 1958). Based on empirical observations, vocal fold oscillations were explained based on the elasticity of the vocal fold tissue and the Bernoulli effect. Shortly after, theoretical studies and observations using high speed image capturing technology revealed that this myoelastic-aerodynamic theory of speech production was not able to fully explain the self-sustaining vocal fold oscillations. Current views support the assumptions that there are mainly two effects enabling phonation; the inertial loading of inner-glottal pressure due to VT on one hand side and the mucosal wave, leading the pressure wave propagating through the larynx and other side.

Subsequently, the theory of the inertial loading due to the VT was first demonstrated using a one-mass model of the vocal fold motion (Flanagan and Landgraf, 1968). In this model, a mass was attached to a rigid lateral boundary using springs and masses to account for tissue elasticity and energy losses. The mass was allowed only lateral movement so as to simulate opening and closing of the vocal folds. Due to its simplicity, this model could not account for the mucosal wave. Therefore it was not long until a two-mass model was proposed (Ishizaka and Matsudaira, 1972; Ishizaka and Flanagan, 1972). In this model, two vertically placed masses represent the lateral movement. An additional spring connecting the two masses allows the modelling of shear forces present between the upper and lower vocal fold tissue

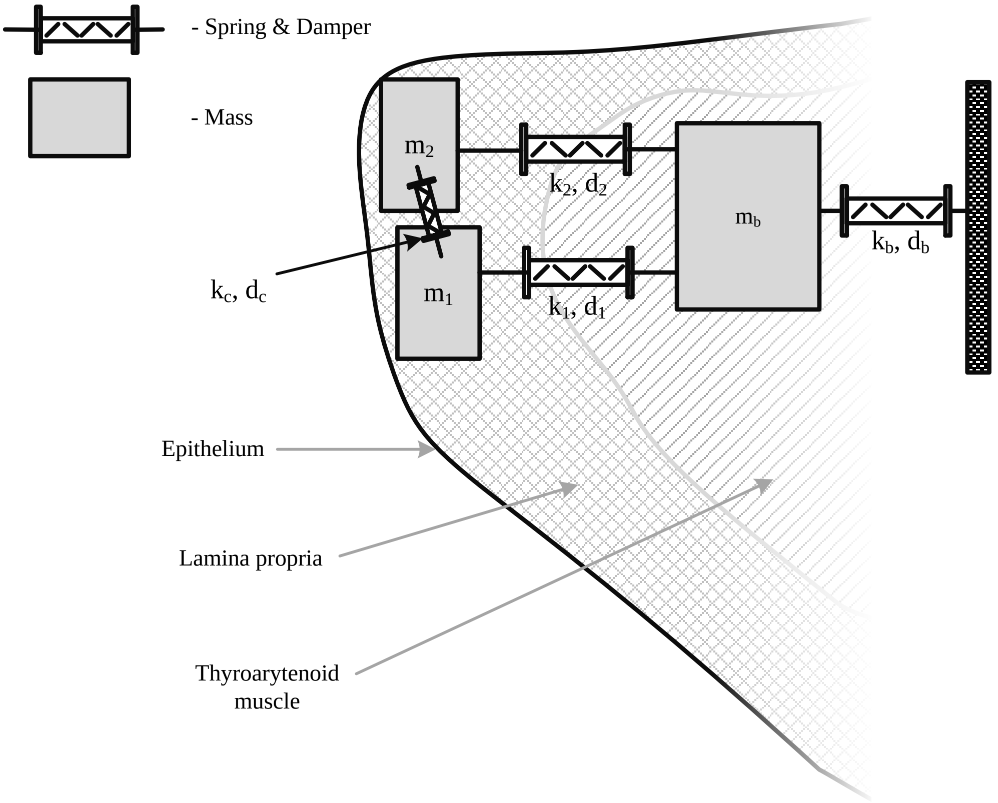


Figure 2.6: Coronal view of a schematic section of a right vocal fold (Story and Titze, 1995) overlaid with the three-mass model (see Section 2.2).

regions. This model gained popularity due to its simplicity and reasonable agreement with physiological observations. A limitation of the two-mass model is that the layered structure of the vocal folds is not captured. For that reason, the two-mass model is mostly a cover model instead of a cover-body model of the vocal folds.

This limitation eventually lead to the development of a three-mass model, where an additional 'body' mass was positioned laterally to the other two masses. This additional mass represents the inner layers of the vocal folds (the thyroarytenoid muscle) as displayed in Fig. 2.6. This additional mass again is coupled to the other two masses and also to a rigid lateral boundary using spring damper elements. This model was found to provide better correspondence between model components and anatomical structure of the vocal folds. For example, a contraction of the thyroarytenoid muscle leads to an increase in the stiffness of the body, which can be represented by adjusting the respective parameters of the body mass (m_b in Fig. 2.6). In the two-mass model this analogy is more difficult to find.

Subsequently, more complex models have been presented in (Titze, 1974a,b) to account for observed vibrations also in the transverse plane. Their ability to capture more details is achieved by lumping smaller anatomical regions into additional, finely distributed masses. Another step up in complexity came with the introduction of the continuum mechanics model (Titze, 1979) and a finite-element model (Alipour et al., 2000). However, it was observed that the additional complexity and increased number of degrees of freedom only marginally increased the accuracy of the correspondence between the models and observed modes of the vocal fold vibrations (Berry and Titze, 1996). Therefore it was concluded in (Titze, 2002) that the lumped-element models seem to capture sufficient details of the vocal fold motion to serve a useful research tool if fine details is unnecessary.

2.2.2 Acoustic Vocal Tract Model

In terms of acoustics, it is the task of the vocal tract to emphasize and attenuate certain frequency ranges of the voice source spectrum. This is mainly achieved by continuously transforming the shape of the vocal cavity. Thereby the acoustic properties of the vocal tract are modified and resonances emerge at different frequencies.

The resonant nature of systems is often described by considering the motion of a mass connected to a rigid boundary through a spring and a damper. Such a mechanical composition inherently forms a resonance frequency that informally can be described as $f_R = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$. The compliance k of the spring is proportional to the resonance frequency, whereas the mass m is inverse proportional. If excited by a periodic forcing function, the system will oscillate with frequency determined by the periodicity of the driving force. In case the frequency of the driving force is near f_R the system will enhance the oscillation (without modifying the frequency). On the other hand, driving forces with a frequency that is distant from f_R will be attenuated. An additional damping element ensures that the oscillation will not grow to infinity by inducing a system inherent resistance. These three factors describing the system, inertia (mass), resistance and capacitance (the reciprocal of spring compliance), are collectively known as *impedance*.

The acoustic properties of the vocal tract can be modelled using such a lumped parameter system. Yet, the vocal tract resembles a distributed system, in which the properties of inertia, capacitance and resistance are evenly distributed through the medium of air. Numerically the behaviour of such a system can be described as a large number of connected, discrete systems (Rabiner and Schafer, 1978). As the size of the smaller systems tends to zero, the model resembles the behaviour of a continuous system. Finite-element methods and boundary element methods have been proposed in the past to model the vocal tract using such methods (Ling, 1976; Lu, 1993).

In a similar manner the vocal tract can be modelled by a tube in which air represents a medium having a mass (intertia or acoustic inductance) and a capacitance due to its compressibility. Furthermore, friction and soft side walls of the vocal tract induce an acoustic resistance to the resonance system. Several factors, such as the speed of sound and temperature have a direct influence on the system's characteristics, but these are varying very slowly in time. On the other hand, the cross sectional area of the tube is directly proportional to the amount of air present in a specific tube segment. Therefore, the area is proportional to the inertia of the particular tube segment and thereby directly influences the overall resonance characteristic in this segment.

Several models that describe the vocal tract as the concatenation of one or more tube segments have been proposed in the past. In these models, the propagation of the acoustic pressure waves is described using a set of partial differential equations (Portnoff, 1973; Maeda, 1982). The vocal tract is split into several segments each with its own cross sectional area as illustrated in Fig. 2.7. In principle, each section of the vocal tract contributes a different inertia to the overall vocal tract resonance system and thereby the different formants (or resonance frequencies) of the vocal tract are shaped.

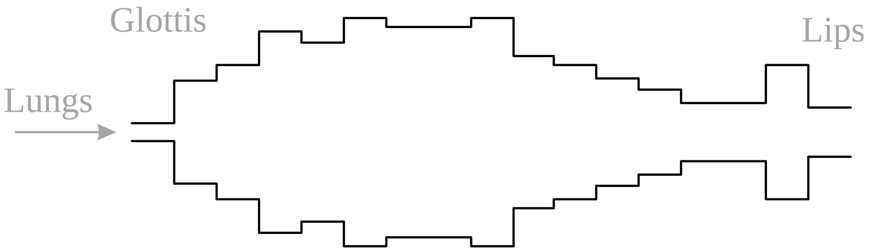


Figure 2.7: Vocal tract model composed of a sequence of joined tubes of differing cross sectional area.

2.3 Source-Filter Model

The model of speech production presented in Section 2.2 is an interesting tool for the analytical or numerical simulation and analysis of the process of speech production. It has enabled researchers to get an advanced understanding of various aspects such as the mechanical behaviour of the vocal folds or different modes of coupling between the vocal tract impedance and the innerglottal pressure. On the other hand, in many applications of speech processing it is desirable to build models that are tractable, yet may be parameterized in a way so as to directly manipulate physiologically and acoustically relevant features. In such a model, a few parameters correlate directly with application-relevant aspects such as voice quality, prosody, VT configuration, etc. For such a scenario, the mechanical and acoustical models are not well suited due to their highly complex nature and the large degree of freedom in their parameters (Ljungvist, 1986). A simpler model is preferable and often also found to be of sufficient accuracy for the application's requirements.

The source-filter model, first described by Fant (Fant, 1960) is currently the most widely used model of this kind. Its general outline is illustrated in the right part of Fig. 2.1. The separation of speech into source and filter adequately represents the mechanisms involved in speech production. Furthermore, it is known that listeners separate their perception of the source in terms of its fundamental frequency from the modified pattern of its harmonics. This separation plays an important role in the perception of many acoustical cues. For instance, the fundamental frequency is thought to be the main acoustic dimension for the perception of prosody, whereas the main dimensions of verbal distinction are based on a combination of the voice quality, voice timbre and the modification by the vocal tract. Therefore, in many applications the source-filter model also represents a reasonable model of perception.

With respect to the acoustic model, there are three simplifying assumptions underlying the source-filter model:

I No non-linear feedback exists between the vocal tract and the glottal flow observed in the source. The resulting speech is always the output of a linear system consisting of the source and the VT filter.

II The VT filter is time-invariant during a period of analysis. Although the articulatory organs vary in their position over time, there is always a time-span that is sufficiently short in order to make this assumption valid.

III In the time-domain, the source and filter components are convoluted, which corresponds to a multiplication in the spectral domain. Speech may therefore be represented by

$$S(j\omega) = G(j\omega) \cdot H(j\omega) \cdot A(j\omega) \cdot L(j\omega), \tag{2.1}$$

where the spectral contribution of the glottal source is represented by $G(j\omega)$ and $H(j\omega)$ stands for the harmonic structure due to the periodicity of the glottal opening and closing. The filter $H(j\omega)$ also accounts for a linear phase component due to the temporal position of a particular glottal cycles relative to some time reference. The filter $A(j\omega)$ models the resonances and anti-resonances of the VT, also known as *formants* and *anti-formants*. Eventually, the radiation occurring at the lips and the nostrils is merged into the filter $L(j\omega)$.

The VTF can be assumed to be time-invariant only during a short period of time. Therefore, in a typical application scenario voiced speech is analysed on a frame-by-frame basis by segmenting the signal into overlapping frames. During each frame, the source may be represented by the transfer function of periodic glottal volume velocity waveforms[1], multiplied with the respective VT transfer function and the lip radiation filter.

For computational and analytical convenience the periodic glottal cycles are often simplified to a train of Dirac pulses. The accuracy of this simplification was sufficient in a surprising number of applications. Nevertheless, as we will describe in more detail in Chapter 3, there

---

[1] Shorter, common names for *glottal volume velocity waveform* are *glottal excitation*, *glottal source* or *glottal cycle*.

are also various limitations to this representation. An alternative to Dirac pulses is to describe the glottal excitation waveform by an analytical voice source model. In the following two sections we will provide an overview of existing glottal source models and an auto-regressive VTF representation. The glottal source models are an important component of currently used source-filter separation methods. Following this overview we will provide the motivation for our choice of a source model for this work.

2.3.1 Glottal Source Models

A glottal model in the framework of a source-filter model is potentially very useful for a parametric manipulation or encoding of the voice source. The glottal excitation can be considered as a mixture of deterministic and non-deterministic components. The latter comprises all source components that are not modeled by the deterministic parts, such as aspiration noise, formant ripples and other phenomena due to the non-linear coupling between vocal tract pressure and glottal volume velocity. Hereafter, we refer to the non-deterministic components as *glottal noise*.

The deterministic part originates from the air flow modulated by the periodic lateral and medial motion of the vocal folds that opens and closes the glottis (see Section 2.1.1). The trans-glottal pressure drives flow through the glottis resulting in a volume velocity waveform. Most of the glottal models describe one period of the glottal waveform $g(t)$ or its time-derivative $\dot{g}(t) = \partial g(t)/\partial t$ in the time-domain (Veldhuis, 1998). They are commonly identified by all or a subset of the following charactertic instances in time, as illustrated in Fig. 2.8:

t_o - The beginning of the glottal cycle

t_i - The maximum of the time-derivative, $\dot{g}(t)$

t_p - The maximum of the glottal flow waveform, $g(t)$

t_e - The minimum of the time-derivative, $\dot{g}(t)$, also the instant of maximum excitation

t_a - The effective return phase duration, which is proportional to the exponential decay of the closing phase.

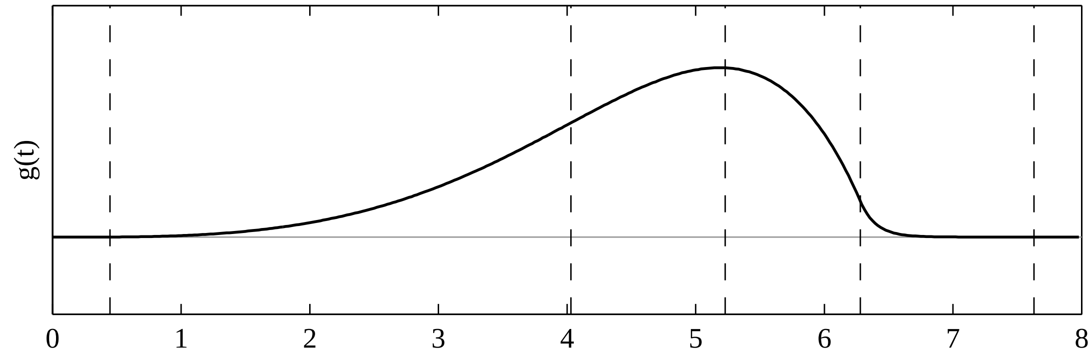
t_c - The instant of closure of the vocal folds, which effectively stops the glottal flow

T_0 - The fundamental period, $T_0 = 1/f_0$

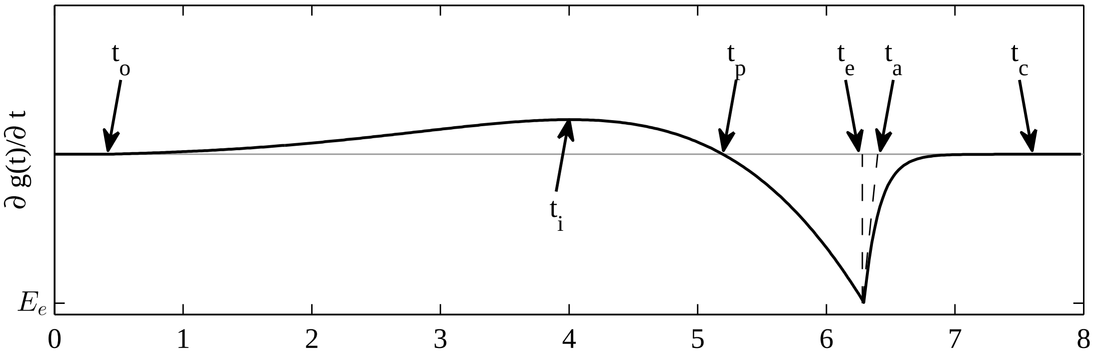
O_q - The open quotient, $O_q = t_e/T_0$

AV - Amplitude of voicing

If the instantaneous fundamental frequency of the glottal cycle k is given by f_0 and the sampling rate is f_s, then the number of samples representing cycle k is defined by $N_0 = \lceil f_s/f_0 \rceil$. Provided that no energy is present in the spectrum of the glottal source signal above the



(a) Time (ms)



(b) Time (ms)

Figure 2.8: Example of a glottal volume velocity waveform (top) and its derivative (bottom) generated by the LF model.

Nyquist frequency $f_N = f_s/2$, the timing parameters can be defined in the discrete-time domain as $N_o = \lceil t_o \cdot f_s/f_0 \rceil$, $N_i = \lceil t_i \cdot f_s/f_0 \rceil$, $N_p = \lceil t_p \cdot f_s/f_0 \rceil$, $N_e = \lceil t_e \cdot f_s/f_0 \rceil$, $N_a = \lceil t_a \cdot f_s/f_0 \rceil$ and $N_c = \lceil t_c \cdot f_s/f_0 \rceil$.

Each model then defines one ore more analytical curves through these points. Other models also exist that define the glottal contribution in the frequency domain. A glottal cycle is divided into an *open phase* (t_o to t_c) and a *closed phase* (t_c to t_o of the next glottal period). The open phase is further defined by an *opening phase* (t_o to t_p) and a *closing phase* (t_p to t_a). Due to the vibratory dynamics of the vocal folds and the inertive properties of the lower vocal tract, the glottal volume velocity waveform typically exhibits an asymmetry in time such that the closing phase is shorter than the opening phase (Rothenberg, 1973; Titze, 1988).

A large variety of glottal source models has been proposed in the past (Cummings and Clements, 1995; Doval and d'Alessandro, 2006). In the following we describe the most commonly used and most known glottal models of different families to provide the reader a non-exhaustive overview and to motivate our choice for one of these models.

Rosenberg: Initially, six different models were proposed in (Rosenberg, 1971). A preference listening test was then conducted to select one of them based on the criterion of which one most closely resembles the sound of a voice source. This particular model then became known as the *Rosenberg* model. It is defined as

$$g(n) = \begin{cases} an^2 - bn^3, & 0 < n \leq N_e = N_c \\ 0, & N_c < n < N_0, \end{cases}$$

with $a = \frac{27}{4}\frac{AV}{O_q^2 T_0}$ and $b = \frac{27}{4}\frac{AV}{O_q^3 T_0^2}$. Note the absence of a closing phase. The instance of maximum flow is proportional to N_e by a constant factor: $N_p = \frac{2}{3}N_e$.

Klatt & Klatt: (Klatt and Klatt, 1990): This model, also known as KLGLOTT88, constructs a glottal flow signal in two steps. In a first step, the same analytical expression as in the Rosenberg model is used to construct the opening phase. Subsequently, a low pass filter is used to dampen high frequencies contained in the signal due to the discontinuity at t_e. An additional parameter TL is introduced that influences the spectral tilt by determining the amplitude in dB at a frequency of 3 kHz. The model has found wide use, for instance in studies such as (Hanson, 1997) or in the speech synthesizer introduced in (Klatt and Klatt, 1990).

Fujisaki: This model uses four polynomials and six shape parameters (Fujisaki and Ljungqvist, 1986) to define $g(n)$. Discontinuities are allowed not only at the moment of glottal closure, but also at t_o and at t_c. It is also worth noting that this model, in contrast to most other models, allows a negative flow. This allows to model the effect of air pressure waves on the closed glottis, which are travelling back and forth in the vocal tract.

Fant: This model was proposed in (Fant, 1979). It is made of two sinusoidal parts and uses two shape parameters, t_p and K:

$$g(n) = \begin{cases} \frac{1}{2}(1 - \cos(\omega_g t)), & 0 < n < N_p \\ K \cdot \cos(\omega_g (t - t_p)) - K + 1, & N_p < n < N_c, \\ 0, & N_c < n < N_0. \end{cases}$$

with $\omega_g = \pi / t_p$. K acts as a symmetry control parameter in that if $K = 0.5$ then the pulse is symmetric and $K \geq 1$ yields a closing phase of duration 0.

LF: The Liljencrants-Fant (LF) model is by large the most widely used glottal source model (Alku, 2011). Instead of modeling the glottal flow waveform like the models presented so far, the LF model describes the time-derivative of the glottal source, $\dot{g}(n)$ (see Fig. 2.8(b)). The opening phase and the first part of the closing phase are described by the product of a growing exponential and a low frequency sinusoid. The remaining part of the closing phase of the glottis from t_e until t_c is modeled by a decaying exponential.

For syntehsis, commonly the direct synthesis parameter set $\{E_0, \alpha, \omega_g, \epsilon\}$ is used in the synthesis equations:

$$\dot{g}(n) = \begin{cases} E_0 e^{\alpha n} \sin(\omega_g n), & 0 \leq n \leq N_e \\ -\frac{E_e}{\epsilon N_a}\left[e^{-\epsilon(n-N_e)} - e^{-\epsilon(N_e-N_c)}\right], & N_e \leq n \leq N_c \\ 0, & N_c \leq n \leq N_0 - 1. \end{cases} \tag{2.2}$$

The transformation between the two parameter sets can be carried out using the following relations and constraints:

$$\begin{aligned} \textstyle\sum_0^{N_0} \dot{g}(n) &= 0 \\ \omega_g &= \frac{\pi}{N_p} \\ \epsilon N_a &= 1 - e^{-\epsilon(N_c-N_e)} \\ E_0 &= -\frac{E_e}{e^{\alpha N_e}\sin(\omega_g N_e)}. \end{aligned} \tag{2.3}$$

The condition defined on the first line of Eq. (2.3) ensures that the glottal flow waveform returns to zero after each glottal cycle and is typically enforced by iteratively optimizing the damping parameter α of the exponential segment in Eq. (2.3) (Gobl, 2003; Childers, 2000). Including the shape-defining temporal parameters N_o, N_e and N_a, the Equations (2.2) and (2.3) have a total of seven free parameters; E_0 (a scaling factor warranting the area balance between the opening and closing phases), α (growth rate of the exponential used in the opening phase), ω_g (low frequency sinusoid used in the opening phase) and E_e (amplitude of the negative peak of $\dot{g}(n)$). Yet, the LF model is referred to as a four-parameter model, which will be further detailed below. In general, the number of parameters is reduced by iteration and by requiring the integral of the pulse over the glottal cycle to be zero (Gobl, 2003). The LF model has received large attention and has been studied in depth (van Dinther, 2003; Henrich, 2001; Doval and d'Alessandro, 2006; Fant and Lin, 1988). An analytical definition of the LF spectrum has been developed in (Doval and d'Alessandro, 1997).

Transformed LF: This model was proposed in (Fant, 1995) in an attempt to reduce the complexity of the LF model. The three parameters defining the shape of the LF model (t_p, t_e and t_a) are reduced to a single curve parameter, R_d. By varying R_d, various glottal source waveform shapes can be obtained ranging from extreme tight addicted phonation to very breathy, abducted phonation. Optimal values for R_d across a range of voice qualities were then found by measurements on various speakers (Fant, 1995).

CALM: The causal-anticausal linear model (CALM) is the only model listed here that is entirely described in the spectral domain, proposed by Henrich, Doval and d'Alessandro (Henrich et al., 1999; Doval et al., 1997, 2003). It is based on the observation that the open phase is a truncated impulse response of a filter having one anti-causal stable pole. The return phase is likened to a dampened exponential. Therefore, defining the time origin at t_e, the open phase can be approximated using an anti-causal conjugate pole pair and the return phase can be approximated using a causal real pole. This model can thus be defined as a pair of an anti-causal and a causal filter.

The list presented here of course is far from exhaustive. Over the years, many other models have been proposed (Krishnamurthy, 1992; Ananthapadmanabha, 1984; Hedelin, 1984; Milenkovic, 1986; Shue and Alwan, 2010). A main divider between the models naturally is their complexity, *i.e.* the number of free model parameters which determine their coverage of the space of real voice source waveforms.

In this work, our main focus is not the evaluation of existing or new voice source models. Instead, we are mainly concerned with methods for estimation of model parameters in order to obtain an optimal set of parameters subject to some previously chosen criterion. From the source models known to us it appeared that the **LF model** was the most suitable one for the following reasons:

- The LF model is widely used and has been well studied in the past (Raitio et al., 2008; Airas, 2008; van Dinther, 2003; Henrich, 2001; Doval and d'Alessandro, 2006; Fant and Lin, 1988).
- Compared to other source models, the LF model is able to cover a relatively wide range of voice types and voice qualitities due to its high dimensionality. In previously published source-filter separation methods (see Section 3.3.7), often trade-offs in favour of simpler source models are made in order to simplify the estimation of the optimal model parameters (see Section 3). In this study we present methods that efficiently and accurately allow to estimate more complex models while being robust to premature convergence to local minima of the error surface. This allows us to use source models of a higher complexity with a potentially higher accuracy in reference to real speech signals.
- It was shown previously that the parameters of the LF model do not form an orthogonal basis and therefore are not mutually independent (Vincent, 2007). Different combinations of parameters may describe very similar or identical glottal source waveforms. Commonly used, gradient-based optimization methods are not able to nicely cope with this kind of parameter dependency due to their tendency to get stuck in local optimal values. The methods proposed in this work are able to overcome this limitation and therefore make the LF model a suitable candidate for being used as a source model in the proposed source-filter separation framework.

Spectral Correlates of the LF model parameters

An analytic and exhaustive study of the influence of the parameters of various glottal source models and their respective frequency-domain representations was presented in (Doval and d'Alessandro, 2006). In the following, we provide a short overview of the impact of the relevant

LF model parameters and their influence in the spectral domain. In Figures 2.9 to 2.12, the glottal flow $g(t)$ and its time-derivative $\dot{g}(t) = \partial g(t)/\partial t$ are displayed in the respective left panels. On the right hand side, the respective magnitude Fourier transforms are depicted, *i.e.* $G(j\omega) = \mathrm{DFT}\{g(t)\}$ and $G_\partial(j\omega) = \mathrm{DFT}\{\dot{g}(t)\}$, where $\mathrm{DFT}(\cdot)$ denotes the discrete Fourier transform (DFT) (Smith, 2007b). Note several characteristic features of the glottal source spectrum. Firstly, there is the *glottal formant*; an emphasized region in the glottal magnitude spectrum of $G_\partial(j\omega)$, typically below 500 Hz. Its name is derived from its shape in the spectral domain, which resembles that of a common VT resonance. If emphasized further, the glottal formant is responsible for a shallow voice sound with relatively low harmonic richness. Secondly worth noting is the *spectral tilt* of the voice source. This term refers to the constant decay of the energy at frequencies above the glottal formant. A stronger spectral decay also leads to an attenuation of higher harmonics and thereby is responsible for the attenuation of high frequencies in soft voices.

As pointed out before, besides the fundamental period, T_0, there are four parameters that mainly determine the characteristics of the LF model, t_p, t_e, t_a and E_e. Further, for reasons of normalization, some helper variables are commonly defined, the effect of which is explored in the following:

E_e: the amplitude at the minimum of the glottal flow derivative, ocurring at t_e. In the spectral domain, E_e acts like a global gain factor and highly correlates with the perceived sound pressure level (SPL) (Holmberg et al., 1988). As can be observed in Fig. 2.9, an increase of E_e leads to a homogeneous amplification of the energy at all frequencies.

O_q: the open qotient, $O_q = (t_e - t_o)/T_0$, relates the duration between the glottal opening instant (GOI) and the instant of maximum excitation to the fundamental period. Note that in fact the name is somewhat misleading, since after t_e, the vocal folds are not yet closed and the open phase in thus not yet terminated. This term has been coined historically though and remains in use. To refer to a normalized open phase, often the term *effective open quotient* is used. As can be seen in Fig. 2.10, a varying open quotient O_q mainly affects the lower frequency regions of $G_\partial(j\omega)$ whereas the higher frequency regions are largely unaffected. In (Doval and d'Alessandro, 2006) it was shown that a linear dependency exists between O_q and the maximum frequency of the glottal formant in that a doubling of the value of O_q leads to a 6 dB increase of the glottal formant amplitude. Furthermore, it may also be observed that O_q changes the position or *centre frequency* of the glottal formant.

α_m: the asymmetry coefficient, defined by the ratio between the opening phase and the open phase durations; $\alpha_m = (t_p - t_o)/(t_e - t_o)$. As illustrated in Fig. 2.11, this parameter also mainly affects the glottal formant whereas the spectral tilt remains largely unaffected by a varying asymmetry coefficient. Besides influencing the amplitude of the glottal formant, α_m mainly plays the role of controlling the *bandwidth* of the glottal formant.

Q_a: the return phase quotient, defined by the ratio between the return phase time constant and the duration between the glottal closure instant (GCI) and the end of the period:

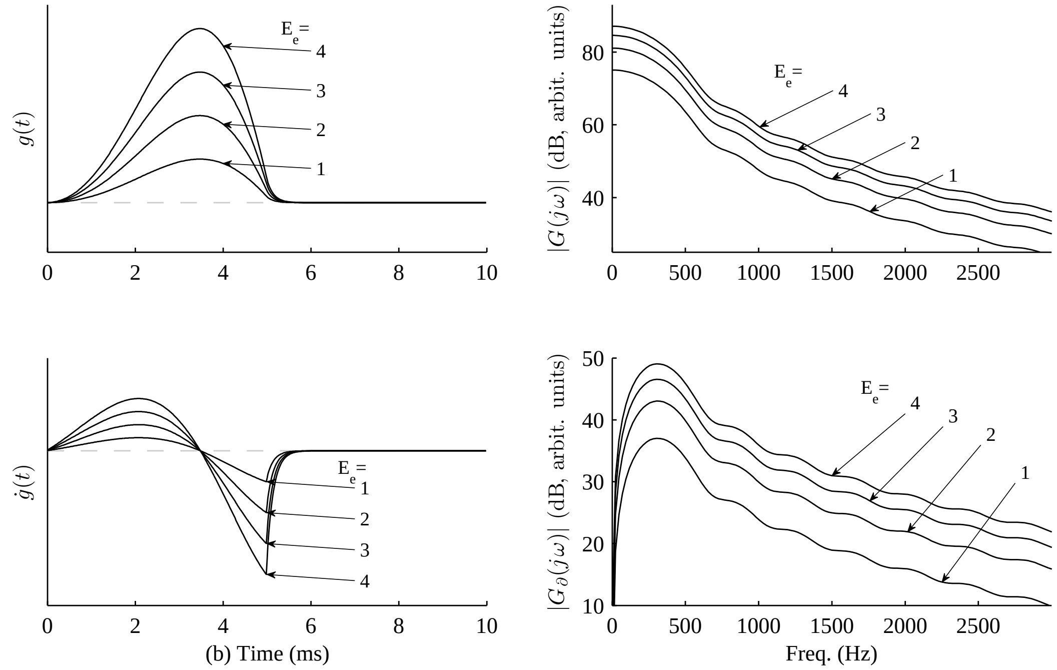


Figure 2.9: Correlates of E_e on the LF model of the glottal flow waveform (top) and its derivative (bottom) in the time-domain (left) and the spectral domain (right).

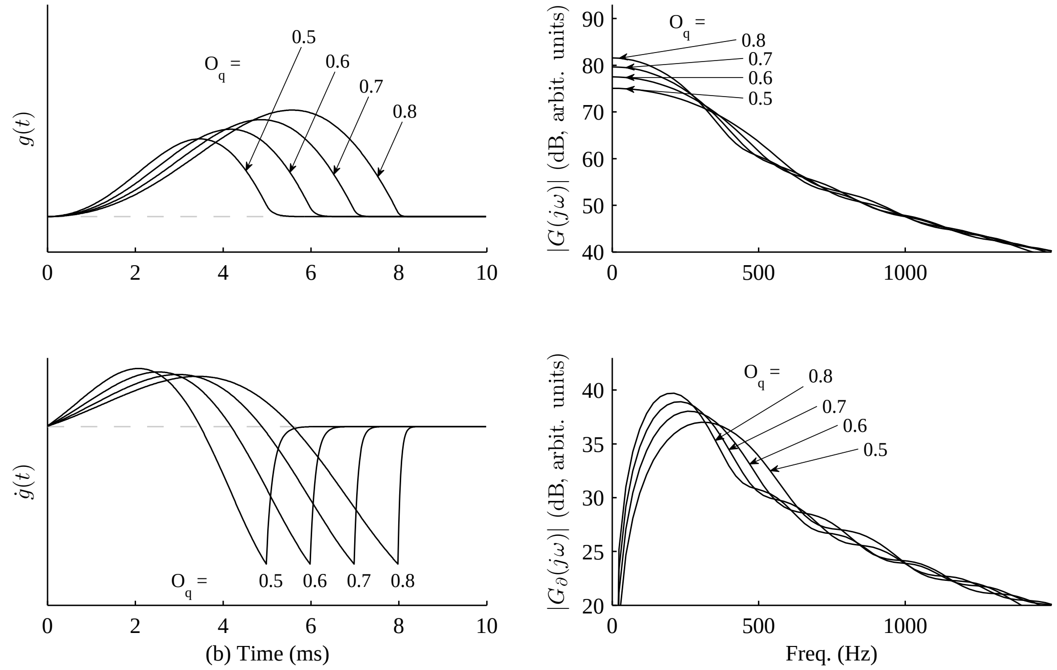


Figure 2.10: Correlates of O_q on the LF model of the glottal flow waveform (top) and its derivative (bottom) in the time-domain (left) and the spectral domain (right).

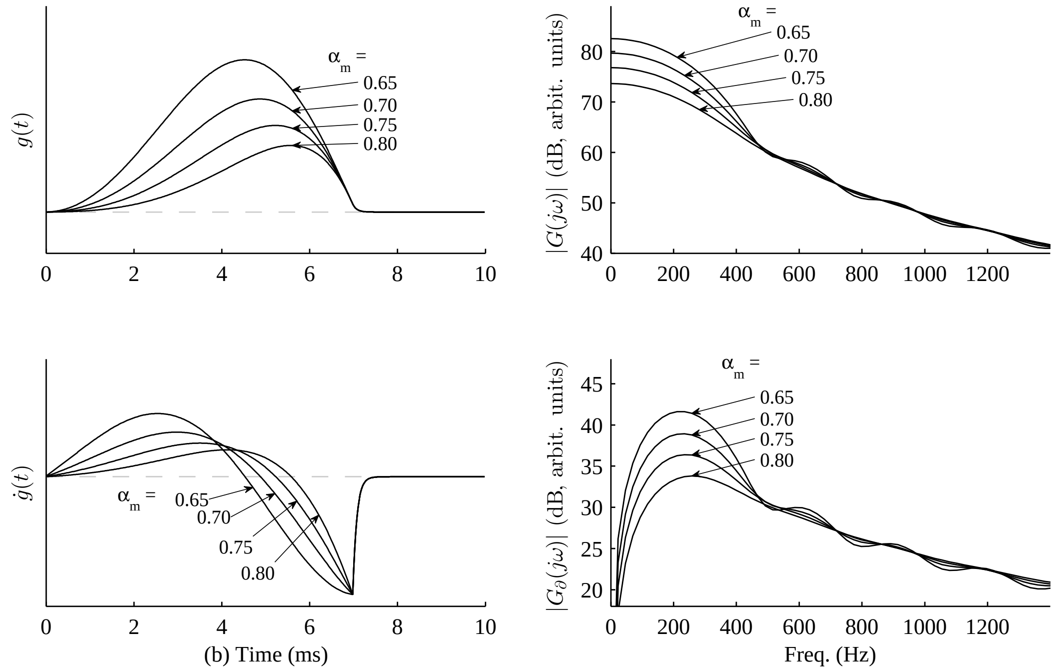


Figure 2.11: Correlates of α_m on the LF model of the glottal flow waveform (top) and its derivative (bottom) in the time-domain (left) and the spectral domain (right).

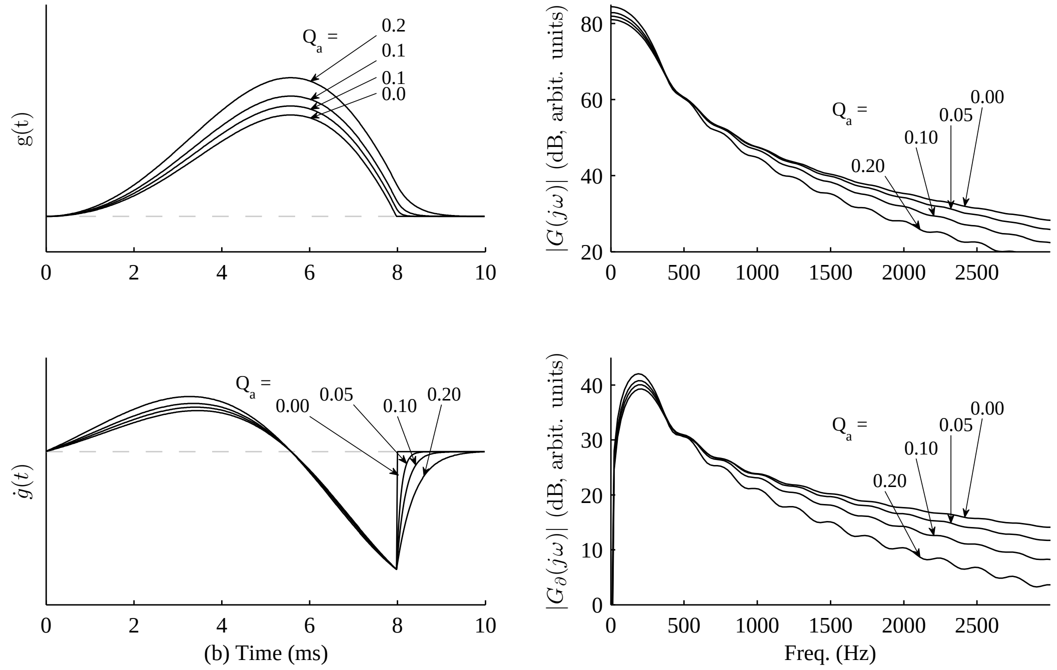


Figure 2.12: Correlates of Q_a on the LF model of the glottal flow waveform (top) and its derivative (bottom) in the time-domain (left) and the spectral domain (right).

$Q_a = t_a/[(1-O_q)T_0]$. This parameter is of high importance for the source-filter separation, as will be further outlined in Chapter 4. The main effect of an increased return phase duration (*i.e.* an increased Q_a value) is illustrated in Fig. 2.12. It largely influences the *spectral tilt* above a certain cutoff frequency and thereby plays an important role in the attenuation and amplification of higher frequency harmonics and formants. A small value of Q_a entails a short return phase, the extreme case being to have no return phase at all and an instant closure of the glottis. In turn, this yields a glottal source excitation signal with large energy at high frequencies and an excitation of the VT over a high range of frequencies. A second order side effect is a slightly influenced centre frequency and bandwidth of the glottal formant. For an analytic way of determining the cutoff frequency for the spectral tilt, please refer to (Doval and d'Alessandro, 2006).

2.3.2 Non-deterministic Voice Components

As described in Sec. 2.3.1, the voice source can be considered as being made of deterministic and non-deterministic components. The latter term refers to all those parts of the source signal that are not directly a result of the volume velocity waveform modulated by the lateral movement of the vocal folds and often modelled by one of the source models described above. These non-deterministic components comprise aspiration noise, glottal jitter and shimmer and effects of the non-linear interaction between the vocal tract inertia and the glottal flow.

Aspiration noise is produced at various, sporadically existing constrictions in the path of the exhaled air. Since in this work we are mainly concerned with the analysis of voiced sounds, we limit the description of aspiration noise to so called *glottal noise*; where the source of the noise is at or near the vocal folds. Aspiration noise often is quantified using the Reynold's number, Re (Flanagan, 1972). This is a measure of the probability of the occurrence of a turbulence given a medium of viscosity μ and of density ρ that is travelling with a volume velocity U through an area A:

$$Re = \frac{2\rho U}{\mu\sqrt{\pi A}}. \tag{2.4}$$

Given that Re surpasses a certain value Re_c, the relationship between the resulting noise sound pressure level P_n and Re is:

$$P_n = \begin{cases} Re^2 - Re_c^2 & Re > Re_c \\ 0 & otherwise \end{cases} \tag{2.5}$$

Note that the glottal flow $g(t)$ and the glottal area are out of phase though, which is one of the causes for self-oscillation of the vocal folds (see Sec. 2.1.1). The opening and closing of the vocal folds always precedes the resulting glottal flow. According to Eq. 2.4, we observe that the ratio $U/\sqrt{A}$ as result of the phase delay is always more skewed than the flow itself. The aspiration noise may always be expected to reach its maximum amplitude around t_e.

Jitter and *shimmer* refer to irregularities in the periodicity of the vocal fold oscillation.

In particular, jitter is a measure of the temporal deviation of glottal cycles from a perfect harmoniousness. It is usually given in percentage of the instantaneous absolute period, T_0. In normal phonation, jitter values between 0.1 % and 1.0 % were reported (Brockmann et al., 2008). Similarly, shimmer on the other hand refers to variations of the amplitude of the glottal excitation around a nominal value, E_e.

In Sec. 2.1.1 the self-sustained oscillation of the vocal fold movements were described. It is known that one major factor contributing to the self-sustainability is the just-in-time arrival of pressure waves at the glottis, which are back-propagating in the vocal tract. Inherently, these subglottal and supraglottal pressure differences lead to a *distortion* in the linear source-filter model of speech production. Such non-linear feedback contributions are not captured in the deterministic voice models. Various effects of this non-linear feedback mechanism were described in the literature (Quatieri, 2001). Most prominently is the deviation from the glottal model during the opening phase, often referred to as *formant ripple*.

2.3.3 The Vocal Tract Filter

In Section 2.2.2 it is illustrated how the acoustic properties of the vocal tract can be modelled using a sequence of connected tube segments and the resulting impedance sections. It can be shown that for any number of tubes, the resulting process is autoregressive (AR) and therefore has a transfer function that can be modelled using an all-pole filter. In particular, the number of tubes representing the vocal tract N given the length of the vocal tract L is determined by the sampling rate, f_s (Deller et al., 1993)

$$\tau = \frac{T}{2} = \frac{1}{2f_s} = \frac{L}{cN}, \tag{2.6}$$

where τ is the time a pressure waves takes to propagate from the glottis to the lips, given the speed of sound, c. Therefore, the number of tubes N, or the number of resonances or formants respectively, is determined by $N = 2f_s \cdot \frac{L}{c} = 2f_s \cdot \frac{0.17\,\mathrm{m}}{330\,\mathrm{m/s}} \approx \frac{f_s}{1000}$. In practical applications it is usually recommended to slightly overestimate the order so as to accommodate for spurious noise or other undesired factors. Hence, the common rule of thumb for determining the order of the VTF is (Markel and Atal, 1976)

$$N_f = \left\lceil \frac{f_s}{1000} \right\rceil + 1. \tag{2.7}$$

In the following we will describe the common all-pole model for the vocal tract transfer function, as it is used widely. The details highlighted here will serve as a basis for the discussions and motivation for the proposed methods in the next chapter.

In digital signal processing, an all-pole filter of order p in the discrete time-domain is expressed using the notation $x(n)$ for the input signal and $y(n)$ for the output signal by the following difference equation

$$y(n) = b_0 x(n) - a_1 y(n-1) - \ldots - a_p y(n-p) = b_0 x(n) - \sum_{k=1}^{p} a_k y(n-k), \tag{2.8}$$

where b_0 and $a_1 \ldots a_p$ are the real-valued *weighting factors* or *filter coefficients*. An important tool for the analysis of discrete signals is the z-transform, which transforms a discrete time-domain signal into a complex frequency-domain representation. For *causal* signals and filters, the unilateral z-transform $X(z)$ of a signal $x(n)$ is defined as

$$X(z) \triangleq \sum_{n=0}^{\infty} x(n) z^{-n}, \tag{2.9}$$

or in operator notation

$$X(z) = \mathcal{Z}\{x(\cdot)\}, \tag{2.10}$$

The amplitude A of the complex variable $z = Ae^{j\phi}$ determines the region of convergence (ROC) and is usually chosen such that the z-transform summation converges to zero. Note that for $A = 1$ the z-transform resembles the Discrete Fourier Transform (DFT). The z-transform of our all-pole filter can be written as

$$\begin{aligned} \mathcal{Z}\{y(\cdot)\} &= \mathcal{Z}\{b_0 x(n) - a_1 y(n-1) - \ldots - a_p y(n-p)\} \\ &= \mathcal{Z}\{b_0 x(n)\} - \mathcal{Z}\{a_1 y(n-1)\} - \ldots - \mathcal{Z}\{a_p y(n-p)\} \\ &= b_0 \mathcal{Z}\{x(n)\} - a_1 \mathcal{Z}\{y(n-1)\} - \ldots - a_p \mathcal{Z}\{y(n-p)\} \\ &= b_0 X(z) - a_1 z^{-1} Y(z) - \ldots - a_p z^{-p} Y(z) \\ &= b_0 X(z) - Y(z)\left[a_1 z^{-1} - \ldots - a_p z^{-p}\right]. \end{aligned} \tag{2.11}$$

Regrouping and defining $A(z) \triangleq \left[1 + a_1 z^{-1} - \ldots - a_p z^{-p}\right]$ lets us write the transfer function $H(z)$ of our all-pole filter

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0}{A(z)} = \frac{b_0}{1 - a_1 z^{-1} - \ldots - a_p z^{-p}} \tag{2.12}$$

The z-transform of a signal x can be seen as a polynomial in z^{-1}. Therefore we can apply the fundamental theorem of algebra and factor the pth order polynomial into a product of p first-order polynomials:

$$H(z) = \frac{b_0}{(1 - q_1 z^{-1}) \ldots (1 - q_p z^{-1})} = \frac{b_0}{\prod_{k=1}^{p} \left(1 - q_k z^{-1}\right)}, \tag{2.13}$$

where q_k represent the complex-valued roots of the polynomial $A(z)$, in polar notation $q_k = r_k e^{j\phi_k}$. Note that it is therefore possible to consider Eq. 2.13 as a series of single-order all-pole filters, $H_k(z)$:

$$H(z) = b_0 \prod_{k=1}^{p} \frac{1}{\left(1 - q_k z^{-1}\right)} = b_0 \prod_{k=1}^{p} H_k(z), \tag{2.14}$$

To obtain the frequency response in factored form, consider we take the factored form of the general transfer function, Eq. 2.13 and we set $z = e^{j\omega T}$ (T is the sampling period, $T = 1/f_s$):

$$H(e^{j\omega T}) = \frac{b_0}{\prod\limits_{k=1}^{p} \left(1 - q_k e^{j\omega T}\right)}. \tag{2.15}$$

The magnitude spectrum $G(\omega) \triangleq \left|H\left(e^{j\omega T}\right)\right|$ can be expressed as

$$\begin{aligned} G(\omega) &= \frac{b_0}{\prod\limits_{k=1}^{p} \left|\left(1 - q_k e^{j\omega T}\right)\right|} \\ &= \frac{b_0}{\left(e^{j\omega pT}\right) \cdot \prod\limits_{k=1}^{p} \left|\left(e^{j\omega T} - q_k\right)\right|} \\ &= \frac{b_0}{\prod\limits_{k=1}^{p} \left|\left(e^{j\omega T} - q_k\right)\right|} \end{aligned} \tag{2.16}$$

From Eq. 2.16 we note that the magnitude response $G(\omega)$ at a particular frequency ω is given by the reciprocal of the product of the distances of each pole q_k to the point $e^{j\omega T}$ on the unit circle.

At this point we may realize that the filter representation in the factorized form is very useful. The factors q are the complex roots of the polynomial $A(z)$. They convey all the information with respect to the spectral characteristics of the filter $A(z)$.

An instructive illustration of the meaning of the poles is provided by the *z-plane.* Consider first the case of a first order all-pole filter:

$$\begin{aligned} y(n) &= b_0 x(n) - a_1 y(n-1) \\ H(z) &= \frac{b_0}{1 - a_1 z^{-1}} = \frac{b_0}{1 - q_1 z^{-1}}. \end{aligned} \tag{2.17}$$

For this simple filter, the root of $(1 - a_1 z^{-1})$ is simply $q_1 = a_1$. Since we require all filter coefficients a to be real-valued, so will be q_1. It comes to lie on the real axis in the z-plane (see Fig. 2.13(a)). Following Eq. 2.16, its magnitude response is given by

$$G(\omega) = \frac{b_0}{\left|\left(e^{j\omega T} - q_1\right)\right|}. \tag{2.18}$$

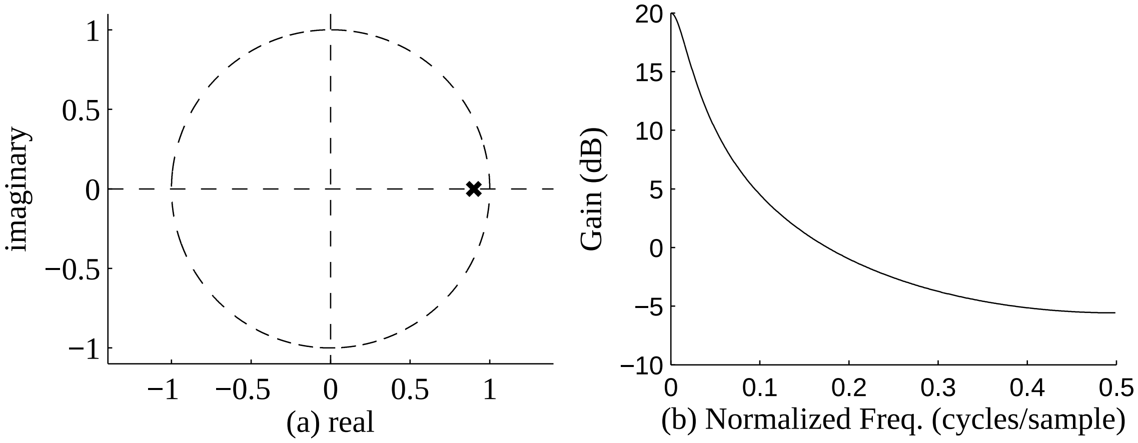


Figure 2.13: First order all-pole filter.

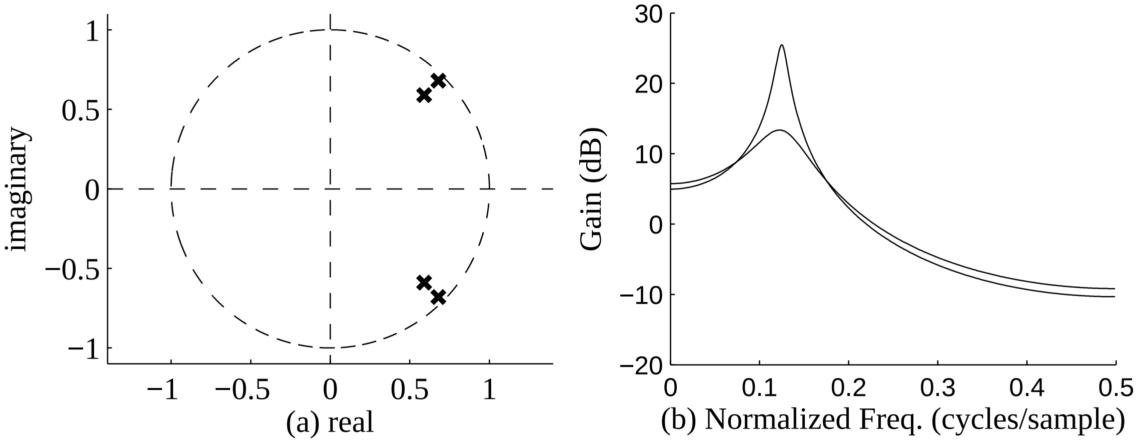
Essentially, if we start to evaluate this equation at different values of ω, we measure the distance between the respective point $e^{j\omega T}$ on the unit circle and the pole q_1. The magnitude response of the filter at a given frequency ω then is the reciprocal of this distance multiplied by b_0 (2.18). Clearly, the distance between $e^{j\omega T}$ and q_1 is the smallest when the angle of the two complex numbers is the same, *i.e.* $\omega = \phi$. Therefore, a resonance peaks at ϕ. Fig. 2.13(b) illustrates the magnitude response of the single all-pole filter over the range of frequencies from 0 to $f_s/2$.

The center frequency of the resonance due to a pole, f_r, is determined by the pole's angle $\angle(q) = \angle(re^{j\phi}) = \phi$ in the z-plane by

$$f_r = \pm 0.5 f_s \phi / \pi. \tag{2.19}$$

In other words, if we introduce an imaginary component to our complex-valued pole, the frequency of the resulting resonance changes. Yet, a single complex-valued pole would breach one of our initial requirement by yielding complex-valued filter coefficients when expanding the roots of $H(z)$. This can be prevented by introducing the non-real-valued poles in complex conjugate pairs. Throughout the expansion, the imaginary parts will cancel each other and we will be left with real-valued filter coefficients. This means that for each resonance frequency, it is necessary to introduce a pair of complex conjugate poles. Note also that the true resonance frequency of a pole pair is not exactly defined by ϕ for complex conjugate pole pairs, since the magnitude response is defined by using the product of all pole-unit-circle distances (Eq. 2.16). Although the main contribution to the final magnitude response near ϕ always results from the closest pole, the eventually resulting peak is slightly offset by the magnitude contributions of the other poles. This fact is commonly neglected in practice though.

In Fig.2.14 the poles in the z-plane (a) and the respective magnitude response (b) of two different second-order all-pole filters are displayed. This figure illustrates the second important property encoded in the complex pole values. The radius of the pole, r controls both the *amplitude* and the *bandwidth* of the resonance. The pole-pair having a smaller radius yields a magnitude response having a lower but wider resonance peak. The 3 dB-bandwidth of a



(a) real (b) Normalized Freq. (cycles/sample)

Figure 2.14: Second order all-pole filter.

formant, b_r, can be approximated by

$$b_r = -\frac{f_s}{\pi}\log_n(r), \tag{2.20}$$

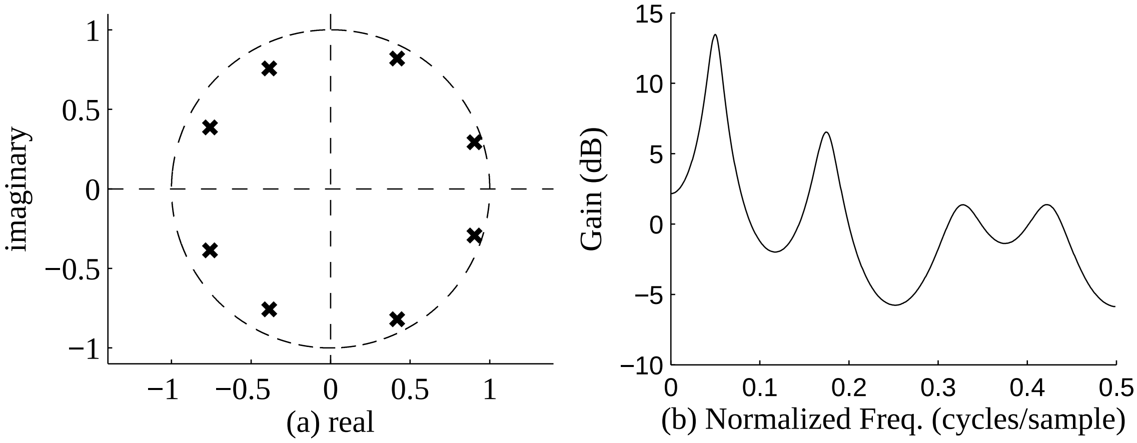
where $r = |q| = |re^{j\phi}|$.

Another important property of all-pole filters that we will meet again later is concerned with the stability of the system. A filter is defined to be *stable* if its impulse response $h(n)$ decays to 0 as n goes to infinity. In terms of an all-pole filter, this means that all its poles must lie strictly within the unit circle, *i.e.* the radius r of each pole must be smaller 1, $|r| < 1$. To see this, consider the causal impulse response of the form $h(n) = r^n e^{j\omega nT}$ for $n = 0, 1, 2, \ldots$. Given $0 < r < 1$, this signal is a damped, complex sinusoid oscillating with a zero-crossing rate of ω/π per second and a decaying amplitude envelope. If r were to be greater than 1, the amplitude envelope would diverge towards infinity. The signal $h(n)$ has the z-transform

$$\begin{aligned} H(z) &= \sum_{n=0}^{\infty} r^n e^{j\omega nT} z^{-n} \\ &= \sum_{n=0}^{\infty}\left(re^{j\omega T}z^{-1}\right)^n \\ &= \frac{1}{1 - re^{j\omega T}z^{-1}}, \end{aligned} \tag{2.21}$$

where the last step follows from the definition of the sum of a geometric series. This last equality holds *iff* $|r| < 1$, otherwise the unit circle is not included in the ROC and the frequency response is not defined. Therefore it is strictly required that all poles of a filter have a radius $|r| < 1$.

In the beginning of this section we looked at the minimum number of formants or resonance regions required to model a vocal tract of length L in a system having a sampling rate, f_s (Eq. 2.7). It is relatively easy to see now that the second-order all-pole filter from above may



(a) real

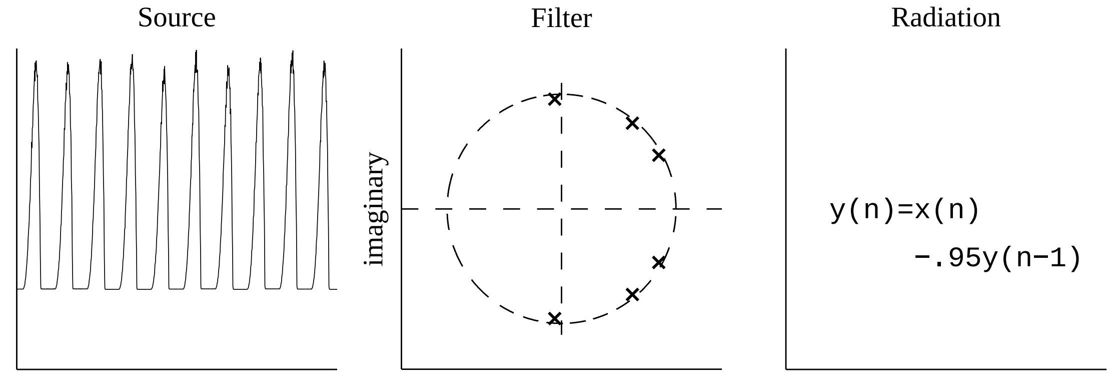
(b) Normalized Freq. (cycles/sample)

Figure 2.15: All-pole filter modelling four resonance regions.

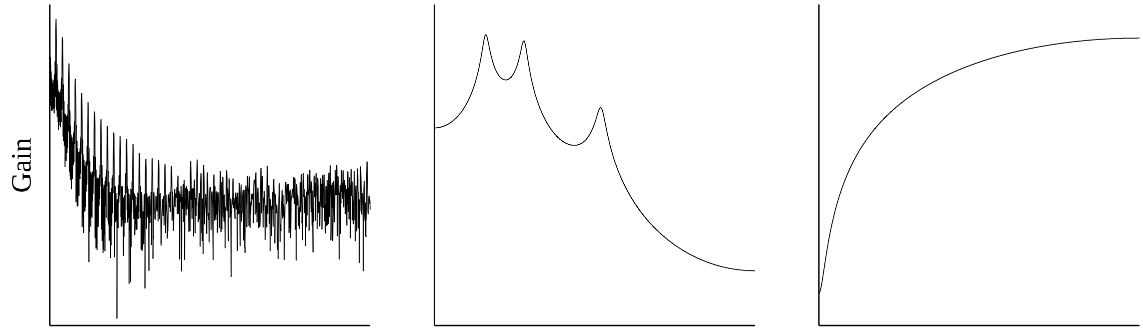
easily be extended to represent more than one resonance by adding a complex conjugate pole pair for each formant. In Fig. 2.15 an order eight all-pole filter is depicted, modelling four formants.

2.3.4 Speech Production Process

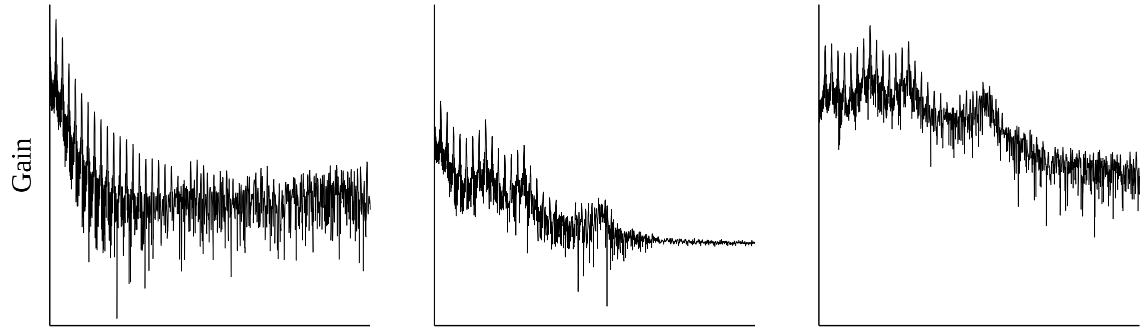
To summarize the process of speech production using the source-filter model, we depict the different steps involved in Fig. 2.16 in different domains (top panels), as well as their Fourier spectra (middle panels) and the resulting combined Fourier spectra (bottom panels). The figure resembles from left to right the terms comprising Eq. 2.1, namely the periodic voice source $G(j\omega) \cdot H(j\omega)$, the vocal tract filter $A(j\omega)$ and the radiation at the nostrils and the lips, $L(j\omega)$. The source comprises both, deterministic and non-deterministic components as presented in Sections 2.3.1 and 2.3.2. In particular, the LF model is used to produce a glottal flow waveform $g(n)$, which is distorted by jitter and to which aspiration noise is added. The resulting periodic waveform is depicted in Fig. 2.16(a). Note the discrete sampling of the glottal spectrum in panel (d) due to the periodicity of the glottal flow waveform. Also, note that the harmonics of f_0 are well below the level of aspiration noise above a certain frequency. The vocal tract filter, simplified to an all-pole filter as presented in Section 2.3.3, is displayed in panels (b) and (e). As can be observed in panel (h), the VTF has the effect of a spectral envelope applied to the glottal source spectrum. Eventually, the radiation of the speech signal from the vocal tract into the open field at the nostrils and the lips is modelled as a first-order differentiator (panel (c)) having a spectral slope of +6 dB per octave. The final speech spectrum is displayed in panel (i).



(a) Time (b) real (c)



(d) Frequency (e) Frequency (f) Frequency



(g) Frequency (h) Frequency (i) Frequency

Figure 2.16: Steps involved in the source-filter model speech production. From left to right: glottal flow waveform, vocal tract filter and lip radiation. The middle panels display the respective spectral contributions and the bottom panels illustrate the combined spectra at the respective stages.

2.4 Conclusion

- To understand the mechanisms of voice production it is important to derive accurate models describing all possible linear and non-linear interactions between the source and filter (Section 2.1). However, for the purpose of this study, we assume that in terms of the perception of speech it is sufficient to model speech production using a linear model without feedback between the vocal tract and the source, the source-filter model (Section 2.3).

- The main objective of this work is the evaluation of a novel estimation method for the separation of the voice source and articulation. Therefore, the well-known and well-studied Liljencrants-Fant (LF) glottal model has selected for this work to represent the voice source from a list of known existing glottal models (Section 2.3.1).
- The resonances of the VTF can be modelled by an all-pole filter, whereas the resonances in the nasal cavity introduce pole-zero pairs. Commonly the latter resonances are neglected due to their lower importance and for mathematical simplicity. Due to its passivity all poles of the VTF reside inside the unit circle and thereby ensure filter stability.
- The radiation at lips and nostrils is modelled using a single, stationary, first-order filter. Although this model is only valid for low frequencies and small mouth opening, we will assume that it is valid for common speech signal.

3 Theory and Tools for Source-Filter Separation

In this chapter we will outline some of the theoretical background underlying the source-filter separation and then motivate the proposed methods by pointing out some of the disadvantages of currently used methods. Thereafter we will describe a computational tool named differential evolution that is used in the proposed methods.

In general, the decomposition of speech into voice source and articulation filter is very interesting because the two components of the speech signal carry rather different and indipendent linguistic information. The source controls the *pitch*, which is the acoustic correlate of the intonation and a strong cue for interpretation of *prosody*. Furthermore, non-semantic information is contained in the voice quality. For instance, hoarseness, breathiness or unsteadiness in the pitch may subconsciously reveal information about a speaker's psychological state. Voice quality can also serve as a marker for certain psychological conditions (Pittam, 1987). On the other hand, the filter mainly carries semantic information encoded in the articulation of vowels. Also in the filter linguistic information may be contained that is more than just semantics. A general shift in the center frequencies of vowels may reveal indications on the mental condition of the speaker such as depressive or manic moods (Hargreaves and Starkweather, 1964). Therefore, the separation of the source and the filter is potentially useful in many areas such as speech coding and analysis (Schröder, 2009), parametric speech synthesis (Raitio et al., 2011), remote and/or non-invasive voice disorder diagnosis (Hartl et al., 2005), restoration of pathological voices (Schleusing et al., 2011) or as front-end processing for classification tasks such as speaker verification (Plumpe et al., 1999). Unfortunately though, the task of separating source and filter is everything but trivial.

3.1 General Considerations for Source-Filter Separation

In most common applications, voiced speech is analyzed on a frame-by-frame basis through *glottal inverse filtering* (GIF) (Miller, 1959). This approach comes with some flaws as we will see in this chapter. GIF attempts to first obtain an estimate of the VT filter and thereafter an estimate of the glottal source by filtering the speech signal with the inverse of the VTF filter. Let us write down those two expressions derived from the general expression of speech production, Eq. 2.1 on page 23. Firstly, in principle, the deterministic component of the glottal source can be estimated from the speech signal $S(j\omega)$ by division of the VTF and the radiation filter:

$$\hat{G}(j\omega)\cdot H(j\omega) = \frac{S(j\omega)}{A(j\omega)L(j\omega)}. \tag{3.1}$$

Similarly, the VTF can be estimated by

$$\hat{A}(j\omega) = \zeta\left(\frac{S(j\omega)}{G(j\omega)L(j\omega)}\right), \tag{3.2}$$

where $\zeta(\cdot)$ is an estimate of a smooth envelope commonly obtained using a method such as LP or DAP (see Section 3.2). Replacing the observed speech with the model of speech production (2.1) allows us to further examine Eq. 3.2:

$$\hat{A}(j\omega) = \zeta\left(\frac{H(j\omega)\cdot G(j\omega)\cdot A(j\omega)\cdot L(j\omega)\cdot}{G(j\omega)L(j\omega)}\right) = \zeta\big(H(j\omega)\cdot A(j\omega)\big), \tag{3.3}$$

Thus it is clear that estimation of the VTF envelope has to cope with the harmonic structure inherent to the glottal source. The envelope $\hat{A}(j\omega)$ is an interpolation of the sampling of the VTF by the harmonics of f_0 of the glottal source, $H(j\omega)$. More generally spoken, the estimation of the VTF in (3.3) is an inverse problem to obtain an estimate of a transfer function of a system (the vocal tract), which is excited by the the glottal source. *System identification* deals with such a problem by employing a known white noise or a known sweep tone as an input to the filter covering the entire frequency range of interest. Given the system is linear, the acquired system output fully describes the spectral characteristics of the system under investigation (Fujimura and Lindqvist, 1971; Ljung, 1999). However, the glottal excitation is an unknown component of the speech production system itself; it is unsteady and does not span all frequencies uniformly:

- The periodic source yields a sampling of the vocal tract and introduces zeros between the harmonic frequencies, thereby leading to a bias in the estimation of the envelope.
- The radiation may be assumed to be stationary, but it generates a zero on the real axis of the z-plane of the z-transform of $S(j\omega)$
- The glottal source does not excite the VTF uniformly over all frequencies. In fact, the glottal source is bandlimited (see Sec. 2.3.1).

These characteristics have a remarkable influence on the performance of most commonly used estimators such as LP. More details of the problems with simplified source models are provided below in Sec. 3.2.4.

The two unknowns, $A(j\omega)$ and $G(j\omega)$, are related to each other through the Eqs. 3.1 and 3.2. Therefore, we have at our hands a joint estimation problem of an unknown filter excited by an unknown source. In currently known methods either the source or the filter is simplified in order to approximate the other, or joint estimation processes are used. In the following we will further examine the problems encountered by common VTF envelope estimators, generally referred to as $\zeta(\cdot)$ in Eq. 3.2. This is then followed by a description of currently known methods of source, filter or joint source-filter optimization methods and a motivation for the approaches presented in the subsequent chapters.

3.2 Autoregressive methods

The standard tool for obtaining an estimate of the spectral VT envelope $A(j\omega)$ is linear prediction (LP), which assumes that the vocal tract can be represented by an all-pole filter and that the input to the vocal tract filter after application of a preemphasis filter (PE) is spectrally white (Makhoul, 1975; Vaidyanathan, 2008). LP is a very versatile tool and has been used extensively not only in the area of speech signal processing, but also in various other fields, such as biomedical signal processing (Gersch, 1970) or geophysics (Robinson, 1967). In the following we provide some insight in the most common method for VTF envelope estimation, linear prediction (LP), and illustrate how the above mentioned problems influence the result.

3.2.1 The Normal Equations

In principle, LP may optimize an all-pole model for autoregressive processes, an all-zero model for moving average (MA) processes or a pole-zero model for autoregressive moving average (ARMA) processes. The linear solutions for AR models presented below cannot easily be extended to pole-zero modelling though, because the solution of ARMA models are non-linear (Steiglitz, 1977). Some methods exist that model the valleys in the VT transfer function (Kopec et al., 1977; Makhoul, 1975). For mathematical convenience though and at the cost of a mostly minor error, in many applications of LP in speech signal processing it is assumed that the vocal tract may in general be modelled by an AR process.

We have already introduced the notation of AR filters in Section 2.3.3. The objective of LP is to find the optimal parameters a_i in Eq. 2.12 on page 34. Let us first define the error $e(n)$ between the discrete-time input signal $s(n)$ and its predicted value, $\hat{s}(n)$. Since $s(n)$ is assumed to be AR, the $\hat{s}(n)$ may be modelled as a linear combination of its p past values of $s(n)$:

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i), \tag{3.4}$$

The error signal $e(n)$ is also known as the *residual*.

The criterion for finding the optimal parameter values a_i is to minimize the mean of the squared error $e(n)$. Since LP is so versatile, different interpretations of the error $e(n)$ and approaches at minimizing it have been presented in different fields. In an algebraic sense, the error is minimized when it is orthogonal to the previous samples $(s(n-i), i \geq 1)$ (Vaidyanathan, 2008). Orthogonality is achieved by setting the expected value of the inner product of the respective vectors to zero to obtain:

$$\mathbb{E}\left[\mathbf{e}(n)\mathbf{s}^*(n-i)\right] = 0. \tag{3.5}$$

In an analytical sense, the mean squared error E for a deterministic signal $s(n)$ can be defined as

$$E = \sum_n e^2(n) = \sum_n \left(s(n) + \sum_{i=1}^{p} a_i s(n-i) \right)^2. \tag{3.6}$$

In a statistical sense, we minimize the expected value of the square of the error for a stochastic signal that is wide-sense stationary (WSS) (Papoulis, 1989):

$$E = \mathbb{E}\left(e^2(n)\right) = \mathbb{E}\left(s(n) + \sum_{i=1}^{p} a_i s(n-i) \right)^2. \tag{3.7}$$

In (3.6) and (3.7) we purposely left the range of the evaluation unspecified for the moment. The optimal values of a_i in (3.6) and (3.7) are found by setting the derivative of E with respect to each parameter $a_i, 1 \le i \le p$ to zero:

$$\frac{\partial E}{\partial a_i} = 0. \tag{3.8}$$

These interpretations all give rise to a set of p equations with p unknowns (a_i):

$$\sum_{k=1}^{p} a_k \sum_n s(n-k)s(n-i) = -\sum_n s(n)s(n-i), \quad 1 \le i \le p, \tag{3.9}$$

which are typically referred to as the *normal equations*, *Yule-Walker* equations or the *Wiener-Hopf* equations.

3.2.2 Solution of Normal Equations

Above, the range of the signal over which $s(n)$ is evaluated to compute the mean squared error E was omitted. In fact, two different approaches that utilize different ranges exist for solving the normal equations. Once this range is found, it is merely a question of applying an appropriate general simultaneous linear equation solving algorithm, *e.g.* Gaussian elimination, Crout decomposition, etc. (Strang, 2003). The two methods each have advantages and disadvantages.

The *autocorrelation* method minimizes the error over all time, *i.e.* from $-\infty$ to $+\infty$. In real applications, signals are not infinitely long though and speech signals also do not have the property to be stationary over an extended period. Therefore the signal is usually windowed using a Bartlett window of length N. Other common windowing functions are the Hamming and Hann windows (Smith, 2011). Eq. 3.9 thus becomes

$$\sum_{k=1}^{p} a_k \mathbf{R}(i-k) = -\mathbf{r}(i), \quad 1 \le i \le p, \tag{3.10}$$

where

$$\begin{aligned} \mathbf{r}(i) &= \sum_{n=-\infty}^{+\infty} s(n)s(n+i) \\ &= \sum_{n=0}^{N-1-i} s(n)s(n+i) \end{aligned} \tag{3.11}$$

is the autocorrelation matrix of the signal $s(n)$. The second equality is valid because we assume that the signal is WSS and therefore ergodic. We may replace the ensemble average with a time average. The matrix **R** has some very useful properties. It is a symmetric *Toeplitz* matrix, where all elements on a diagonal are equal. This means that computationally efficient methods such as the Levinson-Durbin algorithm (Quatieri, 2001) may be employed to inverse $\mathbf{R}(i-k)$ and to solve Eq. 3.10 for a_i.

The *covariance method* differs from the autocorrelation method in the way the autocorrelation matrix is computed. In contrast to Eq. 3.11, it is assumed that the mean squared error E is minimized over a finite interval, say, $0 \le n \le N-1$. Eq. 3.9 then transforms to

$$\sum_{k=1}^{p} a_k \Phi_{ki} = -\Phi_{0i}, \quad 1 \le i \le p, \tag{3.12}$$

where

$$\Phi_{ik} = \sum_{n=0}^{N-1} s(n-i)s(n-k) \tag{3.13}$$

is the covariance of the signal s(n) during the given interval. In the case of the covariance method, the resulting matrix Φ_{ki} does not demonstrate a Toeplitz structure and therefore it is remarkably more difficult to invert by using for instance the Cholesky method or the square-root method (Wilkinson, 1967). In general, the covariance method requires less samples (shorter analysis window sizes) to obtain a reasonable accuracy of the estimated formant coefficients, but on the other hand it is not guaranteed to yield a minimum-phase polynomial for the filter $A(z)$. Therefore, in most applications where the analysis window is not required to be very short, often the autocorrelation method is preferred over the covariance method. On the other hand, the covariance method and its ability to obtain reliable results using short analysis windows is preferred for instance in the closed-phase covariance linear prediction method (see below), in which the vocal tract transfer function is estimated during the very short period of time during which the glottis is closed.

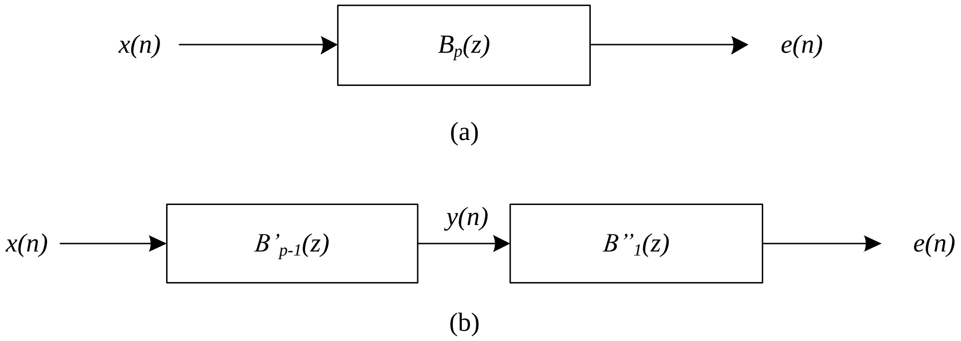


Figure 3.1: The prediction filter $B_p(z)$ in *(a)* and an equivalent drawing with one of the root factors shown in a separate filter *(b)*.

3.2.3 Stability of the Vocal Tract Filter

A very important property of the filter coefficients obtained using the auto-correlation method is their minimum-phase lag property, which implies that the resulting filter is stable. We saw in Eq. 2.14 on page 35 that an all-pole filter can be considered as a series of single-order all-pole filters. Consider the FIR prediction filter $B_p(z)$ of order p obtained by inverting the optimal filter found using LP and the autocorrelation method, $B_p(z) = 1/A_p(z)$. By associativity of multiplication we can rewrite Eq. 2.14 as a series of an order $p-1$ FIR filter, $B'_{p-1}(z)$, and a single-order filter, $B''_1(z) = 1 - qz^{-1}$:

$$\begin{aligned} B(z) &= \prod_{k=1}^{p} \left(1 - q_k z^{-1}\right) \\ &= \prod_{k=1}^{p-1} \left(1 - q_k z^{-1}\right) \cdot \left(1 - q_k z^{-1}\right) \\ &= B'_{p-1}(z) \cdot B''_1(z), \end{aligned} \tag{3.14}$$

as illustrated in Fig. 3.1. Both filters are causal FIR and q in $B''(z)$ is any of the zeros of $B(z)$. By definition, $B''_1(z)$ is the optimal first-order prediction polynomial for the WSS process $y(n)$. Otherwise there would exist another q that would make the mean square error of its output smaller, which would contradict the fact that $B(z)$ is optimal. Therefore, using Eq. (3.10), q can be expressed as

$$q = \frac{R_{yy}(1)}{R_{yy}(0)}, \tag{3.15}$$

where $R_{yy}(i)$ is the autocorrelation of the WSS process $y(n)$. WSS processes have the property that $R_{yy}(0) \leq |R_{yy}(k)|$ for any $k > 0$. Therefore it follows that

$$|q| \leq 1$$

and the resulting filter is guaranteed to be stable. Equality is only achieved for line spectral processes that preserve the input energy without loss, *i.e.* for processes that are perfectly predictable with $e(n) = 0, \forall n$ (Lang and McClellan, 1979; Vaidyanathan et al., 1997).

3.2.4 Limitations of Linear Prediction

The estimation and separation of the source-filter model using LP has several drawbacks. While for most vowels $A(j\omega)$ varies sufficiently slowly to be considered time-invariant during an analysis frame, this is often *not* the case for the glottal source $G(j\omega)$. According to the myoelastic-aerodynamic theory of voice production, the source is mainly affected by the sub-glottal air pressure, the tension of the vocal folds and the physiological configuration of the speech production organs (Titze, 2000). Various combinations of these variables produce diverse glottal waveforms that are perceived as different voice qualities (breathy, modal, pressed, *etc.*). Some voice types, in particular *pathological* voices — those produced by speakers with malfunctioning, partially or even completely excised vocal folds as a result of laryngeal surgery — often exhibit considerable inter-glottal-cycle variations in their period and waveform, which are important acoustical cues often carrying prosodic and idiosyncratic information. These observations imply that the glottal transfer function $G(j\omega)$ is indeed different from the residual of the conventional LP model described above and that the linear source-filter model is a simplification in several aspects:

- Firstly, the glottal source is quasi-periodic in the time domain. This is reflected in the spectral domain by a sampling of the spectral envelope at multiples of the fundamental frequency, f_0. Here, $f_0 = 1/T_0$ is the rate of the vocal fold vibration, i.e. the reciprocal of the fundamental period, T_0. This sampling of the vocal tract envelope and the location of particular harmonics relative to the true formants may have a considerable influence on the formants estimated using LP. Fig.3.2 illustrates this using three examples of a synthetically generated speech signal. The three examples are produced using the method described later in Section 4.5 and are identical except for the value of the fundamental period of the source, T_0. The VT is simulated using an order-six all-pass filter. For the LP estimation, two additional poles (on the real axis in Fig. 3.2(a2-c2)) are used to account for various errors, *e.g.* the mismodelling of the vocal tilt or noise (see below).
  In Fig. 3.2, panels (a1-c1) show DFT spectra of synthetic speech signals generated using slightly different values of f_0. The ground truth vocal tract envelope $A(j\omega)$ (thick grey line) and its estimate $\hat{A}(j\omega)$ (thick black line) obtained using linear prediction (LP) are overlaid on the spectra with a vertical offset for increased clarity. Panels (a2)-(c2)

show the respective roots in sections of the z-plane. We can observe various errors introduced in the estimated formant frequencies and formant bandwidths, depending on the relative distance and energy of the harmonics with respect to the true formants. The errors are introduced by the spectral matching properties of the LP error measure, which Makhoul referred to as the *local property* in (Makhoul, 1975). The ratio between the true power spectrum $P(\omega)$ and the estimated power spectrum $\hat{P}(\omega)$ is defined as the error $E\omega = P(\omega)/\hat{P}(\omega)$. LP minimizes this error such as to yield unity when integrated over the considered spectrum:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} E(\omega)d\omega = 1. \tag{3.16}$$

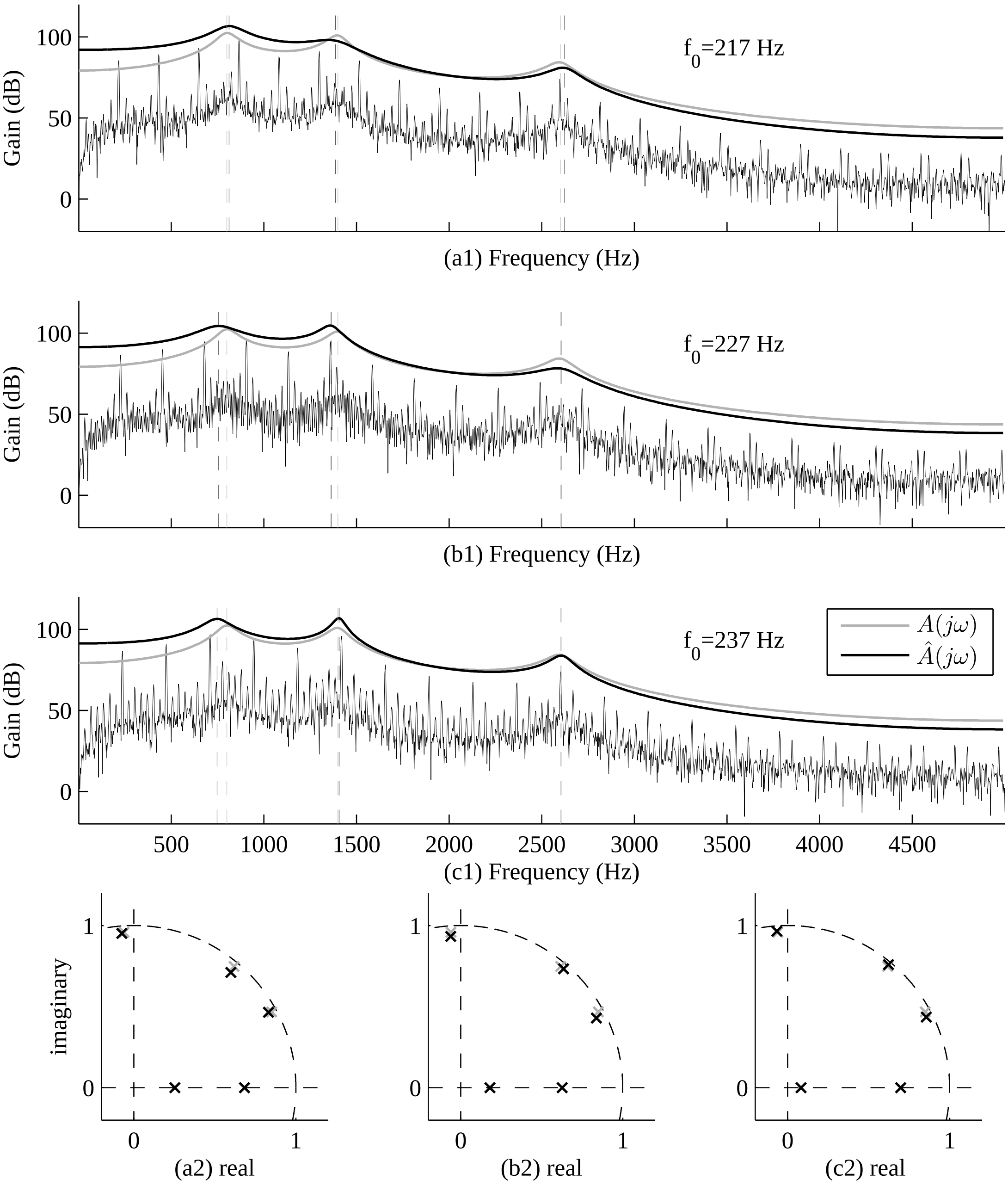
Equation (3.16) can be interpreted as the *arithmetic mean* of $E(\omega)$ that is fixed to a constant value, namely 1. The arithmetic mean has the property to give higher significance to values greater than 1. In other words, such frequencies ω where $P(\omega) > \hat{P}(\omega)$ are contributing more to the total error. Accordingly, on average we expect a better fit of $\hat{P}(\omega)$ to $P(\omega)$ where $P(\omega)$ is greater than $\hat{P}(\omega)$, than where $P(\omega)$ is smaller than $\hat{P}(\omega)$. In general this leads to the tendency of LP to overestimate spectral peaks and to underestimate spectral valleys. More observations can be made though related to the harmonic structure of the speech spectrum.

For instance, the presence of two harmonics with equal distance to a formant frequency increases the bandwidth of the estimated formant, as can be observed in Fig. 3.2(a1) at the location of the second formant or in Fig. 3.2(b1) at the location of the third formant. The mean-squared error criterion of LP leads to a wide distribution of the energy and bandwidth increase of the second formant due to the fifth and sixth harmonics at approximately 1.4 kHz and 1.6 kHz. Accordingly, the pole pair at approximately $\pm\pi/2$ in Fig. 3.2(a2) has a radius estimated to be smaller than the radius of the true pole.

Two harmonics with non-equal distance to a formant introduce a bias in the estimated formant frequency, as illustrated in Fig. 3.2(c1) and (c2) at the location of the first formant. The second source harmonic biases the estimate of the first formant towards a lower frequency.

Furthermore, the coincidence of an f_0-harmonic and a formant at the same frequency may lead to a reduced bandwidth of the estimated formant due to the limited support around the respective harmonic. This can be observed for instance at the location of the second formant in Fig. 3.2(c1) and (c2). Typically, these problems are more severe in voice sources with higher fundamental frequencies, since there the harmonics of the source are more distant. In voice sources with low f_0-values, the denser harmonics provide a better sampling of the spectral envelope and the above mentioned problems are less severe.

These known problems of LP has motivated the development of various methods such as pitch-synchronous LP (PSLP) (Rabiner and Schafer, 1978), closed-phase covariance LP (CPLP) (Wong et al., 1979) or discrete all-pole modelling (DAP) (Roebel et al., 2007; El-Jaroudi and Makhoul, 1991). PSLP uses an analysis window spanning several glottal



(a1) Frequency (Hz)

(b1) Frequency (Hz)

(c1) Frequency (Hz)

(a2) real

(b2) real

(c2) real

Figure 3.2: Illustration of the influence of the harmonic structure of the source signal on the vocal tract estimation. Panels (a1)-(c1) show DFT spectra of synthetic speech signals generated using slightly different values of f_0 overlaid with the ground truth vocal tract envelope $A(j\omega)$ (thick grey line) and its estimate $\hat{A}(j\omega)$ (thick black line) obtained using linear prediction (LP). Panels (a2)-(c2) show the respective roots in sections of the z-plane. The underlying harmonic structure may lead to a formant *bandwith reduction* (*e.g.* F2 in (c1) and (c2)), formant *bandwidth increase* (*e.g.* F2 in (a1) and (a2)) or a *bias* in the formant frequency (*e.g.* F1 in (c1) and (c2)) of the estimated formants.

cycles and the auto-correlation method for the estimation of the LP coefficients. The analysis window is centered on the closed phase of the glottis, between t_e and t_o of the next cycle. CPLP uses a shorter analysis window, spanning exactly the duration of the closed phase of the source. The LP coefficients are then determined using the covariance method. Both methods require a pre-processing step to determine the location of the glottal cycles.

- Secondly, as outlined in Section 2.3.1, the spectral envelope of the actual voice source varies with different parameters such as vocal effort or vocal fold tension. As a consequence, the actual source spectral envelope in some voice types permanently deviates from the non-adaptive PE filter used by LP, possibly to a large degree. This deviation may have several causes. Different voice types exhibit various degrees of spectral tilt. Also, during the glottal open phase, there exists a non-linear feedback of the pressure in the vocal tract to the glottal volume velocity waveform. As a result, the glottal source waveform is modulated by the supraglottal pressure and it exhibits ripples and a glottal formant that is not accounted for by the myoelastic-aerodynamic theory of voice production (Titze, 2000; Quatieri, 2001). Typically the order of the AR model estimated by LP is chosen such as to account for such model deviations. Additional poles of the AR model may model spurious noise peaks in the spectrum or they take on the role of modelling the mismatch between the actual spectral tilt and the tilt represented by the PE-filter. As can be seen in Fig. 3.2(a2-c2), real positive values are assigned to the two additional poles by LP, thereby accounting for a spectral tilt mismatch. As our analysis in Section 2.3.1 has revealed though, the shape of the glottal spectral envelope may be more complex than the envelope of a single-order lowpass filter. For instance, a single real-valued pole may not account for the vocal formant. In those cases where the glottal source envelope shape deviates considerably from the average represented by the preemphasis filter, the poles meant to model the formants are diverted by the glottal source spectrum. A possible solution to overcome this problem is to use an all-pole PE filter of order higher than one and to adaptively adjust the coefficients of this PE filter for each signal frame being analysed. A representative example of such a method is the *iterative adaptive inverse filtering* (IAIF) method (Alku et al., 1991), which iteratively estimates the coefficients of a low-order filter representing the glottal source and an all-pole VT filter.

- Thirdly, the time-invariance of the PE filter inherently poses a problem for glottal sources with large waveform shape variability between consecutive glottal cycles. Typical LP analysis frames comprise several glottal cycles and the variation in the glottal waveform shapes is averaged throughout this duration. A longer analysis frame duration would improve the cancellation of these variations but would also impose a reduced temporal resolution of the time-varying VT envelope. An alternative approach for reducing the influence of the glottal source would be to restrict the LP analysis to the zero-input closed phase (CP) of the glottal cycle, as for example in the closed-phase covariance linear prediction (CPLP) (Wong et al., 1979). However, the performance of

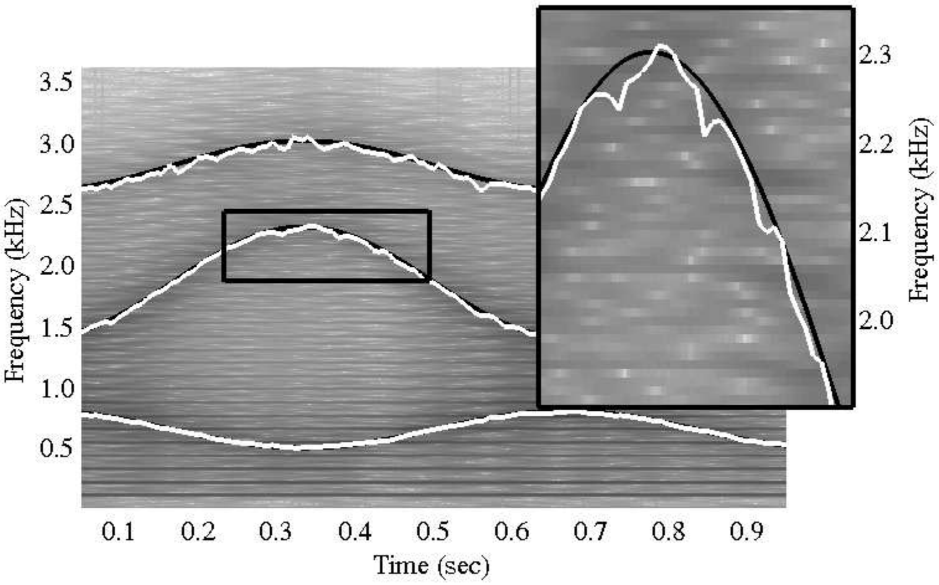


Figure 3.3: Section of a spectrogram of a synthetically generated vowel transition, overlaid with the true formant center frequencies (solid black lines) used for synthesis and their estimates using conventional linear prediction (solid white lines). The estimates, in particular the one of the third formant, are biased towards a lower frequency, and also exhibit a considerable variance in subsequent frames due to the underlying harmonic signal structure.

> this method depends on the duration of the closed phase. Furthermore, although the covariance method of linear prediction usually outperforms the autocorrelation method for short segments, there is no guarantee that the resulting VT filter is *stable, i.e.* that it has all its poles inside the unit circle (Vaidyanathan, 2008).

Following the argumentation above it becomes obvious that LP and especially the simplifications of the voice source model are leading to an insufficient separation of source and filter. Fig. 3.3 illustrates a typical spectrogram of a smooth vowel transition of a synthetic speech signal. Overlaid, in a black solid line are the center frequencies of three formants and their LP estimates as white, solid lines. Clearly, the formant estimates do not merely capture the VTF components, but are also influenced by the source signal components. In the following, we will provide an overview of currently known, more advanced source-filter separation methods and conclude this chapter with the motivation for the chosen approach presented in the subsequent chapters.

3.3 Source-Filter Separation Methods

There are different ways to obtain an indirect measurement of the glottal source properties. Well known is for instance the glottal flow mask allowing neutralization of the vocal tract and a direct measure of the glottal flow (Rothenberg, 1973). Widely used is also the method of electroglottography (EGG) (Lecluse et al., 1975; Fourcin and Abberton, 1971), where the electrical impedance of the throat is measured laterally. The impedance is modulated by the size of the glottis. The obtained signal may be used to detect instants of glottal closure (Thomas and Naylor, 2009; Henrich et al., 2004) or the open quotient (Sturmel et al., 2006; Henrich et al., 2004). These signals often serve as a reference or auxiliary signal for other methods. Contrarily, in this work, we are interested in estimating source and filter from acquired speech signal directly.

3.3.1 Pre-emphasis filtering

The employment of a preemphasis filter to represent an average spectral contribution of the glottal source and lip radiation is very widely used. The spectral tilt of the glottal source $G(j\omega)$ is assumed to be time-invariant. It is approximated using a second order low-pass filter having a spectral slope of -12 dB per octave (Fant, 1960; Markel and Atal, 1976), *i.e.*

$$G(j\omega) = \frac{1}{(1-\mu e^{-j\omega})^2}. \tag{3.17}$$

The common root μ is usually assumed to be real-valued and close to unity, $0 \ll \mu < 1$. As before, also the lip radiation $L(j\omega)$ is time-invariant and approximated by a differentiator with a single zero, ν, yielding a spectral slope of $+6$ dB per octave:

$$L(j\omega) = 1 - \nu e^{-j\omega}. \tag{3.18}$$

Consequently, if one chooses $\mu = \nu$, their joint effect can modelled by a single order preemphasis filter with a net slope of $+6$ dB per octave. From Eq. 3.2, the vocal tract filter estimate $\hat{A}(j\omega)$ can be obtained by

$$\hat{A}(j\omega) = \zeta\left(\frac{S(j\omega)\cdot(1-\mu e^{-j\omega})^2}{1-\nu e^{-j\omega}}\right) = \zeta\left(S(j\omega)\cdot(1-\mu e^{-j\omega})\right). \tag{3.19}$$

The PE filter captures the average of the spectral contributions of the glottal source and the lip radiation. In order to obtain an estimate of the source, the speech signal $S(j\omega)$ is inverse filtered using the filter $B(j\omega) = 1/\hat{A}(j\omega)$.

3.3.2 Analysis-by-Synthesis Methods

Contrarily to the previous approach, this method allows the source to be time-variant. The idea behind analysis-by-synthesis methods is to store a codebook of typical source and filter components and then find the best combination that minimizes some error criterion. This approach was initially proposed in (Stevens, 1960; Bell et al., 1961), where six source spectra and 24 resonance spectra were used as codebooks and the mean squared log amplitude difference served as an error criterion. In a similar manner, an ARMA model was used in (Mathews et al., 1961), assuming the source and the filter consisted of zeros and poles, respectively. Other methods used codebooks only for the source (Shue et al., 2009) and amplitude differences in harmonic models (Oliveira, 1993) or applied constraints on the formant frequencies or bandwidths (Hawks and Miller, 1995) to find optimal codebook entries. All these methods have in common to only utilize the magnitude spectrum for the computation of an error criterion.

3.3.3 Iterative Adaptive Inverse Filtering (IAIF)

IAIF was first introduced by Alku in (Alku et al., 1991) and (Alku, 1992) and has found since many applications, for example in speech synthesis (Alku et al., 2006; Raitio et al., 2011). IAIF uses an autoregressive error minimization method such as LP or DAP (Alku and Vilkman, 1994) to obtain initial estimates of AR models of the source and the VTF. First, the source is estimated from a windowed signal segment spanning several glottal cycles using a low-order (order 1) all-pole model. After cancelling the estimated effect of the source, a preliminary, higher order (order p) estimate of the vocal tract is obtained. These first steps are equivalent to using a preemphasis filter with coefficient μ, only that here, μ is adapted to the spectral tilt instead of being fixed. In a second iteration, a refined estimate of each model is obtained by repeating the first steps. This time the source is estimated after cancellation of the preliminary estimate of the VT and by using a higher order (order g) AR model. Again, the effect of the source is cancelled before estimating a refined version of the VT model.

3.3.4 Adaptive Estimation of the Vocal Tract (AEVT)

The method proposed in (Akande and Murphy, 2005) utilizes the fact that the glottal formant (see Section 2.3.1) can be approximated independently from the vocal tract formants. A high-pass filter designed from this approximation is then used to eliminate the influence of the glottal formant and the glottal tilt from the speech signal. Therefore, this method is similar to using a static preemphasis filter, because the used filter is also single order, but in fact the filter adapts to the time-varying signal characteristics.

3.3.5 Closed-Phase Linear Prediction

An entirely different approach at eliminating the effect of the glottal source is to carry out the formant estimation only during very short segments during which the glottis is closed (Strube, 1974; Wong et al., 1979). Referring back to Fig. 2.8 on page 25, the analysis window for this method is restricted to the period from time instant t_c until t_o of the subsequent glottal cycle. This period may of course be very short in some voices. Consequently, covariance linear prediction is typically used for solving the normal equations. The method has been shown to produce accurate estimates for normal speech with low fundamental frequency and well-defined closed phase (D. and S., 1985; Krishnamurthy and Childers, 1986). Several studies (Larar et al., 1985; Riegelsberger and Krishnamurthy, 1993; Yegnanarayana and

Veldhuis, 1998) have shown it to be crucial to use reliable estimates of glottal opening and glottal closure instants, which have to be estimated *à priori* by utilizing, for instance, EGG signals. Later, Plumpe *et al.*(Plumpe et al., 1999) extended the argumentation for closed-phase linear prediction by pointing out that during the closed phase also the non-linear source-filter interaction is most negligible. The linearity assumption underlying the source-filter model are maximized during this time segment (Moore and Torres, 2008).

3.3.6 Cepstral Analysis

Cepstral analysis was first introduced by (Bogert et al., 1963) and in parallel, Oppenheim developed his homomorphic system's theory comprising the complex cepstrum (Oppenheim, 1965). Cepstral features are the features of choice in many speaker and also speech recognition systems because they form a very compact representation of the spectral filter envelope (Campbell, 1997; Baker et al., 2009). In addition, an accurate statistical distribution of the VTF features is given by their means and variances alone, not requiring their covariances.

Most commonly, the cepstrum is defined as the inverse DFT (DFT^{-1}) of the logarithmic magnitude of the DFT of a time-domain signal:

$$c(n) = DFT^{-1}\Big\{log\big(|DFT\{x(n)\}|\big)\Big\}. \tag{3.20}$$

Since the DFT is defined over a limited number of samples, N, we can write for a frame of speech windowed by an arbitrary windowing function, $w(n)$:

$$c(n) = \frac{1}{N}\sum_{n=0}^{N-1} log\left(\left|\sum_{n=0}^{N-1} w(n)\cdot x(n)e^{-j\frac{2\pi}{N}kn}\right|\right)e^{j\frac{2\pi}{N}kn}. \tag{3.21}$$

Under certain conditions, this has very practical implications for a time-domain signal. Consider the speech signal $s(n)$ generated by convolution of an excitation $(g(n))$ with a VTF impulse response, $a(n)$:

$$s(n) = g(n)\otimes a(n). \tag{3.22}$$

Using Eq. 3.20, we can rewrite 3.22:

$$\begin{aligned} c_s(n) &= DFT^{-1}\Big\{log\big(|S(j\omega)|\big)\Big\} \\ c_s(n) &= DFT^{-1}\Big\{log\big(|G(j\omega)|\big)\Big\} + DFT^{-1}\Big\{log\big(|A(j\omega)|\big)\Big\} \\ c_s(n) &= c_g(n) + c_a(n) \end{aligned} \tag{3.23}$$

Consequently, in the cepstral domain addition is equivalent to the convolution operator $\otimes(\cdot)$ in the time-domain. Referring back to the example presented in Fig. 2.16(i), computing the cepstrum yields a signal displayed in Fig. 3.4(a), with energy peaks at integer multiples of the fundamental frequency $f_0 = 100\,\text{Hz}$ and the VT contribution encoded in the lower samples, up

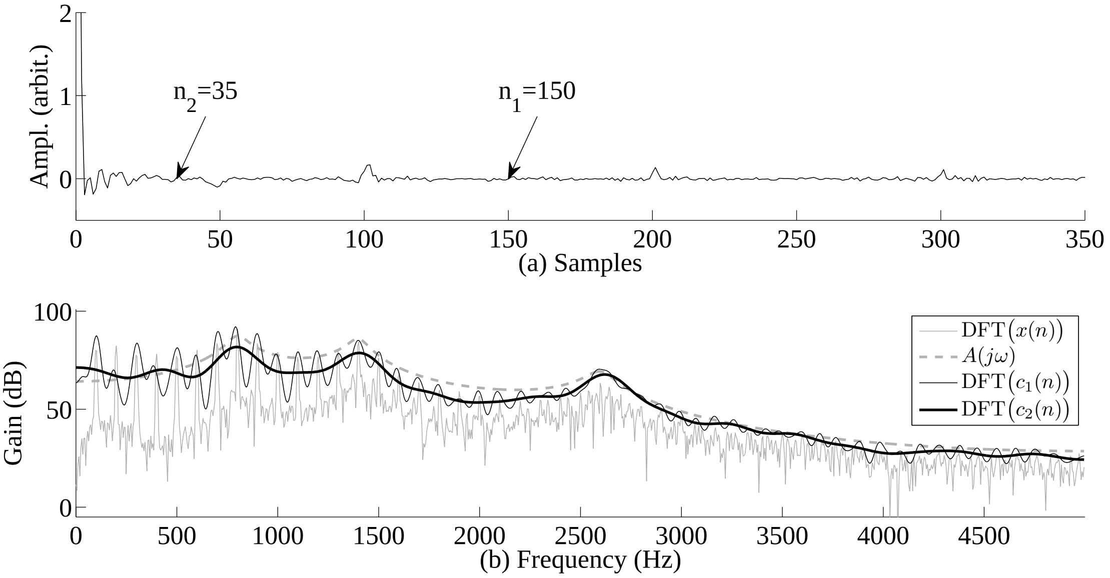


Figure 3.4: Cepstral Smoothing.

to approximately sample 50. Two observations are worth noting.

Zeroing bins above a certain sample (*e.g.* $n_1 = 150$ or $n_2 = 35$) and re-application of the DFT yields a smoothed magnitude spectrum, as displayed in Fig. 3.4(b). The critical aspect is to choose the cutoff bin so as to preserve sufficient VT envelope information but to also discard the harmonic content due to f_0. Clearly, the choice of $n_1 = 150$ is insufficient in this example, while the Fourier spectrum obtained with $n_2 = 35$ more closely resembles the true VTF, $A(j\omega)$. An optimal choice for cepstral analysis $n^* = 0.5 \cdot f_s/f_0$ was given in (Roebel et al., 2007), based on the observation that the Nyquist frequency, $f_N = f_s/2$, is a proper indicator for model order selection.

The second observation concerns the value of f_0. Higher values of f_0 will lead to a denser spacing of the cepstral peaks related to f_0. Depending on the VTF configuration, this may lead to an overlap of these harmonic peaks in the descrete cepstrum with the cepstral coefficients of the vocal tract. In those cases, a clear separation of source and VTF is not possible anymore. This clearly represents a limitation of this method for source-filter separation. Furthermore, there is no intuitive interpretation of the cepstral bins with respect to physiological parameters such as formants. This makes it difficult, at least for a human, to relate cepstral results to perceptual observations from real speech. Machine learning tasks though have successfully exploited cepstral analysis in many areas such as speech or speaker recognition.

The cepstrum has been used for VTF estimation in a method called True-Envelope-Linear-Prediction (Villavicencio et al., 2006). The idea is to use cepstral smoothing to mitigate the impact of the harmonic structure of the source and to obtain higher quality formant estimates.

3.3.7 Joint Estimation Approaches using Parametric Glottal Models

The methods presented so far assumed the source component to be time-invariant for at least an analysis window of several glottal cycles or tried to estimate the VTF during very short speech segments where no source component was assumed to be present. In 1986, Milenkovic presented the first *joint* source-filter optimization (SFO) method (Milenkovic, 1986) that attempted to jointly optimize an all-pole VTF model and a parametric linear source model. The idea is to utilize all speech samples of a glottal cycle instead of restricting the analysis to the short closed phase. Therefore, the deterministic part of the voice source is explicitly modelled using a parametric glottal source model (see Section 2.3.1). This allows a separation of the volatile source and the slowly varying VTF coefficients through the help of two dedicated models. The idea was soon after utilized by other researchers (Fujisaki and Ljungqvist, 1987; Isaksson and Millnert, 1989) and found increasing interest in the following decades. In (Ding et al., 1997; Kasuya et al., 1999) used the Rosenberg-Klatt model (Klatt and Klatt, 1990) described in Section 2.3.1 in their joint estimation methods. In (Shapira and Gath, 1998) a method was presented based on the fuzzy clustering of hyperplanes to estimate the source signal. Södersten *et al.* presented a well-documented overview (Södersten et al., 1999) and assessment of early joint source-filter optimization methods in reference to signals obtained using Rothenberg's mask.

Methods presented later on attempted to use more sophisticated glottal source models in order to improve the accuracy of the estimated parameters. Fröhlich *et al.* were the first to utilize the multi-parametric LF source model (Fant et al., 1985) and a DAP model for the VTF in their method (Fröhlich et al., 2001). The procedure for finding the optimal model parameters is based on multi-dimensional, iterative optimization. Lu (Lu, 2002) presented a convex optimization approach for optimizing a single parameter variant of the LF model for singing voice synthesis. Fu *et al.* presented a method (Fu and Murphy, 2003, 2004, 2006) comprising a two stage optimization, where the initial parameters for a second, more complex stage are found in a primer convex optimization using a simplified glottal model. The second stage then uses a more complex glottal model and the initial parameters are obtained from the simpler model estimated in the first step. Effectively, this mitigates the starting-point problem of single-point optimization methods. An entirely different model for the speech production process called the zeros of z-transform (ZZT) was presented in a series of publications (Bozkurt et al., 2005, 2007; Sturmel et al., 2007; Drugman et al., 2009). ZZT assumes a polynomial representation of the two models and categorizes the source and filter by the locations of the roots of these polynomials with respect to the unit circle. Jinachitra presented an iterative joint estimation approach (Jinachitra, 2007) of the glottal source and vocal tract parameters using Kalman filtering and expectation-maximization algorithm. Recently, in (Ghosh and Narayanan, 2011), the LF model was optimized using an exhaustive combinatorial search over the entire parameter space consisting of both the glottal parameters and the VT parameters. Degottex *et al.* (Degottex et al., 2011) presented a novel method that minimizes the error in the phase spectrum using a single parameter voice model.

Furthermore, we would like to point to several detailed discussions of other existing approaches, in particular (Holmes, 1976; Boves and Cranen, 1982; Cranen and Boves, 1988; Thomson, 1992; Childers and Ahn, 1995; Lu and Smith, 1999; Arroabarren and Carlosena, 2003; Shiga and King, 2003; Moore and Clements, 2004; Deng et al., 2006; Schnell and Lacroix, 2007; Dalsgaard et al., 2008; Gudnason et al., 2009; Perez and Bonafonte, 2009) and (Alku, 2011).

All these methods have in common that they employ a dedicated model to capture the glottal source contribution. A natural line of division between the different SFO methods is the complexity of the employed source models, *i.e.* the number of free source model parameters, which determine the coverage of the space of real voice waveforms. In general, the challenge lies in finding an efficient and reliable optimization method to estimate a non-trivial source model plus the VTF. The presented SFO approaches typically represent a compromise between the complexity of the voice model and the efficiency of the optimization method employed. Voice models with fewer parameters are easier to optimize, but fail to accurately describe voice types observed in real speech. On the other hand, using multi-parameter source models usually prohibits the usage of classical gradient-based optimization methods due to the non-convex nature of the error surface. Instead, computationally demanding methods such as exhaustive combinatorial search of the parameter space were used (Ghosh and Narayanan, 2011). In addition, source models with a higher degree of complexity require the formulation of constraints on the source and the filter model in order to prohibit a mutual inter-dependecy between the two models.

In Chapter 4, we propose a novel joint SFO approach, in which the source is modelled using the multi-parametric LF model. A global, population-based, stochastic direct search method called *differential evolution* (DE) is used to optimize the source and the filter parameters. Before we present the proposed SFO methods, we will give a short overview of the DE methods and provide the motivation for why it was chosen over other methods.

3.4 Global Optimization using Differential Evolution

DE was first introduced in (Storn and Price, 1995) and quickly gained large popularity in many engineering applications (Das and Suganthan, 2011). DE is a generic, population-based, meta-heuristic, global optimization method belonging to the family of evolutionary algorithms (EA). Other examples of EAs are the widely known genetic algorithms (GA) (Holland, 1962; Goldberg, 1989) and evolution strategies (ES) (Rechenberg, 1971; Schwefel, 1994). EAs iteratively explore the parameter space by using a *population* of NP candidate solutions called *parameter vectors* or *agents*. Each agent is a D-dimensional, concrete instantiation of a complete parameter set. At each iteration or generation m, we denote the ith population member as

$$\mathbf{x}_{i,m} = \left[x_{1,i,m} \dots x_{D,i,m}\right], \quad 1 \le i \le NP \text{ and } 1 \le m \le I_{max} \tag{3.24}$$

where D is the vector's dimension, *i.e.* the number of free parameters of the problem statement and I_{max} determines the maximum number of iteration cycles of the optimization process. A cost function $J(\cdot)$ provides a criterion to determine the *fitness* of each agent. Instead of propagating each agent in isolation to an optimal value, the candidate solutions converge to an optimum in an iterative process, based on a heuristic that weighs, combines and mutates the candidate solutions. The pseudo-code in Algorithm 1 illustrates a general template of an EA algorithm. In the following we describe these individual steps and how they are carried out in the case of DE.

Algorithm 1 Evolutionary Algorithm Structure

```
initialize population
while No determination criterion met do
    for each population member do
        Mutate various parents to create a child
        Recombine parents and child's genes
    end for
    Select parents for next generation
end while
```

Initialization: An initial *generation* of parental agents is populated with random values, chosen from a range determined by the *initial bounds*. In order for DE to work, this initial population should be distributed throughout the parameter space. Most important is that the distances are sufficiently broad, since the heuristic to traverse the parameter space in the subsequent iterations is based on the differences between the particular parent generation agents. The kind of *probability distribution function* (PDF) used to seed the initial population appears to actually not affect the convergence or success rate. Experiments (Price et al., 2005) with different distributions such as Halton (Halton and Weller, 1964), guaranteeing a minimum distance between random draws, did not affect the final result of th optimization compared to other initial PDFs such as a regular, uniformly distributed random distribution.

Mutation: At the beginning of each iteration, a second, intermediary population of NP *mutant agents*, $\mathbf{V}_m$, is created. Various agents from the parental population are combined in a heuristic manner to create $\mathbf{V}_m$. In the case of DE, a set of vectors is chosen from the parental population and mutation is performed by adding one or more weighted vector differences to a base vector. For instance, the mth intermediary population member of the ith generation is created by

$$\mathbf{V}_{i,m} = \mathbf{X}_{i,r_1} + F \cdot \left(\mathbf{X}_{i,r_2} - \mathbf{X}_{i,r_3}\right), \quad \text{with } m \neq r_1 \neq r_2 \neq r_3 \tag{3.25}$$

where r_1, r_2 and r_3 usually are disjoint random indices drawn from the range $(1, NP)$. It is ensured that during each iteration no index is used more than once for each of the indices. The scaling factor, $F \in (0, 1+)$, is a positive, real-valued number that controls the rate at which the population evolves. Typically, F is smaller than one and for increased robustness it may also adhere to a random distribution, where a slightly different value is drawn for either each parameter (*jitter*) or just for each vector (*dither*). As such, dithering merely changes the length of the resulting difference vector, whereas jitter also changes the vector's orientation. Transforming F into a random variable effectively prevents *stagnation* and makes it possible to construct a limited convergence proof (Zaharie, 2002).

The base vector, r_1, can be chosen in a variety of ways. In GAs, better vectors are more likely to be chosen for recombination (Holland, 1975), inducing a bias into the selection scheme. In a similar way, some versions of DE select the base vector r_1 based on the objective function value, where for instance only the best-so-far vector is chosen as the base vector for the mutation (Storn, 1996). However, experiments have shown that for DE it is usually preferable to select the base vector randomly, since this increases the probability of success, although at the cost of a slower convergence. Selecting only one vector as the base vector for all mutations of a particular iteration creates a large selection pressure. Note that the random base vector selection scheme nevertheless represents an elitist scheme, because the current best-so-far vector can only be replaced by a better vector (see selection). Several alternatives exist that represent a compromise between the two extremes. In (Price, 1997), the base vectors are randomly selected from a reduced set of the best parent vectors. In (Storn, 1996), an arithmetic combination between the best vector and a randomly selected vector is chosen. Furthermore, other methods for compensating the lost variety in the *best-so-far* scheme attempt to reduce the selection pressure by using a combination of two difference vectors (Storn, 1996; Price, 1996).

Recombination: DE performs *uniform crossover* to produce *trial vectors*, $\mathbf{U}_{i,m}$, by crossing parameters from a mutant vector, $\mathbf{V}_{i,m}$, and the respective, mth, parent vector:

$$\mathbf{U}_{i,m} = u_{j,i,m} = \begin{cases} v_{j,i,m}, & \text{if} \left(\text{rand}_j(0,1) \leq CR \quad \text{or} \quad j = j_{\text{rand}}\right) \\ x_{j,i,m}, & \text{otherwise.} \end{cases} \tag{3.26}$$

The crossover probability, $CR \in [0,1]$, is a real-valued, user-defined value, effectively determining the fraction of parameters that are copied from the mutant vector in order to build the *trial* vector for the selection step. The crossover rate has been shown to play an important role in the presence of mutal dependency between different parameters (see below). The crossover probability CR is usually uniformly distributed. In the case that the random crossover selection happens to choose no parameter to be copied, an exception is made by copying at one single parameter from the mutant vector to guarantee a difference between the mutant vector and parental vector.

Selection: Eventually, each trial vector $\mathbf{U}_{i,m}$ competes with the respective mth parental vector $\mathbf{X}_{i,m}$ from which it inherited parameters during recombination. The vector with the lower cost function value, J, gains a place in the parent population of the next generation,

$\mathbf{X}_{i+1}$:

$$\mathbf{X}_{i+1,m} = \begin{cases} \mathbf{U}_{i,m}, & \text{if } J\big(\mathbf{U}_{i,m}\big) \leq J\big(\mathbf{X}_{i,m}\big) \\ \mathbf{X}_{i,m}, & \text{otherwise.} \end{cases} \tag{3.27}$$

Termination: A termination criterion is used to stop the optimization. Typically, the number of iterations is limited to a maximum, I_{max}, that is deemed sufficient for successfully finding an optimal solution. Alternatively, the optimization process may also be terminated if a previously specified, lowest cost function value is reached or if the best-so-far cost function value has not been changing for a fixed number of iterations.

Algorithm 2 Glottal cycle optimization

```
Step 1: Set control parameters crossover rate CR, difference scale factor F, population size NP
and max. number of iterations, I_max.
Step 2: Initialize the agents $\mathbf{X}_{i,m}$ of the population number $i = 0$ with random values and
subject to the constraints, where $m = [1,2,\dots,\mathrm{NP}]$, $\mathbf{X}_{i,m} = [x_{1,i,m},\dots,x_{D,i,m}]$ and $D$ is the
dimension of the parameter vector.
Step 3:
while $i \leq$ I_max do
  for $m = 1$ to NP do
    Step 3.1: Mutation step
      Create a donor vector:
      $\mathbf{V}_{i,m} = \mathbf{X}_{i,r_1^m} + \mathrm{F} \cdot (\mathbf{X}_{i,r_2^m} - \mathbf{X}_{i,r_3^m})$
      using disjoint random indices $r_1$, $r_2$ and $r_3$
    Step 3.2: Crossover step
      Create a trial vector $\mathbf{U}_{i,m} = [u_{1,i,m},\dots,u_{D,i,m}]$:
        $u_{j,i,m} = \begin{cases} v_{j,i,m}, & \text{if rand[0,1]} \leq \mathrm{CR} \text{ and } j = j_{\mathrm{rand}} \\ x_{j,i,m}, & \text{otherwise,} \end{cases}$
      or reinitizialize $u_{j,i,m}$ if constraints are breached.
    Step 3.3: Selection step
      Evaluate performance and select next generation
      member $\mathbf{X}_{i+1,m}$:
        $\mathbf{X}_{i+1,m} = \begin{cases} \mathbf{U}_{i,m}, & \text{if } J(\mathbf{U}_{i,m}) \leq J(\mathbf{X}_{i,m}) \\ \mathbf{X}_{i,m}, & \text{otherwise.} \end{cases}$
  end for
  Step 3.4: Increase the generation count $i = i + 1$
end while
```

DE stands out from other EA algorithms in several aspects. DE is rather simple and straightforward to implement, yet its performance has been shown to be largely better (Kennedy and Eberhart, 1995; Das and Suganthan, 2011) than that of the (also popular) *particle swarm optimization* (PSO) and its variants over a wide variety of problems (Vesterstrøm and Thomson, 2004). Mostly though, in the context of the SFO approach presented in this work, DE may be recognized for two properties; *contour matching* and its performance in the presence of *parameter dependencies*.

Contour matching refers to the automatic adaptation of the step size and also the step orientation to the error function landscape. We illustrate contour matching with the help of an example, the $2D$ *peaks* function, also obtainable through a Matlab (The Mathworks, 2006) command:

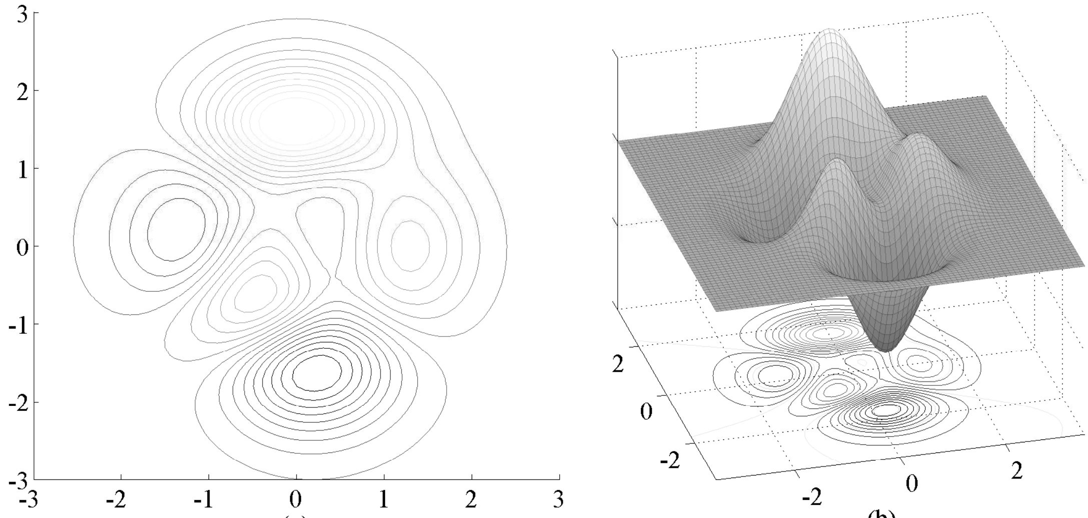
$$f(x_1, x_2) = 3 \cdot (1 - x_1)^2 \cdot e^{x_1^2 + (x_2+1)^2} - 10 \cdot \left(\frac{x_1}{5} - x_1^3 - x_2^5 \right) \cdot e^{x_1^2 + x_2^2} - \frac{1}{3} \cdot e^{(x_1+1)^2 + x_2^2}. \tag{3.28}$$

Fig. 3.5(a) and (b) displays a level contour and a 3D plot of the corresponding error surface. The series of Fig. 3.6 to 3.10 helps to visualize the progress of convergence to a global minimum, despite the presence of a local minimum. In the respective panels *(a)*, the members of the current parental population are drawn, whereas the panels *(b)* illustrate the possible pool of vector differences available for the mutation step. Note that for clarity only the endpoints of the difference vectors are drawn.

Due to the initial wide spread of the agents throughout the error landscape, necessarily also the resulting vector differences exhibit a large range in their magnitude and orientation (Fig. 3.6). With each iteration though, only the best solutions survive, such that the agents coalesce first to valleys in the error function landscape (Fig. 3.7 and 3.8). At that moment, the vector differences comprise local and remote members, the latter pointing from one valley to another. Note that these remote vector differences also help to explore other valleys and even remote, yet unexplored regions. Eventually, the solutions will converge to a valley containing the global minimum (Fig. 3.9) and at that moment the vector difference distribution becomes uni-modal and inherently spans only short distances. The smaller scale of the vector differences makes the population well suited for the local search around the global minimum and increases the probability of finding an optimal value (Fig. 3.10).

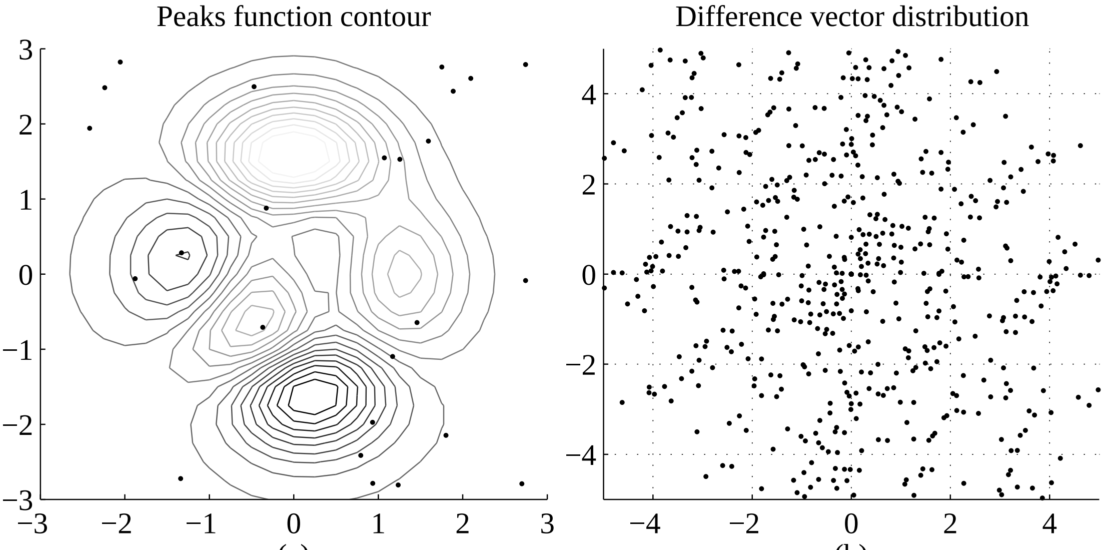
In this scheme, the role of the scaling factor, F, becomes more clear. It acts as a relative step size factor. The actual, current step size during a particular iteration is always determined by the underlying error function surface. For this automatic adaptation of the step size and step orientation, no parameters of the DE method need to be updated. Therefore, a predetermined probability distribution for mutation, often introducing a bias, is not required. Contour matching inherently promotes *basin-to-basin* transfers in the beginning of the optimization process, where search points may move from one basin of attraction, i.e., a local minimum, to another one. Later on, the search space becomes increasingly local, which helps to find an accurate solution in the valley of the global optimum. Contour matching therefore considerably reduces both the starting-point problem of single-point optimizers and the probability of premature convergence to a local minimum.

Another interesting aspect of DE with respect to the problem of source-filter separation is its performance in the presence of dependent parameters. Vincent (Vincent, 2007)



(a) (b)

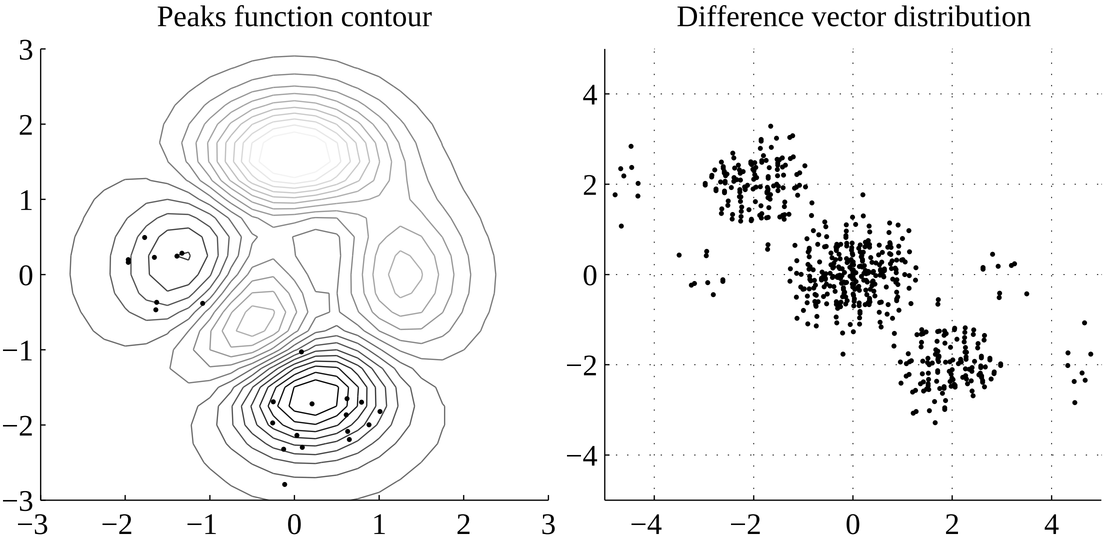
Figure 3.5: Contour (a) and three-dimensional illustration (b) of the peaks function.



(a) (b)

Figure 3.6: *(a)* Initial, uniformly distributed population and *(b)* the respective difference vector distribution.

has shown that the parameters of the LF model are not entirely independent and several solutions describing similar or nearly identical source waveforms may exist. As described in Section 4.1, the resulting error surface is not convex, but may exhibit local minima. This was one of the reasons why in many of the studies presented in Sec. 3.3.7 simple, single-parameter source models were used for the joint SFO. In (Price et al., 2005) and (J. et al., 2005), it was demonstrated that choosing a high crossover rate CR in the range (0.9,1) for DE is a successful strategy for tackling the problem of parameter dependency. A large CR value ensures that

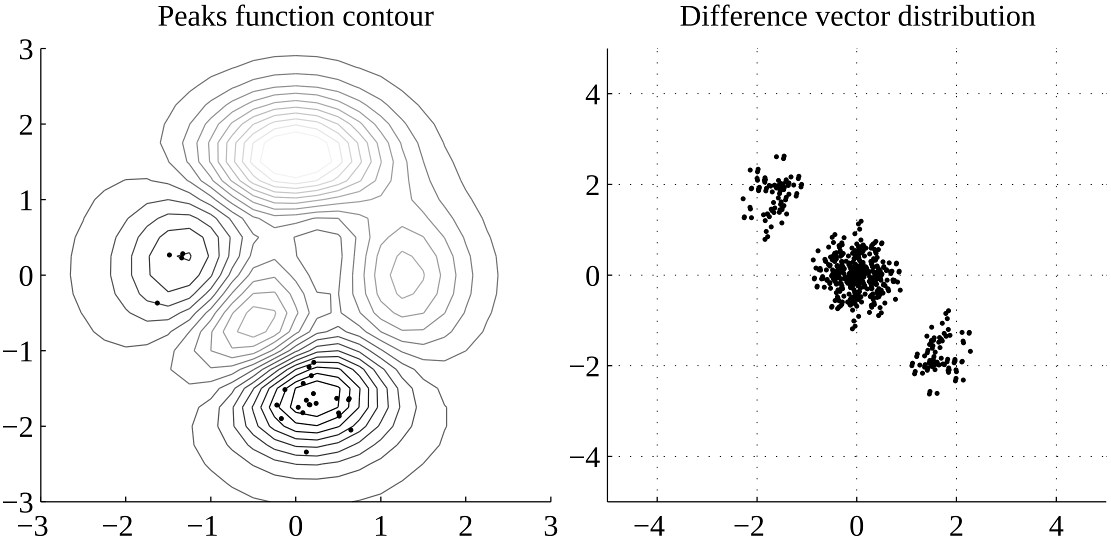

Peaks function contour

Difference vector distribution

(a)

(b)

Figure 3.7: *(a)* Population and *(b)* the respective difference vector distribution after $i = 12$ iterations. Dispersed clusters of far- and near vector differences are observed.


Peaks function contour

Difference vector distribution

(a)

(b)

Figure 3.8: *(a)* Population and *(b)* the respective difference vector distribution after $i = 18$ iterations. The vector differences start to coalesce and form a multimodal distribution.

the parameter space is propagated not only in parallel to the parameter axes. Thereby, the likelihood to get trapped in local minima is reduced.

In summary, DE has only a few control parameters, namely the crossover rate CR, the difference weight F and the population size NP, which makes its application straightforward and easy. Furthermore, its algorithmic nature qualifies DE to benefit well from the current massive trend in hardware development towards parallel computing environments (Sutter,

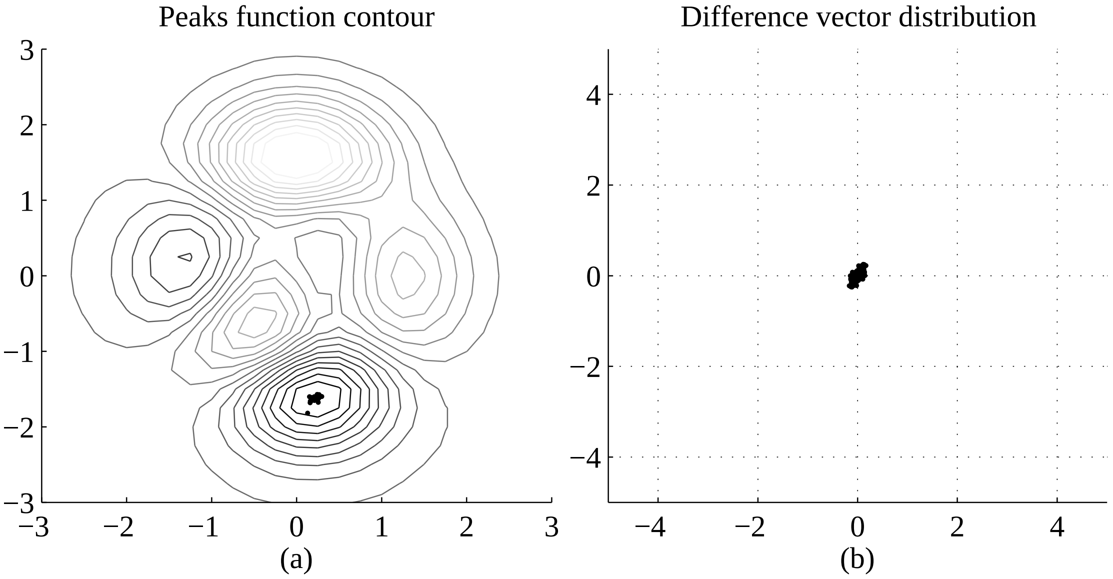


Figure 3.9: *(a)* Population and *(b)* the respective difference vector distribution after $i = 30$ iterations. The vector differences have coalesced to the valley of the global minimum and form a uni-modal distribution.

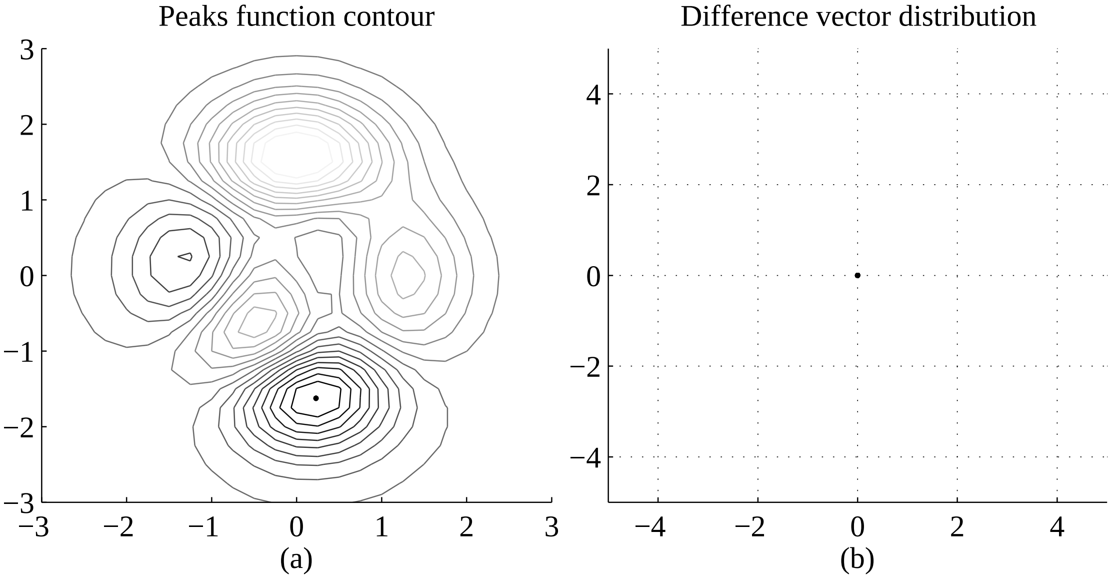


Figure 3.10: *(a)* Population and *(b)* the respective difference vector distribution after $i = 60$ iterations. The optimim has been found (subject to a minor, remaining error).

2012). A summary of the glottal cycle optimization procedure is given in Algorithm 2.

3.5 Conclusion

- In a general context of system identification, the source-filter separation is an ill-posed inverse problem aiming to obtain an estimate of the unknown VTF, which is excited by the unknown glottal source (see Sec. 3.1).
- The glottal source has several characteristics that make the estimation of the VTF very difficult; it exhibits a large volatility, it exhibits harmonic spectral peaks, its power spectrum is non-uniform and time-variant and it sports a time-varying spectral tilt (see Sec. 3.1 and 3.2.4).
- A common and wide-spread analysis tool for source-filter separation is linear prediction. We review commonly used methods, conditions of stability of the resulting filter coefficients and illustrate some of the problems typically encountered with LP in the context of source-filter separation (see Sec. 3.2).
- A review of the state of the art of source, filter and joint source-filter estimation methods is given in Sec. 3.3. Currently known SFO approaches typically represent a compromise between the complexity of the voice model and the efficiency of the optimization method employed. New, computationally efficient methods for the estimation of more descriptive models of the glottal source are needed.
- Differential evolution appears to be a promising computational tool for SFO (see Section 3.4). It has been proven to be a fast and reliably converging optimization technique in many applications. DE has been shown to be a robust tool also in the presence of parameter dependencies and non-convex error surfaces. The efficiency of the DE method allowed us to carry out extensive experiments on different speech signals.

Proposed Methods Part II

4 Source-Filter Separation using Differential Evolution

In this chapter, we propose a novel joint SFO approach, in which the source is modelled using the multi-parametric LF model described in Section 2.3.1. The proposed method is based on a pitch-synchronous analysis-by-synthesis approach. In contrast to traditional analysis-by-synthesis methods, we do not use a codebook to generate reference speech signal patterns. Instead, a time-varying auto-regressive VT model with exogenous input (ARX) (Ljung, 1999) is used to generate candidate solutions. *Differential evolution* serves as a computational tool to optimize the source and the filter parameters. The objective function is constructed so as to reduce the effect of inter-glottal-cycle resonances to increase the effective duration of the analysis window. The efficiency of the DE method allows us to carry out extensive experiments on different speech signals. The proposed optimization method converges reliably under a variety of modifications such as environmental and glottal noise, varying fundamental frequencies, jitter and vowel transitions.

4.1 Formulation of Optimization Problem

The speech production model used in this method is based on the model presented in Sec. 2.3. Instead of carrying out a frame-based analysis though, we obtain new model estimates pitch-synchronously so as to capture the inter-glottal-cycle variations of the glottal source. The speech production model from Eq. (2.1) is therefore modified. The harmonic structure expressed in $H(j\omega)$ is discarded, since the analysis window comprises only a single glottal cycle. The speech signal originating from one particular glottal cycle, k, is then written as

$$S_k(j\omega) = e^{-j\omega t_{o,k}} G_k(j\omega) \cdot A_k(j\omega) \cdot L_k(j\omega), \tag{4.1}$$

where the temporal location of the glottal cycle is determined by the linear-phase component $e^{j\omega t_{o,k}}$, which merely induces a delay of $t_{o,k}$ seconds with respect to an arbitrary time reference.

Crucially, the finite (windowed) glottal source represented by $G_k(j\omega)$ is mixed-phase; it has both zeros with a magnitude greater and smaller than unity. This implies that no stable inverse representation, $1/G_k(j\omega)$, exists and a direct deconvolution for obtaining the vocal tract transfer function $A_k(j\omega)$ from $S_k(j\omega)$, as implied in Eq. 3.2, is impossible. Therefore, the glottal source model and vocal tract coefficients are jointly estimated using a global optimization technique in an analysis-by-synthesis framework.

As mentioned above, speech production is modelled by a linear, time-varying, auto-regressive (AR) model with exogenous input (ARX). The exogenous input is provided by the glottal source signal of cycle k

$$\dot{g}_k(n) = \dot{g}(n) \otimes \mathrm{sinc}(n - t_{o,k}), \tag{4.2}$$

where $\dot{g}(n)$ refers to the LF model defined in Eq. 2.2, sinc represents the cardinal sine function, $\mathrm{sinc}(\cdot) = \sin(\pi\cdot)/(\pi\cdot)$, and $\otimes$ stands for convolution. Note that using the cardinal sine function instead of a dirac pulse enables $t_{o,k}$ to be real-valued and independent of the sampling rate. The speech signal produced during cycle k is represented by the difference equation

$$\hat{s}_k(n) = -\sum_{i=1}^{p} a_{i,k}\,\hat{s}_k(n-i) + \dot{g}_k(n). \tag{4.3}$$

The parameter n is the discrete-time index defined in the range $0 \leq n \leq N_0$, where $N_0 = \lceil f_s/f_0 \rceil$ is the number of samples used to describe one glottal period. Parameter p refers to the order of the time-varying auto-regressive filter representing the vocal tract. The coefficients $a_{i,k}$ of the ARX model are chosen to be real, therefore its complex poles always appear in complex conjugate pairs. Thus, p also corresponds to twice the number of formants and should generally be chosen to be even (see Section 2.3.3). Eq. (4.3) may be expressed in vector notation as

$$\hat{s}_k(n) = -\mathbf{a}_k^{\top}\hat{\mathbf{s}}_k^{-}(n) + \dot{g}_k(n) \tag{4.4}$$

with

$$\mathbf{a}_k = \left[a_{1,k}\ a_{2,k}\ \ldots\ a_{p,k}\right]^{\top}$$

and $\hat{\mathbf{s}}_k^{-}(n)$ representing the past p samples of $\hat{\mathbf{s}}_k$ up to and including $n-1$:

$$\hat{\mathbf{s}}_k^{-}(n) = \left[\hat{s}_k(n-1)\ \hat{s}_k(n-2)\ \ldots\ \hat{s}_k(n-p)\right]^{\top}.$$

The error, or *residual*, between the observed speech $s(n)$ and the modelled speech $\hat{s}(n)$ is defined as

$$e_k(n) = s_k(n) - \hat{s}_k(n). \tag{4.5}$$

Using the definition of the squared Euclidean error $\|\cdot\|^2 = \sum(\cdot)^2$ and by defining the parameter vector

$$\theta = \left[E_e \;\; t_p \;\; t_e \;\; t_c \;\; t_a \;\; f_1 \dots f_{p/2} \;\; b_1 \dots b_{p/2}\right], \tag{4.6}$$

the optimization problem can now be formulated as

$$\begin{aligned} \min_{\theta_k} J(\theta_k) &= \min_{\theta_k} \left\| e(n) \right\|^2 \\ &= \min_{\theta_k} \left\| s_k(n) + \mathbf{a}_k^\top \hat{\mathbf{s}}_k^-(n) - \dot{g}_k(n) \right\|^2, \end{aligned} \tag{4.7}$$

subject both to *inequality constraints* on the order of the temporal LF parameters

$$0 < t_p < t_e < t_c < T_0 \tag{4.8}$$

and *bound constraints* on the temporal LF parameters, formant frequencies f_m and bandwidths b_m. The VT filter coefficients $\mathbf{a}_k$ are obtained by expanding the pairwise roots r_m, which are determined by the formant frequencies and formant bandwidths contained in θ_k through the following relationsips: $\angle(r_m) = \pm 2\pi f_m / f_s$ and $|r_m| = e^{-\pi b_m / f_s}$ (Smith, 2007a).

Several remarks can be made at this point. The error function $J(\theta)$ spans a D-dimensional surface but the cost function (4.7) does not provide a closed-form solution. We also notice that the LF source model (2.2) represents a non-continuous and, thus, non-differentiable function and furthermore, we may also note that the parameters of the LF model (2.2) do not form an orthogonal basis and therefore are *not* mutually independent (Vincent, 2007). Different combinations of parameters may describe very similar or identical glottal source waveforms. As a result, the error surface defined by (2.2) and (4.7) is generally non-convex and may exhibit several local minima. Therefore, classical iterative gradient-based optimization methods are not applicable for finding a solution. The global optimization technique presented in the previous chapter, *differential evolution*, is an ideal candidate for solving this optimization problem in the presence of the aforementioned problems.

4.2 Conditions for Convergence

In any optimization task, it is necessary to ensure that the estimated set of parameters, θ, may converge to the optimal set of parameters, θ^*. In our specific problem, we would like to ensure that the parameters representing the glottal source model cannot compensate for an error in the VT model and vice versa.

Using Eq. 4.1, we can rewrite Eq. 3.1 such that the resulting glottal source is modelled using a particular set of parameters, θ, as

$$\hat{A}_k^\theta(j\omega) = \zeta \left(\frac{e^{j\omega t_o} S_k(j\omega)}{G_k^\theta(j\omega) \cdot L_k(j\omega)} \right). \tag{4.9}$$

Let us assume that the real glottal source $G_k(j\omega)$ may indeed be modelled by the glottal source model used to construct $G_k^\theta(j\omega)$ and the real vocal tract transfer function indeed may be perfectly described by the model chosen for $A_k^\theta(j\omega)$. Then, we can replace the observed speech, $S_k(j\omega)$, by its model (Eq. 4.1) parameterized with the optimal set of parameters, θ^*:

$$\begin{aligned}\hat{A}_k^\theta(j\omega) &= \zeta\left(\frac{e^{j\omega t_o} G_k^{\theta^*}(j\omega)\cdot A_k^{\theta^*}(j\omega)\cdot L_k(j\omega)}{e^{j\omega t_o} G_k^{\theta}(j\omega)\cdot L_k(j\omega)}\right)\\ &= \zeta\left(\frac{G_k^{\theta^*}(j\omega)\cdot A_k^{\theta^*}(j\omega)}{G_k^{\theta}(j\omega)}\right).\end{aligned} \tag{4.10}$$

If we ignore for the time being the position error due to t_o, we may formulate the following error due to the parameter θ:

$$\begin{aligned}E_k^\theta(j\omega) &= S_k(j\omega) - G_k^\theta(j\omega)\cdot\hat{A}_k^\theta(j\omega)\cdot L_k(j\omega)\\ &= G_k^{\theta^*}(j\omega)\cdot A_k^{\theta^*}(j\omega)\cdot L_k(j\omega) - G_k^\theta(j\omega)\cdot\hat{A}_k^\theta(j\omega)\cdot L_k(j\omega)\\ &= G_k^{\theta^*}(j\omega)\cdot A_k^{\theta^*}(j\omega)\cdot L_k(j\omega) - G_k^\theta(j\omega)\cdot\hat{A}_k^\theta(j\omega)\cdot L_k(j\omega)\\ &= L_k(j\omega)\Big(G_k^{\theta^*}(j\omega)\cdot A_k^{\theta^*}(j\omega)\ -\ G_k^\theta(j\omega)\cdot A_k^\theta(j\omega)\Big).\end{aligned} \tag{4.11}$$

In the general equation above, the glottal source $G_k(j\omega)$ and the VTF $A_k(j\omega)$ can compensate each other. It is therefore necessary to provide constraints on the two models that prevent this interference. Two mutually exclusive hypotheses are necessary for the glottal source model used to construct $G_k^\theta(j\omega)$ and for the VT model used for $A_k^\theta(j\omega)$.

4.3 Necessary Constraints on Voice and VT model

Let us first take a closer look at the characteristics of the two models we use; the autoregressive all-pole VTF model and the glottal source LF model. The VTF model was described in Section 2.3.3. It represents a passive system and it is here generally assumed that all its poles are inside the unit circle. It is a minimum-phase lag system.

The LF source model was described in Section 2.3.1. For convenience, we repeat here Eq. 2.2 for generating one glottal cycle using the synthesis parameter set, but we replace the sine function in the opening phase with its exponential equivalent, *i.e.* $\sin(\theta) = \frac{1}{2j}\cdot\left(e^{j\theta} - e^{-j\theta}\right)$:

$$\dot{g}(n) = \begin{cases} \frac{E_0}{2j} e^{\alpha n}\left(e^{j\omega_g n} - e^{-j\omega_g n}\right), & 0 \le n \le N_e\\ -\frac{E_e}{\epsilon N_a}\left[e^{-\epsilon(n-N_e)} - e^{-\epsilon(N_e-N_c)}\right], & N_e \le n \le N_c\\ 0, & N_c \le n \le N_0 - 1,\end{cases} \tag{4.12}$$

For an analysis of the LF model it is probably helpful to study the location of the roots of the z-transform of exponential functions first. The z-transform of an infinite exponential signal of the form $x_1(n) = a^n$ is a geometric series, polynomial in n (Råde and Westergren, 1997):

$$X_1^{\infty}(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \frac{1}{1-\left(\frac{a}{z}\right)}. \tag{4.13}$$

The z-transform $X_1^{\infty}(z)$ has a single pole on the real-axis at $z = a$. If the signal is truncated after N samples, then its z-transform is represented by a finite geometric series:

$$X_1(z) = \sum_{n=0}^{N-1} a^n z^{-n} = \frac{1-\left(\frac{a}{z}\right)^N}{1-\left(\frac{a}{z}\right)}. \tag{4.14}$$

The expression $X_1(z)$ has a pole at the same location as $X_1^{\infty}(z)$, but this pole is cancelled by one zero at the same location. Furthermore, the remaining roots of the nominator of Eq. 4.14 are uniformly distributed on a single circle at radius $r = a$. Therefore, the roots of Eq. 4.14 are:

$$Z_m = ae^{j2\pi m/N}, \quad m = 1,2,\ldots,N-1 \tag{4.15}$$

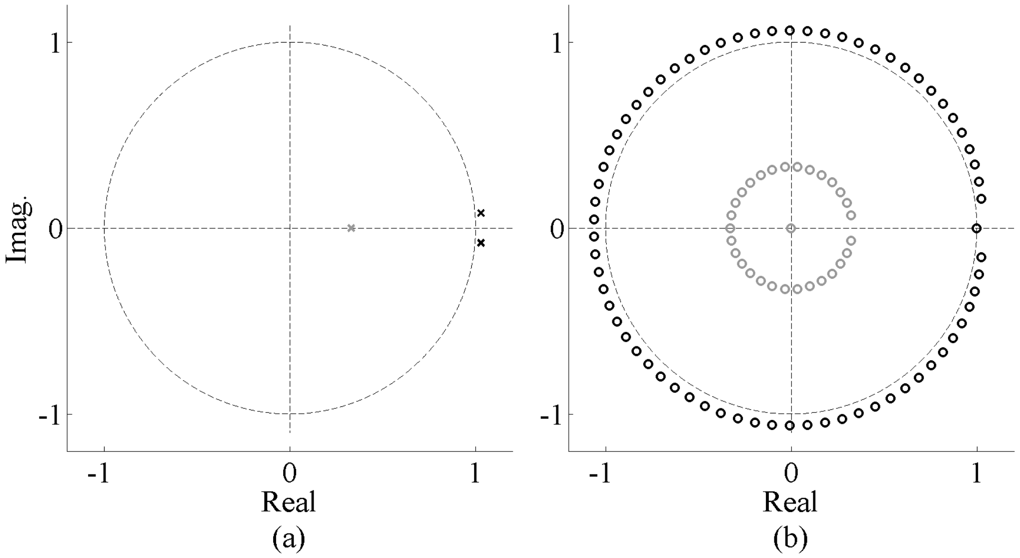
An example of a z-transform of a single, real-valued pole with $|a| < 1$ in the z-plane is illustrated using grey colour in Fig. 4.1(a) for the infinite case and in Fig. 4.1(b) for the finite case. The single pole on the real axis in Fig. 4.1(a) carries all the information describing the spectral characteristics of the underlying signal, an exponentially decaying function.

The z-transform for a signal consisting of two exponentials having a complex-valued exponent is a little different. Assuming the two exponentials have a factor that is also polynomial in n (*i.e.* $x_2(n) = e^{bn}\left(e^{j\omega n} - e^{-j\omega n}\right)$ as in the opening phase of the LF model), then the infinite z-transform is given by

$$X_2^{\infty}(z) = \sum_{n=0}^{\infty} e^{b}\left(e^{j\omega n} - e^{-j\omega n}\right) z^{-n} = \frac{1}{\left(1-\left(e^{b}e^{j\omega}z^{-1}\right)\right)\left(1-\left(e^{b}e^{-j\omega}z^{-1}\right)\right)}, \tag{4.16}$$

and exhibits two poles at angles of $\pm\omega/f_s$ rad and with a magnitude of e^b. Again, in the case of a truncated version of $x_2(n)$, the z-transform will have uniformly distributed zeros on a circle at radius $r = e^b$ with a gap where the two roots of the denominator reside. For an example, refer to the black coloured poles and zeros in Fig. 4.1(a) and (b).

Let us now return to the LF model and let us assume that speech indeed may be modelled by the LF model and the AR model of the VTF. The opening phase of a glottal cycle is determined by the first line in Eq. 4.12. It represents a maximum-phase lag component due to the exponentially growing factor $e^{\alpha n}$, where α is chosen such that $|e^{\alpha}| > 1$. The angle of the two poles is determined by ω_g. The return phase on the other hand, described by the second line in Eq. 4.12, is a minimum-phase lag component in the form of an exponentially decaying function. Its single pole is real-valued and its magnitude by definition is smaller than unity. Fig. 4.1 in fact is an example of a z-transform of a glottal cycle. In many previously published works on glottal source-filter separation, the return phase was ignored and forced



(a)

(b)

Figure 4.1: Illustration of the roots of the LF source model in the z-plane, developed as an *(a)* infinite geometric series and *(b)* a finite geometric series. In the finite case, the series was truncated after one complete glottal cycle. The three poles in *(a)* completely describe the characteristics of the LF model. The magnitude of the single pole on the real axis determines the duration of the return phase. The angles of the two conjugate complex poles determine the frequency of the sinusoidal component of the opening phase and thereby influence the respective timing parameters of the LF model (t_i, t_p and t_e). Eventually, the magnitude of these two conjugate complex poles determines the growth rate of the exponential component of the opening phase. Convergence of the joint source-filter optimization process can be achieved by forcing all the conjugate complex poles of the VTF to be non-real and having a magnitude $|z| < 1$.

to zero (*e.g.* (Lu, 2002)). This simplified the optimization process and a division between the VTF model and the source model could be made based on the phase characteristics of the two models. In other words, in methods using such a simple model it is generally assumed that the VTF is represented by a strictly minimum-phase lag, dampening system, whereas the source model has strictly maximum-phase lag characteristics and injects energy into the system. A compensation between the two models is hindered by the strict separation of the magnitudes of the roots of the two models and the requirement to build a causal system. This simplified model also has the advantage that the resulting error surface is convex and therefore only exhibits one global minimum. However, in this work we are interested in using a source model with a higher number of degrees of freedom in that the model also describes the return phase of the glottal cycle. The return phase is not set to zero, but represented by the decaying exponential in the second line of Eq. 4.12.

Naturally, the return phase is minimum-phase lag, and thereby collides with the hypothesis of the VTF. The question remains how can we incorporate the minimum-phase lag glottal return phase while guaranteeing that it will not compensate for errors in the minimum-phase lag VTF model and vice versa. By observing that the return phase parameter is real-valued, we may formulate the requirement for the VTF model to have all its poles representing the VT formants at angles greater than zero and with a magnitude of less than unity. Thereby we obtain a guarantee that the two models may not compensate an error in the respective other model. We may summarize the conditions for mutually exclusive source and filter models as follows:

Source model: All roots of the opening phase of the source model must occur in conjugate complex pairs and must strictly lie outside the unit circle. In addition, all roots determining the decay rate of the return phase must be real-valued and implicitly have a magnitude smaller than unity.

Filter model: All roots of the VTF model must occur in complex conjugate pairs, complex valued and their magnitudes must strictly be smaller than unity.

4.4 Implementation Details

The actual implementation of the routine to compute the cost function was implemented in C++ in order to reduce the time to compute the results. This routine was then embedded in a Matlab implementation of the differential evolution algorithm. To give the reader an orientation, the optimization of one glottal cycle on a commercial PC platform takes about 400 to 600 times real-time, depending on the actual parameterization. This main part of this duration is largely due to the time to convert the LF model parameters from the time-domain representation into the synthesis representation (see Sec. 2.3.1). For expressing the optimization problem, the time-domain parameters are preferred since it is easier to express the boundary conditions in this way.

4.4.1 Optimization Parameters

The aim is to find the optimal set of model parameters θ_k^* that minimizes (4.7) by using an analysis-by-synthesis approach. The speech signal is first segmented into analysis frames $s_k(n)$, the length of which approximately correspond to the period between successive glottal opening instants (t_o in Fig. 2.8). It is assumed that the fundamental frequency and the approximate location of each glottal cycle are known *a priori*. Numerous methods exist that may assist in finding these values, e.g. (de Cheveigné and Kawahara, 2002; Camacho, 2007; Thomas et al., 2012; Kounoudes et al., 2002)). The exact values of t_o are to be found during the optimization.

In the next step, an initial population $i = 0$ of M candidate solutions $\theta_{i=0}^M$ is populated with random values. The temporal LF model parameters in θ adhere to the inequality constraints

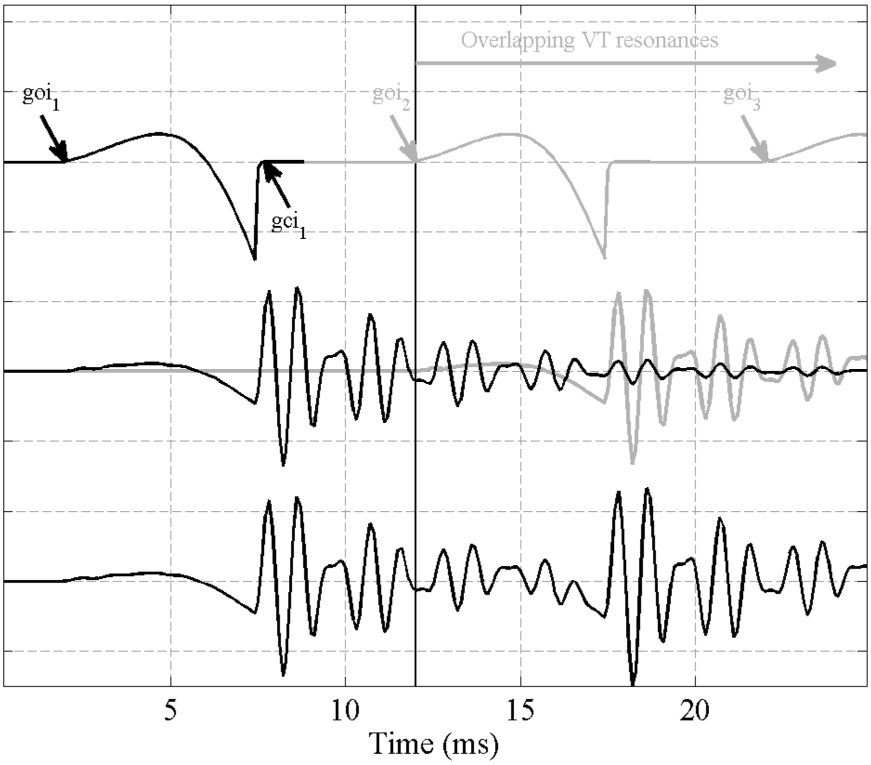


Figure 4.2: Synthetic modal glottal excitations (upper graph) and their respective (middle graph) and joint (bottom graph) vocal tract resonances for a vowel */a:/*. The decaying VT resonances of the first glottal excitation (black solid line), depicted in the middle graph, clearly overlap with the subsequent glottal excitation, resulting in the commonly observed speech waveform shown in the bottom graph.

defined in Eq. 4.8. The boundary constraints for the parameters are listed in Table 4.1 and Table 4.2. The values for the formant frequencies were derived from (Childers, 2000) and the values for the formant bandwidths were taken from (Fant, 1962). The termination criterion used was a maximum number of iterations of I_max = 600. The DE parameter values used for the joint SFO in this work were determined by empirical observations and set to CR = 0.9, F = 0.3 and NP = 120. The optimization for a particular glottal cycle k starts by calculating the cost of each member m of the initial population. The parameter set $\theta^m_{k,i}$ is used to synthesize $\hat{s}^m_k(n)$ as defined in (4.3).

Table 4.1: Boundary constraints of the center frequencies (Freq) and bandwidths (BW) of formants F1 to F3 in Hz.

Boundary	**F1**	**F2**	**F3**
Freq$_{low}$	450	1200	2500
Freq$_{up}$	860	2400	3100
BW$_{low}$	30	30	50
BW$_{up}$	70	80	200

Table 4.2: Boundary constraints of the LF parameters.

Boundary	t_o	t_p	t_e	t_a
lower	0.0	0.0	0.0	0.15
upper	10.0	60.0	90.0	10

4.4.2 Compensation of Overlapping Resonances

Note that the vocal tract, represented by the first term on the right hand side of Eq. 4.3, is an auto-regressive structure. Vector $\mathbf{a}_k$ contains the coefficients of a recursive all-pole filter using its past output as its input. Depending on the bandwidths of the formants, the decay times of this filter are often found to be considerably longer than the intervals between successive glottal cycles. Therefore, it is often found that the decaying resonances of previous glottal cycles are not yet negligible at the instant of the beginning of the next glottal cycle. This can especially be observed in female voice sources with higher average f_0-values. See Fig. 4.2 for an illustration of the overlapping of the resonances across subsequent glottal cycles.

In the formulation of the optimization process in Eq. 4.7 we have not yet considered the effect of this overlapping. Therefore we devise a method that helps to decrease the influence of the resonances of previous cycles. First $\hat{s}^*_{k-1}(n+l)$ is defined to be the synthetic speech generated by the optimal parameter set θ^*_{k-1} found for glottal cycle $k-1$. Here, l corresponds to the number of samples between the beginnings of cycle $k-1$ and cycle k. $\hat{s}^*_{k-1}(n+l)$ is then subtracted from $s_k(n)$ before the optimization of glottal cycle k starts. Eq. 4.7 thus is rewritten as

$$\begin{aligned}\min_{\theta_k} J(\theta_k) &= \min_{\theta_k} \left\| e'(n) \right\|^2 \\ &= \min_{\theta_k} \left\| s_k(n) - \hat{s}^*_{k-1}(n+l) \quad + \mathbf{a}_k^\top \hat{\mathbf{s}}_k^-(n) - \dot{g}_k(n) \right\|^2,\end{aligned} \tag{4.17}$$

where $e'(n)$ stands for the modified residual shown in (4.17).

Following the calculation of the cost of the initial population's members, the DE algorithm heuristics are applied until the termination criterion is met (see Section 4.4.1). Fig. 4.3 illustrates an example of the optimization process for one glottal cycle over $I_{max} = 600$ iterations. In all but the two upper panels, the state of various parameters throughout the optimization process is shown at intervals of seven iterations. The top right panel shows the speech waveform, on which the optimization was performed, as a thick grey line. The waveform was generated using the default variances of glottal jitter, aspiration noise and f_0 as explained in Sec. 4.5.4, but an additional environmental noise resulting in $\mathrm{SNR}_e = 15\,\mathrm{dB}$ was added to the speech signal. Clearly the decaying resonances of the previous glottal cycle can be observed in the approximately first 50 ms. In the same panel, a thin, solid black line, illustrates the final, optimized speech waveform. It was generated using the optimal parameter set θ_k^*, found during the optimization and can be seen to match the original speech waveform quite well. The top left panel shows the cost of the best-so-far parameter set at various iteration cycles. The remaining panels in the lower part of Fig. 4.3 show scatterplots to illustrate the distribution and convergence of the various parameters contained in θ. The true values are represented by the respective solid line in each panel. Some parameters in general converge faster than others. For instance, the two parameters related to the first formant (frequency and bandwidth) generally converge faster than higher formants. Presumably, this is due to the higher SNR in lower frequency bands.

4.5 Experimental Validation

4.5.1 General Considerations and Problematics

A proper evaluation of source-filter separation methods is a difficult task due to the uncertainty regarding the *correct* glottal source and VT. In fact, there exists no method that allows measuring the glottal excitation directly from the human larynx while preserving natural voice production. This makes it impossible to compare a glottal waveform estimated from natural speech with a *ground truth* waveform. Therefore, often, synthetic speech is used in the evaluation of the performance of estimation methods. This approach may be considered problematic though, if both the synthesized samples and the evaluated method are based on the same hypothesis regarding the mechanisms of human speech production. Simple synthetic vowels are certainly useful for a validation of the methodology under changing environmental conditions. In principle they are insufficient though to assess the accuracy of a method that uses the same source-filter model as the one used for generating these vowels.

An alternative to synthetic vowels is the usage of synthetic speech generated by physical models of voice production, as for example used in (Alku et al., 2006) for the same reason and described in Section 5.1. This physical model incorporates a three-mass model of the vocal folds set into self-sustained oscillation while interacting with subglottal and supraglottal pressures, similar to the approach described in Sec. 2.2. Thereby, the resulting glottal volume velocity is not only a function of the sub-glottal pressure. The vocal folds may oscillate when-

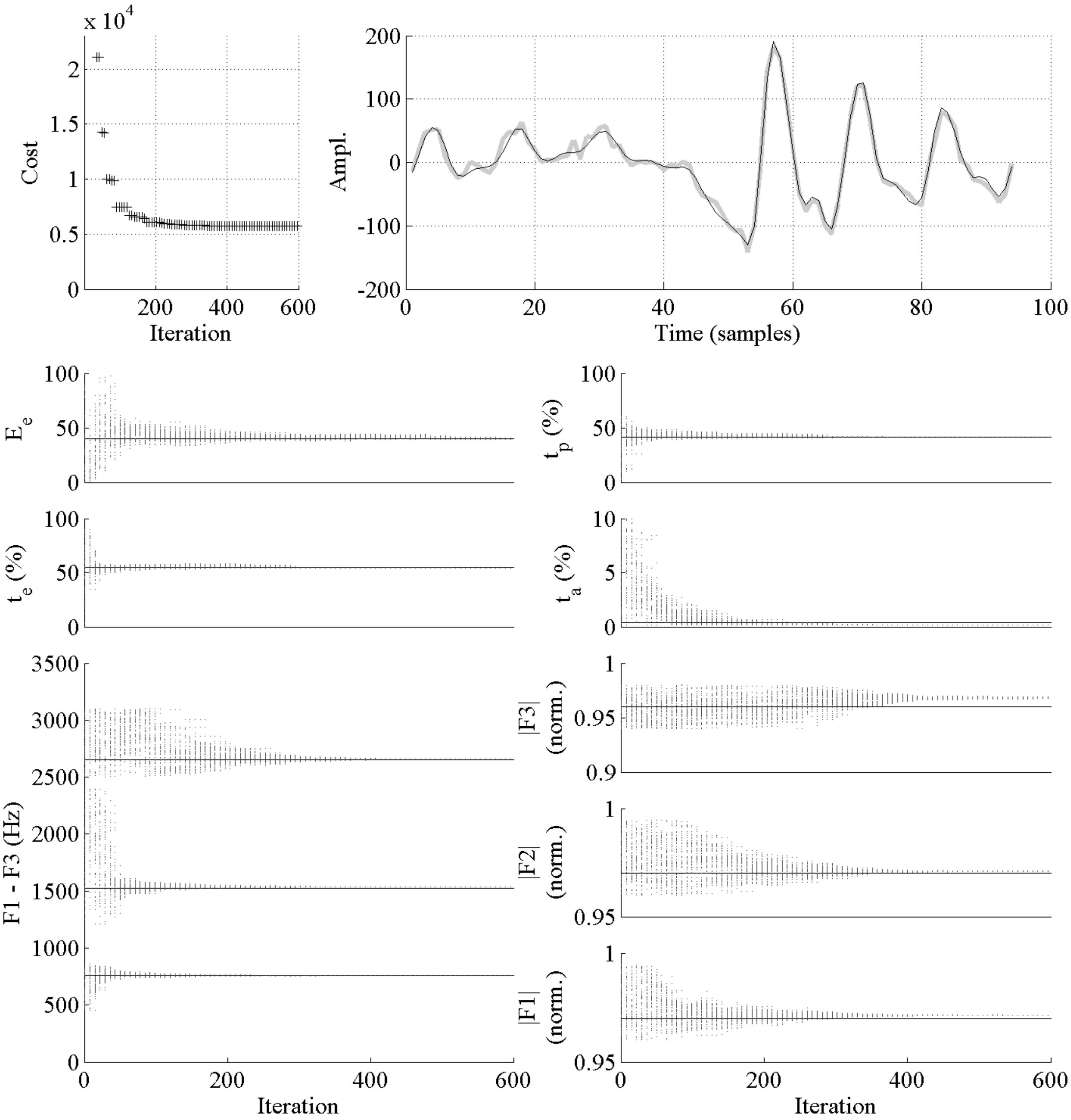


Figure 4.3: Optimization of a glottal cycle of synthetic speech generated by Eq. 4.20, embedded in additive environmental noise of 15 dB SNR level. The thick, grey line in the right, top panel represents the speech signal to be optimized. The thin black line is the signal generated with the optimal parameter set, θ_k^*. One may observe the resonances of the previous glottal cycle during the first 50 ms, which are canceled in the optimization cost function (see Section 4.4.2 and Eq. 4.17). The top left panel illustrates the respective minimum cost found in each iteration. The remaining panels display scatterplots illustrating the evolution of the parameter set throughout the optimization process to the optmal values represented by a solid line, respectively. The dots at a particular iteration represent a subset of the *NP* population members.

ever an asymmetry exists between the aerodynamic driving forces produced within the glottis and the opening and closing phases of the vocal folds (Story, 2002). The resulting speech waveform is thereby generated by a process that is based on a different hypothesis compared to the process used to generate the candidate solutions.

Following the above discussion, the proposed optimization method is first validated in a series of experiments using synthetic speech samples (Section 4.5.4). These experiments aim at investigating the ability of the proposed method to converge to the optimal set of parameters while the underlying ground truth values are actually known. For this evaluation, the reference material is exposed to a variety of potentially harmful alterations such as varying environmental noise, varying fundamental frequency and glottal jitter. Furthermore, since the proposed method utilizes an analytical model for the deterministic part of the voice source, it is interesting to see how the proposed method performs in those cases where this model significantly deviates from the "real" source. Therefore, experiments are conducted that investigate the effect of a glottal source that is altered such that the source model used for optimization is not capable of exactly modelling this glottal source used during synthesis of the speech signal. Eventually, in Chapter 5, the performance of the proposed method is investigated for the case when the evaluation material indeed was produced using a different hypothesis with respect to the speech production model. For this evaluation we are using speech signals generated by a physical model of speech similar to the one described previously. Furthermore, in that chapter, the performance of the proposed method when applied to real speech signals will be qualitatively examined.

4.5.2 Reference Methods

The proposed method is compared to three other widely used methods for inverse filtering, one of them also operating pitch-synchronously.

1. *Iterative adaptive inverse filtering (IAIF)*: IAIF was first introduced by Alku in (Alku et al., 1991) and (Alku, 1992) and has since found many applications, for example in speech synthesis (Raitio et al., 2011). IAIF uses an autoregressive error minimization method such as *discrete all-pole model* (DAP) (Alku et al., 2006) or *LP* to obtain estimates of AR models of the source and the VT. First, the source is estimated from a windowed signal segment spanning several glottal cycles using a low-order (order 1) all-pole model. After cancelling the estimated effect of the source, a preliminary, higher order (order p) estimate of the vocal tract is obtained. In a second iteration, a refined estimate of each model is obtained by repeating the first steps. This time the source is estimated after cancellation of the preliminary estimate of the VT and by using an AR model with an order higher than during the first iteration (order $q > 1$). Again, the effect of the source is cancelled before estimating a refined version of the VT model. In this paper, the choice of parameters was based on the values used in (Alku et al., 2006). In particular, as autoregressive estimation methods we use the DAP method, a window length of 200 ms, $q = 2$ and $p = 10$. The windowed segments are positioned on a glottal cycle and shifted

pitch-synchronously.

2. *Linear prediction (LP)*: Linear prediction is probably the most widely used method for the estimation of vocal tract coefficients (Quatieri, 2001). For the purpose of our experiments, a pre-emphasis ($b_1 = -0.98$) filter is applied to jointly represent the average effect of lip radiation and glottal source. The LP window length is chosen to be 51.2 ms and the LP model order is $p = 10$. As with the IAIF method, the windowed segments are positioned in time so as to be centered on a glottal cycle and shifted pitch-synchronously.

3. *Closed-Phase Linear Prediction (CPLP)*: The CPLP method performes an autoregressive error minimization, similar to the IAIF and LP methods. In contrast to these methods though, CPLP carries out the analysis in a pitch-synchronous manner and over a significantly shorter analysis window. This window is placed exactly over the closed phase of one glottal cycle (ranging from t_c of that cycle until t_o of the next cycle). The size of this window is usually very short (a few milliseconds) and therefore the covariance method is employed for solving the normal equations. As explained in chapter 3.2.4, this method therefore does not always result in minimum-phase lag AR filter coefficients, but often exhibits roots having a magitude greater than unity. This was also observed in our experiments. The order of the AR model was chosen to be $p = 10$.

In the following, all signals are sampled at 10 kHz.

4.5.3 Generation of Synthetic Speech

A glottal source signal is controlled by the glottal opening instant $t_{o,k}$, the LF model parameters contained in θ_k defined in Eq. 4.6 on page 71 and a glottal noise $w^{\sigma_g}(n)$ with standard deviation σ_g added to the glottal source $g(n)$:

$$g(n) = \sum_{k=0}^{K} v_{g,k}(n) + w^{\sigma_g}(n). \tag{4.18}$$

The glottal source $v_g^{\theta_k}(n)$ is generated using (2.2) and $w^{\sigma_g}(n)$ is a high-pass filtered (f_c=2 kHz) white Gaussian noise that was pitch-synchronously amplitude modulated in order to create a perceptionally coherent aspiration noise, as proposed in (Hermes, 1991). A clean speech signal $s_c(n)$ is then generated using

$$s_c(n) = -\sum_{i=1}^{p} a_i(n) s_c(n-i) + g(n). \tag{4.19}$$

Eventually, environmental noise $w^{\sigma_e}(n)$ is added to $s_c(n)$ in order to emulate the conditions of real world speech recordings. The final synthetic speech signal is represented by

$$s(n) = s_c(n) + w^{\sigma_e}(n). \tag{4.20}$$

Table 4.3: Formant frequencies and bandwidths (Bw) in Hz used for synthesizing the test material for the first two experiments.

Vowel	**F1** (Bw)	**F2** (Bw)	**F3** (Bw)
/a:/	800 (65)	1400 (68)	2600 (128)
/i/	500 (63)	2300 (78)	3000 (129)
/a:/ /i/	repeated transition through above vowels		

The noise $w^{\sigma_e}(n)$ has standard deviation σ_e and was chosen to be a white Gaussian noise for mathematical convenience. The amplitudes of both Gaussian noise are set so as to obtain a particular signal-to-noise ratio (SNR_g and SNR_e) with respect to the glottal source or the speech signal, respectively. The VT coefficients $a_i(n)$ are obtained by expanding the polynomial roots determined by the formant frequencies f_p and formant bandwidths b_p, contained in θ_k, and interpolated to find a set of coefficients at each sample n.

4.5.4 Performance Comparison

Experiment Description

Using the synthesized speech as described above, the accuracy of the proposed method is first assessed with respect to variations in model-independent variables, *i.e.* *(a)* environmental noise w^{σ_e}, *(b)* fundamental frequency and *(c)* glottal jitter. While focusing on the effect of varying one particular variable, the respective *other* variables were fixed to the following values: $f_0 = 108\,\mathrm{Hz}$, $\mathrm{SNR}_e = 80\,\mathrm{dB}$ and jitter = 0.3% (with respect to the fundamental period $T_0 = 1/f_0$). This jitter value was reported to be commonly found in normal phonation (Brockmann et al., 2008). In addition, the intensity of the glottal noise, $w^{\sigma_g}(n)$, was set to a value of $\mathrm{SNR}_g = 80\,\mathrm{dB}$. As test material, six samples of 2 s in duration were generated. These samples cover the range of the combinations of two voice types (see Table 4.4) and three vowel configurations (see Table 4.3). The vowel configurations are two sustained vowels and a vowel transition. For an example of such a vowel transition, see Fig. 3.3 on page 51. In addition, the LF parameters used for generating the glottal source obey a normal distribution with standard deviation of 2 % around the nominal values listed in Table 4.4, varying from glottal cycle to glottal cycle, as described in (Childers, 2000). The results for each experiment and each test configuration were averaged from 100 glottal cycles.

Choice of Measurement Metric

In the literature, the Itakura-Saito (IS) distance (Itakura and Saito, 1970) is often used as a measure of the similarity between two AR spectra. IS is computed as the mean-squared

Table 4.4: LF parameters used for synthesizing the test material for the first two experiments.

Voice Type	$\mathbf{t_p}$(%)	$\mathbf{t_e}$(%)	$\mathbf{t_a}$(%)	$\mathbf{E_e}$
Modal	41.21	54.93	0.42	40.03
Harsh	25.01	29.89	0.99	39.98

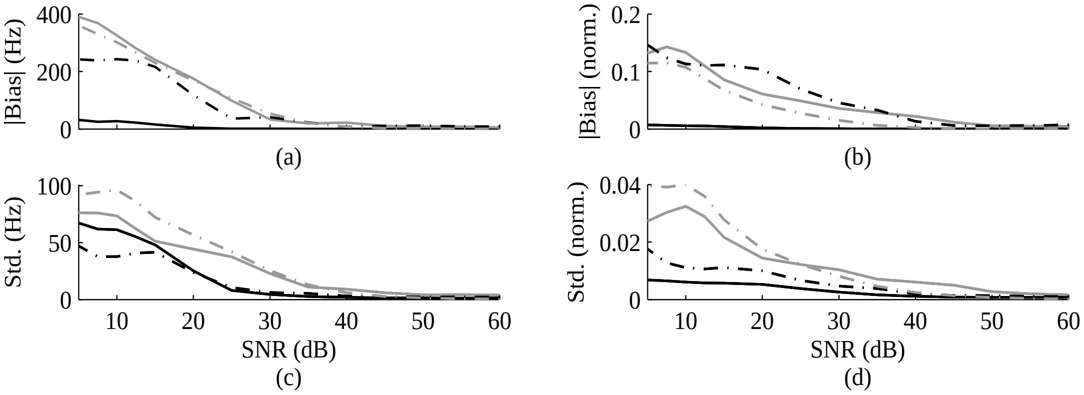
difference of two AR spectra evaluated at equidistant, discrete frequencies. This poses a problem though for the three reference methods, IAIF, LP and CPLP. All three methods assume an underlying auto-regressive process and use an optimality criterion that maximizes the whiteness of the residual spectrum. A small error in the bandwidth of a formant may then lead to an unproportionally large error in the IS measure. Consider, for instance, the case where a significant amount of noise distorts the signal. In frequency bands with a low SNR, poles will tend to compensate the error that is due to the noise. Therefore a large error will be introduced in regions that are dominated by the noise. In that case the IS distance is not a fair measure and does not reflect the ability of the tested method to estimate the formant frequencies and formant bandwidths.

Instead, we decided to report two types of errors related to the VT formants and to the glottal source, directly. First, two errors related to each formant are computed and averaged over all voice types and the three formants. In particular, we report the relative errors of the estimated formant frequencies and formant radii with respect to the ground truth for each of the three vowel configurations. An error with respect to the glottal source mostly influences the shape of the extracted glottal waveform. Therefore we report the relative errors of the estimated instants of the maximum of the glottal flow waveform (t_p) and the minimum of the glottal flow derivative waveform (t_e), with respect to the ground truth. Although also the parameter t_a, related to the return phase of the LF model, influences the shape of the glottal waveform, we do not report this value here. This value typically is so small that it is highly influenced by any kind of noise and our experiments showed that no statistically significant estimation was possible.

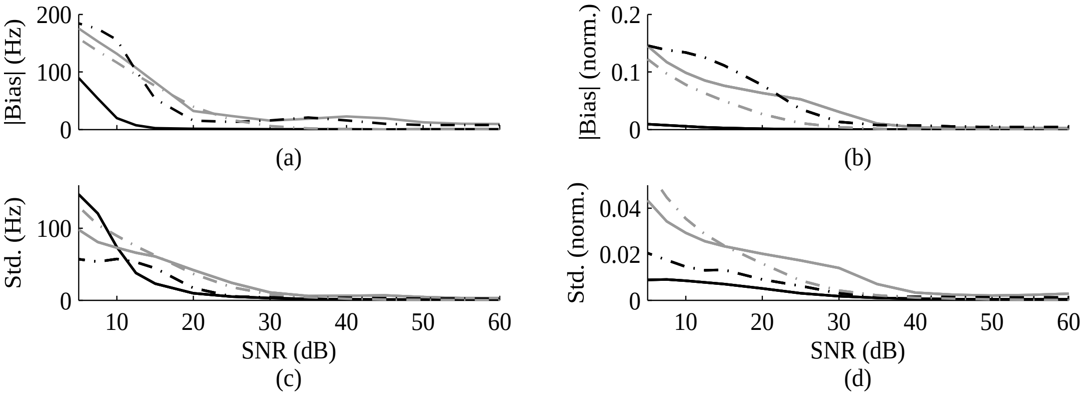
Similarly to the formant errors, the source-related errors are averaged over all voice types and also over all vowel configurations. The LP and CPLP methods are excluded from this second result, since the residual of these methods are not meant to extract the glottal waveform. The source-related values extracted by the IAIF method were obtained from the respective inverse filter residual, using methods found in the *Aparat* toolbox (Airas et al., 2005; Helsinki University of Technology (HUT), TKK Laboratory of Acoustics and Audio Signal Processing).

Error related to Environmental Noise

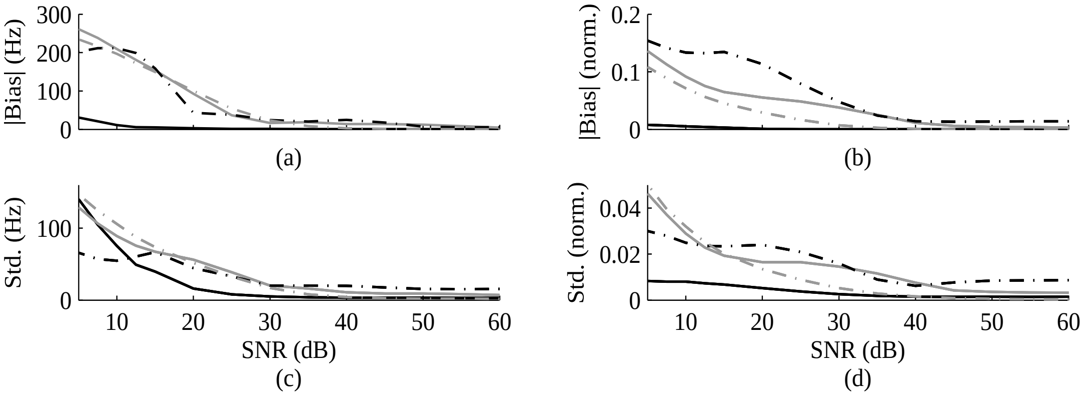
In the first experiment, we assess the influence of the presence of environmental noise on the reliability of the estimated parameters. Environmental noise is added to the speech as described in Eq. 4.20. Since the additive noise is spectrally white, it implicitly has the greatest



I: Vowel */a:/*

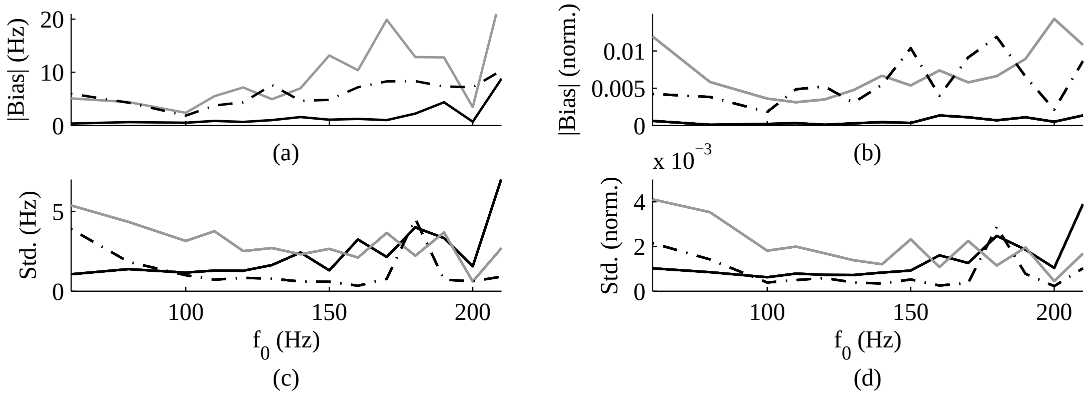


II: Vowel */i:/*

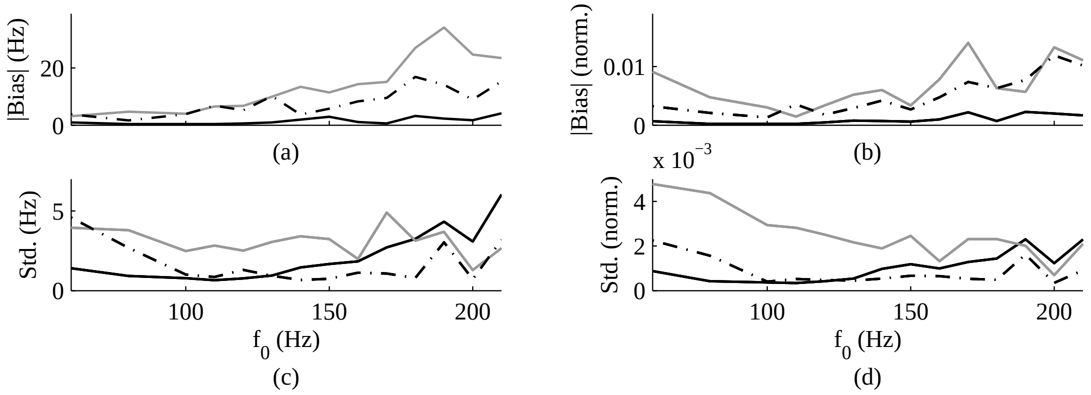


III: Vowel Transition

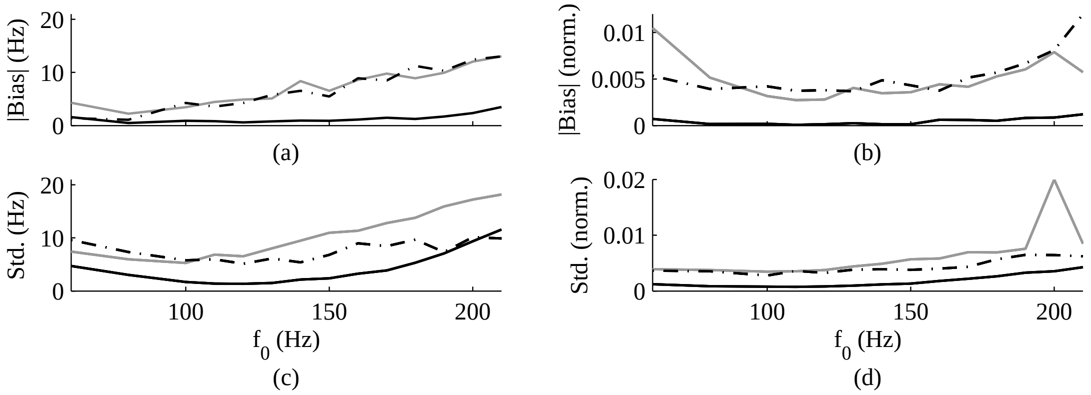
Figure 4.4: Absolute value of bias (top) and variance (bottom) of the estimation errors regarding the formant frequencies (left) and the formant radii (right) as measured over a range of different levels of *environmental noise* from three different vowels. The proposed method (black solid line) exhibits a reduced bias in all examples compared to the LP method (grey solid line), IAIF method (black dash-dot line) and the CPLP method (grey dash-dot line).



I: Vowel */a:/*

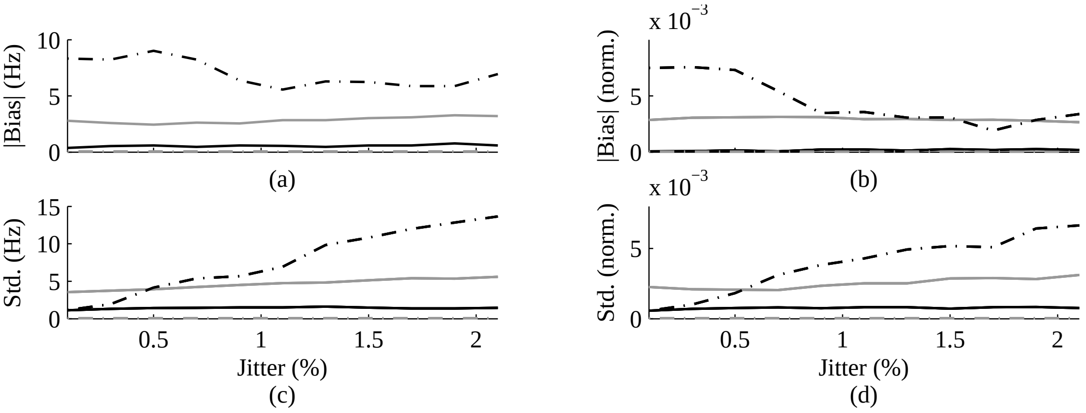


II: Vowel */i:/*

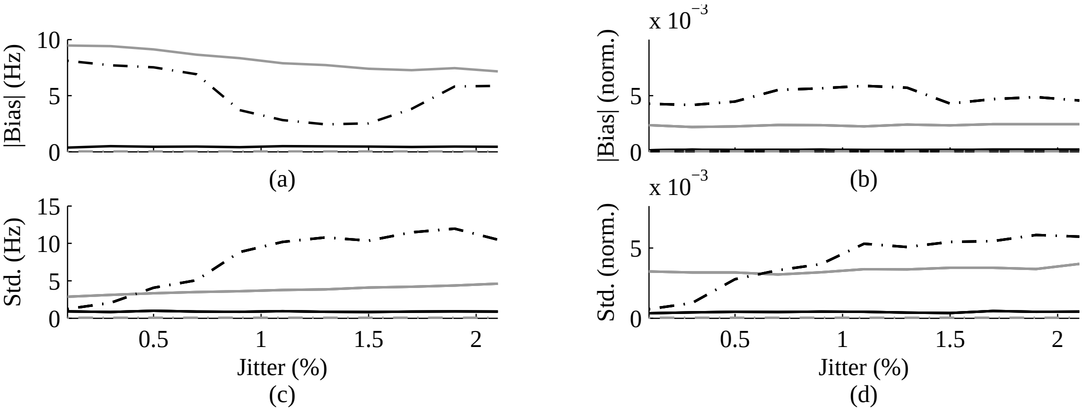


III: Vowel Transition

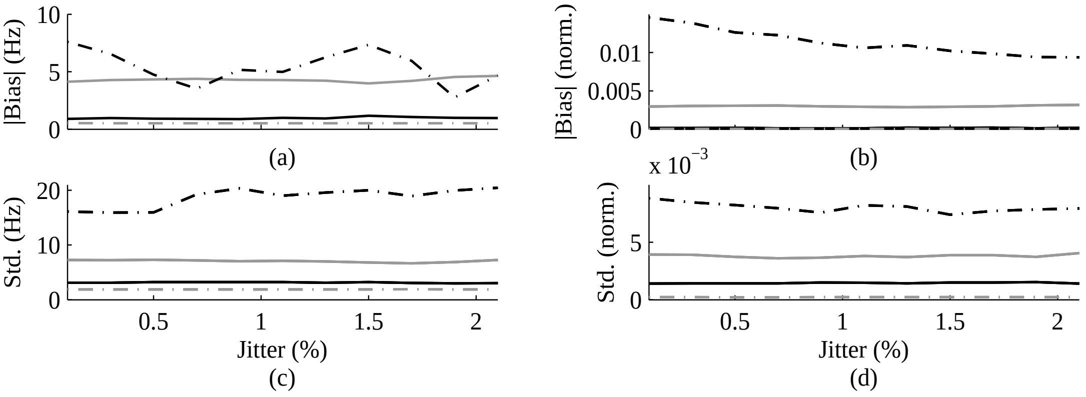
Figure 4.5: Absolute value of bias (top) and variance (bottom) of the estimation errors regarding the formant frequencies (left) and the formant radii (right) as measured over a range of different values of *fundamental frequency* from three different vowels. The proposed method (black solid line) exhibits a reduced bias in all examples compared to the LP method (grey solid line), IAIF method (black dash-dot line).



I: Vowel */a:/*



II: Vowel */i:/*



III: Vowel Transition

Figure 4.6: Absolute value of bias (top) and variance (bottom) of the estimation errors regarding the formant frequencies (left) and the formant radii (right) as measured over a range of different values of *glottal jitter*. The proposed method (black solid line) exhibits a reduced bias variance in all examples compared to the LP method (grey solid line), and IAIF method (black dash-dot line), whereas the CPLP method (grey dash-dot line) performs similar or slightly better.

effect on the higher formants due to their relatively low SNR.

Each of the figures combined in Fig. 4.4 represents the result with respect to a different vowel configuration. In each of these figures, the absolute value of the bias (upper panels) and standard deviation (lower panels) of the estimated formant frequencies (left panels) and radii (right panels) are displayed. For each vowel, the bias of both, the formant frequencies and radii, estimated by the proposed method is clearly smaller compared to that of the other three methods. From examining individual examples, it is observed that especially the value of the lower formant frequencies estimated using the proposed method exhibit a high accuracy at all SNRs. For higher formants, in frequency bands with lower SNR values, the estimates occasionally got trapped in local minima. This resulted in sporadic outliers of the estimated third formant. This explains the increased standard deviation of the average formant frequency estimates for SNR values below 15 dB.

Qualitatively, the results of all three formants estimated by the proposed method agree with each other, except that the bias of the formant frequencies estimated from the vowel /i/ (see Fig. 4.4 I) deteriorates faster for low SNR values. From Table 4.3 we see that the frequencies of the second and third formant of that vowel are higher. This result is therefore a confirmation of above observation that the proposed method performs less well in frequency bands where the SNR value is low. The estimated formant radii exhibited a lower bias and standard deviation throughout all SNR values compared to the other methods. The relatively worse performance of the other three methods at low values of SNR may be explained by the tendency of these methods to model the noise instead of formants in frequency bands with low SNR. All three other methods performs relatively similar, with the bias of the formant frequencies starting to deteriorate at an SNT value of about 20 dB.

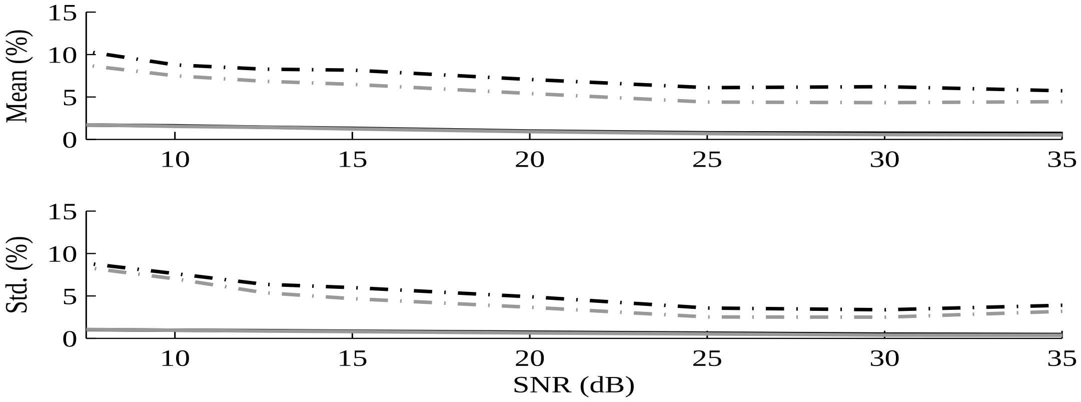
The errors related to the glottal source temporal parameters are displayed in Fig. 4.7 I. The error of the proposed method is relatively small at high SNR values and steadily increases for lower SNR values. In comparison, the error of the IAIF method is generally higher and appears to be more affected by the increasing noise level.

Error related to Fundamental Frequency

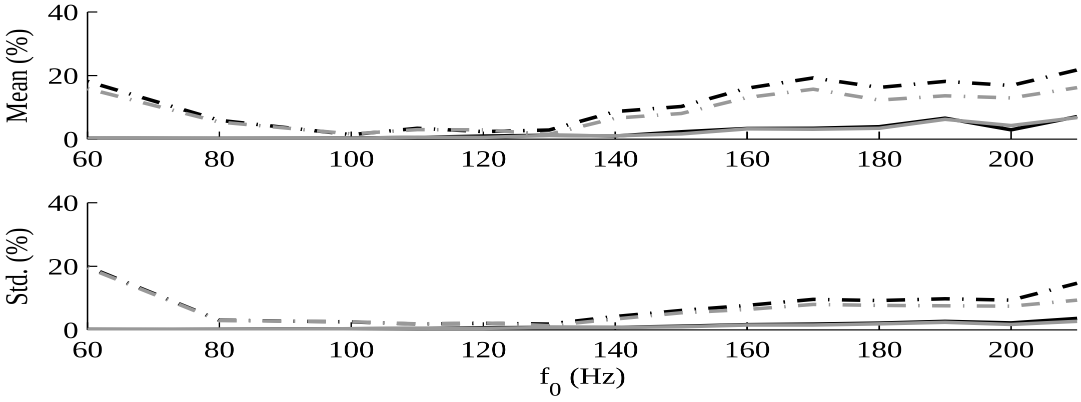
For this experiment, synthetic vowels with different fundamental frequencies were generated. With an increasing value of f_0, the duration of the analysis window for the CPLP methods becomes very short very quickly. The results for this method for frequencies above $f_0 = 140\,\mathrm{Hz}$ largely exceeded the visible range of the axes displayed in Fig. 4.5. For this reason we do not include the results for this method in this experiment.

As illustrated in Section 3.2.4, frame-based analysis methods (here, IAIF and LP) may be influenced by the harmonics of the source fundamental frequency. At lower values of f_0, the estimated poles form a well-defined spectral envelope over the densely distributed f_0-harmonics. At higher values of f_0, the harmonics are sparser and thus represent single points of attraction for the poles. Thus, with rising f_0, it becomes more likely that a pole models a harmonic instead of a formant, resulting in a higher probability of formant estimation errors.

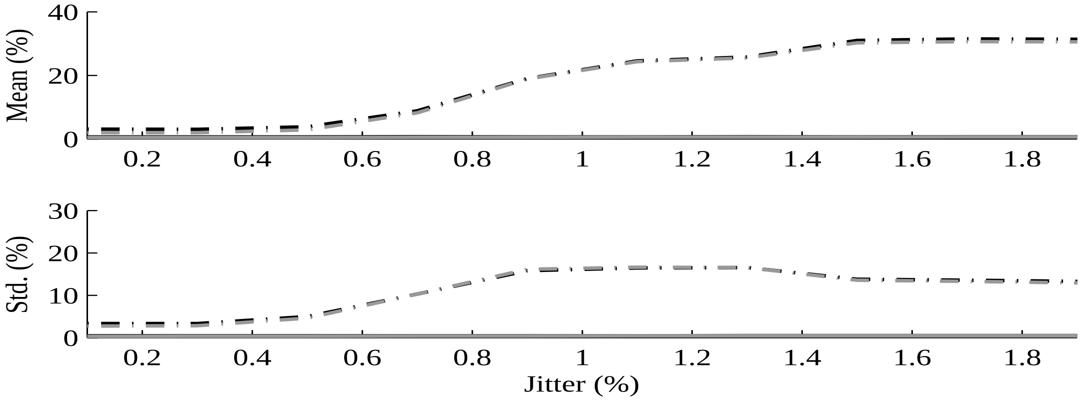
This is what can be observed in the figures summarized in Fig. 4.5. Similarly to the ex-



I: Environmental noise



II: f_0



III: Jitter

Figure 4.7: Relative error of the estimated LF model parameters t_p and t_e with respect to the instantaneous glottal period. The estimates of the IAIF method are displayed using dashed lines, estimates of the proposed method are displayed using solid lines. The black color represents the error in t_p, while the grey color refers to the error in t_e.

periment above, the bias of the error in the formant frequencies and radii estimated by the proposed method is generally lower compared to the other methods. A trend is also observable in that with increasing values of f_0 the error of all methods increases. Relatively speaking, the error due to the the proposed method remains unaffected up to a certain value. Above $f_0 = 200\,\text{Hz}$, the error of the proposed method starts to rise sharply. This may be explained by the considerably shortened analysis window in voice sources with higher a value of f_0.

Note that in the example of the vowel */a:/*, the error of the formant frequencies and radii of all methods at a frequency of $f_0 = 200\,\text{Hz}$ are much lower compared to the errors at the surrounding values of f_0. Interestingly, this was observed because incidentally all the formant frequencies of that particular vowel are at exact multiples of $f_0 = 200\,\text{Hz}$. Therefore, the estimated poles were in fact not deviated from the correct formant frequency values, but the underlying harmonic structures even attracted the poles to the correct values. For the same reason, a certain unsteadiness of the measured error values may be observed across different values of f_0 for the LP and the IAIF methods. In some vowel-f_0 combinations, the harmonics increase the accuracy of the estimated formants of the frame-based methods, but not in all cases. This is evidence of the advantage of pitch-synchronous methods.

The results with respect to the glottal source timing parameters are displayed in Fig. 4.7 II. Notably, the error of the proposed method is less affected across different f_0 values and is also smaller compared to the error of the IAIF estimates.

Error related to Glottal Jitter

This experiment investigates the error induced by different values of jitter in the fundamental period of the source. Jitter is a measure of deviation from perfect harmoniousness, *i.e.* how much a particular glottal cycle deviates from an averaged, instantaneous glottal period, T_0. Jitter here is measured in per cent, relative to T_0.

The results are displayed in the figures contained in Fig. 4.6 and in Fig. 4.7 III. As one may expect, the two pitch-synchronous methods are not affected at all by a variation in the value of jitter. The estimates of the formants and those of the glottal parameters measured by these two methods are constant and independent of the value of jitter. Also, their formant related estimates are of the same low value for all the three different vowels, whereas the estimates of the two frame-based analysis methods are different for the different vowel configurations; a clear sign of the incomplete separation of source and filter components in these methods.

Another observation concerns the standard deviation of the VT measures of the IAIF method. It appears that this value varies with a varying amount of jitter, while the estimate of the LP method is not affected by jitter. This indicates that the VT estimates obtained by the IAIF method are influenced by a variation in the glottal source. This observation was also made in other experiments on real speech data, described in Chapter 5. From experimental observations it appeared to us that the IAIF method requires a careful choice of its parameters, which sometimes needed adjustment for particular examples.

The source related errors (t_p and t_e) are very similar in both methods (IAIF and the proposed method), but IAIF is affected by higher values of jitter. This is to be expected from a frame-based analysis methods.

Glottal Source Distortion

This experiment addresses two issues. On one hand, glottal noise is largely composed of aspiration noise carrying idiosyncratic and semantic cues. On the other hand, glottal noise represents a distortion in the glottal source, because it is not captured in the LF model that represents only the deterministic source components (see Section 2.3.1). Hence, this experiment can be seen as validation against glottal noise and source mis-modelling.

The results were computed in the same manner as in the previous experiment. The errors of the estimated formants and LF model parameters across a range of glottal noises w^{σ_g} are displayed in the figures contained in Fig. 4.8 and in Fig. 4.9, respectively. For all four methods, the influence of the glottal distortion on the formant estimates is negligible up to $\mathrm{SNR}_g = 20\,\mathrm{dB}$. For SNR values lower than that, the performance of all but the CPLP method starts to deteriorate. The proposed method is strongly affected, in particular stronger than in the experiment with the environmental noise. Clearly, this indicates that the performance of the proposed method depends on the ability of the source model to correctly represent the excitation of the source-filter model. When compared to the experiment with environmental noise it may also be noticed that the performance of the frame-based analysis methods, IAIF and LP, is less affected in terms of the absolute value of the errors. An explanation for this might be the low-pass nature of the glottal noise, leaving the lower formants unharmed. Further, it is remarkable how the CPLP method is completely independent of any glottal noise or distortion, in particular when compared with its performance under the influence of the environmental noise. This can of course be expected, since the motivation for this method is exactly to stop variations in the glottal source from influencing the formant estimation results.

As in the previous experiments, the LF model parameters estimated by the proposed method exhibit a smaller bias and a smaller standard deviation compared with IAIF, as shown in Fig. 4.9. As can be seen there as well, it may be observed that the magnitude of the error in terms of its bias and its variance is largely independent of the amount of glottal noise present. A possible explanation is that the low frequency characteristic of the glottal LF model is not greatly influenced by the glottal noise having a high-pass characteristic.

A general remark concerning the performance of the proposed method in the presence of noise shall be made. In the previous experiments we investigated the influence of noise added to the glottal source before articulation on one hand and additive, white noise simulating environmental distortions. In real conditions, both types of noise may be expected to have magnitude spectra differing from the idealized flat characteristics of white noise. Such deviations may have the most influence when occuring at the voice source, since the LF voice source model used for the experiments is not able to represent a glottal roll-off other than the smooth decays illustrated in Fig. 2.9 to Fig. 2.12. Any narrow-band attenuation or amplification due to non-white glottal noise will lead to a bias on the estimated formant parameters. With respect to non-white environmental noise it may be expected though that the proposed method performs better than the averaging-based methods relying on auto-regressive error minimization. While these methods make no assumptions on the kind of signal estimated, the proposed method omposes a pre-condition on the analyzed system to be excited by a deterministic signal. This pre-condition should favor the separation of resonances due to the deterministic voice source from correlations in the observed signal due to colored noise.

4.6 Conclusion

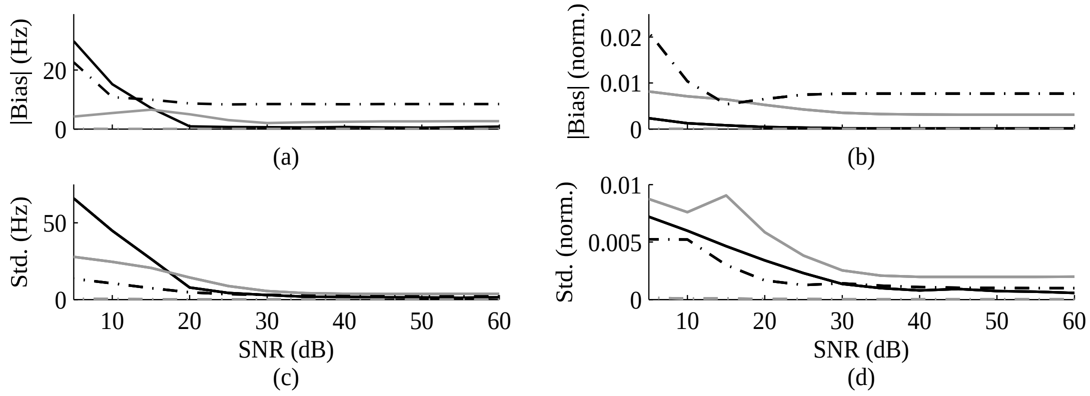
- We formulated an optimization scheme that pitch-synchronously fits a multi-parametric source model and an auto-regressive VT model to observed speech signals. Differential evolution serves as a computational tool for the optimization process (Sec. 4.1).

- Given that the observed speech signal is modelled by the source and filter models, criteria were formulated to guarantee that the two different models may not compensate for errors in the respective other model. In particular, the optimization process has to guarantee, through boundary conditions and model definitions, that the poles of the AR VTF model are conjugate complex, complex-valued and with a magnitude $|z_p| < 1$. Furthermore, the opening phase of the glottal source model necessarily is required to have one pair of complex conjugate poles with magnitude $|z_p| > 1$. The return phase of the glottal model has a single, necessarily real-valued pole with magnitude $|z_p| < 1$ (Sec. 4.3).

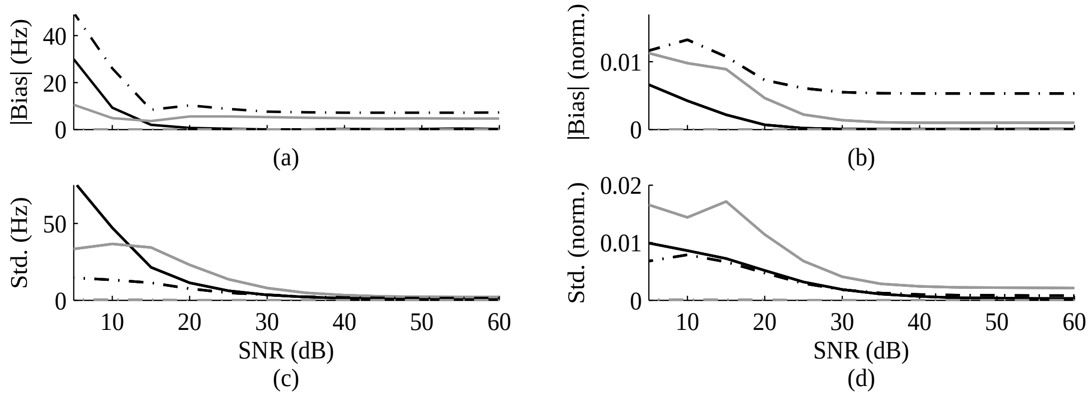
- The glottal source model parameters are expressed using the time-domain representation, but their conversion to the synthesis paramter representation for each parameter set evaluated during the optimization proves computationally very costly. An alternative representation might further speed up the optimization process in the future.

- The optimization problem is formulated in such a way as to reduce the effect of previous VTF resonances on the currently optimized glottal cycle (Sec. 4.4). This is achieved by subtracting the estimated resonances of previous cycles before the estimation of the current cycle begins. Effectively, this increases the analysis window of the proposed method.

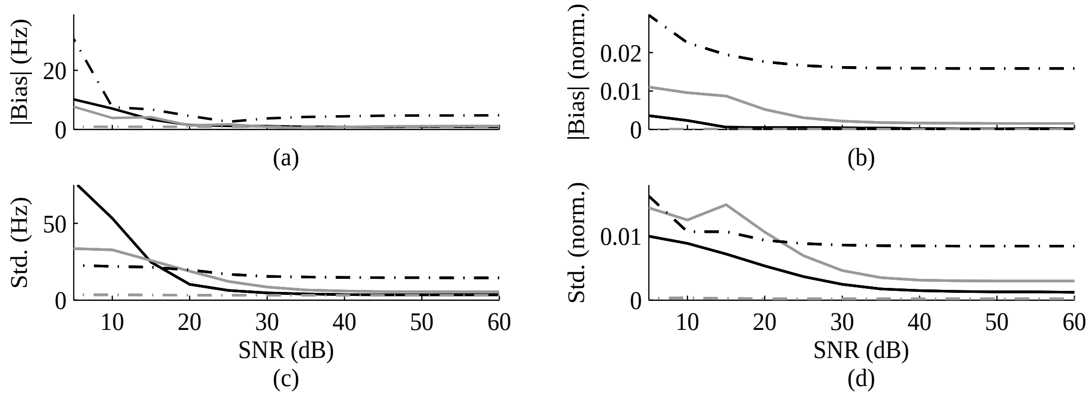
- The convergence characteristics of the proposed method are examined in a variety of modifications such as varying f_0, glottal jitter, environmental noise and glottal source distortions (Sec. 4.5). The performance of the proposed method was evaluated in comparison to three other, state-of-the-art methods, two of them frame-based and one of them pitch-synchronous. Generally, the following observations can be concluded:
  - The proposed method performs better than all the other three methods in the presence of environmental noise. In particular, the bias of the estimated formant frequencies and bandwidths is largely reduced in these conditions.
  - As error of the CPLP method, also the error of the proposed method is insensitive to glottal jitter and to a large degree also to variations in the fundamental frequency.



I: Vowel */a:/*



II: Vowel */i:/*



III: Vowel Transition

Figure 4.8: Absolute value of bias (top) and variance (bottom) of the estimation errors regarding the formant frequencies (left) and the formant radii (right) as measured over a range of *glottal noise*. The proposed method (black solid line) exhibits a reduced bias in the formant radii in all examples compared to the LP method (grey solid line) and IAIF method (black dash-dot line). The CPLP method (grey dash-dot line), as expected, is not affected by glottal noise.

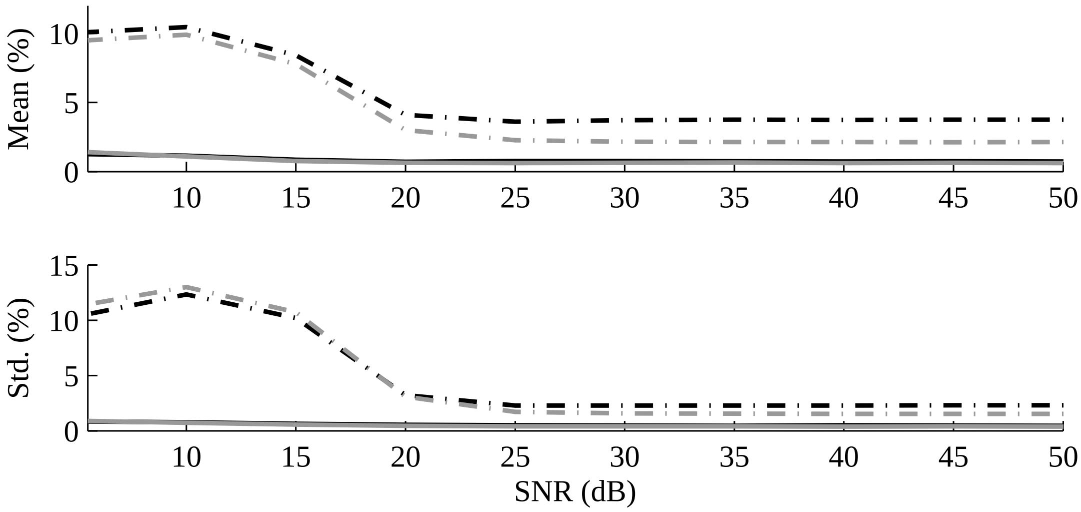


Figure 4.9: Relative error of the estimated LF model parameters t_p and t_e with respect to the instantaneous glottal period. The estimates of the IAIF method are displayed using a dashed line, estimates for the proposed method are displayed in a solid lines. The black color represents the error in t_p, whereas the grey color refers to the error in t_e.

Above a frequency of 200 Hz, the performance of the proposed method is penalized significantly by the short duration of the analysis window.

- The performance of the proposed method with respect to extracted formants is unharmed by deviations of the glottal model up to a certain noise level (≈ 15 dB). Deviations incurring a higher noise than this level have a significant influence on the performance. This is contrast to the other pitch-synchronous method, which is completely untroubled by glottal variations.
- A general observation from the examination of individual results indicates that the performance of the proposed method with respect to the estimated formants is very high as long as the deterministic portion of the glottal source has significant energy with respect to the noise in the frequency band of the formant concerned.

5 Application of Source-Filter Separation

5.1 Source-Filter Separation on Physically Modelled Speech

In a first of two series of experiments, we assess the performance of the proposed method on synthetic vowel samples, representative of an adult male speaker, generated using a physical, computational model of the speech production system. The voice source component of the model consists of a kinematic representation of the medial surfaces of the vocal folds (Titze, 2006, 1984; Samlan and Story, 2011). The output of this source model is controlled by vocal fold parameters such as surface bulging, adduction, length, and thickness as fundamental frequency. Vocal fold length and thickness were set to 1.6 cm and 0.3 cm, respectively. As the vocal fold surfaces are driven in vibration the model produces a time-varying glottal area that is coupled to the acoustic pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations (Titze, 2002). The resulting glottal volume velocity was determined by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis.

The vocal tract shape, which extends from glottis to lips, was specified by area functions representative of */i/* and */a/* vowels, or as a transition from */i/* to */a/*, and were based on data reported by Story (Story, 2008). The tracheal shape was also specified by an area function that extended from the glottis to bronchi (Story, 1995). Acoustic wave propagation in the subglottal and supraglottal airspaces was computed with a wave-reflection model (Liljencrants, 1985; Story, 1995) that included energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips (Story, 1995). This form of the computational model was similarly used to generate synthetic speech samples for (Samlan and Story, 2011); a more extensive description of the model can be found there.

The test material consists of nine speech samples, each 0.7 s long. Three vowel configurations were used (*/a/*, */i/*, transition */i/* to */a/*). Of each vowel, three different realizations were synthesized using three different voice types (*pressed*, *modal* and *breathy*) and a constant fundamental frequency of $f_0 = 105\,\text{Hz}$. Along with the synthesized speech, a true glottal flow signal generated by interaction with trachea and VT, as well as true formant frequencies are available. All speech samples were low-pass filtered ($f_c = 4\,\text{kHz}$) and downsampled to a sam-

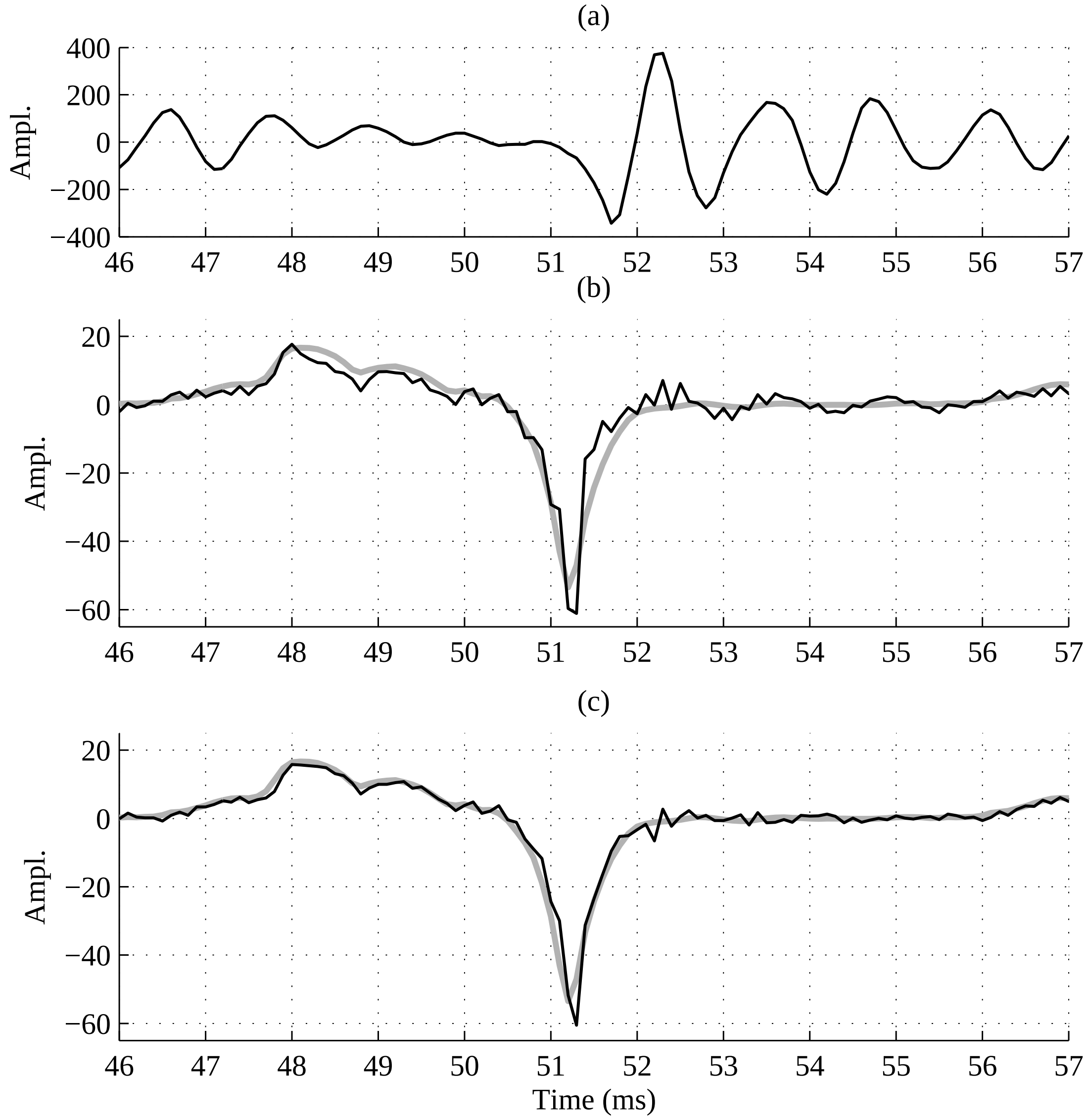


Figure 5.1: Example of a speech segment of a vowel /a/ (top) synthesized with the physical model of speech. In the middle and bottom panel, the grey, solid line represents the true glottal flow derivative, as simulated by the model. Note the formant ripple in the middle of the opening phase (around 48 ms). Overlaid as a black, solid line is the inverse filter residual of the IAIF method in the middle panel and the respective residual of the proposed method in the bottom panel.

pling rate of $f_s = 10\,\mathrm{kHz}$.

A qualitative analysis of the error related to the glottal source estimation can be done using an illustration of the inverse filtered glottal derivative waveform, as shown in Fig. 5.1. Fig. 5.1(a) shows a section of the speech waveform used for the analysis. It represents the exact size of the analysis window used for by the proposed method. For the IAIF method, a larger window of

200 ms duration was used. The panels 5.1(b) and 5.1(c) show the derivative of the true glottal flow, as simulated by the physical model, in a thick, grey line. The inverse filter residuals of the proposed method and IAIF are shown in the same panels, as thin, black lines, respectively. Both, the proposed method and IAIF, are able to retain the general waveform of the glottal source including low frequency glottal distortions such as formant ripples (observable in the first 4 ms of the example). From visual inspection it is also observable that in this particular example there remains slightly more high-frequency noise in the IAIF residual. The IAIF method did not capture all the VT resonance components in the estimated VT filter. These remaining spectral components were thus not removed by inverse filtering. A possible explanation is temporal averaging in the IAIF method. The glottal VT coefficients estimated for a speech segment may well represent the average spectra of the observed respective components, but individual glottal cycles may diverge considerably from this average. No parametric representation of the true glottal source is available, thus no objective results are reported.

In Tables 5.1-5.3, the errors related to the estimated formant frequencies are presented. In virtually all examples, the bias of the first formant (F1) estimated by the two pitch-synchronous methods (proposed method and CPLP) is smaller compared to that of the other two methods (IAIF and LP). The standard deviation of the F1 estimate of all methods varies from example to example but is largely similar between the four methods. The low bias found in the lower formants estimated by the proposed method confirms the findings in Chapter 4.5.4 during the validation of the method.

In general, the proposed method performs best for pressed voice and worst for breathy voice sources. This is expected, for two reasons. In pressed voice sources, the instant of greatest excitation (t_e) occurs relatively early in the glottal cycle. Therefore, a relatively large portion of the analysis window contains the VT resonances and a larger proportion of samples contributes to the minimization of the error criterion. Furthermore, the proposed method relies on an excitation of the VT by a deterministic signal. Pressed voices have a short return phase, t_a, which is strongly correlated with the spectral tilt of the source. As a result, the deterministic part of the excitation in pressed voice yields a low spectral tilt and results in higher energy in higher frequency bands. The ratio between deterministic and non-deterministic energy in pressed voice is therefore greater compared to other voice types. Consequently, one can expect a more reliable estimation of higher formants in pressed voices. These results are in line with the observations made during the validation of the proposed method in Sec. 4.5.4. For the breathy voice types, no estimates of the CPLP method were obtained because in that voice type the closed phase of the glottal cycle is too short for a regular analysis.

Another interesting observation concerns the error found for *higher formant* estimates. The performance of the proposed method appears to deteriorate when compared to the other methods for some configurations (*e.g.* pressed and modal voice of vowel */i/*). By inspection of the results of individual glottal cycles it was discovered that these errors were mostly introduced by outliers in formant estimation. For occasional glottal cycles the estimated formants were far off the ground truth values, while the majority of the values were much closer to correct values. Further inspection of glottal cycles exhibiting estimation outliers revealed that their glottal spectra in frequency ranges corresponding to higher formants (above 2 kHz)

Table 5.1: Formant frequency estimation results using speech with a *pressed* voice synthesized by the physical model. The values represent the absolute value of the bias (in Hz) followed by the error standard deviation in parentheses.

Vowel	Method	F1	F2	F3
/a/	LP	48.4 (0.7)	74.5 (1.2)	56.4 (4.9)
	IAIF	21.7 (0.5)	31.3 (0.6)	18.1 (1.5)
	CPLP	**1.6** (0.9)	4.1 (8.9)	5.6 (13.1)
	DE	3.6 (0.4)	16.2 (1.2)	9.1 (9.2)
/i/	LP	41.0 (0.3)	45.9 (1.9)	74.1 (3.0)
	IAIF	7.3 (0.4)	6.2 (0.7)	14.8 (0.8)
	CPLP	**0.8** (1.1)	3.9 (9.0)	9.9 (15.0)
	DE	**1.0** (1.0)	14.9 (13.3)	25.9 (19.1)
trans. /a/ to /i/	LP	43.4 (5.0)	59.8 (13.0)	64.0 (8.9)
	IAIF	10.1 (13.3)	10.0 (25.6)	15.6 (8.2)
	CPLP	1.9 (8.2)	8.9 (24.2)	15.1 (30.3)
	DE	**0.3** (5.1)	24.1 (16.2)	9.3 (22.5)

show considerable, cycle-specific, attenuations and amplifications in relatively narrow frequency bands. In other words, the high frequency spectral characteristics of some glottal cycles show large frequency-dependent deviations from the constant spectral decay assumed by the LF model. The LF model used in the proposed method is not capable of describing such fine details due to its constant decay in high frequencies. It is currently not clear whether these spectral ripples are a specific phenomenon of the physical speech production model or whether this is a universal phenomenon, observable in real speech signals as well.

A possible explanation and interesting finding is that the LF model, despite its relatively high degree of freedom, lacks in its ability to represent the details of high frequency components of the glottal source. Narrow-band deviations from the general spectral decay of the LF model at high frequencies may lead to a bias in the formant estimation results since these spectral amplifications and attenuations are captured by the VT model instead of the voice source model. Therefore, errors in the higher formant frequency estimates are introduced.

Table 5.2: Formant frequency estimation results using speech with a *modal* voice synthesized by the physical model. The values represent the absolute value of the bias (in Hz) followed by the error standard deviation in parentheses.

Vowel	Method	F1	F2	F3
/a/	LP	**0.5** (1.4)	10.0 (4.8)	39.4 (24.9)
	IAIF	10.0 (0.8)	10.2 (5.5)	46.0 (11.6)
	CPLP	1.5 (6.3)	4.2 (13.5)	23.4 (35.1)
	DE	**0.6** (0.5)	29.5 (4.7)	26.7 (19.7)
/i/	LP	6.7 (0.4)	28.5 (7.1)	12.0 (20.1)
	IAIF	5.1 (0.9)	7.8 (2.9)	18.4 (11.7)
	CPLP	4.3 (6.8)	5.1 (9.8)	18.4 (28.9)
	DE	**3.7** (2.0)	20.8 (16.7)	40.6 (30.2)
trans. */a/* to */i/*	LP	2.1 (3.7)	14.5 (11.3)	17.3 (28.7)
	IAIF	6.5 (9.3)	8.0 (23.5)	32.4 (10.6)
	CPLP	4.7 (7.4)	9.2 (23.1)	26.6 (45.2)
	DE	**1.5** (2.3)	26.9 (15.6)	39.5 (33.3)

5.2 Source-Filter Separation of Real Speech Signals

5.2.1 Test material description

In a second series of experiments we applied the proposed method to real speech signals. Since no reliable ground truth is available from such signals, we report here only a qualitative comparison and discussion of the estimated formant trajectories and temporal source parameters.

Three different speech signals were used, all resampled to a sampling rate of $f_s = 10\,\text{kHz}$. Their temporal waveforms are shown in the panels *(a)* of Figs. 5.2, 5.4 and 5.6. The respective spectrograms are illustrated in the panels *(b)* of the same figures.

Signal A: A sustained vowel */a:/* articulated by a male speaker and with an average fundamental frequency of $f_0 \approx 120\,\text{Hz}$. The modal voice quality of this voice inflicts a relatively strong harmonic richness, which can be observed in Fig. 5.3. Even in relatively high frequency bands of up to 3 kHz harmonics of f_0 are visible.

Signal B: A female voice articulating the English word *"foul"* with an American accent and with an average fundamental frequency of $f_0 \approx 150\,\text{Hz}$. For the processing, the speech

Table 5.3: Formant frequency estimation results using speech with a *breathy* voice synthesized by the physical model. The values represent the absolute value of the bias (in Hz) followed by the error standard deviation in parentheses. Note that for the CPLP method no estimates were obtained due to the very short duration of the glottal closed phase of the glottal cycle, severly limiting the duration of the analysis window.

Vowel	Method	F1	F2	F3
/a/	LP	**8.3** (14.7)	84.4 (53.5)	88.9 (58.4)
	IAIF	39.5 (5.9)	43.7 (27.5)	124.5 (23.7)
	CPLP	xx (xx)	xx (xx)	xx (xx)
	DE	**9.6** (4.9)	48.1 (22.7)	36.1 (29.7)
/i/	LP	23.6 (1.4)	70.0 (46.0)	100.1 (67.0)
	IAIF	14.3 (5.2)	10.5 (13.7)	178.8 (39.1)
	CPLP	xx (xx)	xx (xx)	xx (xx)
	DE	**2.3** (3.7)	53.6 (262.9)	94.9 (0.2)
trans. */a/* to */i/*	LP	14.6 (12.4)	63.1 (57.1)	113.3 (67.8)
	IAIF	11.6 (23.7)	58.1 (45.6)	187.2 (56.7)
	CPLP	xx (xx)	xx (xx)	xx (xx)
	DE	**9.3** (9.3)	39.9 (32.1)	80.4 (49.7)

sample was trimmed in the time domain to contain only voiced phonetic units, *i.e.* the diphtong */aù/* followed by the lateral approximant */l/*. The signal was recorded after it was transmitted over an analogue telephone line and contains a significant amount of background noise.

Signal C: A male, alaryngeal voice performing a sustained vowel */a:/* with an average fundamental frequency of $f_0 \approx 150\,\text{Hz}$. The speaker had undergone a hemilaryngectomy, which is a partial excision of the larynx for invasive cancer. This voice-conserving procedure consists of dividing the thyroid cartilage in the midline and resecting in continuity the thyroid cartilage with the corresponding true and false vocal cords and ventricle. The speaker is therefore missing one lateral half of the larynx and his voice source deviates significantly from that of laryngeal speakers. In particular, the voice source is found to be breathy, hoarse at times and of low harmonic richness. The glottal waveform exhibits an extended open phase and a relatively long return phase. The closed phase is very short, often so short that it simply does not exist and the end of the return phase coincide with the beginning of the next glottal cycle. Therefore, the distribution of the spectral energy of the deterministic voice source portions is rather confined to the lower frequency bands. As can be observed in the spectrogram in Fig. 5.7(b), no harmonics of f_0 are present above 1.5 kHz.

5.2.2 Discussion of Formant Estimation

Similar to the results obtained for the synthetic speech signals we also compared the proposed method with estimates obtained by LP, IAIF and CPLP (see Sec. 4.5.2). With respect to the differential evolution optimization procedure it was observed that the convergence to the optimal parameters took more iterations than it did with the synthetic signals. Therefore we slightly adjusted the DE parameters (see Section 3.4). In particular, we used I_max = 1000 as a termination criterion. The other DE parameters (CR = 0.9, F = 0.3 and NP = 120) were left unchanged since they were found to work well for real speech signals as well. The respective spectrograms with overlays of estimated formant trajectories are illustrated in the panels *(b)* of Figs. 5.3, 5.5 and 5.7. The estimated LF model parameters are shown in the panels *(b)* of Figs. 5.8-5.10.

Signal A

The wide-band excitation of signal A provides a rich harmonic structure and clearly benefits the proposed method in the estimation of higher formants. Throughout the first 0.5 s, the consecutive estimates of the third formant appear on a virtually straight line, which indicates a robust estimation. The LP estimate of the third formant shows some fluctuations, supposedly due to the underlying harmonic signal structure. The F3 estimates of the IAIF method appear to fluctuate to a larger extent. The fact, that the transitions for these fluctuations are rather smooth over consecutive estimates indicates that this deviation has a deterministic cause. A possible reason could be minor fluctuations in the glottal source spectrum that lead to varying estimates of the glottal spectrum captured in the dedicated coefficients of the model underlying the IAIF method. As a consequence this could lead to the observed deviations in the estimates of higher formants.

With respect to the lower formants it can be observed that all methods appear to yield estimates with a low variance. In this particular example, the second formant, F2, seems to have a higher gain compared to the first formant. All three methods estimate this formant to occur at a frequency between the seventh and eighth harmonic. When taking a closer look one may observe that the IAIF estimates of F2 tend to be slightly below the LP F2 estimates on average, with the F2 estimates of the proposed method sandwiched in between these two, also on average. The estimates of the CPLP method indicate a degraded performance with respect to what one would have expected after observing the results on the synthetic signals. Naturally occuring environmental noise inherent to the recorded signal might be an explanation for this degradation, since this appeared to affect this method severly in previous experiments (see Sec. 4.5.4).

The results of the estimation of the source related LF parameters, t_p, t_e and t_a from signal

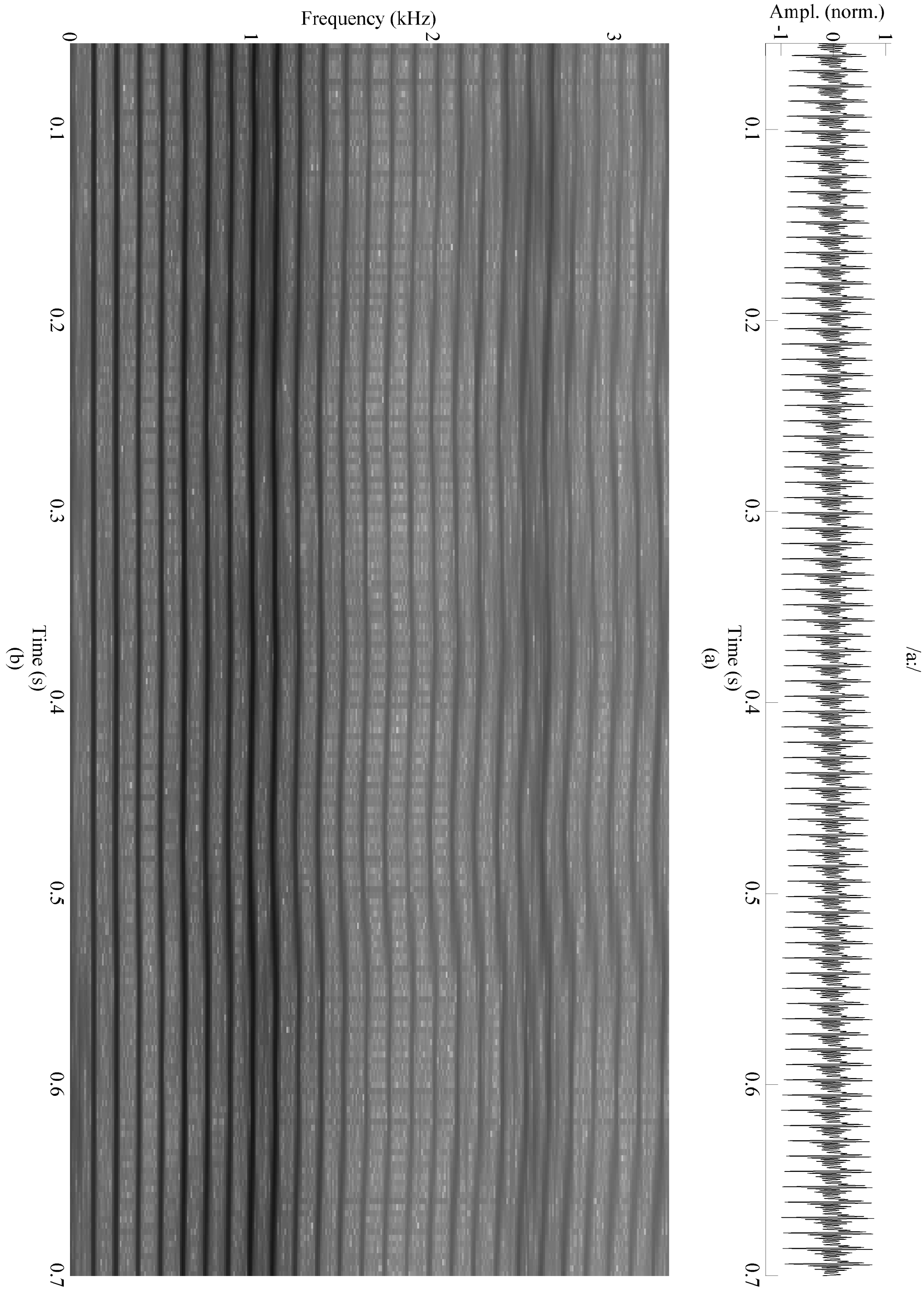


Figure 5.2: Time-domain signal *(a)* and spectrogram *(b)* of *Signal A*.

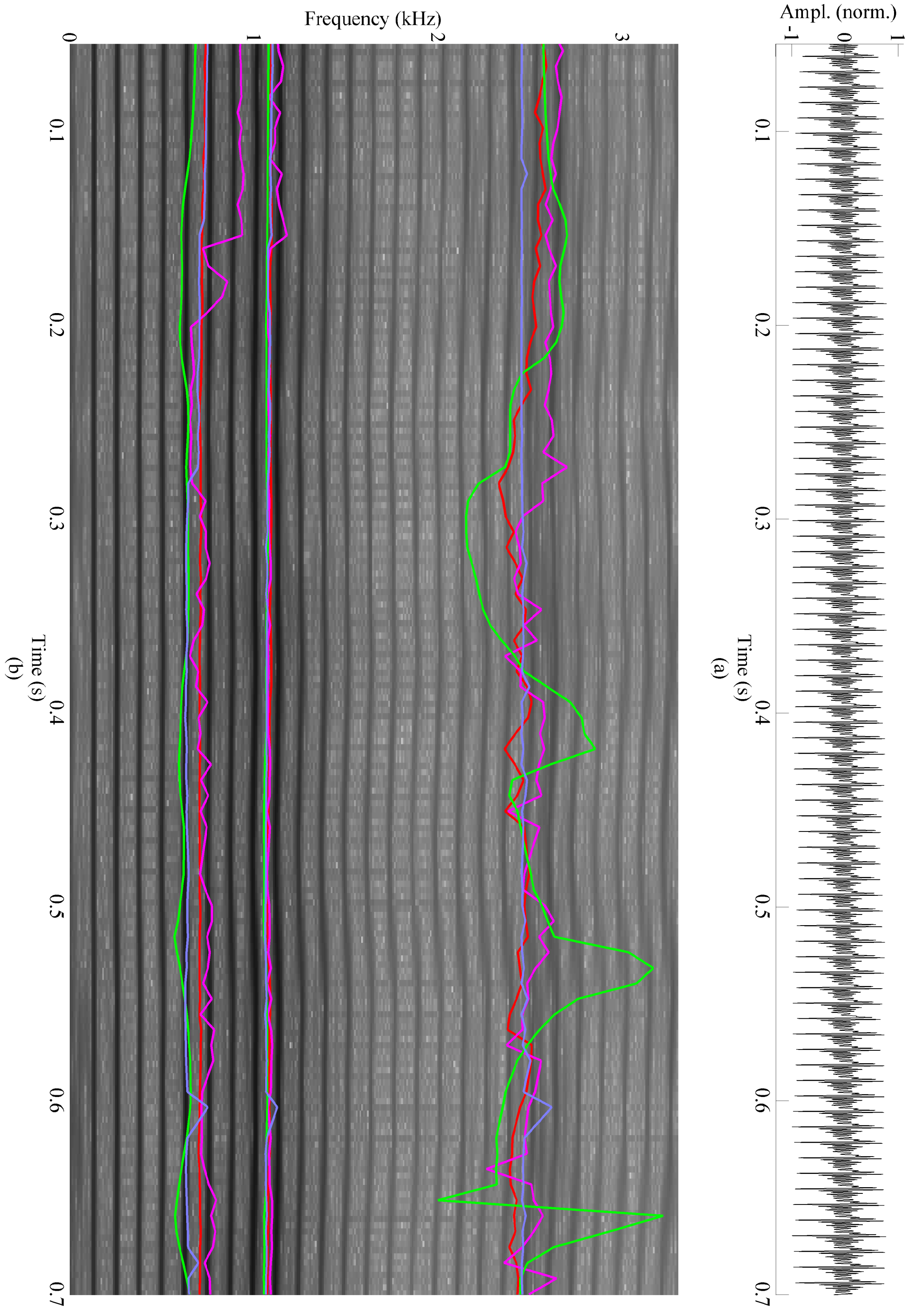


Figure 5.3: Time-domain signal *(a)* and spectrogram *(b)* of *Signal A*, overlaid with the formant trajectories of the first three formants, estimated using the proposed method (blue lines), LP (red lines), IAIF (green lines) and CPLP (magenta line).

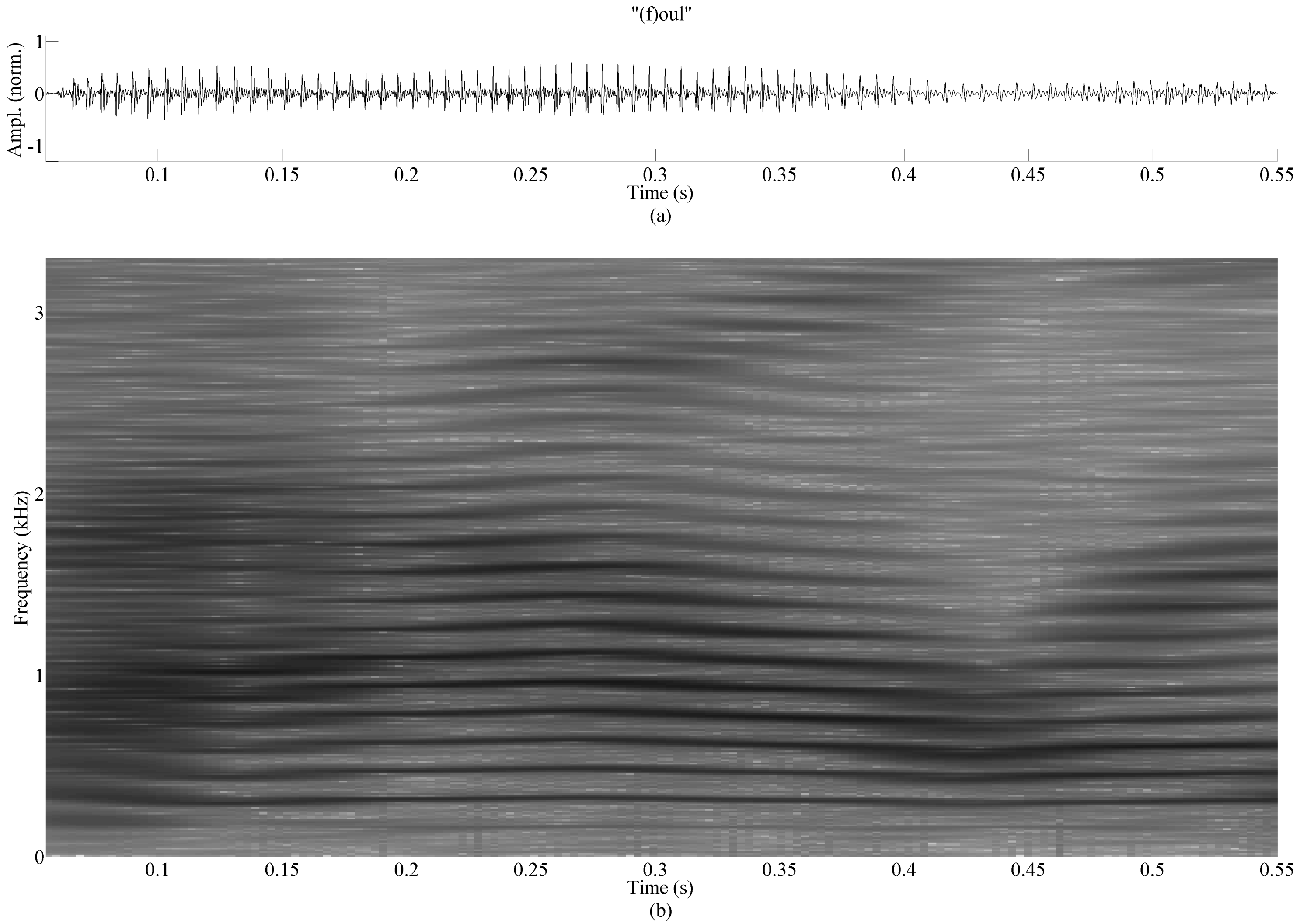


Figure 5.4: Time-domain signal *(a)* and spectrogram *(b)* of *Signal B.*

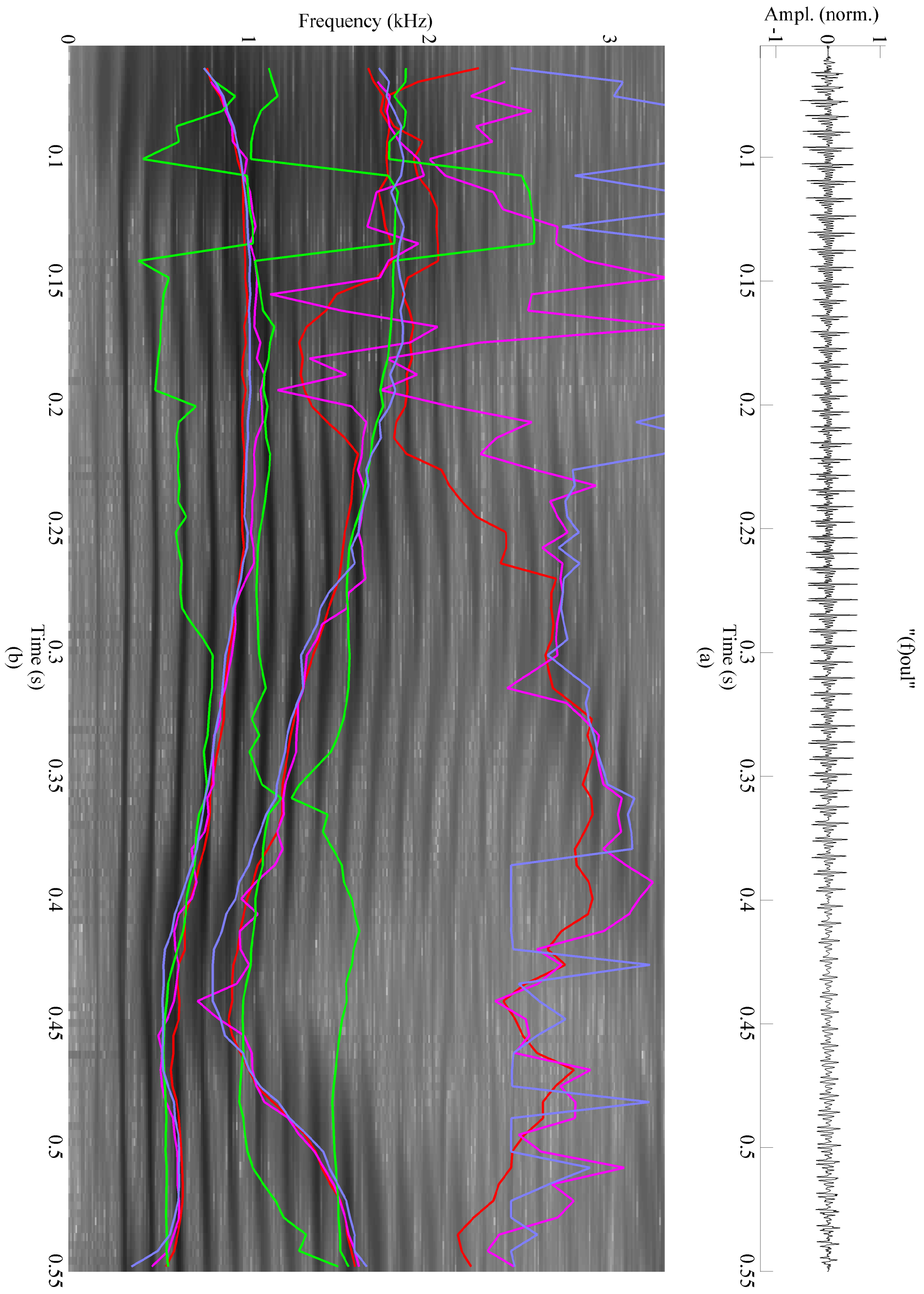


Figure 5.5: Time-domain signal *(a)* and spectrogram *(b)* of *Signal B*, overlaid with the formant trajectories of the first three formants, estimated using the proposed method (blue lines), LP (red lines), IAIF (green lines) and CPLP (magenta line).

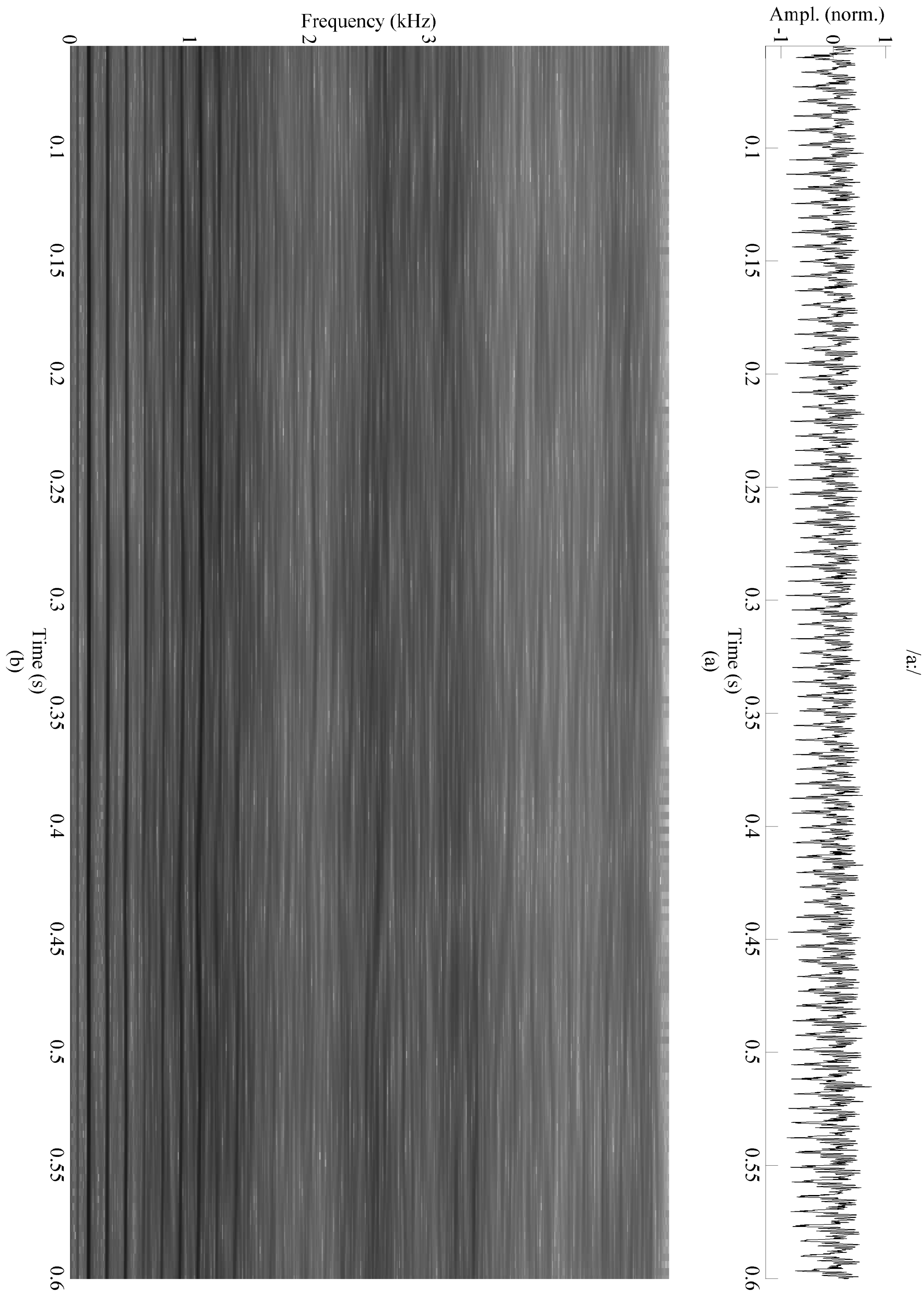


Figure 5.6: Time-domain signal *(a)* and spectrogram *(b)* of *Signal C.*

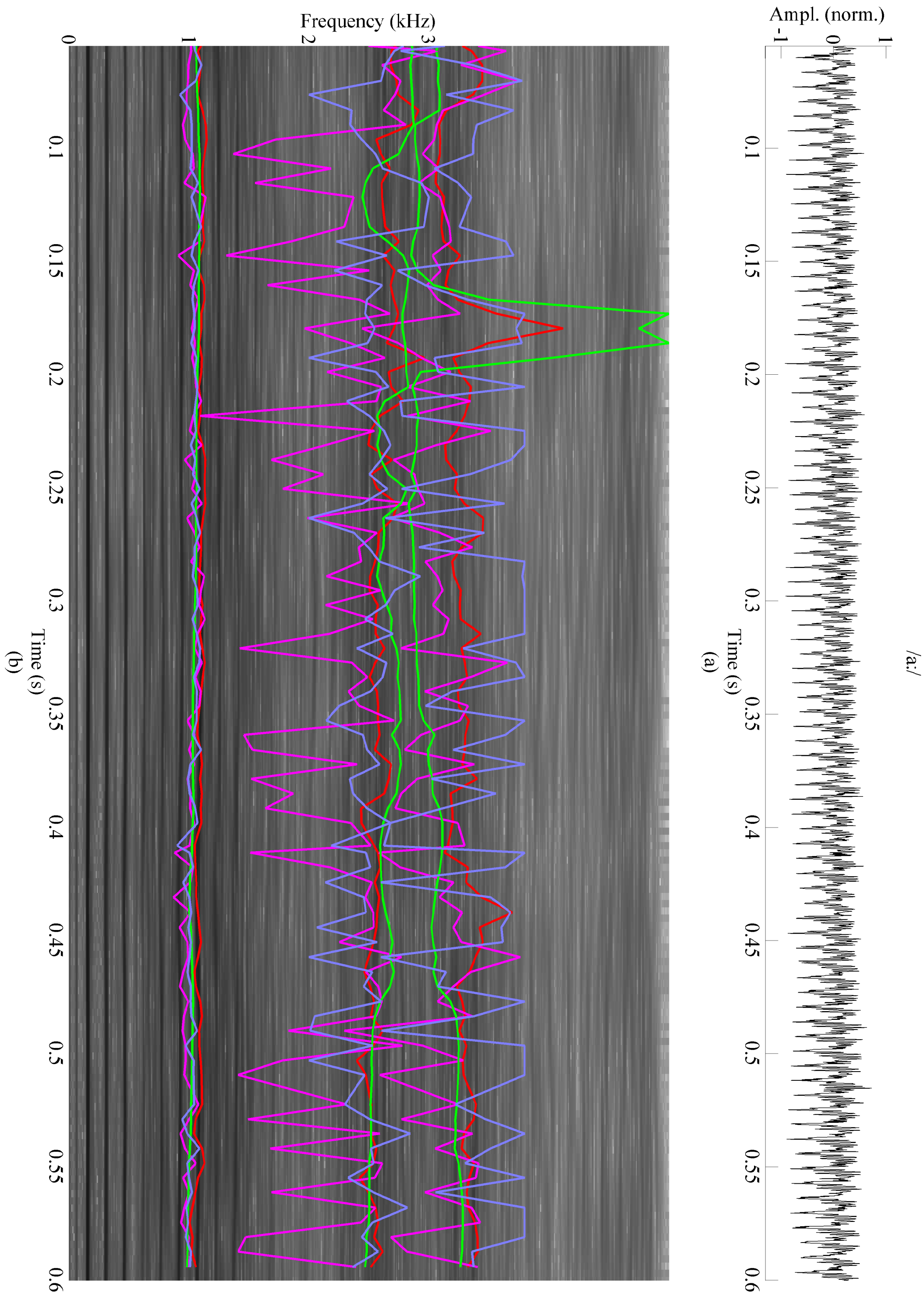


Figure 5.7: Time-domain signal *(a)* and spectrogram *(b)* of *Signal C*, overlaid with the formant trajectories of the first three formants, estimated using the proposed method (blue lines), LP (red lines), IAIF (green lines) and CPLP (magenta line).

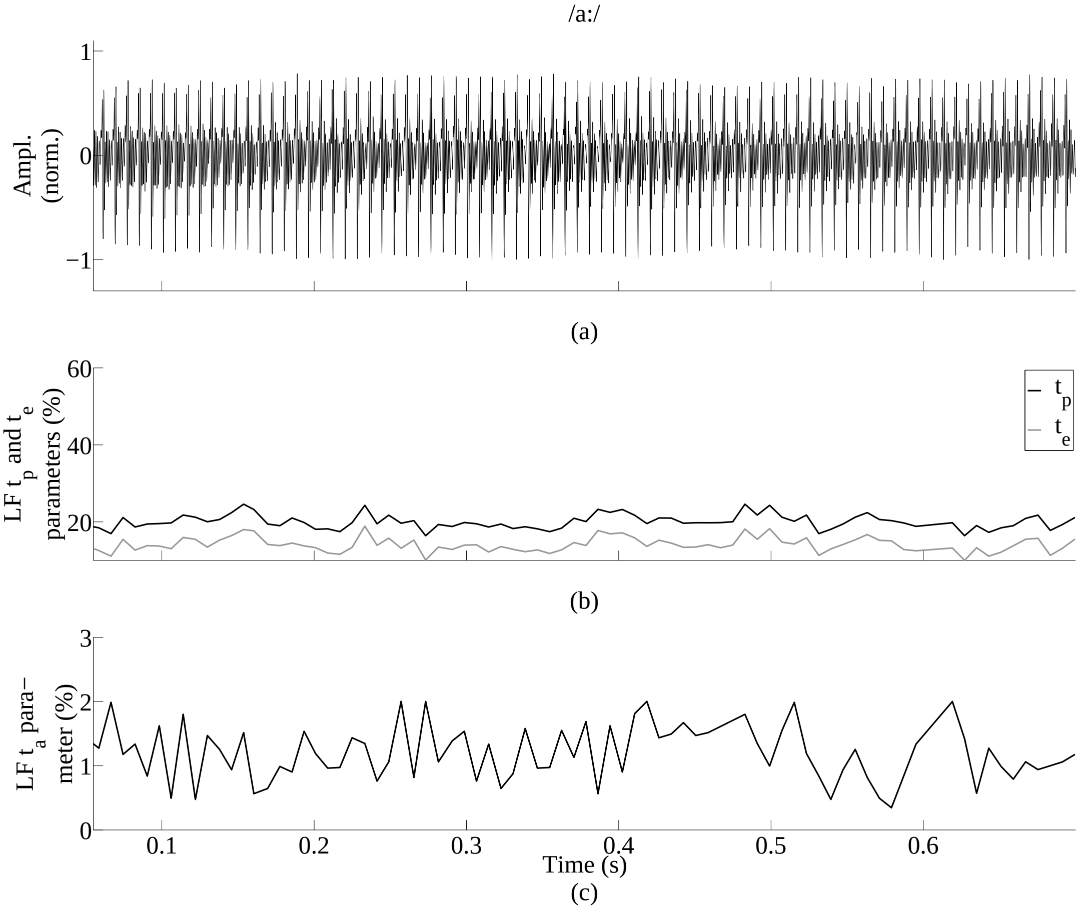


Figure 5.8: Time-domain signal *(a)*, LF model open phase parameters *(b)* and LF model return phase parameter *(c)* as estimated from *Signal A*. In panel *(b)*, LF parameter t_p always has a smaller value than LF parameter t_e, conforming to the inequality constraints formulated in Eq. 4.8.

A are shown in Fig. 5.8. The two parameters defining the shape of the open phase of the glottal cycle, t_p and t_e, both are continuously estimated in a relatively narrow range, indicating a relatively reliable estimation for this example. From examining the speech waveform in panel *(a)* it appears that the glottal open phase in this particular example is rather short, which is well reflected in the estimated values. Both values are relatively low, averaging around $t_p \approx 10\,\%$ and $t_p \approx 20\,\%$, which are typical values for voice sources with a short open phase. The qualitative analysis indicates a similar result for the parameter t_a, although higher variance in the estimation may be observed. In fact, this high variance has also been observed in other studies (Fu and Murphy, 2006) and may be expected due to the very short duration of t_a. The average value $t_a \approx 1.5\,\%$ is a typical value for a normal, modal voice source.

Signal B

Signal B represents a very interesting example for the evaluation of formant estimation methods. Over the course of the 0.55 s, the *phonetic vowel target position* changes twice, while also a slight alteration of f_0 can be observed. The shading in the spectrogram in Fig. 5.3 clearly indicates the transitions of the first and second formants. Notably, the second formant descends from a value of about 2 kHz to 1 kHz during the pronunciation of the dipthong */aů/* (approximately from 0 s to 0.45 s). Thereafter, the formant aims for a new *phonetic target position* resembling the lateral approximant */l/*. The first formant follows the general trend of F2, but to a reduced extent. Note that in this example, the third formant visually only appears during a limited duration ranging from 0.22 s to 0.4 s. Before and after that time range, the third formant is covered by the environmental noise and is not visible in the spectrogram.

Regarding the estimation results, one may observe that formant frequency estimates of the proposed method appear to follow very smoothly the hypothetical trajectories of the first two formants. The estimated values show a largely monotonic increase or decrease, following the hypothetical trajectories. No major outliers may be spotted. This clearly supports the observations of the experiments with synthetic signals. The proposed method yields reliable results for signals with a strong deterministic source. With respect to the third formant it appears that the proposed method picks up the trajectory of the resonances during the time range mentioned above, during which also a harmonic pattern is visible in the spectrogram at the location of F3 (0.22 s to 0.4 s). During the time before and after this time-span, the estimate seems to randomly pick up values or it runs into the boundary constraints. This can be expected, since also in the spectrogram no resonances are visible during these time instances.

The other three methods clearly struggle more in estimating the formants. The LP method for a large part also estimates values close to the hypothetic trajectories of the first two formants, but it exhibits a higher variance, that possibly can be attributed to the underlying harmonic structure of the source. Note, for instance, how the LP estimates of F1 and F2 around 0.45 s are attracted to harmonics of the source. Also, around 0.18 s, the third formant estimate models the actual second formant and the F2 estimate models an f_0 harmonic between F1 and F2. The IAIF appears to be even more affected by its affinity to model f_0 harmonics. Various different settings of its parameters were tried improve the estimates, but it proved to be rather difficult to find better settings for this particular example. The CPLP method again appears to suffer largely from the environmental noise.

The results of the estimation of the source related LF parameters from signal B are shown in Fig. 5.9. Also for this example, the two parameters related to the open phase exhibit a smoother trajectory compared to the trajectory of the parameters t_a. In all trajectories, varying trends may be observed. The duration of the open phase (related to parameters t_p and t_e) generally seems to be longer compared to that found in signal A and eventually gets further increased during the pronunciation of the lateral approximant */l/*. The return phase, determined by the parameter t_a, is longer throughout the first part of the diphtong */aů/* and remains very short ($t_a < 1\,\%$) throughout the rest of the example. In the spectral domain, this corresponds to an

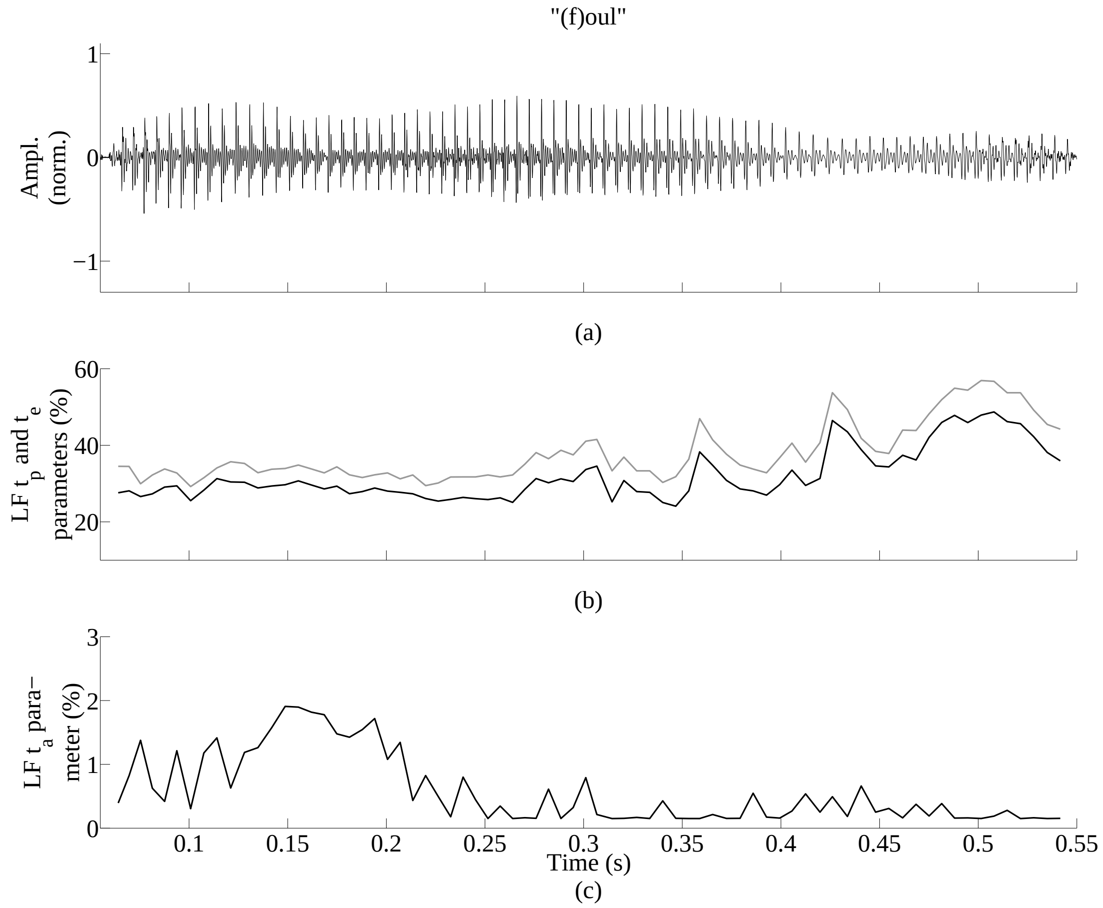


Figure 5.9: Time-domain signal *(a)*, LF model open phase parameters *(b)* and LF model return phase parameter *(c)* as estimated from *Signal B*.

increase in spectral tilt throughout this example.

Signal C

Signal C, with its band-limited harmonic excitation, proved to be a very difficult case for the proposed method. The alaryngeal glottal source generally exhibits more variance in its glottal period and in the glottal waveform shape of subsequent glottal cycles. This can be observed in the estimate of the first formant, where the proposed method shows a significant variance around a hypothetical F1 trajectory. Due to their inherent averaging, the other two methods estimate a smoother trajectory, possibly with a larger bias as sees in our experiments using the synthetic signals. Interestingly, it appears that the first and second formant of the vowel */a:/* are merged into a single formant.

Two formants appear at frequencies just below and above 3 kHz, respectively. All the four methods have difficulties in obtaining a reliable estimate of these formants, with the IAIF method exhibiting the smoothest trajectory, but the LP method apparently following closer the hypothetic trajectory (in particular visible at the third formant in the last mid-segment of the example). The glottal excitation in these high frequency bands is largely made of aspiration noise, since no energy due to the deterministic source is present at these frequencies. As one may expect after the previous experiments, the proposed method has large difficulties in obtaining reliable estimates of the higher formants. It relies mainly on a reasonably strong deterministic glottal excitation.

The results of the estimation of the source related LF parameters from signal C are shown in Fig. 5.10. The estimation apparently exhibits outliers due to estimation errors, which may be explained by the reduced SNR found in alaryngeal voice sources. Qualitatively it appears that all source related values are globally increased compared to the other two, laryngeal voice source examples. The relatively large difference between the average values of t_p and t_e in combination with the large average value of t_a also hints at a slowly varying glottal source waveform, just as one would expect in an alaryngeal voice source. The waveform often is observed to have a sinusoidal shape.

5.3 Conclusion

- The accuracy of estimated formants of the proposed method was compared to three other methods, LP, IAIF and CPLP. The evaluation was carried out objectively using speech produced by a physical model of speech production and qualitatively using healthy and pathological, real speech signals. In addition, the glottal source waveform obtained using inverse filtering was compared qualitatively for the speech simulated by the physical model.

- The glottal source waveform obtained using the proposed method exhibits a shape captured by the glottal source model used for the optimization, but also shows features commonly attributed to non-linear feedback mechanisms between the vocal tract and the inner-glottal air flow. The proposed method is capable of preserving the general structure of the glottal waveform, but also retains its fine details.

- The objective comparison of the formant estimates using physically modelled speech revealed a sensitivity of the proposed method to source modelling errors. It was observed that glottal source spectra, which show large frequency-dependent deviations from the constant spectral decay assumed by the LF source model in high frequency bands, may lead to spurious formant frequency estimation errors. Such large, cycle-specific glottal source spectra fluctuations were not observed in the real speech signal examined. The question, whether this phenomenon and the resulting formant estimation errors may be observed in real speech signals, remains unanswered by this work for the moment.

- A qualitative examination of the estimated formants on real speech signals confirmed the results from the previous chapter. The proposed method is capable to extract smooth and presumably accurate formant trajectories given that the deterministic source component is of sufficient energy in frequency bands where formants are to be estimated.

The proposed method appears to perform better than the frame-based state-of-the-art methods in those cases and also better than the other pitch-synchronous method, CPLP. This method in particular appears to suffer significantly from environmental noise, as observed also previsouly.

- The proposed method appears especially reliable in the presence of environmental noise.
- The proposed method is not capable to reliably estimate formants if the excitation is constituted of non-deterministic signal components only.
- Temporal parameters of the glottal source model extracted from all three examples indicate that their estimation is consistent and mostly reliable. Trends in their trajectories may be observed that correlate with visually observable cues from the speech waveform and also correspond to values that one would expect theoretically in the respective examples.

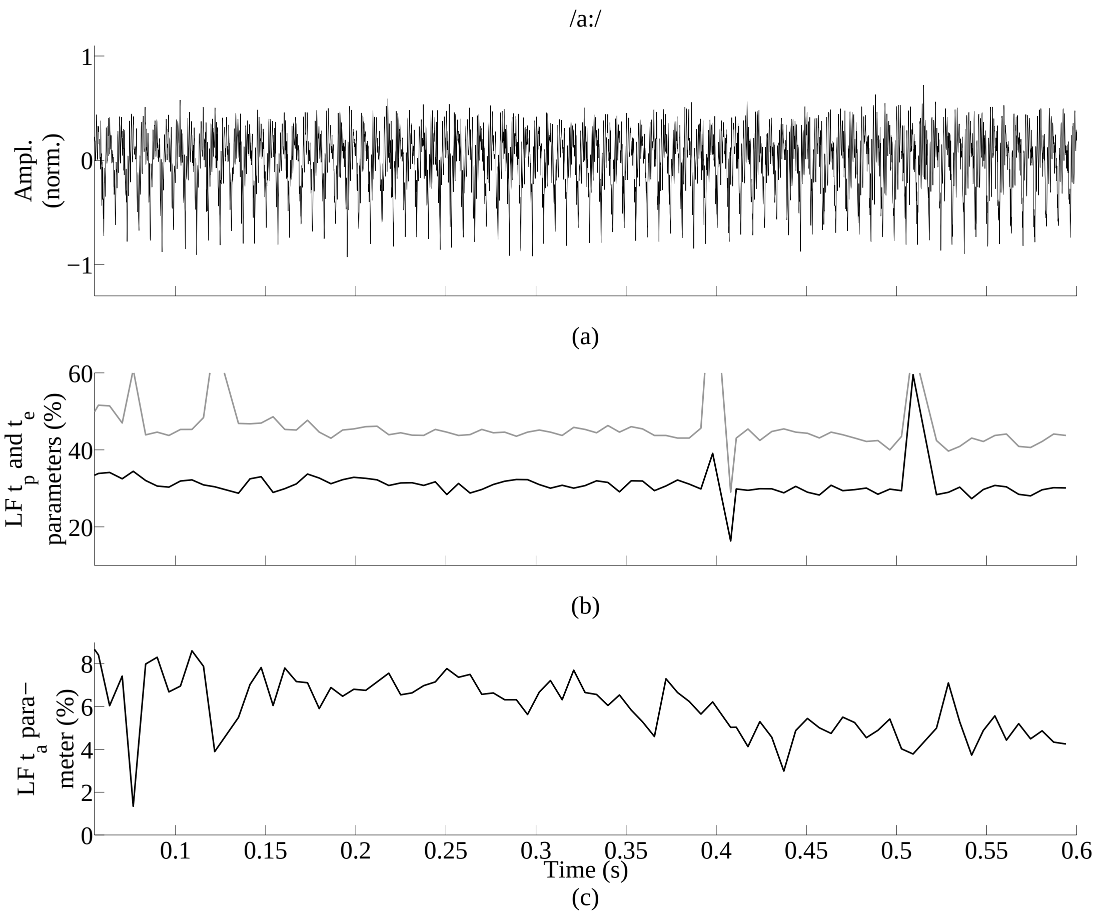


Figure 5.10: Time-domain signal *(a)*, LF model open phase parameters *(b)* and LF model return phase parameter *(c)* as estimated from *Signal C*.

6 Prosody Restoration

In this chapter we describe a method that was developed in an early phase of this thesis. The method is intended for use in a real-time scenario in a device for the restoration of authentic, f_0-related characteristics in pathological speech uttered by subjects with laryngeal disorders. The original speech signal is acquired and analysed by the device and a speech signal with improved, healthy-like features is reconstructed. For the reconstruction, different cues of the original, acquired signal are used.

In order to obtain a perceptionally superior voice source in the reconstructed speech signal, the pathological excitation is replaced by a concatenation of healthy, glottal waveform patterns, which are randomly chosen from a reference database. Furthermore, to increase the naturalness of the f_0-variability in the reconstructed voice source, a multi-resolution approach is used to determine the instantaneous intervals between subsequent reference patterns. In particular, f_0-variability is reconstructed using different cues for its reconstruction at different time scales. The long-term f_0-trend is estimated by a method called adaptive wavetable oscillator (AWO), a novel, reliable and computationally efficient f_0-estimation method adapted to the particularities of pathological voice sources. Furthermore, the middle-term f_0-variability is restored through its correlation with speech intensity or loudness. For the reconstruction of short-term f_0-variability, a statistical noise model is used to induce jitter based on the instantaneous loudness of the speech signal. Two authentic features are used to assess the performance of the method, namely breathiness and prosody. Preliminary results indicate that breathiness of the restored signal is reduced and prosody related features are improved. On the other hand, it also became apparent that better methods for source-filter separation were required to obtain more reliable VTF estimates, which was the main motivation for the methods presented in the previous chapters.

6.1 Introduction

The degree of degradation in pathological voices often engenders a decrease in a patient's speech intelligibility and thereby a severe limitation in his social life and oral interaction (Weinberg, 1986). A particularly severe degradation of natural vocal excitation may be observed in subjects who have undergone laryngectomy (Williams and Barber Watson, 1987; Most et al., 2000; Moerman et al., 2004). Laryngectomy is the common treatment after diagnosis of larynx cancer in an advanced stage and constitutes the partial or total excision of the larynx. This significantly reduces the patient's ability to produce voiced sounds due to the reduced or missing vocal fold functionality (van As, 2001; Pindzola and Cain, 1988). During speech rehabilitation, patients are encouraged to learn alternative voicing methods, but the result usually is a noisy and intermittently obstructed voice source. It lacks power and f_0-variability and typically has an unnaturally low value of f_0. Alaryngeal voice sources are often also perceived as not gender-discriminative and to have a largely breathy voice quality. Also, alaryngeal speakers often find it difficult to express prosody. In accordance with the widely accepted source-filter-model in healthy speech processing (Fant, 1981), the vocal folds are an essential component of the speech production process. They provide an excitation signal with distinctive, periodical energy concentrations. This signal undergoes further spectral shaping due to resonances in the oral and nasal cavities as well as the lip radiation function (see Section 2.1). In contrast, the alaryngeal voice excitation consists of a flawed, distorted excitation signal where the glottal peaks are much less concentrated in the time domain. This results in an unpleasant and unnatural voice with a fluctuating and often intermitted periodicity. In addition, due to the lack of control over or even absence of laryngeal muscles, the speaker loses most of its control over f_0-variability.

In the past, several advanced voice source restoration systems and methods have been presented aiming at the improvement of the quality and intelligibility of alaryngeal speech. In (Qi et al., 1995) methods based on linear prediction for analysis and synthesis were used to enhance the perceived, subjective voice quality. In (Bi and Qi, 1997), modified voice source conversion methods combined with formant enhancement were utilized to reduce the pathological speech signal's spectral distortions. In (del Pozo and Young, 2006) a voice restoration system is described that synthesizes speech using f_0 information obtained by an electroglottograph (EGG) and a jitter reduction model. In (Vetter et al., 2006), a system is presented that reconstructs healthy speech from alaryngeal speech by replacing its pathological excitation with a concatenation of glottal reference waveforms randomly chosen from a database extracted from healthy speakers. There, the intervals between successive healthy glottal waveforms are determined by instantaneous f_0-values extracted from the original, pathological speech signal. Promising performances have been obtained in terms of reduction of breathiness and increase of the average f_0-value, but the resulting speech lacks authenticity due to the significantly reduced f_0-variability in pathological speech.

To overcome these deficiencies, we propose a speech restoration approach based on a multi-resolution method with the aim of increasing the variability of the restored f_0. Natural prosody is restored by obtaining the intervals between subsequent glottal waves through a multi-resolution approach. The long-term variability is deduced from the f_0-trend in the original speech signal. Middle-term variability is restored with the help of the instantaneous signal intensity estimated from the acquired speech signal. The idea is based on the hypothesis that f_0 and the signal intensity envelope in natural speech are highly correlated (Rosenberg and Hirschberg, 2006) and an improved f_0-value may be reconstructed from the signal intensity.

Indeed, alaryngeal speakers are well able to modulate the intensity of their voice source using the pulmonary air pressure. In healthy subjects, this goes along with a modulation of the rate of vocal fold oscillation. Not so in the case of alaryngeal speakers, where the laryngeal physiology is largely altered. Therefore, additional f_0-variability is deduced from the instantaneously estimated voice intensity.

Short-term variability is restored using a statistical variation model, influenced by the signal intensity. The speech signal is reconstructed subsequently with the enhanced excitation and can be deployed in manifold applications such as voice enhancement systems or interactive support systems for voice rehabilitation and tutoring.

6.2 Speech Rehabilitation for Alaryngeal Speakers

During laryngectomy, the larynx including the vocal folds and the laryngeal muscles is partially or totally removed (van As, 2001). Generally, postlaryngectomy patients may regain means of verbal communication in two ways.

On one hand there exist electro-mechanical devices called electrolarynx that use a membrane to generate an external, synthetic speech excitation when held against the neck. This sets the air volume in the vocal tract into vibration and the patient can articulate in a natural manner. Unfortunately, the voice quality achievable with electrolarynx devices is low since there is no intuitive control over the fundamental frequency and voice quality parameters such as breathiness. The resulting speech sounds very monotonic and robot-like. Advantages of this method are its simplicity and short learning phase. The patient does not need to undergo additional surgery and can start communicating verbally almost immediately.

On the other hand, postlaryngectomy speakers may learn to use other tissues called *neoglottis* to substitute the functionality of the vocal folds. Commonly, during laryngectomy the remaining tissue is sutured in such a way as to promote oscillatory behaviour. In tracheoesophageal speech, the speaker utilizes pulmonary air to produce voicing with the substitutional tissues. The speaker may retain intuitive control over aspects of its voice source, but only to a very limited extent. The expression of prosody such as variations of the fundamental frequency or modulation of the voice quality is greatly reduced compared with healthy speakers. In addition, the aptitude of the remaining tissue to produce a rich, harmonic sound is very limited and its physical properties vary greatly among speakers and differ significantly from those of the vocal folds.

6.3 Characteristics of Pathological Speech

In healthy speech production, subglottal air pressure leads to a sudden, non-symmetric opening of the vocal folds and a release of this pressure. Various aspects of the glottal physiology and the air flowing through the glottis induce a self-sustained oscillation of the vocal folds, as described in Section 2.1.1. Varieties in the glottal physiology amongst humans lead to speaker-specific patterns for the opening and closing process as well as to the introduction of jitter in the period between subsequent glottal cycles. These effects amongst others lead to speaker-specific voice characteristics.

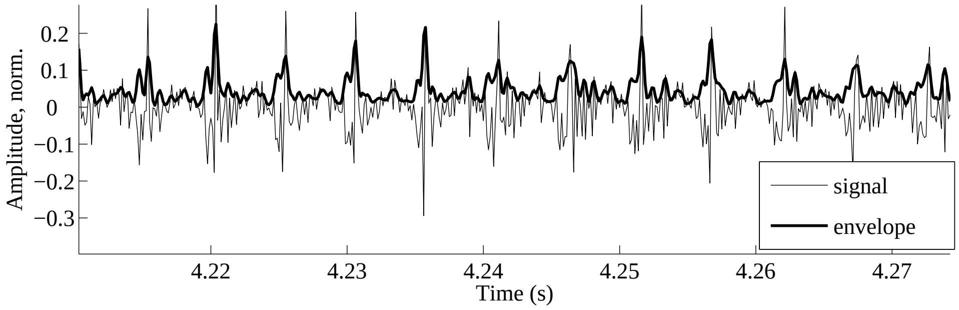
In comparison to healthy voice sources, voice production processes are not very well studied in alaryngeal speech (van As, 2001). Alaryngeal voice characteristics have been found to differ remarkably from that of healthy voice sources. Among subjects, the position and shape of the neoglottis vary significantly (Qi et al., 1995). Often incomplete glottal closure can be observed. Furthermore, the flexibility and controllability of the neoglottis lacks greatly when compared to a healthy glottis, especially due to the absence of the laryngeal musculature. The high mass of the neoglottis and low resistance to mucus aggregation influence the absolute value and stability of the fundamental frequency in a disadvantageous manner. The alaryngeal oscillator tends to break down intermittently (Kasuya et al., 1986). For example, observe the irregularities in the harmonics above 500 Hz in the spectrogram of the example of a sustained vowel in Fig. 6.5 I. Eventually, the resulting voice source has an unnaturally low and unstable f_0 and often is found to have a hoarse, croaky and breathy voice quality (Verma and Kumar, 2005). Figure 6.1 depicts segments of residual signals of laryngeal and alaryngeal speech inverse filtered with LP estimates of the VTF. This figure highlights the alteration of the produced harmonic excitation due to the changed physiologic conditions. The glottal wave patterns in the excitation of the healthy speaker are well focused in the time domain, whereas the excitation of the alaryngeal speaker appears merely as a modulated noise signal.

6.4 Multi-Resolution Voice Restoration

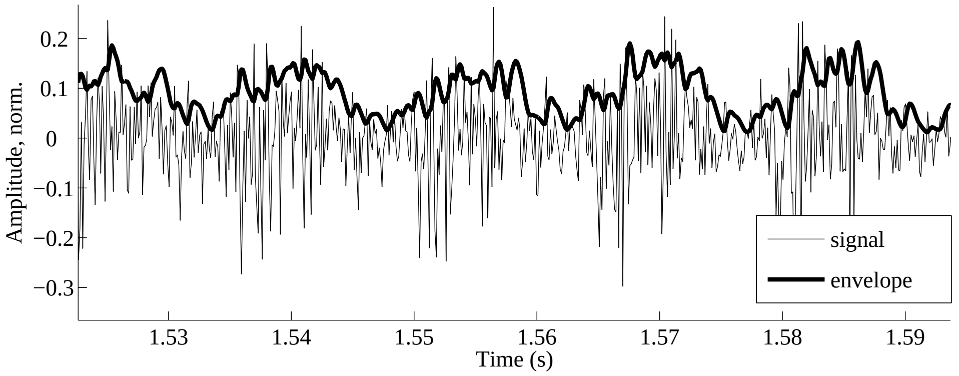
6.4.1 Method

A block scheme of the proposed method is depicted in Figure 6.2. The articulation information and the voice excitation are separated in a primer LPC-based signal analysis. Since we are only interested in restoring the voiced excitation signal, the obtained excitation signal is divided into voiced ($g_v(n)$) and unvoiced ($g_u(n)$) segments. Then the voiced excitation segments are replaced by concatenating healthy reference glottal cycle waveforms. These reference waveforms were previously extracted from laryngeal speakers and are randomly chosen from a respective database. The intervals between successive reference patterns determine the fundamental frequency of the reconstructed speech signal. As pointed out above, the fundamental frequency in pathological speech is degraded in terms of variability and stability and thus insufficient for a successful restoration of an authentic speech signal. To increase authenticity, the intervals between subsequent glottal waves are obtained through a multi-resolution approach on three different time scales:

- The long-term f_0 trend, $f_{0,LT}$, is estimated from the alaryngeal voice excitation by an instantaneous f_0-estimation method called adaptive wavetable oscillators, which is subsequently low-pass filtered ($f_c = 2\,\mathrm{Hz}$).



I: Healthy Excitation



II: Pathological Excitation

Figure 6.1: Segments of a laryngeal (a) and an alaryngeal (b) voice excitation signal (thin solid lines), obtained by inverse filtering using an VTF estimated using linear prediction, and their respective envelopes (bold solid lines).

- The middle-term f_0-variability, $f_{0,MT}$, is strongly related to prosody and is reconstructed by exploiting the correlation between f_0 variations and instantaneous signal energy. The trajectory $f_{0,MT}$ contains energy in the frequency band from $f_{c_1} = 2\,\text{Hz}$ to $f_{c_2} = 8\,\text{Hz}$.

- Short-term f_0-variability $f_{0,ST}$ is introduced through a random signal $e_{ST}(n)$ and a statistical model to mimic the presence of f_0-dependent jitter, as in healthy speech. The trajectory $f_{0,ST}$ contains energy at frequencies above $f_c = 8\,\text{Hz}$.

Finally, the improved excitation signal is recombined with the unmodified unvoiced speech segments and then filtered with the previously estimated articulation information to form a reconstructed speech signal, $s'(n)$.

6.4.2 Long-Term f_0-Estimation

The objective of the long-term f_0-estimation is to grasp what remains of the f_0-trend in the alaryngeal speech. The selection of method for the extraction of the fundamental frequency of a specific signal depends on different characteristics of the signal itself:

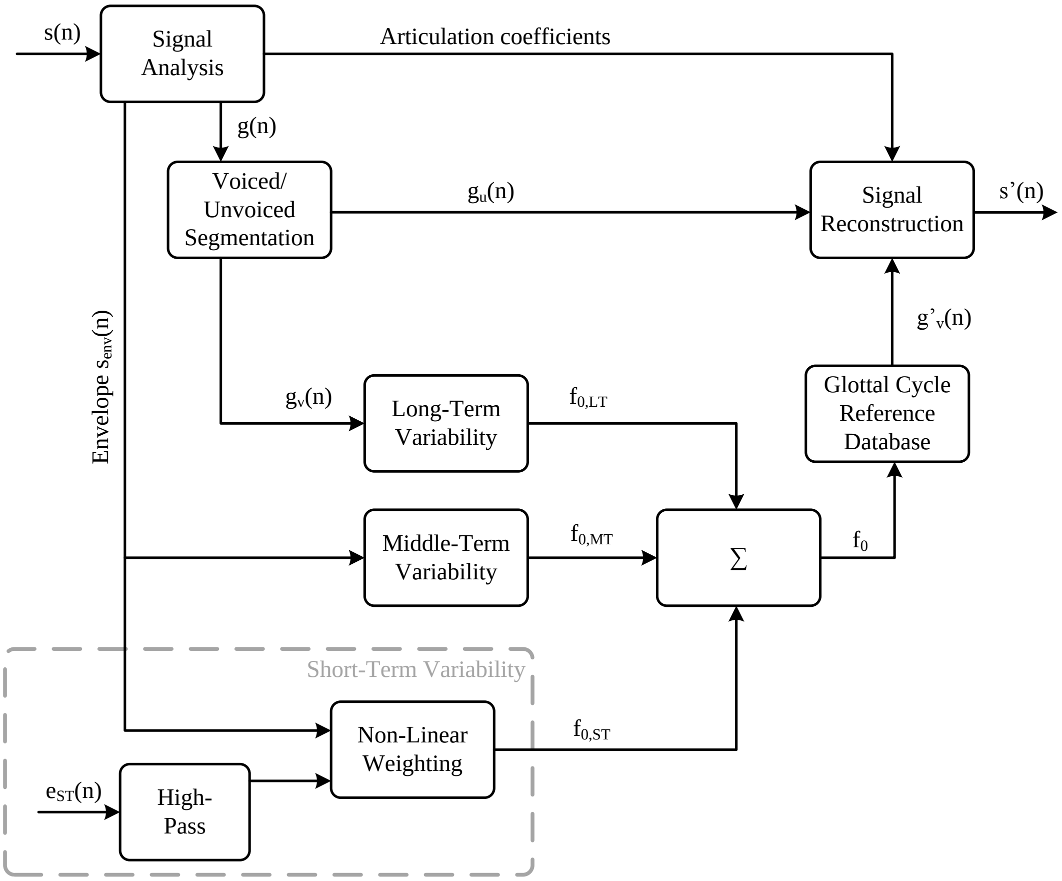


Figure 6.2: Block diagram of the multi-resolution f_0-restoration method.

- The nature of the signal in terms of time-frequency distribution
- The amount and characteristics of additional harmful background noise
- The affordable computational complexity.

In general, f_0-estimation methods can be classified into event detection methods and short-term averaging methods (Kepesi and Weruaga, 2005). Event detection methods such as for example zero-crossing (Gerhard, 2003) or threshold-guided maxima localization (Gerhard, 2003) are computationally inexpensive and yield high performance for well-shaped signals in low-noise environments. Signals with higher harmonic complexity or increasing noise level require more advanced methods such as the matched filter method (Turin, 1960) or auto-correlation method (Un and Yang, 1977). They are based on short-term averaging and have generally a higher computational complexity. More advanced methods with yet increased computational complexity, decompose the signal into its eigenspace components (Murakami and Ishida, 2001). Joint approaches (Mitev and Hadjitodorov, 2003) combine three different methods, namely in time, frequency and cepstrum domain.

For the method presented in this chapter, the focus is on the *efficient* utilization of the given computational resources. We propose to use a new f_0-estimation method taking into account the demand for low computational load and for the pertinence and simplicity of fixed-point real-time implementation. The method is based on adaptive wavetable oscillators, a method recently published in (Arora and Sethares, 2007). An evaluation of the method comparing it with other state-of-the-art methods for fundamental frequency estimation was presented in (Schleusing et al., 2009).

The AWO constitutes a time-frequency method combining wavetables and adaptive oscillators. Wavetables generate periodic output signals by cyclic indexing of a lookup table that stores a single period of the waveform. Adaptive oscillators synchronize their output to both frequency and phase of the input signal. The indexing parameters of the AWO are determined by optimizing a well-defined cost function such that the error between the wavetable output and an incoming, periodic signal is minimized.

The first step in the design of an AWO requires the selection of an appropriate pattern. This pattern should represent a high similarity with the signal pattern to be extracted and is stored in a wavetable as numerical, digital information. With respect to the above consideration, we use the energy distribution of the glottal excitation envelope as input (see Figure 6.1 II) and a Gaussian function as wavetable pattern. As one can observe, the envelope of energy during glottal patterns of the excitation signal has a high similarity with a Gaussian shape. A Gaussian function is easily controllable with only a few parameters such as a time index n, a phase offset in samples β and a temporal width σ:

$$w(n) = e^{-\frac{1}{2}\frac{(n-\beta)^2}{\sigma^2}} \tag{6.1}$$

Cyclic sampling is used to generate a periodical reference signal $v(n)$:

$$v(n) = w(k(n) \mod N) \tag{6.2}$$

where $k(n)$ is the cyclic sampling index

$$k(n) = (k(n-1) + \alpha) \mod N \tag{6.3}$$

$k(0)$ is initialized to 0, $x \mod N$ is the remainder operator, and α is the sampling step size determining the sub-sampling rate of the wavetable pattern. The control parameters of the periodic output of Equation 6.1 are adaptively updated by using well understood gradient techniques (Haykin, 2001). The output of the wavetable oscillator thus is locked to the input signal and the parameter α is related to the fundamental frequency of the alaryngeal excitation signal by $\alpha/(NT_s)$, with $T_s = 1/f_s$ being the sampling period. The phase of the resulting signal is determined by an offset β of the sampling index. The adaption of the indexing parameters is achieved by minimizing a well-defined cost function that gauges the error between the wavetable output and an incoming, periodic signal:

$$J(n) = s(n)v(n) \tag{6.4}$$

where $s(n)$ is the envelope of the extracted speech excitation signal.

Assuming that the phase and frequency of the input signal vary slowly over time one can follow these changes by moving the argument of the cost function slowly in the direction of the derivative:

$$\alpha(n+1) = \alpha(n) + \mu_\alpha \left.\frac{\partial J}{\partial \alpha}\right|_{\alpha=\alpha(n)} \tag{6.5}$$

and

$$\beta(n+1) = \beta(n) + \mu_\beta \left.\frac{\partial J}{\partial \beta}\right|_{\beta=\beta(n)} \tag{6.6}$$

It can easily be seen that the gradients $\frac{\partial J}{\partial \alpha}$ and $\frac{\partial J}{\partial \beta}$ are similar up to a constant. Indeed, they include the partial derivative of w,

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial w}\frac{\partial w}{\partial x}, \tag{6.7}$$

which is typically stored in a wavetable of N samples to minimize the computational load. The learning gains μ_α and μ_β should be chosen such that the oscillator can change rapidly enough to follow changes in the fundamental frequency and minimize noise influences. Since the adaptation of the frequency is much more sensible than that of the phase, μ_α should be much smaller than μ_β.

In (Schleusing et al., 2009) the AWO method was compared to two other common f_0-

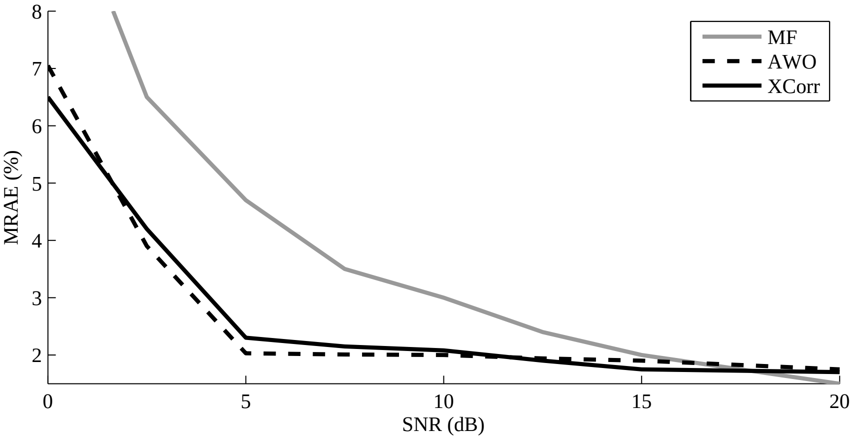


Figure 6.3: Mean relative average error of the f_0-estimation methods for the healthy speech signal with different levels of AWGN.

estimation methods; the correlation method and the matched filter method. A quantitative validation was performed on healthy, phonetically equilibrated French sentences with additive white Gaussian noise at SNRs ranging from $-10\,$dB to $20\,$dB. To obtain an objective performance assessment we evaluated first the most likely fundamental frequency during voiced segments as the median value of the estimations from the three different methods at each sample. Quantitative performances were then assessed as the Mean Relative Absolute Error (MRAE) between an estimation result from a specific estimator and the most likely value. The results displayed in Fig. 6.3 confirm the results on healthy speech signals. The AWO and XCorr perform better than the the MF method, particularly at low levels of SNR. For pathological voice sources, very likely instantaneous fundamental frequencies to serve as a reference are not available. Thus, a subjective validation by listeners was carried out, based on the Mean Opinion Score (MOS) (Virag, 1996) of a signal generated by the presented method.

The results of the subjective evaluation displayed in Table 6.1 highlight that XCorr outperforms AWO and MF in the mean with respect to listener specific subjective evaluation. However, an analysis of the variance of the XCorr and AWO methods yields a p-value of 0.61, which suggests that from a statistical point of view this result is not significant (Papoulis, 1989). The performance of the matched filter method drops due to insufficient reliable support points for its restoration, which made it impossible to follow the changes in the fundamental frequency.

The above results indicate that the proposed AWO method performs similar to the correlation method under a variety of experimental conditions. In fact, the correlation method is outperformed by the AWO method when applied to healthy voice sources and when significant amounts of additive background noise is present. When applied to pathological voice sources, the performances of AWO and the correlation method are nearly the same. Both methods reconstruct f_0 in the speech analysis and restoration system to a quality, where it is rated by listeners between fair and good. However, as the computational load of the AWO method is much lower than that of the correlation method, AWO appears to be a more promising

Table 6.1: Mean Opinion Score (mean ± standard deviation) of seven listeners assessing the performance of specific f_0-estimation methods as a preprocessing unit to a voice restoration method for pathological voice sources. Applied MOS-scale: bad-1, poor-2, fair-3, good-4 and excellent-5.

Method	XCorr	AWO	MF
MOS	3.7 ± 0.49	3.6 ± 0.53	1.7 ± 0.49

f_0-estimation method for real-time fixed point implementation on embedded platforms.

6.4.3 Middle-Term f_0-Restoration

Middle-term f_0-variability, $f_{0,MT}$, is restored by exploiting the correlation between f_0 and the signal envelope at this time scale. It has been shown that prosody is not only strongly related to variations in f_0, but also to variations in the envelope of the speech signal (Rosenberg and Hirschberg, 2006). Figure 6.4 shows a segment of a healthy speech signal *(a)*, the estimated f_0-trajectory *(b)* and variations in the signal's intensity envelope *(c)*. Clearly, the correlation between the signal's intensity and the resulting f_0 can be observed. The key point of the proposed method is to infer variations in f_0 from variations in the signal envelope of the alaryngeal speech signal. This will allow an alaryngeal speaker to modulate the fundamental frequency of the restored voice source signal by varying the intensity or loudness of his speech and to regain some of the lost dynamic range of f_0. To implement this correlation between intensity and f_0, the segment-wise estimated signal envelope is bandpass-filtered ($2-8\,\mathrm{Hz}$) and then used to construct the middle-term f_0-variability. Thereby, the pathological speaker is given a means to intuitively manipulate f_0 of the restored speech signal by manipulating the intensity of the produced speech.

6.4.4 Short-Term f_0-Restoration

An important characteristic of natural voice sources are small imperfections, such as non-deliberate variations in f_0. Human perception expects this short-term variability in natural speech and the lack of it is perceived as unnatural, unpleasant and buzz-like. In the proposed approach, short-term f_0-variability $f_{0,ST}$ is induced through high-pass-filtered ($f_c = 8\,\mathrm{Hz}$) and weighted additive white Gaussian noise (AWGN), added to the f_0-trajectory. The weighting of this AWGN is determined through a signal-envelope-dependent, nonlinear weighting inspired by recent findings in healthy subjects (Brockmann et al., 2008). In healthy voices, jitter in f_0 was found to be constantly low during voicing with sound pressure levels of $70-75\,\mathrm{dB}$ and above. At lower intensity levels though, jitter was found to steadily and sharply increase with falling sound pressure levels. The resulting f_0 is increasingly unsteady with decreasing intensity of the produced voice source. We model this non-linear behaviour with a piecewise

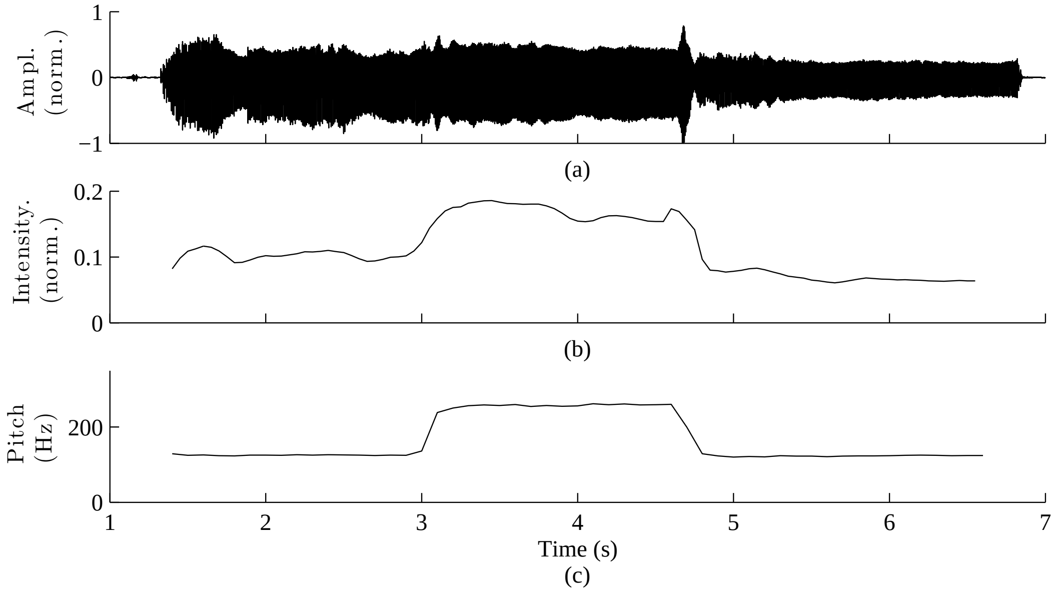


Figure 6.4: Illustration of the correlation between the speech signal envelope, which corresponds to the signal intensity (middle panel), and f_0 (bottom panel) in a healthy speech signal (top panel).

linear function.

$$f_{PL}(s_{env}(n)) = \begin{cases} 0.1 & \text{if } s_{env}(n) \geq \gamma \\ 0.1 + 0.9\frac{\gamma - s_{env}(n)}{\gamma - \kappa} & \text{if } \gamma > s_{env}(n) > \kappa \\ 1 & \text{if } \kappa \geq s_{env}(n) \end{cases} \tag{6.8}$$

where $s_{env}(n)$ is the logarithm of the averaged instantaneous signal envelope normalized with respect to the given acoustical configuration, γ and κ have been adjusted with respect to subjective listening tests.

6.5 Results

An evaluation was performed to assess the successful restoration of authentic characteristics from pathological speech to a higher quality. For the evaluation, a sustained sound of a vowel */a:/* of a pathological, male speaker with varying f_0 was recorded at a sampling rate of 8kHz. From this signal, a reconstructed speech signal was produced using the method as described in Sect. 6.4.1, implemented in the Matlab programming language (The Mathworks, 2006). Twelve amateur listeners quantified the perceived improvement in terms of prosody and breathiness using a mean opinion score (MOS) by listening to the restored speech signals. All listening tests were performed using ordinary headphones. The relative contributions of short-term, middle-term and long-term f_0-variabilities to the improved speech quality were assessed using three different system configurations, where f_0 was restored from:

Table 6.2: Mean Opinion Score (mean ± standard deviation) of twelve listeners that assessed the quality of three different restored voice sources. Applied MOS-scale: highly increased-1, no alteration-3, highly decreased-5.

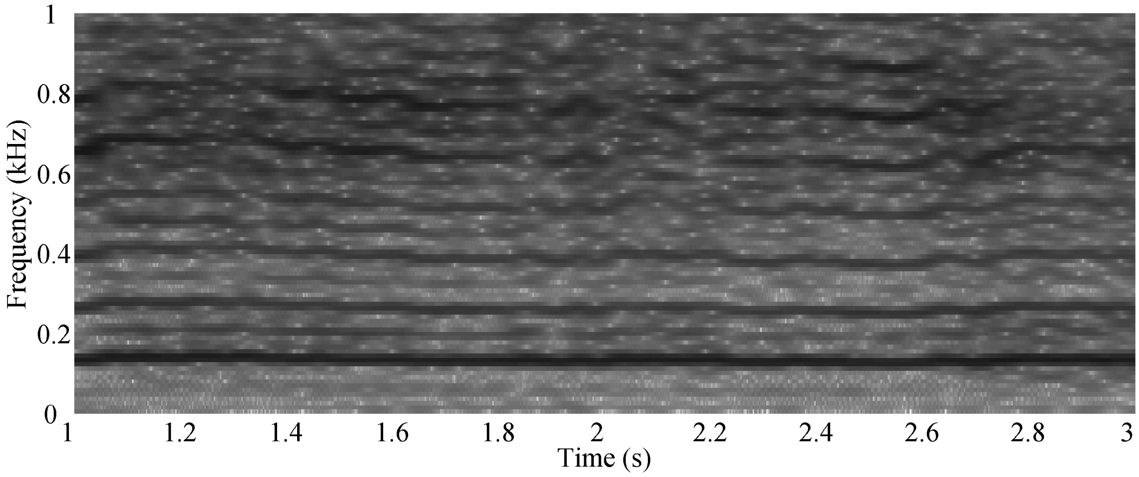
Method	Improved Feature	
	Prosody	Breathiness
LT	2.9 ± 0.7	1.7 ± 0.8
LT+MT	1.9 ± 0.7	1.6 ± 0.5
MR	2.0 ± 1.2	2.0 ± 1.2

- LT: long-term f_0-variability
- LT-MT: long-term and middle-term variability
- MR: multi-resolution approach.

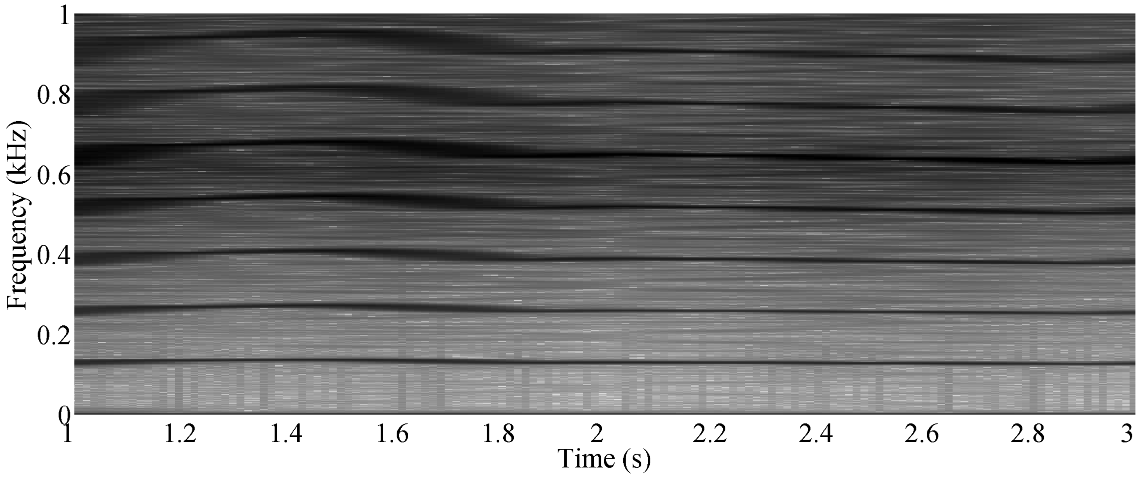
The results displayed in Table 6.2 indicate that the proposed restoration approach improves the perceived quality with respect to both criteria. The contribution of the f_0-variability restored at the middle-term scale appears to be most significant (1.1 points of improvement compared to 0.1 points by LT-variability alone). This seems to emphasize our assumption that additional f_0-variability at the middle-term scale (Fig. 6.5) can contribute to the restoration of prosody. The contribution of the MR approach yields no significant improvement to the perceived prosody. The high amount of standard deviation and the relatively small amount of listeners prohibits drawing general conclusions. Nevertheless, a positive trend can be recognized.

Regarding the breathiness of the restored voice source, a clear improvement (1.0 to 1.4 points) in all voice can be observed. Due to the low SNR in alaryngeal voice sources, higher frequency harmonics are submerged in noise, leading to the perception of a less harmonic, breathy voice. For voice sources restored with the MR approach, the additionally induced short-term variability seems to imply a degradation in terms of increased breathiness compared to the result of the LT+MT system configuration. This could be due to the fact that short-term variability is related to the glottal jitter. Indeed, jitter may be perceived as a desired, authentic feature at a very low intensity level but becomes certainly harmful over a given threshold. This threshold depends on the subject's idiosyncrasies and may be adjusted to the alaryngeal speaker's desire. We suggest that a more carefully designed non-linear model for the short-time variability contribution or a spectrally shaped noise instead of the AWGN may reduce this undesired effect of the short-time f_0-variability.

Another observation concerns the quality of the estimated VTF coefficients. In comparison to healthy subjects it was observed that the VTF coefficients estimated in subsequent frames exhibited an increased variance. This observation eventually lead to the hypothesis of insufficient separation of source and filter due to the used VTF estimation approach, in this case linear prediction and was the motivation for the SFO methods presented in the earlier



I: Alaryngeal voice



II: Healthy voice

Figure 6.5: Lower frequencies of a spectrogram of a sustained vowel */a:/* of an alaryngeal speaker (top panel) and restored with the LT+MT system configuration (bottom panel). The reduced f_0 fluctuations and increased f_0-variability in the restored speech yielded a perception of improved prosody.

chapters.

6.6 Conclusion

- In this chapter, we proposed a multi-resolution approach for the restoration of authentic, f_0-related features in pathological voice sources.
- The aim of the proposed method is to provide alaryngeal speakers the possibility to influence prosodic characteristics of the voice source such as the f_0 in an intuitive manner.
- The Adaptive Wavetable Oscillator has evolved as a sound method for the estimation of the long-term f_0-variations in pathological speech and stands out due to its low computational complexity while performing similar to computationally more complex methods.
- The multi-resolution approach for the restoration of the f_0-variability at different time scales improved the perceived prosody and breathiness of reconstructed pathological voice sources.
- The implementation of this method on an embedded device can be regarded as an attractive alternative to currently used electro-larynx devices due to its hands-free mode of operation and superior acoustic quality.
- An important observation made during the examination of the results concerns the general perceived quality of the produced speech. The simple vocal tract estimation methods used for the separation of voice source and articulation yielded insufficient results and eventually led to the development of the previously proposed method, presented in Chapter 4.

7 General Conclusions

In this thesis, methods and models were developed and presented aiming at the estimation, restoration and transformation of the characteristics of human speech. The initial motivation for this work was to develop a method and a device that allows restoring intelligibility and natural sound in pathological speech while reconstructing authentic and prosodic features at the same time. For this purpose, a multi-resolution approach was proposed. The method allows the speaker to influence prosodic characteristics of the voice source such as f_0 in an intuitive manner. This control over the reconstructed fundamental frequency is achieved using a method that works on three different temporal scales. Long-term trends of the f_0 trajectory were determined from the original speech signal using an f_0-estimation method called adaptive wavetable oscillator. This estimation method stands out from other methods due to its low computational complexity while performing similar to computationally more complex methods. Middle- and short-term f_0-variability were restored using correlations of the intensity of the original speech signal with f_0 variability at the respective temporal scales. Improved speech was then reconstructed using the newly synthesized source signal and the vocal tract coefficients estimated from the original, pathological speech signal.

In subjective experiments it was shown that the proposed approach improved the perceived prosody and breathiness of reconstructed pathological voice sources. It can also be constituted that the implementation of this method on portable devices such as smartphones is an attractive alternative to currently used electro-larynx devices due to its hands-free mode of operation and superior acoustic quality. With respect to the overall quality it was observed that variance in the estimated vocal tract coefficients deteriorated the perceived quality of the reconstructed speech signal. This variance led to the perception of a speech signal that may be described as harsh, rough, mulled and also unnatural. The variance was found to be caused by an incomplete separation of the voice source and the vocal tract during the estimation of the vocal tract filter coefficients. This observation was the motivation for the main part of this thesis, the development of a reliable method for the source-filter separation of speech.

In a general context of system identification, the source-filter separation is an ill-posed inverse problem aiming to obtain an estimate of the unknown VTF, which is excited by the unknown glottal source. In real voice sources, the glottal source exhibits a large volatility, its power spectrum is non-uniform and time-variant. The source characteristics are oversimplified in most commonly used source-filter estimation methods, which are reviewed in this work. This simplification has been a necessity due to practical compromises between the complexity of the voice model and the efficiency of the optimization method. The proposed method presented in Chapter 4 addresses this issue, by using a multi-parametric voice model in combination with a computationally efficient optimization method. We extend the state-of-the-art by formulating an optimization scheme that pitch-synchronously fits a multi-parametric source model and an auto-regressive vocal tract model to observed speech signals. An observed speech signal is modelled by two independent models, a source model and a vocal tract filter model. Criteria were formulated to guarantee that the two different models may not compensate for errors in the respective other model. Furthermore, a scheme is devised so as to reduce the effect of previous VTF resonances on the currently optimized glottal cycle (Sec. 4.4) and to effectively increase the size of the analysis window.

The computational efficiency of the proposed approach allowed us to carry out a large number of experiments. The convergence characteristics of the proposed method were first examined using synthetic speech signals in a variety of modifications such as varying f_0, glottal jitter, environmental noise and glottal source distortions. The accuracy of estimated formants of the proposed method was compared to three other methods, LP, IAIF (both frame-based) and CPLP (pitch-synchronous). Generally, the proposed method largely reduces the bias of estimated formant parameters over a large range of the evaluated conditions and reliably estimates the glottal source parameters with respect to the reference methods (Sec. 4.5), in particular in the presence of significant amounts of environmental noise. The proposed method was shown to be reliable and accurate in the task of separating glottal source and vocal tract filter characteristics in comparison to the other methods. Formant estimation using the proposed methods proved to be unreliable though in frequency bands, in which the glottal source consisted mainly of noise instead of deterministic signal components. Furthermore it is observed that the proposed method performs well in the presence of glottal source modelling errors up to a certain degree of distortion, but deteriorates fast above that threshold.

In a second series of experiments, the proposed method was applied to physically modelled speech as well as real speech signals. The evaluation was carried out objectively using speech produced by the physical model of speech production and qualitatively using healthy and pathological speech signals. In addition, the glottal source waveform obtained using inverse filtering was compared qualitatively for the speech simulated by the physical model. The glottal source waveform obtained using the proposed method exhibits a shape captured by the glottal source model used for the optimization, but also shows features commonly attributed to non-linear feedback mechanisms between the vocal tract and the inner-glottal air flow. The proposed method is capable of preserving the general structure of the glottal waveform, but also its fine details. The objective comparison of the formant estimates using physically modelled speech revealed a sensitivity of the proposed method to source modelling errors. It was observed that glottal source spectra, which show significant, frequency-dependent deviations from the constant spectral decay assumed by the LF source model in high frequency bands, may lead to spurious formant frequency estimation errors. Notably though, such large, cycle-specific glottal source spectra fluctuations were *not* observed with the real speech signals. The question, whether this phenomenon and the resulting formant estimation errors are limited to occur only with these simulated speech data, remains unanswered by this work for the moment.

A qualitative examination of the estimated formants on real speech signals showed that the proposed method is capable to extract smooth and presumably accurate formant trajectories given that the deterministic source component is of sufficient energy in frequency bands where formants are to be estimated. The extracted formant trajectories exhibit low variance, no dependency on the underlying glottal source and align well with the supposed formant trajectories in real speech signals. Also in the case of the real speech signals we observed that the proposed method becomes less reliable when the energy of the deterministic glottal source components are reduced. The proposed method is not capable to reliably estimate formants if the excitation is constituted of non-deterministic signal components only.

Temporal parameters of the glottal source model extracted from all real speech examples indicate that their estimation is consistent and mostly reliable. Trends in their trajectories may be observed that correlate with visually observable cues from the speech waveform and also correspond to values that one would expect theoretically in the respective examples.

There are several paths that appear as promising and logical continuations of the presented work. A first issue that needs to be addressed is a further reduction in the computational complexity. As implemented at the moment, the methods approximately are 400-600 times slower than real-time on a commercial PC. A large part of the processing is spent on converting the LF parameters from their temporal to the synthesis representation. Several order of magnitude of processing time could be gained with a description of the glottal LF model that does not require this conversion, yet allows meaningful boundary conditions to be described.

Another interesting direction for future work is the enlargement of the database used for the evaluation of this method. In the presented experiments, the performance of the proposed source-filter optimization method was shown to be useful for the estimation of a number of relevant speech modeling parameters, but the range of the investigated signals needs to be expanded to include different vowels, other voice quality types and other types of pathological voices. These experiments would allow to get a better judgement of the strengths and limitations of the proposed method.

Another very interesting route for future research are glottal models as a research subjects themselves. The proposed joint source-filter separation method allows the evaluation of new models in a very efficient manner. It would provide a very suitable framework for trying out new and possibly better fitting glottal models. In particular, one could extend the evolutionary concept of finding optimal parameters to the automatic exploration of new models, a concept known as evolutionary programming.

A fourth promising and potentially very useful future direction for a continuation of this work addresses the lack of accuracy of the proposed joint-source filter optimization method in the absence of a deterministic glottal source component. A subband approach may be an interesting route in order to combine the present method with a new method that is capable of reliably estimating formants in frequency bands exhibiting a low glottal source SNR.

Bibliography

M. Airas. *Methods And Studies Of Laryngeal Voice Quality Analysis In Speech Production.* PhD thesis, Helsinki University of Technology, Finland, 2008.

M. Airas, H. Pulakka, T. Bäckström, and P. Alku. A toolkit for voice inverse filtering and parametrisation. In *Proc. of INTERSPEECH*, pages 2145–2148, Lisbon, Portugal, Sep. 2005.

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

F. Alipour, D.A. Berry, and I.R. Titze. A finite-element model of vocal fold vibration. *J. Acoust. Soc. Am.*, 108(6):3003–3012, Dec. 2000.

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

P. Alku. Glottal inverse filtering analysis of human voice production — a review of estimation and parameterization methods of the glottal excitation and their applications. *Sādhanā - Academy Proceedings in Engineering Sciences*, 36(5):623–650, Oct. 2011.

P. Alku and E. Vilkman. Estimation of the glottal pulseform based on discrete all-pole modeling. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, pages 1619–1622, 1994.

P. Alku, E. Vilkman, and U.K. Laine. Analysis of glottal waveform in different phonation types using the new IAIF-method. In *Proc. 12th lnternat. Congress Phonetic Sciences*, volume 4, pages 362–365, Aix-en-Provence, Aug. 1991.

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

T.V. Ananthapadmanabha. Acoustic analysis of voice source dynamics. Technical report, STL-QPSR, 1984.

R. Arora and W. A. Sethares. Adaptive wavetable oscillators. *IEEE Trans. on Signal Processing*, 55 (9):4382–4392, 2007.

I. Arroabarren and A. Carlosena. Glottal spectrum based inverse filtering. In *Proc. of Eurospeech*, pages 57–60, Geneva, CH, Sep. 2003.

J.M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy. Research developments and directions in speech recognition and understanding, part 1. *IEEE Signal Processing Magazine*, 26(3):75–80, 2009.

C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House. Reduction of speech spectra by analysis-by-synthesis techniques. *J. Acoust. Soc. Am.*, 33(12):1725–1736, 1961.

G.S. Berke, D.M. Moore, D.R. Hantke, D.G. Hanson, B.R. Gerratt, and F. Bernstein. Laryngeal modeling: Theoretical, in vitro, in vivo. *Laryngoscope*, 97:871–881, 1987.

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

N. Bi and Y. Qi. Application of speech conversion to alaryngeal speech enhancement. *IEEE Trans. Speech Audio Processing*, 5(2):97–105, Mar 1997.

B.P. Bogert, M.J.R. Healy, and J.W. Tukey. *Proc. Symposium on Time Series Analysis*, chapter (15) The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking, pages 209–243. Wiley, New Yirk, NY, 1963.

L. Boves and B. Cranen. Evaluation of glottal inverse filtering by means of physiological registrations. In *Proc. IEEE ICASSP*, pages 1988–1991, 1982.

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 Lett.*, 12(4):344–347, Apr. 2005.

B. Bozkurt, L. Couvreur, and T. Dutoit. Chirp group delay analysis of speech signals. *Elsevier Speech Communication*, 49(3):159–176, Mar. 2007.

M. Brockmann, C. Storck, P.N. Carding, and M.J. Drinnan. Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language and Hearing Research*, 51:1152–1160, Oct 2008.

A. Camacho. *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music.* PhD thesis, Univ. of Florida, Gainsville, USA, Dec. 2007.

J.P. Campbell. Speaker recognition: A tutorial. *Proc. IEEE*, 85(9):1437–1462, 1997.

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

D.G. Childers. *Speech Processing and Synthesis Toolboxes.* J. Wiley & Sons, Inc., New York, 2000.

B. Cranen and L. Boves. On the measurement of glottal flow. *J. Acoust. Soc. Am.*, 84(3):888–900, Sep. 1988.

K.E. Cummings and M.A. Clements. Glottal models for digital speech processing: A historical survey and new results. *Digital Signal Processing*, 5(1):21–42, 1995.

Veeneman D. and BeMent S. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 33(2):369–377, Apr. 1985.

P. Dalsgaard, C. Pedersen, O. Andersen, and B. Yegnanarayana. Using zeros of the z-transform in the analysis of speech signals. In *Proc. ISCA Tutorial and Research Workshop, Speech Analysis and Processing for Knowledge Discovery*, page CD, 2008.

S. Das and P.N. Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE Trans. Evol. Comput.*, 15(1):4–31, Feb. 2011.

A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

G. Degottex. *Glottal Source and Vocal-Tract Separation.* PhD thesis, UPMC-Ircam, France, 2010.

G. Degottex, A. Roebel, and X. Rodet. Phase minimization for glottal model estimation. *IEEE Trans. Audio, Speech, Language Processing*, 19(5):1080–1090, Jul. 2011.

A. del Pozo and S. Young. Continuous tracheoesophageal speech repair. *Proc. of EUSIPCO*, 2006.

D.D. Deliyski. Clinical feasibility of high-speed videoendoscopy. *Voice and Voice Disorders*, 17(1):12–16, 2007.

J.R. Deller, G. Proakis, and H.L. Hansen. *Discrete-Time Processing of Speech Signals.* Prentice Hall, New Jersey, 1993.

H. Deng, R. Ward, M. Beddoes, and M. Hodgson. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. *IEEE Trans. Audio, Speech, Language Processing*, 14(2):445–455, Mar. 2006.

W. Ding, N. Campbell, N. Higuchi, and H. Kasuya. Fast and robust joint estimation of vocal tract and voice source parameters. In *Proc. IEEE ICASSP*, pages 1291–1294, Apr. 1997.

B. Doval and C. d'Alessandro. Spectral correlates of glottal waveform models: an analytic study. In *Proc. IEEE ICASSP*, pages 1295–1298, Munich, Germany, Apr. 1997.

B. Doval and C. d'Alessandro. The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92:1026–1046, 2006.

B. Doval, C. d'Alessandro, and B. Diard. Spectral methods for voice source parameters estimation. In *Proc. of Eurospeech*, pages 533–536, 1997.

B. Doval, C. d'Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. In *Proc. ISCA Voice Quality (VOQUAL)*, pages 16–20, 2003.

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

A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans. on Signal Processing*, 39(2):411–423, 1991.

G. Fant. Formant bandwidth data. *Quarterly Status Report, KTH*, 3(1):1–2, 1962.

G. Fant. Vocal source analysis — a progress report. Technical report, STL–QPSR, 1979.

G. Fant. The source filter concept in voice production. *Speech Research Summary Report STL-QPSR*, 22:21–37, 1981.

G. Fant. The LF-model revisited. transformations and frequency domain analysis. *Quarterly Status Report, KTH*, 36(2–3):119–156, 1995.

G. Fant and Q. Lin. Frequency domain interpretation and derivation of glottal flow parameters. Technical report, STL–QPSR, 1988.

G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. Technical report, STL–QPSR, 1985.

Gunnar Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

James L. Flanagan. *Speech analysis; synthesis and perception.* Springer Verlag, Berlin, New York,, 2 edition, 1972. ISBN 0387055614 3540055614.

J.L. Flanagan and L. Landgraf. Self-oscillating source for vocal tract synthesizers. *IEEE Trans. Audio Electroacoust.*, 16(1):57–64, Mar. 1968.

A.J. Fourcin and E. Abberton. First applications of a new laryngograph. *Med. Biol. Illus.*, 21(3): 172–182, Jul. 1971.

M. Fröhlich, D. Michaelis, and H.W. Strube. SIM — simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *J. Acoust. Soc. Am.*, 110(1):479–488, Jul. 2001.

Q. Fu and P. Murphy. Adaptive inverse filtering for high accuracy estimation of the glottal source. In *Proc. ISCA Tutorial and Research Workshop, Non-Linear Speech Processing*, page paper 018, 2003.

Q. Fu and P. Murphy. A robust glottal source model estimation technique. In *Proc. of INTERSPEECH*, pages 81–84, Jeju Island, Korea, October 2004.

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

O. Fujimura and J. Lindqvist. Sweep-tone measurements of vocal-tract characteristics. *J. Acoust. Soc. Am.*, 49(2B):541–558, 1971.

H. Fujisaki and M. Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *Proc. IEEE ICASSP*, pages 1605–1608, Tokyo, Japan, Apr. 1986.

H. Fujisaki and M. Ljungqvist. Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. In *Proc. IEEE ICASSP*, pages 637–640, 1987.

D. Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical report, University of Regina, CA, 2003.

W. Gersch. Spectral analysis of EEG's by autoregressive decomposition of time series. *Math. Biosci*, 7:205–222, 1970.

P.K. Ghosh and S.S. Narayanan. Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. *Elsevier Speech Communication*, 53(1): 98–109, Jan. 2011.

C. Gobl. *The voice source in speech communication. Production and perception experiments involving inverse filtering and synthesis.* PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 2003.

D.E. Goldberg. *Genetic algorithms in search optimization and machine learning.* Addison-Wesley Professional, Reading, MA, 1 edition, 1989.

J. Gudnason, M. Thomas, P. Naylor, and D. Ellis. Voice source waveform analysis and synthesis using principal component analysis and gaussian mixture modelling. In *Proc. of INTERSPEECH*, pages 108–111, Aalborg, DE, Jun. 2009.

J. Halton and G. Weller. Algorithm 247: radical inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, Dec. 1964.

H.M. Hanson. Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Am.*, 101(1):466–481, Jan. 1997.

W. Hargreaves and J. Starkweather. Voice quality changes in depression. *Language and Speech*, 7(2):84–88, Apr. 1964.

D.M. Hartl, S. Hans, L. Crevier Buchman, O. Laccourreye, J. Vaissiere, and D. Brasnu. Dysphonia: current methods of evaluation. *Ann Otolaryngol Chir Cervicofac.*, 122(4):163–172, 2005.

M.H. Hast. *Vocal fold physiology: Biomechanics, acoustics and phonatory control*, chapter Comparative anatomy of the larynx: Evolution and function, pages 3–14. Denver Center for the performing arts, Denver, CO, 1983.

J.W. Hawks and J.D. Miller. A formant bandwidth estimation procedure for vowel synthesis. *J. Acoust. Soc. Am.*, 97(2):1343–1344, 1995.

S. Haykin. *Adaptive Filter Theory.* Prentice Hall, 2001.

P. Hedelin. A glottal LPC-vocoder. In *Proc. IEEE ICASSP*, pages 21–24, Mar. 1984.

Helsinki University of Technology (HUT), TKK Laboratory of Acoustics and Audio Signal Processing. TKK Aparat. Online Resource: http://www.acoustics.hut.fi/software/aparat/.

N. Henrich. *Etude de la source glottique en voix parlée et chantée.* PhD thesis, Université Pierre et Marie Curie - Paris VI, 2001.

N. Henrich, B. Doval, and C. d'Alessandro. Glottal open quotient estimation using linear prediction. In *Proc. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 12–17, 1999.

N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *J. Acoust. Soc. Am.*, 115(3):1321–1332, 2004.

D.J. Hermes. Synthesis of breathy vowels: some research methods. *Speech Communication, Special issue on speaker characterization in speech terminology*, 10(5–6):497–502, Dec. 1991.

M. Hirano. *Clinical examination of voice.* Springer Verlag, New York, 1981.

J.H. Holland. Outline for a logical theory of adaptive systems. *Association for Computing Machinery*, 9(3):297–314, Jul. 1962.

J.H. Holland. *Adaptation in natural and artificial systems.* MIT Press, Cambridge, MA, 1975.

E.B. Holmberg, R.E. Hillman, and J.S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *J. Acoust. Soc. Am.*, 85(4):511–529, Apr. 1988.

J. Holmes. Formant excitation before and after glottal closure. In *Proc. IEEE ICASSP*, pages 39–42, Apr. 1976.

E.C. Inwald, M. Döllinger, M. Schuster, U. Eysholdt, and C. Bohr. Multiparametric analysis of vocal fold vibrations in healthy and disordered voices in high-speed imaging. *Journal of Voice*, 25(5):576–590, 2011.

A. Isaksson and M. Millnert. Inverse glottal filtering using a parameterized input model. *Elsevier Signal Processing*, 18(4):435–445, Dec. 1989.

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

K. Ishizaka and M. Matsudaira. Fluid mechanical considerations of vocal cord vibration. *SCRL Monograph*, 8:28–72, 1972.

F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electron. Commun. Japan*, 53(A):36–43, 1970.

Ronkkonen J., S. Kukkonen, and K.V. Price. Real parameter optimization with differential evolution. In *Proc. IEEE CEC*, pages 506–513, 2005.

P. Jinachitra. *Robust Structured Voice Extraction for Flexible Expressive Resynthesis*. PhD thesis, CCRMA, Stanford University, 2007.

H. Kasuya, S. Ogawa, Y. Kikuchi, and S. Ebihara. An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology. *Speech Communication*, 5 (2): 171–181, 1986.

H. Kasuya, K. Maekawa, and S. Kiritani. Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics. In *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 2505–2512, San Francisco, CA, 1999.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proc. IEEE Intern. Conf. Neural Networks*, volume IV, pages 1942–1948, 1995.

M. Kepesi and L. Weruaga. High-resolution noise-robust spectral-based pitch estimation. *Proc. of INTERSPEECH*, pages 313–316, 2005.

A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*. Springer Verlag, 2006.

D.H. Klatt and L. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857, 1990.

G. Kopec, A. Oppenheim, and J. Tribolet. Speech analysis by homomorphic prediction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 25(1):40–49, 1977.

A. Kounoudes, P.A. Naylor, and M. Brookes. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In *Proc. IEEE ICASSP*, pages 349–352, Orlando, FL, 2002.

A. Krishnamurthy and D. Childers. Two-channel speech analysis. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34(4):730–743, Aug. 1986.

A.K. Krishnamurthy. Glottal source estimation using a sum-of-exponentials model. *IEEE Trans. on Signal Processing*, 40(3):682–686, Mar. 1992.

P. Ladefoged. Some physiological parameters in speech. *Language and Speech*, 6:109–119, 1963.

S.W. Lang and J.H. McClellan. A simple proof of stability for all-pole linear prediction models. *Proc. IEEE*, 67(5):860–861, 1979.

J. Larar, Y. Alsaka, and D. Childers. Variability in closed phase analysis of speech. In *Proc. IEEE ICASSP*, pages 1089–1092, Apr. 1985.

F.L.E. Lecluse, M.P. Brocaar, and J. Verschurre. The electoglottography and its relation to glottal activity. *Folio Phoniatrica et Logopaedica*, 27():215–224, 1975.

P. Lieberman. *Speech physiology and acoustic phonetics.* Macmillan, New Your, NY, 1977.

J. Liljencrants. *Speech Synthesis with a Reflection-Type Line Analog.* PhD thesis, Royal Inst. of Tech., Stockholm, Sweden, 1985.

S. Ling. *A finite element method for duct acoustic problems.* PhD thesis, Purdue University, West Lafayette, IN, 1976.

L. Ljung. *System Identifcation.* Prentice Hall, Upper Saddle River, NJ, 1999.

M.G. Ljungvist. *Speech Analysis-Synthesis based on modeling of voice source and vocal-tract characteristics.* PhD thesis, University of Tokyo, Japan, 1986.

C.X. Lu. *A numerical simulation of sound production in the vocal tract.* PhD thesis, Shizuoka University, JPN, 1993.

H.L. Lu. *Toward a High-Quality Singing Synthesizer with Vocal Texture Control.* PhD thesis, CCRMA, Stanford University, 2002.

H.L. Lu and J.O. Smith. Joint estimation of vocal tract filter and glottal source waveform via convex optimization. In *WASPAA*, pages 79–82, New Paltz, NY, Oct. 1999.

S. Maeda. A digital simulation method of the vocal-tract system. *Elsevier Speech Communication*, 1(3–4):199–229, Dec. 1982.

J. Makhoul. Linear prediction: a tutorial review. *Proc. IEEE*, 63(4):561–580, Apr. 1975.

J. Markel and B. Atal. *Linear Prediction of Speech.* Springer Verlag, Berlin, 1976.

M.V. Mathews, J.A. Miller, and E.E. David Jr. Pitch synchronous analysis of voiced sounds. *J. Acoust. Soc. Am.*, 33(2):179–186, 1961.

L. Max, W. Steurs, and W. De Bruyn. Vocal capacities in esophageal and tracheoesophageal speakers. *Laryngoscope*, 106(1):93–96, 1996.

P. Milenkovic. Glottal inverse filtering by joint estimation of an AR system with a linear input model. *IEEE Trans. Acoust., Speech, Signal Processing*, 34(1):28–42, Feb. 1986.

R.L. Miller. Nature of the vocal cord wave. *J. Acoust. Soc. Am.*, 31(6):667–677, 1959.

P. Mitev and S. Hadjitodorov. Fundamental frequency estimation of voice of patients with laryngeal disorders. *Information Sciences*, 156 (1–2):3–19, 2003.

M. Moerman, G. Pieters, J.P. Martens, M.J. van der Borgt, and P. Dejonckere. Objective evaluation of quality of substitution voices. *Eur Arch Otorhinolaryngol*, 261:541–547, 2004.

B.C.J. Moore. *An Introduction to the Psychology of Hearing*. Emerald Group Publishing Ltd., Bingley, UK, 6 edition, 2012.

E. Moore and M. Clements. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In *Proc. IEEE ICASSP*, pages 101–104, 2004.

E. Moore and J. Torres. A performance assessment of objective measures for evaluating the quality of glottal waveform estimates. *Elsevier Speech Communication*, 50(1):56–66, 2008.

T. Most, Y. Tobin, and R.C. Mimran. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *J. of Communication Disorders*, 33(2):165–180, 2000.

T. Murakami and Y. Ishida. Fundamental frequency estimation of speech signals using MUSIC algorithm. *Acoust. Sci. Technol.*, 22 (4):293–297, 2001.

L.C. Oliveira. Estimation of source parameters by frequency analysis. In *Proc. of Eurospeech*, pages 99–102, 1993.

A.V. Oppenheim. *Superposition in a class of nonlinear systems*. PhD thesis, Res. Lab. Electronics, MIT, Cambridge, MA, 1965.

A. Papoulis. *Probability and Statistics*. Prentice Hall, 1989.

J. Perez and A. Bonafonte. Towards robust glottal source modeling. In *Proc. of INTERSPEECH*, pages 68–71, Aalborg, DE, Jun. 2009.

R.H. Pindzola and B.H. Cain. Acceptability ratings of tracheoesophageal speech. *Laryngoscope*, 98(4):394–397, 1988.

J. Pittam. The long-term spectral measurement of voice quality as a social and personality marker: A review. *Language and Speech*, 30(1):1–12, Jan. 1987.

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

M.R. Portnoff. A quasi one-dimensional digital simulation for the time-varying vocal tract. Master's thesis, M.I.T, Electrical Engineering Department, Boston, MA, 1973.

J.J. Pressman. Physiology of the vocal cords in phonation and respiration. *Archives of Otolaryngology*, 35:355–398, 1942.

K.V. Price. Differential evolution: a fast and simple numerical optimizer. In *Proc. 1996 biennial conference of the North American fuzzy information processing society - NAFIPS*, pages 524–527, Berkeley, CA, Jun. 1996.

K.V. Price. Differential evolution vs. the functions of the second ICEO. In *Proc. IEEE international conference on evolutionary computation*, pages 153–157, Indianapolis, IN, 1997.

K.V. Price, R.M. Storn, and J.A. Lampinen. *Differential Evolution: A Practical Approach to Global Optimization.* Springer, 2005.

Y. Qi, B. Weinberg, and N. Bi. Enhancement of female esophageal and tracheoesophageal speech. *J. Acoust. Soc. Am.*, 98(5–1):2461–2465, Nov. 1995.

T.F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hall, Upper Saddle River, NJ, 2001.

L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, Englewood Cliffs, NJ, 1978.

T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku. HMM-based finnish text-to-speech system utilizing glottal inverse filtering. *Proc. of INTERSPEECH*, pages 1881–1884, Sep. 2008.

T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, and P. Vainio, M. Alku. HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Trans. Audio, Speech, Language Processing*, 19(1):153–165, Jan. 2011.

I. Rechenberg. *Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* PhD thesis, Technische Universität Berlin, GER, 1971. reprinted in 1973 by Fromman-Holzboog.

E. Riegelsberger and A. Krishnamurthy. Glottal source estimation: Methods of applying the LF-model to inverse filtering. In *Proc. IEEE ICASSP*, pages 542–545, 1993.

E.A. Robinson. Predictive decomposition of time series with application to seismic exploration. *Geophysics*, 32(3):418–484, 1967.

A. Roebel, F. Villavicencio, and X. Rodet. On cepstral and all-pole based spectral envelope modeling with unknown model order. *Pattern Recognition Letters*, 28(11):1343–1350, 2007.

A. Rosenberg and J. Hirschberg. On the correlation between energy and pitch accent in read english speech. *Proc. of INTERSPEECH*, 1294-Mon2A3O.2, 2006.

A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels,. *J. Acoust. Soc. Am.*, 49:583–590, 1971.

M. Rothenberg. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *J. Acoust. Soc. Am.*, 53(6):1632–1645, 1973.

B. Roubeau, N. Henrich, and M. Castellengo. Laryngeal vibratory mechanisms: The notion of vocal register revisited. *Journal of Voice*, 23(4):425–438, 2009.

L. Råde and B. Westergren. *Springers mathematische Formeln.* Springer, Berlin, Heidelberg, 2 edition, 1997.

R.A. Samlan and B.H. Story. Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling. *J. Speech, Lang and Hear. Res.*, 54(5):1267–1283, Oct. 2011.

O. Schleusing, R. Vetter, Ph. Renevey, J. Krauss, F.N. Reale, V. Schweizer, and J.-M. Vesin. Restoration of authentic features in tracheoesophageal speech by a multi-resolution approach. *Proc. of SPPRA 2009*, pages 643–042, Feb. 2009.

O. Schleusing, R. Vetter, Ph. Renevey, J.-M. Vesin, and V. Schweizer. *CCIS: Biomedical Engineering Systems and Technologies*, volume 127, chapter Prosodic Speech Restoration Device: Glottal Excitation Restoration using a Multi-Resolution Approach, pages 177–188. Springer, 2011.

K. Schnell and A. Lacroix. Time-varying pre-emphasis and inverse filtering of speech. In *Proc. of INTERSPEECH*, pages 530–533, Antwerp, BE, Aug. 2007.

M. Schröder. *Affective Information Processing*, chapter Expressive Speech Synthesis: Past, Present, and Possible Futures, pages 111–126. Springer, London, 2009.

H.P. Schwefel. *Evolution and optimum seeking*. Wiley-Interscience, New York, NY, 1 edition, 1994.

Y. Shapira and I. Gath. A geometrical fuzzy clustering-based solution to glottal wave estimation. *J. Acoust. Soc. Am.*, 104(5):3070–3079, Nov. 1998.

Y. Shiga and S. King. Estimation of voice source and vocal tract characteristics based on multi-frame analysis. In *Proc. of Eurospeech*, pages 1749–1752, 2003.

Y.-L. Shue and A. Alwan. A new voice source model based on high-speed imaging and its application to voice source estimation. In *Proc. IEEE ICASSP*, pages 5134–5137, Dallas, TX, Mar. 2010.

Y.-L. Shue, J. Kreiman, and A. Alwan. A novel codebook search technique for estimating the open quotient. In *Proc. of INTERSPEECH*, pages 2895–2898, 2009.

D.H. Slavit and T.V. McCaffrey. Regulation of phonatory efficiency by vocal fold tension and glottis width in the excised canine larynx. *Annals of Otology, Rhinology & Laryngology*, 100(8):668–677, 1991.

J.O. Smith. *Introduction to Digital Filters with Audio Applications*. W3K Publishing, http://www.w3k.org/books/, 2007a. ISBN 978-0-9745607-1-7.

J.O. Smith. *Mathematics of the Discrete Fourier Transform (DFT)*. W3K Publishing, http://www.w3k.org/books/, 2007b.

J.O. Smith. *Spectral Audio Signal Processing*. W3K Publishing, http://www.w3k.org/books/, 2011.

M. Södersten, Håkansson A., and B. Hammarberg. Comparison between automatic and manual inverse filtering procedures for healthy female voices. *Logoped. Phoniatr. Vocol.*, 1: 26–38, 1999.

K. Steiglitz. On the simultaneous estimation of poles and zeros in speech analysis. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 25(3):229–234, 1977.

K.N. Stevens. Toward a model for speech recognition. *J. Acoust. Soc. Am.*, 32(1):47–55, 1960.

K.N. Stevens. Physics of laryngeal behavior and larynx modes. *Phonetica*, 34:264–279, 1977.

R. Storn. On the usage of differential evolution for function optimization. In *Proc. 1996 biennial conference of the North American fuzzy information processing society - NAFIPS*, pages 519–523, Berkeley, CA, Jun. 1996.

R. Storn and K.V. Price. Differential evolution a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical report, ICSI, 1995.

B.H. Story. *Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract.* PhD thesis, University of Iowa, Iowa City, 1995.

B.H. Story. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoust. Science and Technology*, 23(4):195–206, 2002.

B.H. Story. Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. *J. Acoust. Soc. Am.*, 123(1):327–335, Jan. 2008.

B.H. Story and I.R. Titze. Voice simulation with a bodycover model of the vocal folds. *J. Acoust. Soc. Am.*, 97(2):1249–1260, Feb. 1995.

G. Strang. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, 3 edition, 2003.

H. Strube. Determination of the instant of glottal closure from the speech wave. *J. Acoust. Soc. Am.*, 56(5):1625–1629, 1974.

N. Sturmel, C. d'Alessandro, and B. Doval. A spectral method for estimation of the voice speed quotient and evaluation using electroglottography. In *Proc. Advances in Quantitative Laryngology (AQL)*, page 6p, Groningen, NL, 2006.

N. Sturmel, C. D'Alessandro, and B. Doval. A comparative evaluation of the zeros of z transform representation for voice source estimation. In *Proc. of INTERSPEECH*, pages 558–561, 2007.

H. Sutter. Welcome to the parallel jungle! *Dr. Dobb's Journal*, Jan. 2012. URL http://drdobbs.com/parallel/232400273.

The Mathworks. Matlab 2006b, 2006.

M.R.P. Thomas and P.A. Naylor. The sigma algorithm: A glottal activity detector for electroglottographic signals. *IEEE Trans. on Audio, Speech and Language Processing*, 17(8):1557–1566, Nov. 2009.

M.R.P. Thomas, J. Gudnason, and P.A. Naylor. Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. *IEEE Trans. Audio, Speech, Language Processing*, 20(1):82–91, January 2012.

M. Thomson. A new method for determining the vocal tract transfer function and its excitation from voiced speech. In *Proc. IEEE ICASSP*, pages 37–40, Mar. 1992.

I. Titze. Parameterization of the glottal area, glottal flow, and vocal fold contact area. *J. Acoust. Soc. Am.*, 75(2):570–580, Feb. 1984.

I. Titze. *Principles of voice production.* National Center for Voice and Speech, 2000.

I. Titze. Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *J. Acoust. Soc. Am.*, 111(1):367–376, Jan. 2002.

I. Titze. *The myoelastic aerodynamic theory of phonation.* Iowa City: National Center for Voice and Speech, 2006.

I.R. Titze. The human vocal cords: A mathematical model, part I. *Phonetica*, 28(3):129–170, 1974a.

I.R. Titze. The human vocal cords: A mathematical model, part II. *Phonetica*, 29(1):1–21, 1974b.

I.R. Titze. On the mechanics of vocal fold vibration. *J. Acoust. Soc. Am.*, 60:1366–1380, 1976.

I.R. Titze. A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *J. Acoust. Soc. Am.*, 66(1):60–74, 1979.

I.R. Titze. The physics of small amplitude oscillation of the vocal folds. *J. Acoust. Soc. Am.*, 83(4):1536–53, Apr. 1988.

George L. Turin. An introduction to matched filters. *IRE Transactions on Information Theory*, 6 (3):311–329, 1960.

C. Un and S.-C. Yang. A pitch extraction algorithm based on LPC inverse filtering and AMDF. *IEEE Trans. ASSP*, 25 (6):565–572, 1977.

P.P. Vaidyanathan. *The Theory of Linear Prediction.* Morgan & Claypool, 2008.

P.P. Vaidyanathan, J. Tuqan, and A. Kirac. On the minimum phase property of prediction-error polynomials. *IEEE Signal Processing Letters*, 4(5):126–127, May 1997.

C.J. van As. *Tracheoesophageal Speech: A multidimensional assessment of voice quality*. PhD thesis, University of Amsterdam, 2001.

J. W. van den Berg. Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research*, 1:227–244, 1958.

R. van Dinther. *Perceptual aspects of voice-source parameters*. PhD thesis, Technical University of Eindhoven, The Netherlands, 2003.

R. Veldhuis. A computationally effcient alternative for the Liljencrants-Fant model and its perceptual evaluation. *J. Acoust. Soc. Am.*, 103(1):566–571, 1998.

A. Verma and A. Kumar. Introducing roughness in individuality transformation through jitter modelling and modification. *Proc. IEEE ICASSP*, 1:5–8, Mar 2005.

J. Vesterstrøm and R.A. Thomson. Comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. In *Proc. IEEE Congr. Evol. Comput.*, pages 1980–1987, 2004.

R. Vetter, J. Cornuz, P. Vuadens, I.C.J. Sola, and P. Renevey. Method and system for converting voice. European Patent, November 2006. EP1710788.

F. Villavicencio, A. Roebel, and X. Rodet. Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. In *Proc. IEEE ICASSP*, pages 869–872, 2006.

D. Vincent. *Analyse et controle du signal glottique en synthese de la parole (in French)*. PhD thesis, ENST, Paris, France, 2007.

N. Virag. *Speech Enhancement based on Masking Properties of the Human Auditory System*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 1996.

B. Weinberg. *Laryngectomee Rehabilitation*, chapter Acoustical properties of esophageal and tracheoesophageal speech, pages 113–127. College-Hill Press, San Diego, CA, 1986.

J.H. Wilkinson. *Mathematical Methods for Digital Computers*, chapter The solution of ill-conditioned linear equations, pages 65–93. John Wiley and Sons, 1967.

S.E. Williams and J. Barber Watson. Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope*, 97(6):737–739, 1987.

D.Y. Wong, J.D. Markel, and A.H. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(4):350–355, Aug. 1979.

A. Yanga, M. Stingl, D.A. Berry, J. Lohscheller, D. Voigt, U. Eysholdt, and M. Döllinger. Computation of physiological human vocal fold parameters by mathematical optimization of a biomechanical model. *J. Acoust. Soc. Am.*, 130(2):948–964, Aug. 2011.

B. Yegnanarayana and N. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *IEEE Trans. on Speech and Audio Processing*, 6(4):313–327, Jul. 1998.

D. Zaharie. Critical values for the control parameters of differential evolution algorithms. In *Proc. MENDEL 2002, 8th international conference on soft computing*, pages 62–67, Jun. 2002.

List of Publications

Publications Related to this Work

O. Schleusing, T. Kinnunen, B. Story, and J.-M. Vesin. Joint source-filter optimization for accurate vocal tract estimation using differential evolution. *IEEE Trans. on Audio, Speech and Language Processing*, submitted, 2012

O. Schleusing, R. Vetter, Ph. Renevey, J.-M. Vesin, and V. Schweizer. *LNCS CCIS: Biomedical Engineering Systems and Technologies*, volume 127, Chapter Prosodic Speech Restoration Device: Glottal Excitation Restoration using a Multi-Resolution Approach, pages 177-188. Springer, 2011.

O. Schleusing, R. Vetter, Ph. Renevey, J.-M. Vesin, and V. Schweizer. Device for prosodic speech restoration: A multi-resolution approach for glottal excitation restoration. In *Proc. BioDevices 2010*, pages 38-43, Jan. 2010. *(rec. Best Paper Award)*

O. Schleusing, R. Vetter, Ph. Renevey, V. Schweizer, and J.-M. Vesin. Pitch estimation in pathological voice using adaptive wavetable oscillator. *Poster presented at annual meeting of Swiss Society for Biomedical Engineering*, 2009. *(rec. Best Poster Award)*

O. Schleusing, R. Vetter, Ph. Renevey, J. Krauss, F.N. Reale, V. Schweizer, and J.-M. Vesin. Restoration of authentic features in tracheoesophageal speech by a multi-resolution approach. In *Proc. SPPRA 2009*, pages 643-042, Feb. 2009.

Publications Not Related to this Work

O. Schleusing, Ph. Renevey, M. Bertschi, St. Dasen, and R. Paradiso. Detection of mood changes in bipolar patients though monitoring of physiological and behavioral signals. In *Proc. MBEC*, Budapest, HUN, Sep. 2011.

O. Schleusing, Ph. Renevey, M. Bertschi, St. Dasen, J.-M. Koller, and R. Paradiso. Monitoring physiological and behavioral signals to detect mood changes of bipolar patients. In *Proc. ISMICT*, Montreux, CH, Mar. 2011.

O. Schleusing, B. Zhang, and Y. Wang. Onset detection in pitched non-percussive music using

warping compensated correlation. In *Proc. IEEE ICASSP*, pages 117-120, Las Vegas, NV, 2008.

Y. Wang, B. Zhang, and O. Schleusing. Educational violin transcription by fusing multimedia streams. In *Proc. international workshop on Educational multimedia and multimedia education, EMME 07*, pages 57-66, 2007

Olaf B. Schleusing

Weissenbühlweg 40D
3007 Berne, Switzerland

+41 78 766 05 07
olaf@schleusing.de

Highlights

- EPFL PhD graduate (expected 07/2012)
- Explorative, analytical and goal-oriented mindset
- Interested in performance and parallel computing, signal processing and distributed computing
- International working experience, various languages
- Believe in working smart, but never mind working hard

Academic Experience

PhD Student 11/2007 – present
EPFL, Lausanne, Switzerland (employed by CSEM, Neuchâtel, Switzerland)

- Subject: Speech feature estimation in harsh conditions using evolutionary optimization methods.
- Application of these methods to speech with highly degraded voice.
- Best Paper and Best Poster Awards at international and national conferences.
- Besides PhD research:
  - Responsible for management of EU project subtasks and conduction of EU project applications.
  - Developed accessory firmware and prototype demonstration apps for iPhone and Android devices (Objective C, Java).
  - Designed, implemented and optimized various DSP algorithms on an Android device using the JNI interface (C++).

Research Assistant 01/2007 – 10/2007
NUS - National University of Singapore, School of Computing

- Developed and published novel signal processing methods for music information retrieval (Matlab).
- Designed and developed graphical tools for convenient music annotation, special focus on productivity of prospective users (C#).
- Responsible for installation of new music recording infrastructure for Music Computing Lab.
- Supervised several lab activities of undergrad students.

Professional Experience

R/D Engineer 01/2003 – 12/2006
Studer Professional Audio AG, Zurich, Switzerland
01/2003 – 12/2004 - Embedded software developer

- Responsible for DSP algorithm implementation and intensive performance optimization on proprietary, embedded hardware (C and assembler).
- Developed real-time operating system (RTOS) functionality for proprietary embedded DSP environment (C and assembler).

01/2005 – 12/2006 - High-level software developer

- In a team of seven software engineers, developed and implemented various features of next generation professional digital audio mixing desks (C++).
- Took the lead in development of several features and internal tools for build automation.
- Took initiative for close collaboration with product management for optimization of user experience.

R/D Engineer 07/2002 – 01/2003
Fraunhofer IDMT, Ilmenau, Germany

- Implemented and optimized an auditory filter bank for music information retrieval (C++).
- Contributed to the development of a music identification system using an audio fingerprint database.

R/D Intern 05/2001 – 09/2001
BDTi, Berkeley, CA, USA

- Developed and optimized ARM CPU assembler implementation of a suite of signal processing benchmarking algorithms (BDTi Benchmarks$^{\text{TM}}$).
- Contributed to performance optimization of an audio codec C implementation for three different RISC platforms.

Student Assistant 07/1999 – 04/2001
Fraunhofer IIS, Ilmenau, Germany

- In a team of three students, ported an off-line implementation of an audio watermark decoder to a real-time embedded platform.
- Contributed to various research projects by implementing signal processing algorithms on embedded hardware.

Student Assistant 07/1999 – 04/2001
Ilmenau Technical University, Germany, Dept. of Digital Signal Processing

- Implemented and optimized audio signal processing algorithms in digital signal processors (ADSP floating point DSPs, assembler).
- Gained valuable practical hands-on experience by applying theoretical knowledge.

Education

Diplom-Ingenieur, Media Technology
Ilmenau Technical University, Ilmenau, Germany, July 2002

- Focus: Signals and systems, audio signal processing
- Graduated three months ahead of regular schedule with a thesis on comparison of auditory filterbanks, carried out at Fraunhofer IDMT, Ilmenau, Germany

Associations

ACCU 02/2005 – present
Member of the Association of professional C/C++ programmers.

IEEE 02/2012 – present
Student member of the Institute of Electrical and Electronics Engineers.

ProSchule Bangalore 01/2012 – present
Volunteering member of executive board in an association supporting a school for children in need of financial aid in India.

Main Skills

Languages: German (native), English (fluent), French (good), Swiss German (basic)

IT: C, C++, Matlab, C#, Objective C, Assembler on various processors, (Python), profound hands-on experience with various computing architectures and optimization, mainly Windows, but also Linux and Mac OS.

Hobbies

Running, Sports in general, Traveling, Photography, Programming and Computer Technology in general.