

# **Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals**

Baris Bozkurt

Supervisor: Prof. Dr. Ir. Thierry Dutoit



*Dissertation submitted to the Faculté Polytechnique de Mons  
for the degree of Doctor of Philosophy in applied sciences*

---



# Faculté Polytechnique de Mons



Dissertation originale soumise à la Faculté Polytechnique de Mons  
en vue de l'obtention du grade de docteur en sciences appliquées par

**Ir. Baris Bozkurt**

---

**Zeros of the z-transform (ZZT) representation and chirp group delay  
processing for the analysis of source and filter characteristics of  
speech signals**

---

**Membres du jury :**

Professeur Marc Pirlot, Faculté Polytechnique de Mons , Président  
Professeur Thierry Dutoit, Faculté Polytechnique de Mons , Promoteur  
Professeur Christophe d'Alessandro, CNRS-LIMSI/France  
Professeur Boris Doval, Université de Paris II/France  
Dr. Vincent Pagel, Acapela S.A.  
Professeur Paul Lybaert, Faculté Polytechnique de Mons, Doyen

Thèse préparée au laboratoire de Théorie des Circuits et Traitement du Signal de la Faculté Polytechnique de Mons et dans le groupe 'Perception Située' de LIMSI-CNRS/Orsay.



*To Thierry Dutoit*



# Abstract

**T**his study proposes a new spectral representation called the Zeros of Z-Transform (ZZT), which is an all-zero representation of the  $z$ -transform of the signal. In addition, new chirp group delay processing techniques are developed for analysis of resonances of a signal. The combination of the ZZT representation with the chirp group delay processing algorithms provides a useful domain to study resonance characteristics of source and filter components of speech. Using the two representations, effective algorithms are developed for: source-tract decomposition of speech, glottal flow parameter estimation, formant tracking and feature extraction for speech recognition.

The ZZT representation is mainly important for theoretical studies. Studying the ZZT of a signal is essential to be able to develop effective chirp group delay processing methods. Therefore, first the ZZT representation of the source-filter model of speech is studied for providing a theoretical background. We confirm through ZZT representation that anti-causality of the glottal flow signal introduces mixed-phase characteristics in speech signals. The ZZT of windowed speech signals is also studied since windowing cannot be avoided in practical signal processing algorithms and the effect of windowing on ZZT representation is drastic. We show that separate patterns exist in ZZT representations of windowed speech signals for the glottal flow and the vocal tract contributions. A decomposition method for source-tract separation is developed based on these patterns in ZZT.

We define chirp group delay as group delay calculated on a circle other than the unit circle in  $z$ -plane. The need to compute group delay on a circle other than the unit circle comes from the fact that group delay spectra are often very noisy and cannot be easily processed for formant tracking purposes (the reasons are explained through ZZT representation). In this thesis, we propose methods to avoid such problems by modifying the ZZT of a signal and further computing the chirp group delay spectrum. New algorithms based on processing of the chirp group delay spectrum are developed for formant tracking and feature estimation for speech recognition. The proposed algorithms are compared to state-of-the-art techniques. Equivalent or higher efficiency is obtained for all proposed algorithms.

The theoretical parts of the thesis further discuss a mixed-phase model for speech and phase processing problems in detail.

**Index Terms**—spectral representation, source-filter separation, glottal flow estimation, formant tracking, zeros of  $z$ -transform, group delay processing, phase processing

# Acknowledgements

First of all, I am deeply indebted to Prof. Thierry Dutoit who provided unlimited support not only for all parts of this work but also for all aspects of my life in Belgium. It was the greatest pleasure during this thesis period to work with him, to be able to profit from the positive energy he radiated at all times and to be his friend. He is like a sun, always there with energy and support.

During seven months, I had the pleasure to work with Prof. Christophe d'Alessandro and Prof. Boris Doval in Limsi-CNRS/Orsay/France. Most of the novel ideas in this thesis were born under their guidance. I am very thankful to them for welcoming me in their research team, sharing expertise, friendship and tea. Special thanks to Dr. Vincent Pagel for his friendship, guidance and support.

I also would like to thank Francois Severin and Laurent Couvreur for their help in testing some of my ideas and the Multitel-TCTS family for providing me a peaceful working environment. Special thanks to: Devrim Unay, Nicolas d'Allessandro, Francois Meers and Michel Bagein for the refreshing breaks; Oytun Turk for his friendship and collaboration from the other side of the cable; Olivier Pietquin for the thesis template and Prof. Marc Pirlot for accepting to be part of the jury.

I would like to thank the 'Région Wallonne' and the 'Direction Générale des Technologies, de la Recherche et de l'Energie' (DGTRE) for their financial support (grant FIRST EUROPE #215095) as well as the 'Faculté Polytechnique de Mons' (FPMs), Multitel ASBL and the Acapela S.A. society for their administrative support.

Finally, I would like to express my sincere gratitude to my wife Betul and my parents for their encouragement and love.



# Contents

<b>CHAPTER I: INTRODUCTION .....</b>	<b>14</b>
I.1. <i>Motivations</i> .....	14
The (hi)story of this study.....	14
I.2. <i>Original contributions of the thesis</i> .....	15
ZZT Representation of signals .....	15
Chirp group delay processing .....	16
Applications of ZZT and chirp group delay.....	16
I.3. <i>Plan</i> .....	16
<b>CHAPTER II: STATE-OF-THE-ART .....</b>	<b>18</b>
II.1. <i>Introduction</i> .....	18
II.2. <i>Glottal flow estimation and voice quality analysis</i> .....	19
Glottal flow signal estimation methods.....	21
Glottal flow parameter estimation methods .....	22
Applications of glottal flow estimation in voice quality analysis for concatenative TTS.....	23
II.3. <i>Formant Tracking</i> .....	25
II.4. <i>Phase Processing of Speech</i> .....	25
Phase processing in sinusoidal/harmonic modeling .....	26
Phase processing in speech perception .....	27
Phase processing in speech analysis .....	28
Phase processing in automatic speech recognition.....	28
 <i>FIRST PART SPECTRAL REPRESENTATION OF SPEECH BY ZEROS OF THE Z-TRANSFORM (ZZT) AND CHIRP GROUP DELAY.....</i>	
<b>29</b>	
<b>CHAPTER III: ZEROS OF THE Z-TRANSFORM (ZZT) REPRESENTATION OF SPEECH .....</b>	<b>30</b>
III.1. <i>Introduction</i> .....	30
III.2. <i>Definition</i> .....	30
Finding the roots of high degree polynomials .....	31
III.3. <i>ZZT representation of speech signals</i> .....	32
III.3.1. ZZT of some basic signals.....	32
ZZT of an exponential time series.....	32
ZZT of a damped sinusoid.....	33
III.3.2. ZZT of the glottal flow signal .....	34
Contribution of the first phase to the ZZT of LF model glottal flow signal .....	35
Contribution of the return phase to the ZZT of the LF model glottal flow signal .....	38
III.3.3. ZZT representation and source-filter model of speech .....	38
III.3.4. ZZT of windowed synthetic speech signals.....	41
Effect of window location on ZZT patterns .....	42
Effect of window function on ZZT patterns.....	44
Effect of window size on ZZT patterns .....	45
III.3.5. ZZT of aperiodic components in speech .....	46
III.3.6. Conclusion.....	47
<b>CHAPTER IV: CHIRP GROUP DELAY PROCESSING OF SIGNALS .....</b>	<b>48</b>
IV.1. <i>Introduction</i> .....	48

IV.2.	<i>Methods proposed by Yegnanarayana and Murthy for group delay processing</i> .....	49
	Terminology.....	49
	Difficulties in group delay processing.....	52
	Processing group delay of the minimum-phase version of a signal.....	53
IV.3.	<i>Phase processing of mixed-phase signals</i> .....	54
IV.4.	<i>Mixed-phase speech model</i> .....	56
IV.5.	<i>Effects of windowing on group delay functions</i> .....	59
	Effects of window location on group delay functions.....	59
	Effects of window size on group delay functions.....	60
	Effects of window function on group delay functions.....	61
	Group delay spectrogram.....	62
	Conclusion.....	62
IV.6.	<i>Chirp group delay processing of speech</i> .....	63
IV.7.	<i>Conclusion</i> .....	67

## ***SECOND PART APPLICATIONS OF ZZT AND CHIRP GROUP DELAY PROCESSING IN SPEECH ANALYSIS..... 68***

### **CHAPTER V: APPLICATIONS OF ZZT AND CHIRP GROUP DELAY PROCESSING IN SPEECH ANALYSIS... 69**

V.1.	<i>ZZT-decomposition for source-filter separation of speech</i> .....	69
	V.1.1. The ZZT-decomposition algorithm.....	69
	V.1.2. Examples and evaluation of the decomposition algorithm.....	70
	Synthetic speech example.....	70
	Real speech example.....	73
	Robustness tests.....	74
	V.1.3. Mixed-phase decomposition using complex cepstrum.....	77
	Links between ZZT and complex cepstrum.....	77
	V.1.4. Conclusions.....	79
V.2.	<i>Application to glottal flow parameter estimation</i> .....	79
	V.2.1. Testing the $F_g$ estimation algorithm.....	80
	Tests with synthetic speech.....	80
	Tests with real speech.....	82
	V.2.2. Conclusions.....	83
V.3.	<i>Application to formant tracking</i> .....	83
	V.3.1. Formant tracker – first version.....	84
	V.3.2. Formant tracker – second version (DPPT).....	84
	Tests.....	85
	V.3.3. Formant tracker – third version (Fast-DPPT).....	87
	Tests.....	88
V.4.	<i>A Linear Prediction (LP) algorithm to estimate the glottal flow component from speech signals</i> 90	
	V.4.1. The MixLP algorithm.....	90
	Tests.....	91
V.5.	<i>Application to speech recognition</i> .....	93
	V.5.1. Group delay based features.....	93
	V.5.2. ASR experiments.....	96
	V.5.3. Discussion and conclusion.....	97

### **CHAPTER VI: CONCLUSION AND FUTURE WORKS ..... 98**

VI.1.	<i>Conclusions</i> .....	98
VI.2.	<i>Future works</i> .....	100

# *T*ABLE OF FIGURES

Fig. 1: Source-filter model of speech [Fant, 1960].....	18
Fig. 2: LF model of glottal flow signal.....	19
Fig. 3: Glottal flow contribution in speech signals.....	20
Fig. 4: Spectral variations due to variations in phonation.....	23
Fig. 5: ZZT plots on z-plane.....	31
Fig. 6: ZZT of two exponential functions.....	33
Fig. 7: ZZT of a damped sinusoid.....	34
Fig. 8: ZZT pattern of a typical differential LF signal.....	35
Fig. 9: Effect of parameter variation on the ZZT of glottal flow signals.....	37
Fig. 10: Effect of spectral tilt variation to ZZT of glottal flow signals.....	38
Fig. 12: ZZT patterns for a single excitation speech frame.....	42
Fig. 13: Effect of windowing location on ZZT of a real speech signal.....	43
Fig. 14: GCI synchronous windowing.....	43
Fig. 15: Effect of window function to ZZT patterns.....	44
Fig. 16: GCI synchronous Hanning-Poisson asymmetric windowing.....	45
Fig. 17: Effect of window size on ZZT of windowed speech signals.....	46
Fig. 18: Aperiodic component contribution to ZZT of speech.....	47
Fig. 19: Geometric interpretation for spikes in the group delay function at frequency locations close to a zero.....	52
Fig. 20: Effects of zeros to the spectrum of a signal.....	52
Fig. 21: Modified group delay function.....	54
Fig. 22: Effect of windowing to phase characteristics of a signal.....	55
Fig. 23: Effect of anti-causality on the time-domain waveform.....	57
Fig. 24: The mixed-phase speech model.....	58
Fig. 25: Effect of windowing location to group delay of a real speech signal.....	60
Fig. 26: Effect of windowing size to group delay of a synthetic speech signal.....	61
Fig. 27: Effect of windowing function to group delay.....	62
Fig. 28: Group delay spectrogram.....	63
Fig. 29: Remaining problems for group delay of GCI synchronously windowed data.....	65
Fig. 30: First method of chirp group delay processing.....	65
Fig. 31: Effect of ZZT on chirp group delay.....	66
Fig. 32: Second method of chirp group delay processing.....	67
Fig. 33: The ZZT-decomposition algorithm.....	70
Fig. 34: Synthetic speech signal for testing ZZT-decomposition.....	71
Fig. 35: ZZT-decomposition results of synthetic speech.....	71
Fig. 36: ZZT-decomposition example on a synthetic speech frame (comparison of ZZT-decomposition and PSIAIF).....	72
Fig. 37: ZZT of the real speech frame (/a/) windowed synchronously with GCI.....	73
Fig. 38: ZZT-decomposition result for a real speech frame.....	74
Fig. 39: Tests for robustness to GCI estimation errors.....	75
Fig. 40: Tests for robustness to $F1$ variations.....	75
Fig. 41: Tests for robustness to additive noise and return phase variations.....	76
Fig. 42: Comparison of complex cepstrum and ZZT decomposition results.....	78
Fig. 43: $Fg$ estimation results for two excitation signals.....	81
Fig. 44: Magnitude spectrums of ZZT-decomposition for the worst glottal formant frequency estimation with peak picking.....	82
Fig. 45: Comparing glottal formant frequency estimate with inverse of open quotient estimate and $f0$ estimate ( $k=115$ ).....	83
Fig. 46: The Differential-Phase Peak Track (DPPT) algorithm.....	84
Fig. 47: Parameter space of the synthetic speech stimuli.....	85
Fig. 48: DPPT formant tracks (dots) and formant synthesis parameters (solid lines).....	86
Fig. 49: Differential phase spectrum peaks indicated on the spectrogram of a speech signal.....	87
Fig. 50: The final version of the formant tracking algorithm using peak picking on chirp group delay computed outside the unit circle.....	88
Fig. 51: Histogram of $R$ values obtained by the iterative procedure.....	88

Fig. 52: Formant tracking example 1 .....	89
Fig. 53: MixLP algorithm flow diagram.....	91
Fig. 54: Re-synthesized glottal flow signals obtained with the MixLP and PSIAIF algorithms ....	92
Fig. 55: MixLP open quotient estimation results.....	92
Fig. 56: ASR process .....	93
Fig. 57: Time-domain signal of a 30 ms speech frame and its group delay function. ....	94
Fig. 58: Power spectrum (PowerS) and group delay based functions for the speech signal frame in Fig. 2.....	94
Fig. 59: Spectrogram plots of a noise-free utterance. ....	95
Fig. 60: Window functions .....	103
Fig. 61: Formant tracking example 2 .....	106
Fig. 62: Formant tracking example 3 .....	107
Fig. 63: Formant tracking example 4 .....	107
Fig. 64: Formant tracking example 5 .....	108
Fig. 65: Formant tracking example 6 .....	108
Fig. 66: Formant tracking example 7 .....	109
Fig. 67: Formant tracking example 8 .....	109
Fig. 68: Formant tracking example 9 .....	110
Fig. 69: Formant tracking example 10 .....	110

# Table of Acronyms

ASR	Automatic Speech Recognition
CGD	Chirp Group Delay
CGDGCI	CGD of the GCI-synchronously windowed speech
CGDZP	CGD of the zero-phase version of a signal
DFT	Discrete Fourier Transform
EGG	Electro-glottograph
FFT	Fast Fourier Transform
FT	Fourier Transform
Fg	Glottal formant frequency
F1, F2, F3, F4	Vocal tract formant frequencies
GCI	Glottal closure instant
GF	Glottal Flow
GD	Group Delay
GDGCI	Group Delay of GCI-synchronously windowed speech
LF	Liljencrants-Fant
LP	Linear Prediction
MFCC	Mel frequency cepstrum coefficients
OLA	Overlap Add
PDA	Pitch Detection Algorithm
PS	Product Spectrum
TTS	Text-To-Speech
ZZT	Zeros of the Z-Transform

---

# Chapter I:

## Introduction

---

---

### I.1. Motivations

Starting from 60's speech processing has attracted many researchers. The speech research community and the literature have grown largely in the last 40 years. Today, yearly conferences are dedicated to speech processing (ICSLP and Eurospeech) which gather thousands of researchers from all around the world. Speech technology has taken important steps especially in the last 10-15 years. For single speaker-language settings, very high quality speech synthesis and recognition are achieved. With existing tools and large speech corpora available to many people, even small research labs can develop high quality systems. It is interesting to note that the tools and approaches multiplied in time but most of today's high quality systems have high similarity in terms of structure and performance. We seem to have reached a "plateau" in technology for many speech processing problems where break-through paradigm changes are necessary for further development. For that purpose, we need to re-ask our selves the fundamental questions, try to find gaps in our understanding of the speech phenomenon, try to have access to the "unused" information available in our data. This thesis research tries to follow such a direction and mainly deals with very basic speech processing issues: source-tract separation, speech modeling, spectral processing and tries to propose new ways out of the mainstream to some of these problems.

#### The (hi)story of this study

The initial goals of this research was studying voice quality variations in speech: an area which we believe to be one of promising speech signal processing topics. We immediately faced the well-known basic problem of source-tract decomposition for such a study. Although high quality decomposition is indispensable, the state-of-the-art algorithms have very limited quality. We think that source-tract decomposition stays at the very heart of many speech processing problems and this rather fundamental research topic deserves more attention than the existing efforts. Therefore we have updated our targets and decided to study this basic problem.

For decomposing two components from a given signal, one has to ask the questions: what is the main difference between the two components? For which features of the signal the components have different characteristics? One of the discriminating features we could find is the "causality": the glottal flow (source) component exhibit anti-causal spectral characteristics [Gardner, 1994, Doval & d'Alessandro, 1997] in contrast to the vocal tract filter component. The path to follow was obvious from that point on: we needed to find ways of tracking anti-causal and causal components/information from a given signal. It is a common practice to study the Fourier Transform (FT) spectrum of signals for almost all signal processing problems, and it was also our first direction. In the FT spectrum domain, the causality information is coded in the phase part of the spectrum. (Un) fortunately, phase processing is known to be another difficult topic for which tools/algorithms/advances are rather limited and many problems exist. We were left with two basic problems: source-tract separation and phase processing.

The further questions were: why phase processing is difficult? What are the obstacles? Some answers we could find in literature were: phase information is in wrapped form and unwrapping is

necessary and difficult; phase derivative includes peaks for frequency bins where a zero/root of the  $z$ -transform polynomial occurs very close to the unit circle that make the unwrapping operation difficult [Yegnanarayana et al, 1984]. Therefore we decided to study the zero/root locations of signals'  $z$ -transform on the  $z$ -plane.

Once basic questions and possible paths to be followed are available what a researcher needs more is perhaps some luck. I was in VOQUAL03-Geneva [www-Voqual03] when luck hit me: Prof. Kawahara from Wakayama University/Japan came after my presentation to show me the matching points between his observations and my points/discussions and showed me a one minute movie which created the “butterfly effect”. It was obvious from that point on that I had to study windowing effects to phase spectrum and zero locations. If you continue reading this thesis I think you'll agree that it was “the butterfly”. To keep some interest of the reader to Prof. Kawahara's work (and since it is a bit off-topic), I do not mention in detail what I have seen in that movie.

In short, this thesis issues two fundamental speech processing problems: source-tract separation and phase processing; it proposes new spectral representations to study these problems: the zeros of the  $z$ -transform (ZZT) representation and various chirp group delay functions; it proposes algorithms using these representations in the applications: glottal flow parameter estimation, formant estimation and feature extraction for speech recognition.

---

## I.2. Original contributions of the thesis

### ZZT Representation of signals

In this thesis, we introduce a new spectral representation for a signal: the zeros of the  $z$ -transform (ZZT) representation. There are mainly two useful points of the ZZT representation for speech signals: i) it sheds light into many difficulties involved in phase spectrum processing and for this reason provides us with the opportunity to design better methodologies, ii) patterns exist in the ZZT of speech signals which make it possible to design a new spectral decomposition method for source-tract separation.

Being a form of  $z$ -transform representation, ZZT representation is especially useful for studying some properties of the Discrete Fourier Transform (DFT) of a signal. Through a systematic study of ZZT of windowed speech, we show that windowing lies at the very heart of the problem of spikes in the derivative of phase spectrum due to zeros close to the unit circle. To obtain spike-free group delay functions, all zeros should be at some distance from the unit circle and we can guarantee existence of a zero-free region around the unit circle only for a very special case of windowing: glottal closure instant synchronous windowing with a size of two pitch periods and with one of the three windowing functions: Blackman, Gaussian or Hanning-Poisson.

The fact that we can obtain spike-free group delay functions is an important step for phase processing. There are actually plenty of signal processing applications, which can benefit from the results of this study. In many signal processing studies, phase estimation is considered to be a difficult problem and discarded. However, for some applications, the phase information is essential or at least is an important factor of the efficiency of the algorithms. The methods defined in this thesis provide hopefully a potential to remove some of the obstacles in the phase estimation problem. Some of these applications are listed in Section II.4 dedicated to review of state-of-the-art for phase processing of speech.

The systematic study of the ZZT of windowed speech signals has one more important output: separate patterns for the glottal flow and vocal tract contributions can be observed. The ZZT representation includes two lines/groups of zeros: one outside the unit circle and one inside the unit circle with gaps creating formant peaks on the spectrum. These observations have led us to design a spectral source-tract separation algorithm based on ZZT-decomposition. Our methodology involves no modeling but direct separation in the spectral domain. In addition, such an observation both supports the studies discussing anti-causality of the glottal flow component [Jackson, 1989, Gardner, 1994, Doval *et al*, 2003] and suggests a mixed-phase model for speech signals. For

completeness of the theoretical background, we also discuss a mixed-phase model for speech signals through ZZT patterns of source-filter model of speech.

### Chirp group delay processing

In this thesis, we propose use of new ‘chirp group delay functions’ for speech analysis. The chirp group delay simply corresponds to the group delay function computed on a circle in  $z$ -plane other than the unit circle. It is the negative derivative of the phase component of the chirp  $z$ -transform.

The necessity of such a representation is due to the difficulties involved in group delay processing. We first discuss these difficulties, show that windowing is very important through ZZT representation and we propose criteria for optimum windowing. Although the group delay one can obtain after appropriate windowing is incomparably smoother and spike-free, there are still some problematic issues for some special cases like: reliable glottal closure instant (GCI) detection for synchronicity of the window is not always easy and it is not possible to guarantee absence of zeros close to the unit circle for noisy speech even when windowing is appropriately performed. Therefore we have developed chirp group delay processing as an alternative, for which spike-freeness can be guaranteed. Robust spectral processing can be achieved using this representation.

### Applications of ZZT and chirp group delay

Using the proposed representations and functions, various algorithms are developed in this thesis for formant tracking, glottal flow parameter estimation and feature extraction for speech recognition. Due to time constraints the application areas are kept limited to these topics however we believe that the two representations can be further used in many other speech analysis applications. Some of these potential areas are discussed throughout the manuscript in various sections.

---

## I.3. Plan

The thesis manuscript starts with this introduction chapter dedicated to presentation of motivations and achievements together with a state-of-the-art review for the applications targeted. Following the introduction chapter, the main contributions of this thesis are presented in two parts: theory and applications.

The first part is titled “**Spectral representation of speech by zeros of the  $z$ -transform (ZZT) representation and chirp group delay**” and it presents the two representations developed within this thesis work. The first chapter of this part (chapter III) is dedicated to the ZZT representation. First, the ZZT representation is presented as a form of  $z$ -transform representation of discrete time signals. Then source-filter model of speech is studied through the ZZT representation. Finally the ZZT of windowed speech signals is studied (which is very important for real-life applications). The second chapter (chapter IV) of this part is dedicated to chirp group delay processing. First, a review of group delay processing theory is presented. Then difficulties of group delay processing are discussed and finally chirp group delay function is proposed as an alternative phase based representation for which problems of group delay processing are avoided. This chapter also includes study of group delay characteristics of a mixed-phase speech model.

The second part, “**Applications of ZZT and chirp group delay processing in speech analysis**”, is dedicated to application of the two representations. First, we describe a source-filter decomposition algorithm based on ZZT representation (the algorithm is named as ZZT-decomposition). The second application presented is a glottal flow parameter estimation algorithm using the ZZT-decomposition. Third application is in formant tracking and both the ZZT representation and chirp group delay processing theory is utilized. The fourth algorithm presented is based on linear predictive (LP) modeling of mixed-phase speech signals. This algorithm does not include ZZT or chirp group delay but the mixed-phase speech model discussed in the second part



and the well-known LP-covariance approach to modeling/analysis from literature. Finally the fifth application is in speech recognition. In this part we show that chirp group delay functions can be effectively used in speech recognition systems to improve recognition rates.

Finally, the manuscript closes with the “**Conclusion and future works**” part that summarizes the outputs of the study, lists unanswered questions and proposes future works using ZZT representation and chirp group delay processing theory.

---

# Chapter II:

## State-of-the-art

---

### II.1. Introduction

According to the well-known source-filter model for speech (Fig. 1), speech signals are produced by exciting the vocal tract system by periodic source (glottal flow) signals. Speech analysis studies mainly target analysis of short-time and long-time variations in the characteristics of these two components.

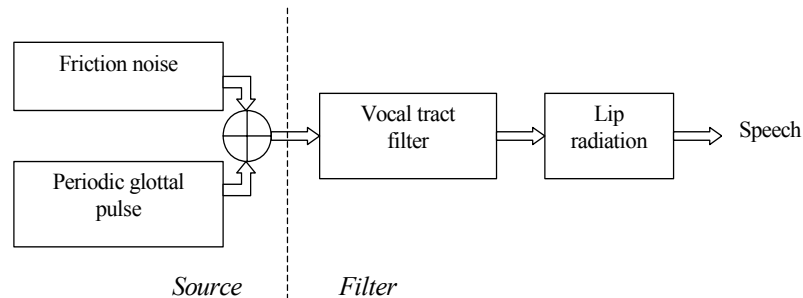


Fig. 1: Source-filter model of speech [Fant, 1960]

Estimation of the vocal tract filter properties and glottal flow estimation are the two problems addressed in most studies in the area of speech analysis. Studies on glottal flow and vocal tract components of speech have the potential to extend our knowledge and understanding of dynamics of speech signals and get more in-depths to natural human communication system. Although these topics have been extensively studied during last 40 years it is very likely that both of the topics will continue to be open for the years to come.

*Glottal flow estimation* is important in many speech applications. In pathological voice processing for diagnosis and therapy, reliable glottal flow estimation is of great importance since perturbations in glottal flow component are considered to be one of the main sources of speech disorders. Glottal flow estimation is also important for voice quality (the auditory ‘coloring’ of a person’s voice) analysis, which is a topic gaining popularity in especially speech synthesis for generation of affect. Some other application areas for glottal flow estimation are: prosodic annotation of speech (stress labeling), expressive or emotional speech synthesis, speaker identification, emotion recognition, high quality parametric speech synthesis.

Natural resonances of the vocal tract filter are named as *formants*, which are the representative features of the vocal tract system. Many researchers have studied ways of tracking formant frequencies and the literature issuing this problem is growing constantly. One of the main applications of formant tracking is parametric speech synthesis in which speech is synthesized by exciting a time-varying vocal tract filter. Formant tracking is important for understanding/modeling vocal tract filter dynamics and for further designing rules needed for computation of filter coefficients during synthesis time. Another application area of formant tracking is speech recognition [Deng & Sun, 1994, Welling & Ney, 1998].

Spectral analysis techniques have been used in speech processing for many years, in many applications. In most of those applications, estimation of some global characteristics is aimed and a smoothed form of the magnitude spectrum obtained from Fourier Transform (FT) spectra is

utilized. Due to difficulties in processing phase spectrum, researchers have preferred working on the magnitude spectrum most of the time. However, amplitude carries only a part of the spectral information and studying phase characteristics, we may get access to some information not completely available in the magnitude spectrum.

This thesis discusses these three issues (source and vocal tract parameter estimation and spectral processing) in speech analysis. A detailed literature review on all of these topics is likely to end up in being a complete PhD thesis due to the size of the literature and the variety of approaches. To be able to concentrate on the topics that are original in this thesis, we first present a general view of the state-of-the-art in these application topics in the following section. Detailed reviews of the state of the art of some specific topics (like group delay processing theory) are presented in the sections where we present our original approaches.

## II.2. Glottal flow estimation and voice quality analysis

Glottal flow (i.e. source) estimation from recorded speech signals is a problem extensively studied by many researchers in the last 40 years. It is considered as one of the important and difficult problems of speech processing. Various techniques have been proposed for the estimation and various models have been introduced describing the glottal flow signal for voiced speech. Among the glottal flow models (a review of the popular models is available in [Doval & d'Alessandro, 1999] and in [Cummings & Clements, 1995]), the Liljencrants-Fant (LF) model [Fant, 1985] is the most frequently used one.

In Fig. 2, the periodic LF model glottal flow derivative signal ( $dUg(t)$ ) is presented together with the glottal flow ( $Ug(t)$ ) signal (scaled for better comparative viewing). The time parameters ( $Tp$ ,  $Te$ ,  $Tc$ ,  $Ta$ ) indicate the zero-crossings and function change points, which serve as landmarks on the waveform.  $Ee$  defines the amplitude of excitation and  $T0$  defines the fundamental pitch period. The LF model signal is composed of two waveform segments concatenated at glottal closure instant (GCI), the instant of maximum negative peak of the glottal flow derivative signal. The first segment is referred as the first phase (Eq. 2.1). The length of the first phase is defined by the  $Te$  parameter. The first phase includes the maximum peak of the glottal flow signal and the location of the maximum peak is defined by the  $Tp$  parameter (or by  $a_m$ , the asymmetry coefficient which indicates the location with respect to the length of the first phase). The open quotient ( $Oq$ ) is used as a measure of glottis open duration within a pitch period. The second segment is named as the return phase (Eq. 2.2) and characterizes the closure of the glottis. The effective duration of the return phase is indicated by the  $Ta$  parameter. The duration of the LF model signal is  $Tc$  and together with a zero-valued segment (Eq. 2.3) a complete pitch cycle is obtained with length  $T0$ .

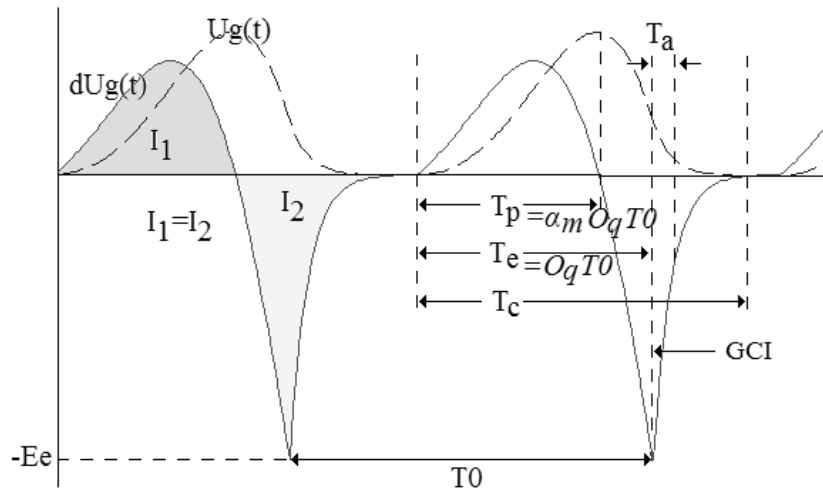


Fig. 2: LF model of glottal flow signal.

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t), 0 \leq t \leq T_e \quad (2.1)$$

$$g(t) = -\frac{E_e}{\varepsilon T_a} \left[ e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)} \right] T_e \leq t \leq T_c \leq T_0 \quad (2.2)$$

$$g(t) = 0, T_c \leq t \leq T_0 \quad (2.3)$$

As time-domain signals are concerned, the differential glottal flow component can very vaguely be observed on the speech pressure signal recorded (see Fig.3a) and once lip radiation (which is often modeled as a differentiator) is compensated by integration of the speech signal, the glottal flow signals can be observed more apparently (see Fig.3c). In the spectral domain, the glottal flow component contributes to the speech spectrum with two components [Doval & d'Alessandro, 1999]: the so-called glottal formant (due to the first phase, Eq. 2.1) and the spectral tilt (due to the return phase, Eq. 2.2). As the spectrum is concerned, visual observation of the glottal flow contribution is more difficult. The glottal formant can be observed (Fig.3b) for some vowels and phonation types for which the first vocal tract formant frequency is relatively high (for example for low pitch /a/).

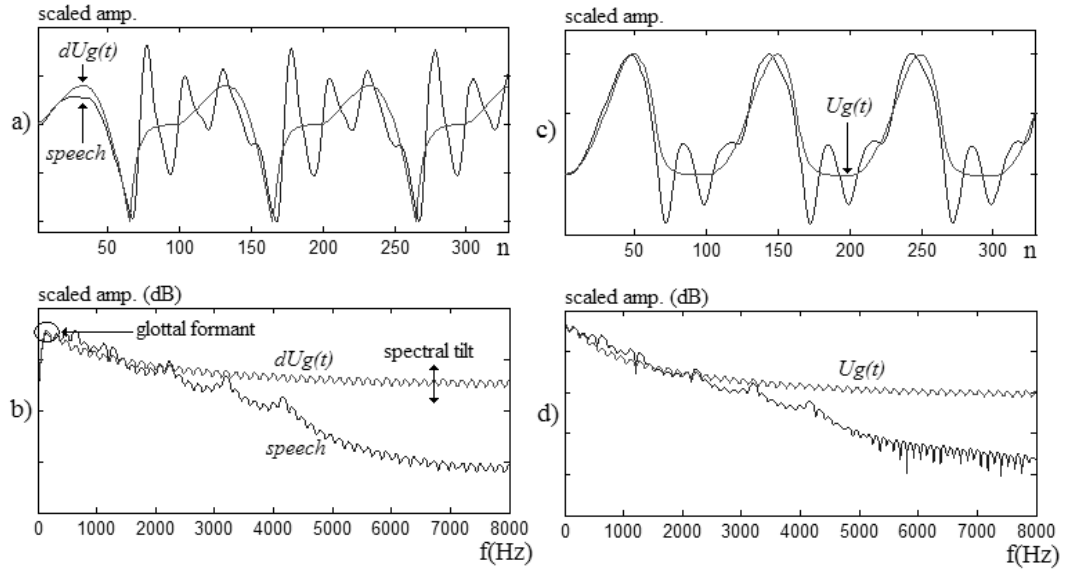


Fig. 3: Glottal flow contribution in speech signals.

- a) time-domain synthetic speech signal (with formants at 600Hz, 1200Hz, 2200Hz, 3200Hz and 4200Hz) and LF model glottal flow derivative, b) magnitude spectrum of synthetic speech and glottal flow derivative, c) time-domain integrated synthetic speech signal and glottal flow signal, d) magnitude spectrum of integrated synthetic speech and glottal flow.

Most (if not all) glottal flow estimation (from recorded speech signals) methods in literature suffer from robustness problems and the techniques, that are known to be best, work only for very limited conditions (normal phonation, sustained vowels with high first formant frequency and moderate or low pitch). Glottal flow estimation is usually accompanied by parameter estimation techniques, which compute the parameters indicated in Fig. 2 that describe the LF model signal. Therefore, there are actually two problems to be solved: estimating the glottal flow signal from speech signals and extracting the parameters of the model from the estimated glottal flow. Below we present a short review of the state-of-the-art in glottal flow estimation. For a more detailed review of the theory and techniques, the reader is referred to the recent PhD dissertation by Christer Gobl [Gobl, 2003].

## Glottal flow signal estimation methods

In the basic source-filter model, speech production is a “forward only” system without interaction between source and vocal tract filter. Many studies are based on the property that linear decoupling of vocal tract (called ‘inverse filtering’) is possible, which is not completely true [Rothenberg, 1981]. However, with the existing approximations and estimations, we are still able to understand some phenomenon and build high quality formant synthesis, which proves the potential of currently existing linear techniques. For this reason and also for its low complexity, linear techniques are still much more popular than non-linear techniques.

One of the earliest ways of inverse filtering is recording the volume velocity waveform by the help of a flow mask [Rothenberg, 1973] and removing formants manually by anti-resonance circuits [Gauffin & Sundberg, 1989]. This technique needs utilization of a specially designed mask and manual inverse filter design. Manual inverse filtering techniques are reported to be sometimes more reliable than automatic inverse filtering techniques [Gobl & Chasaide, 2001], but they are time consuming and subjective. Therefore analysis of large speech databases by manual inverse filtering is inappropriate and robust automatic tools are needed. Manual inverse filtering is more appropriate for studies aimed at investigating certain phenomenon of speech with a small set of data.

One automatic technique that influenced many researches in all areas of speech processing is the Linear Predictive (LP) modeling of speech [Makhoul, 1975], which assumes that each speech sample can be expressed as a linear combination of the past samples (referred to as all-pole modeling). Both the glottal excitation signal and the vocal tract can be modeled by all-pole systems [Jackson, 1989, Gardner, 1994, Doval *et al.*, 2003] and LP analysis applied directly on speech signals provides an all-pole system that is a combination of the two systems. The main difficulty is obtaining the vocal tract part of the all-pole system required for inverse filtering to get the glottal flow signal.

The techniques of glottal flow estimation by LP inverse filtering can be classified by the approach they use to get the vocal tract part. One class of studies uses only some part of the speech waveform called the closed-phase (referring to portion of the pitch period glottis being closed), which is supposed to include only the vocal tract filter impulse response [Wong *et al.*, 1979]. Therefore, such systems need to detect the closed-phase and apply LP analysis to estimate directly the vocal tract filter. Such an approach suffers from the difficulties in finding the closed-phase correctly and with enough duration (i.e. the closed-phase needs to be long enough to be able to estimate the vocal tract filter coefficients accurately). For breathy phonation for example, the existence of closed-phase cannot be guaranteed at every pitch period.

A second class of studies jointly estimates vocal tract and source [Milenkovic, 1986, Lu & Smith, 1999]. These studies utilize a linearly separable source-filter model and a linear model for the source signal. The estimation procedure jointly determines auto-regressive (AR) model of the vocal tract response together with the parameters of the glottal flow model.

Another class of studies cover the iterative approaches [Alku, 1992 a], which start estimation with a rough initialization for the source and tract components and try to separate the two systems in an iterative procedure. There is actually one well-known problem in LP analysis, which is general to all types of LP-based analysis: the dependency on the analysis order (i.e. the number of past samples used in the model). Once the number of resonances available in the speech signal does not match the number of pole-pairs<sup>1</sup> in the model, either spurious resonances are detected or some are missed. In the iterative approaches, the number of poles in the glottal signal and in the vocal tract filter are fixed to constant values (usually 3 for the glottal flow component, corresponding to a real pole and a complex pole-pair) and it is hard to obtain a reliable pole separation with fixed number of poles. In addition, the detection of a glottal formant at an equal or higher frequency than the frequency of the first vocal tract formant ( $F1$ ) pole is very difficult and remains as an important obstacle to source-tract separation by LP analysis.

One of the main difficulties of inverse filtering studies is the lack of available reference data; therefore there is difficulty in comparing various methods. Fortunately, recently some efforts have been made to collect-provide common data (in the voice quality workshop VOQUAL03 [www-Voqual03], a useful data collection is provided for public use). However, this serves as the first

---

<sup>1</sup> Please refer to Appendix B for a short description of the LP model and relation between pole-pairs and resonances.

step only. Because given a speech signal, it is very difficult to judge which inverse filtered version is closer to its real glottal flow. A common agreed (among experts) glottal flow estimate data is of ultimate need for future improvement in inverse filtering techniques for comparing existing methods. Most often controlled tests on synthetic speech are performed, which are currently the only way to have reference data. However, the robustness of methods varies a lot when we switch from synthetic speech to real speech (see our tests for comparing robustness of three formant tracking algorithms on real and synthetic speech signals in section V.3.2). An alternative path to follow is to compare results of glottal flow estimation from speech signals and results obtained using other modalities (like video data [Alku *et al*, 2000] or EGG [Henrich *et al*, 2000]) and check correlation of results. We think this is currently the most appropriate way of testing algorithms, therefore in our tests we used this approach.

## Glottal flow parameter estimation methods

Once an estimate of the glottal flow signal is obtained, parameters of a glottal flow model should be calculated to be able to continue with higher level studies on detection of phonation type, studying voice quality variations, etc.

In [Strik, 1998], some of the available parameter estimation methods are compared. We enlarge his classified list of methods below.

Parameter estimation can be performed on the time domain signal by detecting landmarks of the signal (minima, maxima, zero crossings) [Alku, 1992 b]. These methods are not very robust to noisy data since the time domain waveforms and landmarks vary a lot with noise. More robust time-domain methods also exist like estimation of  $NAQ$  (or  $Rd$ ) parameter [Fant, 1997, Bäckström *et al*, 2002].  $NAQ$  (or  $Rd$ ) parameter is a single parameter representation of the glottal flow signal and may provide good initial estimates of parameters for iterative techniques.

Fitting a source model to time domain data is maybe the most frequently used approach [Riegelsberger *et al*, 1993, Plumpe & Quatieri, 1999, Childers, 1995]. Most of the glottal flow models are nonlinear models. Therefore the parameter estimation problem is actually a nonlinear curve-fit problem. Many studies include nonlinear least squares estimation methods (Gauss-Newton, gradient algorithms, simplex search, etc.) in their procedures. While being more robust to noise, these methods have difficulty with phase distorted data since the waveform shape is the main source of information for this type of analysis.

Frequency domain methods seem to be more robust in handling both noisy and phase distorted data [Oliveira, 1993, Alku *et al*, 1997, Ding & Kasuya, 1996]. Equally, the glottal flow signal is modeled by an all-pole filter and parameters are estimated using LP analysis in some studies [Jackson, 1989, Gardner, 1994, Childers, 1995, Henrich *et al*, 1999], which can be considered to be a frequency-domain method.

Apart from parameter estimation from estimated glottal flow signals, there exist some parameter estimation methods directly on the speech signal spectrum. Doval and d'Alessandro's study [Doval & d'Alessandro, 1997], showed that glottal flow characteristics contribute to the speech signal spectrum with two components: the so-called glottal formant and spectral tilt. Considering the difficulties and robustness problems of inverse filtering, estimation of spectral parameters directly from speech spectrum related with phonation changes is potentially a good direction to follow (i.e. tracking variations on those two components in speech spectrum). In [Hanson & Chuang, 1999], several spectral parameters based on harmonic amplitudes are proposed (like  $H1-A3$  as a measure of spectral tilt) and have been tested in many studies. However, very few studies report successful utilization of such parameters in real applications. We believe that this area is still open for further research and potentially will result in advanced methods with higher robustness than the existing methods.

One of the main application areas of glottal flow signal and parameter estimation is in voice quality research. We find it important at this point to provide a short review of the current state-of-the-art in voice quality analysis in concatenative TTS since some outputs of this research are directly applicable in some problems in this topic and it is within our primary motivations for future works of this thesis.

## Applications of glottal flow estimation in voice quality analysis for concatenative TTS

Voice quality in speech synthesis research usually refers to the perceived degree of characteristics like breathiness, creakiness and loudness. Voice quality variations are considered to be mainly due to the variations in the phonation (production of the glottal excitation signal) process at the glottis. Voice quality issues have often been studied within the context of formant synthesis in the speech synthesis area since in this type of parametric approach, glottal excitation signals are synthesized parametrically and therefore can rather easily be controlled. Few studies address voice quality issues in concatenative synthesis [Alessandro *et al*, 1998, Kawai & Tsuzaki, 2004, Stylianou, 1999, Campbell & Marumoto, 2000]. But voice quality analysis/synthesis is drawing more and more attention in the domain of concatenative speech synthesis since it is one of the most important features of naturalness in speech. In addition, it is especially important for emotional/expressive speech synthesis (an area which is gaining popularity) since voice quality codes as much information about the state of the speaker as does the prosodic information [Campbell & Marumoto, 2000]. There is a strong correlation between voice quality variations and prosodic variations in speech since both are features of the phonation process. Therefore, advances in one of the fields would potentially result in advancement in the other. However, current state of the art in voice quality analysis/modification/synthesis of recorded speech is not yet advanced enough to be widely used since tools for estimating and modifying features of the glottal excitation are necessary; this is a challenging problem due to non-linear processes in speech production.

From a signal processing point of view, voice quality variations mainly correspond to variations in spectral tilt, in the relative amount of aperiodic components in speech, and in some spectral variations in the low frequency part of the spectrum (like variations in the glottal formant frequency ( $F_g$ ), in the first formant bandwidth, in amplitudes of the first few harmonics, etc.) [Alessandro *et al*, 1998]. In Fig. 4, we demonstrate (on synthetic speech signals) how variations in two spectral features (glottal formant frequency,  $F_g$ , and high frequency band energy) correspond to a variation of phonation in the tense-soft dimension. The time-domain signals presented at the bottom figures include the glottal excitation (glottal flow derivative) signal and the speech signal obtained by filtering this glottal excitation signal with an all-pole vocal tract filter with resonances at 600Hz, 1200Hz, 2200Hz and 3200Hz. A variation from tense phonation to soft phonation corresponds to a decrease both in glottal formant frequency and high frequency band energy (and vice versa). For a detailed study of acoustic feature variations due to voice quality variations, see [Klatt & Klatt, 1990]. In the following paragraphs we provide a list of current-future issues in voice quality for concatenative text-to-speech (TTS).

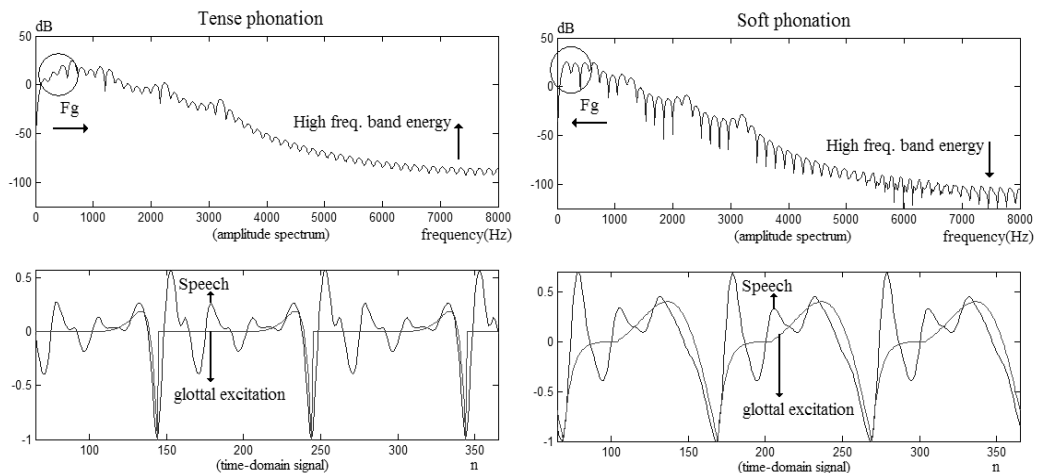


Fig. 4: Spectral variations due to variations in phonation.

Top figures show the magnitude spectrum of the speech signals and bottom figures show the time domain signals for glottal excitation and speech.

Voice quality studies in concatenative speech synthesis research are concentrated mainly on voice quality labeling of speech corpora. In corpus construction for concatenative synthesis, one of the main difficulties is the need for long recording sessions and this brings; as a matter of fact, voice

quality changes during sessions (for example the voice fatigue effect at the end of sessions). Once long sessions are split into smaller sessions to avoid fatigue, then rises the problem of matching voice qualities between sessions. Voice quality variations may be induced by many other factors, which are related to physical properties of the vocal folds or even psychological factors (when the speaker gets bored, he/she may start speaking fast-tense to be able to finish the recording quickly). Such variations in speech corpora potentially result in an increased degree and frequency of acoustic discontinuities in synthetic speech. This problem has been addressed in some recent studies [Campbell & Marumoto, 2000, Kawai & Tsuzaki, 2004] and is an open research area.

Most often, voice quality variations are treated as very slow variations (for example variations between two recording sessions of several hours) but actually, voice quality differences do not only exist between segments from two different phrases. Voice source characteristics may show quite fast variations and voice quality discontinuities may exist even between two segments of the same sentence. One of the reasons why voice quality discontinuities are not seen as a major concern by current TTS systems is that unit selection is based on phonological structure matching and therefore implicitly selects segments with substantial voice quality continuity. However, voice quality and prosody are not completely interdependent and the realization of various voice quality effects for the same prosodic pattern and phonological segment is possible, and even desirable for emotional speech synthesis.

There are mainly two ways to avoid the discontinuities introduced by voice quality mismatches: using off-line signal processing to equalize some voice quality variations, or including voice quality features in concatenation costs for unit selection to guarantee voice quality variation continuity. These are among our future goals of research.

The advantage of off-line processing for voice quality equalization is that it does not introduce additional complexity in unit selection or to run time signal processing. The second solution stays rather as an unstudied subject apart from studies investigating which acoustic features are correlated with voice quality variations. One such study is [Kawai & Tsuzaki, 2004] which investigates the correlation of various measures (MFCC, Spectral tilt, band limited power) to perception scores of voice quality discontinuities constructed by concatenating phrases recorded over a long period of time (up to 202 days) and conclude that band limited power (8-16KHz) is a good measure for detecting voice quality discontinuities. This area is likely to be studied in the concatenative emotional speech synthesis area in the upcoming years.

One other possible solution to the voice quality discontinuity problem is voice quality smoothing on the fly, which is another unstudied subject. For such an operation, some means of voice quality modification without audible degradation in the segmental quality of speech are clearly needed. Most of the concatenative synthesis algorithms, which include some means of spectral smoothing, smooth some of the voice quality discontinuities automatically (for example, spectral envelope smoothing operation results in smoothing of some spectral tilt discontinuity automatically). However such smoothing is too local for removing all voice quality variations. Higher quality smoothing can be possible once the separate components of discontinuity are understood (for example the discontinuity in the aperiodic component of speech).

Voice quality modification is not only useful for smoothing voice quality variations but also for synthesizing variations on purpose. A potentially very useful and interesting approach in emotional speech synthesis is to use some hybrid synthesis paradigm: concatenative synthesis with more control parameters (like energy, spectral tilt, relative energy of the aperiodic component, glottal formant frequency) than the conventional concatenative synthesizers (which mostly include only duration and intonation information). Such an approach would stretch the limits of what is possible with the available data while keeping naturalness at a higher level compared to formant synthesis.

Natural voice quality modification is one of the unaccomplished goals of speech synthesis research. Defining high level rules that will drive low-level parameter modifications is too complex with our current understanding of the phenomenon. Even, signal processing for low-level parameter modification on real speech signals is being addressed in just a few studies.

As in many parameter modification schemes, there are mainly two classes of approaches: spectral modification techniques and parametric methods performing some form of decomposition-modification-synthesis. The latter is usually known to introduce some audible artifacts and spectral modifications are preferred when possible. Spectral techniques are especially advantageous for voice quality modification since the perceptual relevance of spectral parameters is high, the



representations are rather simplified (various time-domain models can be unified into a single spectral representation, as presented in [Doval & d'Alessandro, 1999]) and less sensitive to phase distortions (while time-domain voice quality modification applied on phase distorted speech signals often results in very low quality speech). d'Alessandro and Doval [d'Alessandro & Doval, 1998] draw guidelines for spectral modification for voice quality modification. Their method includes modification on three dimensions: glottal formant, spectral tilt and *PAPR* (periodic to aperiodic ratio). However, the implementation of their ideas in a speech synthesizer is still not tested and is potentially a promising research area.

---

## II.3. Formant Tracking

Automatic tracking of acoustic resonance frequencies of the vocal tract filter, the formant frequencies, has been another important speech analysis problem for many years. Many applications exist for formant tracking including parametric speech synthesis and speech recognition. The proposed algorithms for formant tracking in literature show large variety. We provide a list of approaches below without discussing them in detail to avoid large off-topic discussions.

Formants are observed as smooth peaks on the envelope of magnitude spectrum of short-time speech signals. Most of the formant tracking algorithms perform processing of magnitude spectrum of speech to detect these smooth peaks on the envelope [Schafer & Rabiner, 1970, Sun, 1995, Zolfaghari & Robinson, 1996, Chen & Loizou, 2004]. Another type of formant trackers use LP analysis where the formants are defined as poles in the all-pole vocal tract system function [McCandless, 1974, Snell & Milinazzo, 1993, You, 2004]. Potamianos and Maragos proposed a formant tracking method based on AM-FM demodulation of speech signals [Potamianos & Maragos, 1996, Potamianos & Maragos, 1999]. Some dynamic programming algorithms, which match multiple resonator responses to the speech spectrum, are proposed [Talkin, 1987, Chatwal & Constantinides, 1987, Welling & Ney, 1998, Xia & Epsy-Wilson, 2000, Laprie, 2004]. Algorithms based on hidden Markov models are also available [Kopeck, 1986, Yan *et al*, 2004]. The algorithm by Rao and Kumaresan [Rao & Kumaresan, 2000] uses adaptive filters to decompose speech into modulated components. Various modifications/implementations of this algorithm are derived thereafter [Bruce *et al*, 2002, Mustafa, 2003]. Watanabe proposed an algorithm using notch inverse filters mutually controlled to separate speech into single resonances [Watanabe, 2001].

A rather unpopular way for formant tracking is group delay processing. This is mainly due to the fact that the group delay function has a noisy structure. Early in 1985, it has been shown that formant tracks can be observed on the group delay spectrogram with appropriate post-processing [Friedman, 1985] and theoretically the group delay functions should provide better resolution for formant peaks compared to the magnitude spectrum [Murthy & Yegnanarayana, 1991 a]. Yegnanarayana and Murthy have studied the characteristics of group delay spectra [Yegnanarayana & Murthy, 1992] and their application to formant tracking [Murthy *et al*, 1989 a, Murthy & Yegnanarayana, 1991 a], and drawn the theoretical background for group delay processing. The formant tracking method we propose is also based on group delay processing. Therefore studies of Yegnanarayana and Murthy are discussed more in detail in Section IV.2.

Moderate-good quality of formant tracking has been achieved with most of the presented technologies however further improvement is still needed for applications like formant synthesis of speech. There actually is a lack of evaluation platform. Robustness of methods change from utterance to utterance and from synthetic speech to real speech, so it is difficult to draw conclusions out of tests with few speech examples.

---

## II.4. Phase Processing of Speech

Most of speech processing algorithms use spectral methods, i.e. some processing of the Fourier transform (FT) of speech signals. The magnitude spectrum has been the preferred part of Fourier

Transform (FT) spectrum in most of the speech processing methods although it carries only part of the available information. This is mainly due to the difficulties involved in phase processing. In speech processing research, phase processing is often cited as very difficult and most of the algorithms, which also need to reconstruct speech signals after modifying certain parameters, try to avoid phase related difficulties by “keeping the original phase information”. However, recent studies on speech perception report the importance of phase information [Paliwal & Alsteris, 2003]. Phase processing stands as one of unsolved, less studied topics and deserves much more attention for next generation speech processing applications. Many of the phase processing studies are based on trial-error due to obscurity of the phase related problems. There is an important lack in the understanding of the problems involved and in systematic ways of avoiding them. This thesis addresses some of these issues.

Phase processing is not only necessary for speech processing. It is essential for many fields of signal processing like: radar signal processing [Costantini *et al*, 1999, Chen & Zebker, 2002], medical imaging [Chavez *et al*, 2002, Frolova & Tact, 1996], source localization [Andersen & Jensen, 2001, Li & Levinson, 2002], etc. as well as many other research fields like optics, solid state physics, geophysics, holography, etc [Vyacheslav & Zhu, 2003].

Below, we summarize the phase processing studies in the speech processing area. Again detailed reviews of some of the methods that have similarity to our work are presented in the main text of the thesis where necessary.

## Phase processing in sinusoidal/harmonic modeling

One of the topics where phase processing is essential is sinusoidal/harmonic coding-modification-synthesis of speech signals. The sinusoidal/harmonic modeling literature is quite large; for detailed description the reader is referred to [Stylianou, 1996 b, Quatieri, 2002]. Here, we simply mention the phase related issues.

In the sinusoidal representation of speech, short-time speech signals ( $s(t)$ ) are considered to be composed of harmonically related sinusoids

$$s(t) = \text{Re} \left\{ \sum_{k=1}^{K(t)} A_k(t) e^{j(\theta_k(t))} \right\} \quad (2.4)$$

where  $A_k$  is the amplitude and  $\theta_k$  is the phase of the  $k$ -th harmonic. Plenty of methodologies for estimating these parameters have been proposed [McAulay & Quatieri, 1986, Griffin & Lim, 1988, Macon 1996, Stylianou, 1996 b].

The sinusoidal/harmonic modeling is largely and effectively used in speech coding applications [Marques, 1989, Marques *et al*, 1990, McAulay & Quatieri, 1991]. Reliable estimation and proper coding of harmonic phases is essential for speech coding applications and is reported to be a difficult problem. Often the phase related problems are tried to be avoided by various methods: using the zero-phase or minimum-phase phase spectrum obtained from magnitude spectrum information [Oppenheim, 1969], deriving the mixed-phase signal phase spectrum through complex cepstrum [Quatieri, 1979], phase compensation with some all-pass filtering at the speech reconstruction stage [Hedelin, 1988, Sun, 1997], etc. Most of such methodologies target avoiding the mismatch between what is expected and what is measured, using compensation methods.

High quality phase processing is also essential when sinusoidal/harmonic modeling is to be used in the context of concatenative synthesis, which is synthesis of speech signals by concatenation of pre-recorded speech segments. For such a task, recorded speech segments (the speech database), needs to be transformed into a parametric database by sinusoidal analysis (i.e. each short-time speech frame is represented by harmonic amplitude and phase parameters). During speech synthesis, speech segments are re-constructed with modified prosody (only pitch and duration are modified in most of the concatenative synthesis systems) and concatenated in such a way that no audible discontinuity exists at concatenation points. The phase related issues in concatenative speech synthesis can be considered in two dimensions: synchronicity of speech frames that are concatenated and phase spectrum discontinuities (or inter-frame incoherence and system phase incoherence [Stylianou, 2001]).

The concatenation of speech frames can be performed in many different ways (that can be grouped into two classes: overlap-addition (OLA) concatenation [Charpentier, 1988] and direct concatenation with phase continuity criterion [McAulay & Quatieri, 1986]). In all concatenation methods using sinusoidal representation, the continuity of harmonic phases of consecutive frames is a very important problem. Discontinuities in the harmonic phases result in reverberant, noisy and garbled speech. The most common way to guarantee the least possible audible discontinuity is to perform concatenation synchronously with the glottal closure instants (GCI) or removing some linear phase component from harmonic phases [Pollard, 1997, Stylianou, 2001].

Phase spectrum discontinuities are part of spectral discontinuities, which appear at concatenation points. The existence of spectral discontinuities at concatenation is well-known and the most preferred way to handle this problem today is to find an optimized way to record a large variety of speech segments [Klabbers, 2000] and guarantee the selection of segments with minimum spectral discontinuities [Moebius, 2000]. In systems where the coverage of the recorded speech database is smaller than the target domain of synthesis, spectral discontinuities are likely to occur; therefore the signal concatenation tools need to be equipped with some spectral smoothing methodologies. Magnitude spectrum discontinuities are usually removed by linear interpolation applied on a few concatenation frames around the concatenation point [Stylianou, 1996 b, Dutoit & Leich, 1993, Dutoit & Gosselin, 1996]. Although similar interpolation techniques are also proposed for the phase spectrum discontinuities [McAulay & Quatieri, 1986, Macon 1996, Stylianou, 1996 b], these type of discontinuities are considered to be too difficult to remove since a smooth reliable phase spectrum is needed which is difficult to estimate. One alternative way of handling the phase envelope discontinuities is resetting phase spectrum of all speech frames to a fixed phase spectrum estimate as in the MBROLA algorithm [Dutoit & Leich, 1993]. The payback is some reduction in naturalness of speech (discussed in detail in [Bozkurt *et al*, 2004 f]).

In addition, for high quality pitch modification of speech frames using the sinusoidal model, spectral envelope resampling is performed, that is: a smooth spectral envelope is obtained from the set of harmonic amplitudes which is supposed to include peaks and valleys due to the particular formant structure of the speech frame. To synthesize speech at a new pitch frequency, the modified set of harmonic parameters is calculated by resampling the spectral envelope at harmonic frequencies of the new pitch frequency. This operation ensures that the formant structure in spectrum is not altered (pitch modification, which is considered to correspond to changes only on the glottal excitation frequency, should not alter vocal tract information: the formant peaks and valleys on the spectral envelope). Theoretically, resampling of the spectrum needs to be performed both in phase and in amplitude domains. Therefore, reliable resampling of the phase spectrum is also important for high quality prosody modification.

## Phase processing in speech perception

Early investigations on the perceptual relevance of phase information came up with the conclusion that the human ear is phase-deaf [Helmholtz, 1875]. After a long period of time, the issue was restudied and falsifying results were obtained [Schroeder, 1959, Schroeder & Strube, 1986, Patterson, 1987]. Recently, although in very limited number, more and more researches study the perceptual relevance of phase information and show evidences about the importance of phase information in speech perception.

Through human perception experiments, Liu *et al* [Liu *et al*, 1997] and Paliwal and Alsteris [Paliwal & Alsteris, 2003] showed that the short-time phase spectrum contributes to speech intelligibility as much as the corresponding power spectrum. Pobloth and Kleijn [Pobloth & Kleijn, 1999] showed in a speech coding and psycho-acoustic research that human beings are able to distinguish between different phase spectra much better than often assumed. The studies of Kawahara *et al* showed that phase information plays an important role in high quality speech synthesis [Kawahara *et al*, 2001]. Banno *et al* [Banno *et al*, 2001] proved that the human auditory system is sensitive to the difference between zero and non-zero phase signals.

The studies listed showed that phase plays an important part in perception. However due to the difficulties in analyzing the phase content of signals, perception studies can be conducted within very limited boundaries. Without reliable phase estimation, conducting reliable perceptual experiments using real speech data is not possible. Therefore, phase analysis studies are essential for speech perception studies.

## Phase processing in speech analysis

Phase spectrum has been used in some studies for parameter estimation purposes, though quite limited in number compared to amplitude based parameter estimation methods. Charpentier [Charpentier, 1986] proposed a pitch detection algorithm based on short-time phase spectrum processing. Several studies address GCI detection using group delay processing [Smits & Yegnanarayana, 1995, Murthy & Yegnanarayana, 1999, Kawahara *et al*, 2000]. Yegnanarayana and Murthy have studied the characteristics of group delay spectra and the application to formant tracking in several papers [Murthy *et al*, 1989 a, Murthy & Yegnanarayana, 1991 a].

The main problem with phase based parameter estimation is robustness since phase analysis is very sensitive to windowing effects and noise. However we show in this thesis that it is possible to avoid most of the problems and estimate reliable phase information potentially useful for parameter estimation.

## Phase processing in automatic speech recognition

In most state-of-the-art automatic speech recognition (ASR) systems, again the amplitude/power spectrum has been preferred. Two recent studies address the possibility of using phase information in ASR systems [Hegde *et al*, 2004 b, Zhu & Paliwal, 2004]. The topic is rather very new and currently we have no clues that phase spectrum can bring complementary information to amplitude based systems for improving ASR quality. In the applications part of this study, we propose new group delay based functions and compare our group delay functions with the two group delay based functions in [Hegde *et al*, 2004 b, Zhu & Paliwal, 2004] in an ASR experiment. We show that the results are encouraging and indeed there exists some potential for using phase based features in ASR systems.

*First Part*

# SPECTRAL REPRESENTATION OF SPEECH BY ZEROS OF THE Z-TRANSFORM (ZZT) AND CHIRP GROUP DELAY

**I**n this part of this thesis, we introduce two spectral representations for speech analysis: the Zeros of the Z-Transform (ZZT) representation, which we define as the set of roots of the Z-transform polynomial for a discrete time signal and the chirp group delay, which refers to group delay computed from the chirp  $z$ -transform of a signal. This first part is dedicated to the theoretical study of these two representations and to a mixed-phase model for speech.

We start with presenting a theoretical study of the ZZT of speech signals in chapter three. We discuss both the ZZT of the source-filter model for speech and the ZZT of windowed real speech data. Chapter four is dedicated to the chirp group delay functions and to the mixed-phase model. In addition to the presentation of the theory, we also present the links between ZZT and chirp group delay processing and show that study of ZZT is essential for phase processing of speech signals.

---

# Chapter III:

## Zeros of the z-transform (ZZT) representation of speech

---

### III.1. Introduction

Spectral analysis techniques that aim at estimating some global characteristics and are based on processing a smoothed form of the magnitude spectrum have been used in speech processing for many years, in many applications. Spectral details are also studied (although not so often as smoothed spectral envelope) for example for estimating aperiodic contributions [Yegnanarayana *et al*, 1998] and maintaining the fine structure of speech spectrum for high quality re-synthesis [Spanias, 1994]. These details are assumed to be mainly due to noise or to aperiodic components of speech, or due to windowing effects. Most often, some form of filtering (for example cepstral smoothing [Rabiner & Schafer, 1978]) is used to get rid of their effect on the spectral envelope.

Here we introduce a new spectral representation for a signal: the zeros of the z-transform (ZZT) representation, which can be thought of as an “all-detail” representation. It is simply defined as the set of zeros of the z-transform polynomial of a discrete time signal. For most of the zeros in the ZZT set, the effect of a single zero on the FT spectrum is very local and can be considered to be a detail. However it is the sum of the contributions of all details, which constitute the spectrum of the signal.

A systematic study of all such details (zeros) shows us that some patterns exist for the position of zeros in the z-plane. Studying such patterns observed on ZZT representations of speech signals has several interesting outputs. Among those, the most interesting is that separate patterns for the glottal flow and vocal tract contributions can be observed on ZZT representation of speech signals. Studying the ZZT representations of signals has some important by-products. It is useful for studying the local contributions of zeros of the z-transforms to Fourier Transform (FT) spectra. These effects are especially important for phase spectra since their influence (such as spikes on the derivative of the phase) is so high that they mask important information, like formant peaks. We show, using ZZT representations, that formant peaks can clearly be observed on group delay functions (negative derivative of phase spectrum) of speech signals once windowing is appropriately performed.

### III.2. Definition

For a series of  $N$  samples  $(x(0), x(1), \dots, x(N-1))$  taken from a discrete time signal  $x(n)$ , the Zeros of the Z-Transform (ZZT) representation is defined as the set of roots (zeros),  $\{Z_1, Z_2, Z_3, \dots, Z_m\}$ , of the corresponding z-transform polynomial  $X(z)$  (where  $N$  is the length of the time series):

$$X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} = x(0)z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \quad (3.1)$$

provided that  $x(0)$  is non-zero.

ZZT representation can be presented on the  $z$ -plane in cartesian or polar coordinates. In most of the representations we prefer polar coordinates since visual comparison with amplitude or phase spectrum is much easier as shown in Fig. 5.

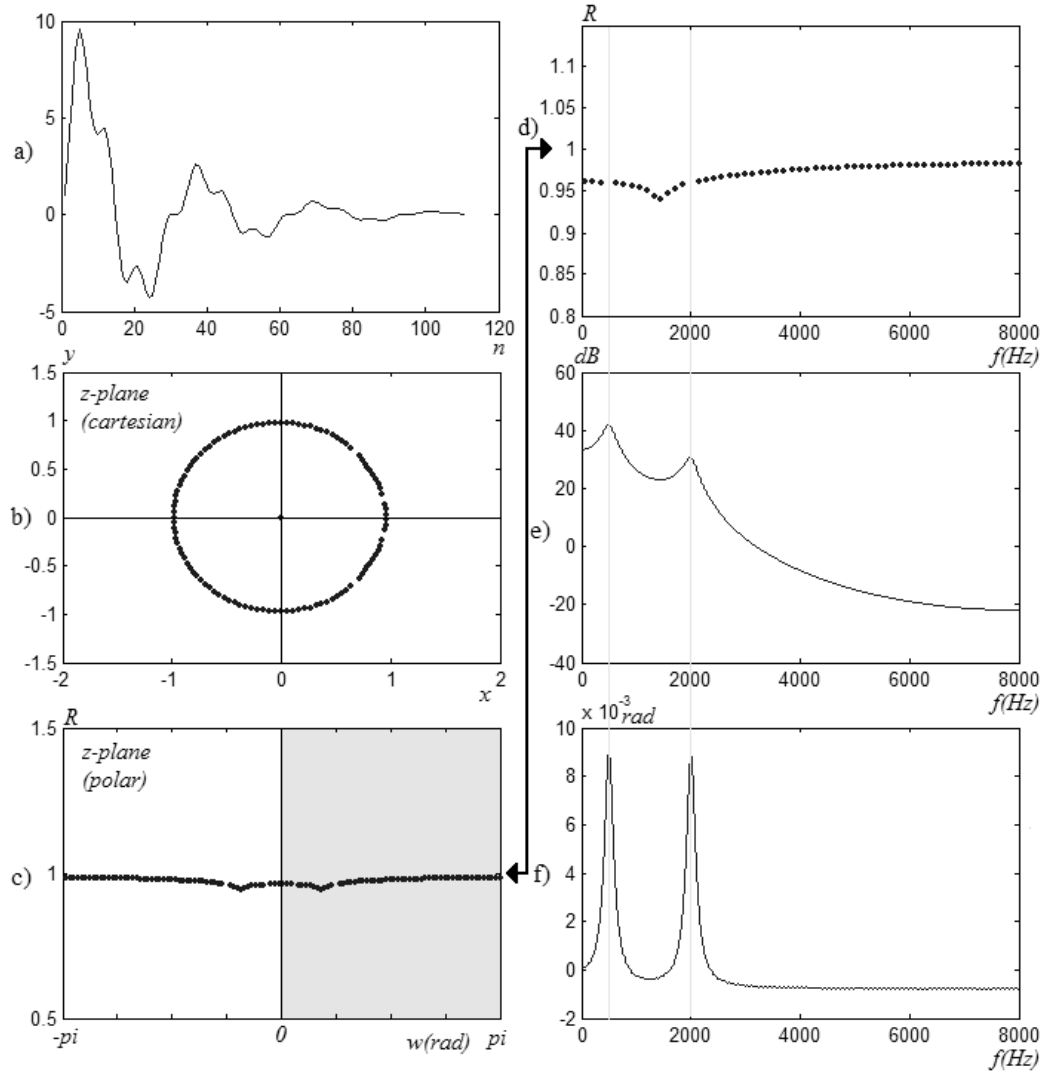


Fig. 5: ZYT plots on  $z$ -plane

a) The time-domain signal, b) ZYT plot in cartesian coordinates, c) ZYT plot in polar coordinates, d) ZYT plot in polar coordinates (right half-plane), e) magnitude spectrum, f) group delay function

For the given signal (a truncated all-pole filter impulse response) in Fig. 5a, the ZYT are located closely to the unit circle. When the ZYT are plotted on polar coordinates, as in Fig. 5d, visual comparison to magnitude spectrum and group delay function is easier (we can observe that the peaks in the spectra correspond to zero-gaps on the ZYT plot). Throughout this thesis, we prefer labeling the frequency axis in (Hz) for the ease of referencing with actual resonance frequencies in speech.

### Finding the roots of high degree polynomials

Computation of the ZYT of a discrete-time signal necessitates finding the roots of high degree polynomials. For example, the  $z$ -transform polynomial of a 30msec. data window at 16000Hz has a degree of 480. There is no known method to derive the roots of such high degree polynomials

analytically (it has been shown that there can be no general formula for the roots of polynomials with degree higher than four (Abel-Ruffini theorem, [Abel, 1826])). Using numerical methods to find the roots is our only alternative way. Twenty years ago, when today's powerful computers were not yet available, this would cause a speech analysis study based on roots of  $z$ -transforms to stop just at the first step: we would maybe need years to handle a systematic study on the roots locations of various speech signals. Today, an ordinary computer with a CPU at 2GHz can find roots of a 480 degree polynomial in 3 seconds. We no more need months but only hours to get a spectrogram-like view of roots for a given speech signal.

This part of the thesis study contains an experimental approach: we use numerical methods to find the roots (which we cannot really explain why they are located such) and then observe the root locations for various signals and draw conclusions from those observations.

In all the root calculations in this thesis, ROOTS function of MATLAB is used, which finds the Eigen values of the associated companion matrix [Edelman & Murakami, 1995]. The ROOTS function is based on the property that for a given polynomial,

$$c_5x^4 + c_4x^3 + c_3x^2 + c_2x^1 + c_1 = 0 \quad (3.2)$$

the roots of the polynomial are the same as the Eigen values of the matrix.

$$A = \begin{bmatrix} -c_2/c_1 & -c_3/c_1 & -c_4/c_1 & -c_5/c_1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (3.3)$$

More efficient algorithms also exist. In [Sitton *et al*, 2003], a review of some of the efficient root finding algorithms can be found.

### III.3. ZZT representation of speech signals

Studying ZZT of some basic signals is useful to understand the patterns in the ZZT of speech signals. Therefore in the following subsection, we first study the ZZT of two basic signals: the exponential and the damped sinusoid. Based on the patterns of these two signals, we study the ZZT of the LF model glottal flow signal. We continue with discussing the source-filter model of speech through the ZZT representation, which concludes the theoretical aspects on this topic. In further sections, we consider the ZZT of windowed speech for real-life applications of the ZZT representation of signals.

#### III.3.1. ZZT of some basic signals

##### ZZT of an exponential time series

Analytically, for a simple exponential function, all the roots,  $Z_m$  (Eq. 3.6), of the  $z$ -transform polynomial  $X(z)$  (Eq. 3.5) calculated for the signal  $x(n)$  (Eq. 3.4) are equally spaced on a single circle at radius  $R=a$  (and the zero on the real axis is cancelled by the pole at the same location resulting in what we call a “zero-gap” in the remainder of this thesis). An example is presented in Fig. 6.

$$x(n) = a^n, n = 0, 1 \dots N-1 \quad (3.4)$$

$$X(z) = \sum_{n=0}^{N-1} a^n z^{-n} = \frac{1 - \left(\frac{a}{z}\right)^N}{1 - \left(\frac{a}{z}\right)} \quad (3.5)$$



$$Z_m = ae^{j2\pi m/N}, m = 1, 2, \dots, N-1 \quad (3.6)$$

For an increasing exponential,  $a > 1$ : the zeros are outside the unit circle. For a decreasing exponential,  $a < 1$ : the zeros are inside the unit circle (Fig. 6).

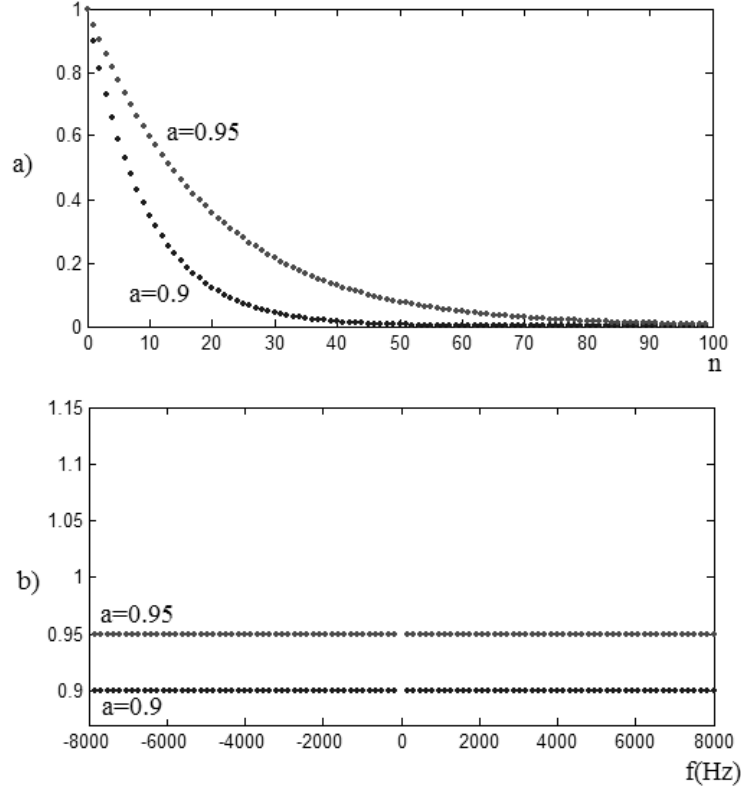


Fig. 6: ZZT of two exponential functions  
a) Time domain waveforms, b) the ZZT plots

### ZZT of a damped sinusoid

A damped sinusoid can be expressed as in Eq. 3.7 where  $k$  is the decaying coefficient of the exponential and  $\omega$  is the frequency of the sinusoid.

$$x(n) = e^{kn} \sin(\omega n), n = 0, 1, \dots, N-1 \quad (3.7)$$

The ZZT pattern of a damped sinusoid can be explained through the ZZT pattern of a sinusoid. Therefore, we start with studying the ZZT pattern of a sinusoid.

In Fig. 7a and Fig. 7b, we present two truncated sinusoids and their corresponding ZZT representations. The ZZT pattern of a sinusoid depends very much on the truncation point. When the truncation points for the sinusoidal signal are at zero-crossings, all the zeros appear on the unit circle (i.e. the ZZT pattern is a line at  $R=1$  in  $z$ -plane in polar coordinates) and two zero-gaps exist at the frequency of the sinusoid (one at a negative frequency and one at a positive frequency) (Fig. 7a) which create two peaks in the magnitude spectrum. The effect of a discontinuity at the truncation point on the ZZT of a sinusoidal function is a wing-like pattern, which can be observed by comparing figures Fig. 7a and Fig. 7b. To observe the dependency on truncation points, the reader is invited to watch the movie "sinusZeros.avi" available on <http://www.tcts.fpms.ac.be/demos/zzt/index.html>. The movie presents variation of ZZT patterns with the truncation point for a discrete time sinusoidal signal.

The damped sinusoid in Fig. 7c can be obtained from Fig. 7b by term-wise multiplication with an

exponential signal. The ZZT pattern of a damped sinusoid is just the shifted version of the ZZT pattern of a sinusoid since a term-wise multiplication of the sinusoid with an increasing exponential simply shifts the ZZT pattern out of the unit circle (Fig. 7c). As a matter of fact, a term-wise multiplication of a discrete time signal by an exponential of form  $e^{kn}$  shifts the ZZT of a signal in the radial direction by an amount equal to  $e^k$ . This can be easily shown by variable substitution in  $X(z)$  and its representation in ZZT form:

$$X(z) = \sum_{n=0}^{N-1} x(n) e^{kn} z^{-n} = \sum_{n=0}^{N-1} x(n) (e^{-k} z)^{-n} \quad (3.8)$$

$$X(z) = x(0) z^{-N+1} \prod_{m=1}^{N-1} (e^{-k} z - Z_m) \quad (3.9)$$

$$Z'_m = Z_m e^k \quad (3.10)$$

The reader is invited to watch the movie ‘expCoeffInDampedSinusoid.avi’ available on <http://www.tcts.fpms.ac.be/demos/zzt/index.html>, for a demonstration of the shift introduced by the exponential component. The movie presents variation of ZZT patterns with the linear variation of the exponential decay coefficient.

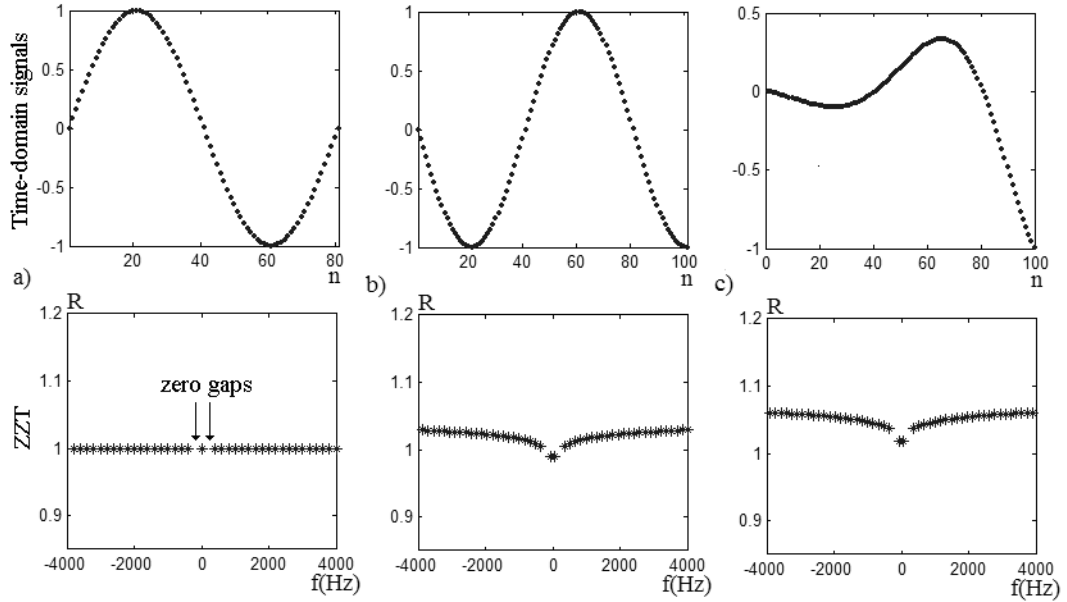


Fig. 7: ZZT of a damped sinusoid.

Waveform and ZZT of a) a sinusoid truncated at zero crossings, b) a sinusoid truncated at a non-zero value at right boundary, c) the sinusoid in b multiplied with an increasing exponential

### III.3.2. ZZT of the glottal flow signal

According to the well-known source-filter model for speech, voiced speech signals are produced by exciting the vocal tract system by periodic glottal flow signals. The most widely accepted model for the derivative of the glottal flow signal is the LF model [Fant, 1995], where the signal is supposed to be composed of two non-overlapping parts: an increasing exponential multiplied by a sinusoid (the first phase, Eq. 3.11) and a decreasing exponential function (the return phase, Eq. 3.12) (both functions are truncated to obtain a one pitch period size data).

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t), 0 \leq t \leq t_e \quad (3.11)$$

$$g(t) = -\frac{E_e}{\mathcal{E}_a} \left[ e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)} \right], t_e \leq t \leq t_c \leq T_0 \quad (3.12)$$

$$g(t) = 0, t_c \leq t \leq T_0 \quad (3.13)$$

In Fig. 8, the ZZT representation of an LF model (glottal flow derivative) signal is presented.

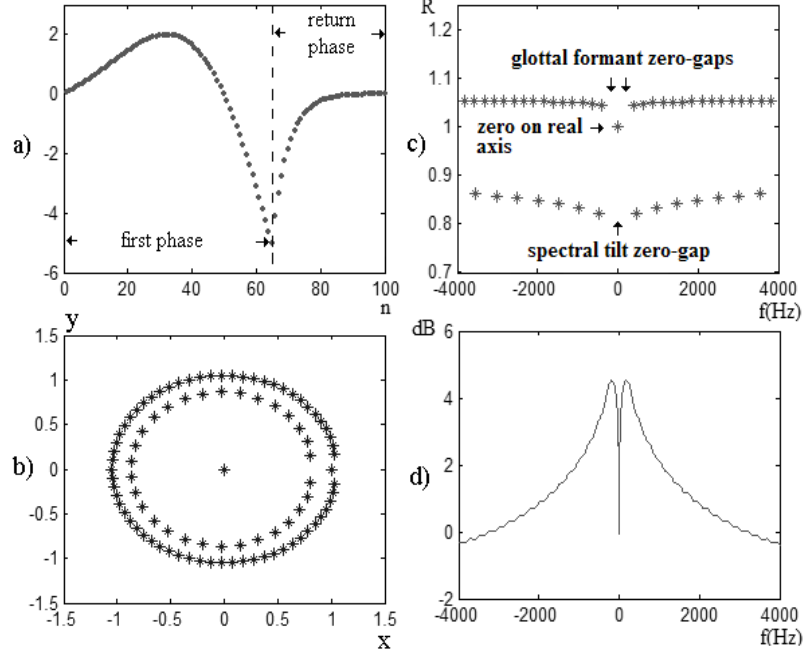


Fig. 8: ZZT pattern of a typical differential LF signal.

a) waveform, b) ZZT representation in cartesian coordinates, c) ZZT representation in polar coordinates, d) magnitude spectrum.

The ZZT representation of the LF signal, shown in Fig. 8 contains two groups of zeros: a circle inside the unit circle and a circle outside the unit circle in cartesian coordinates (Fig. 8b) or a line below  $R=1$  and a line above  $R=1$  in polar coordinates (Fig. 8c). The group of zeros inside the unit circle is due to the return phase and the group outside the unit circle is due to the first phase (Eq. 3.11) of the LF signal.

It is interesting to note at this point that the LF signal appears to be a mixed-phase signal<sup>2</sup> since some zeros occur on both sides of the unit circle and the zeros outside the unit circle are due to its first part. This output/observation is in accordance with those presented in [Jackson, 1989, Gardner, 1994, Doval *et al*, 2003] about anti-causal (maximum-phase) components in glottal flow signals. We further discuss these issues in the next section and in Section IV.4.

### Contribution of the first phase to the ZZT of LF model glottal flow signal

As explained for the damped sinusoid ZZT in the previous sub-section, the sinusoidal component of the first phase is responsible for the zero gaps located outside the unit circle on the wing-like ZZT pattern (Fig. 7c). These gaps create in turn an anti-causal resonance like spectral peak, which can be observed on the magnitude spectrum (Fig. 8d, at around 200Hz for this signal) as discussed in [Doval & d'Alessandro, 1997], and on the group delay as a negative peak. This is like the effect of an anti-causal pole at the frequency of the gap. This resonance-like peak on the spectrum carries

<sup>2</sup> Please refer to Section IV.2 for the description of the minimum-phase, the maximum-phase and the mixed-phase characteristics for a signal.

all information about the first phase of the LF signal and is called the glottal formant ( $F_g$ ) [Doval *et al*, 2003]. The main excitation of the glottal flow signal is produced by the discontinuity at truncation end point of the first phase which also creates the wing-like pattern instead of a straight line as explained in the previous subsection.

When the spectrum of glottal flow signal derivative is considered, there are mainly two important parameters for the first phase of the glottal flow: the asymmetry coefficient ( $\alpha_m$ ) and the open quotient ( $O_q$ ) [Doval & d'Alessandro, 1999]. The asymmetry coefficient controls the location of the peak in the glottal flow signal and the open quotient controls the duration of the open phase. In Fig. 9, we show the effect of variations of these two parameters of the first phase on the ZZT of the glottal flow signal. The first column in Fig. 9 presents the effect of variations of  $\alpha_m$  and the second column in Fig. 9 presents the effect of variations of  $O_q$ .

Three glottal flow signals are presented in Fig. 9a. All parameters of the LF signal are set to constant values ( $O_q=0.75$ ,  $f_0=160$  Hz) except  $\alpha_m$  is varied: 0.6, 0.7, and 0.8 to construct these signals. The differential glottal flow and glottal flow signals obtained are presented in Fig. 9a and Fig. 9b. ZZT and amplitude spectra of the differential glottal flow are presented in Fig. 9c and Fig. 9d. An increase in  $\alpha_m$  corresponds to shifting of the wing pattern of ZZT further away from the unit circle ( $R=1$ ) which corresponds to blurring of the glottal formant peak without much change in its frequency since the center frequency of the zero-gap on the ZZT pattern is almost unchanged. As  $\alpha_m$  increases, the zeros and the zero-gap moves further away from the unit circle. Increase of the steepness of the exponential results in concentration of more energy on the right-hand side of the signal, i.e. the signal becomes more “anti-causal” as zeros and the zero gaps move away from the unit circle.

Four glottal flow signals are presented in the second column of Fig. 9. All parameters of the LF signal are set to constant values ( $\alpha_m=0.6$ ,  $f_0=160$  Hz) except  $O_q$  is varied: 0.3, 0.5, 0.7 and 0.9 to construct these signals. The differential glottal flow and glottal flow signals obtained are presented in Fig. 9e and Fig. 9f. ZZT and amplitude spectra of the differential glottal flow are presented in Fig. 9g and Fig. 9h. An increase in  $O_q$  corresponds to narrowing of the zero-gap, which corresponds to a decrease in the frequency and bandwidth of the glottal formant peak. These observations are in accordance with the results of [Doval *et al*, 2003].

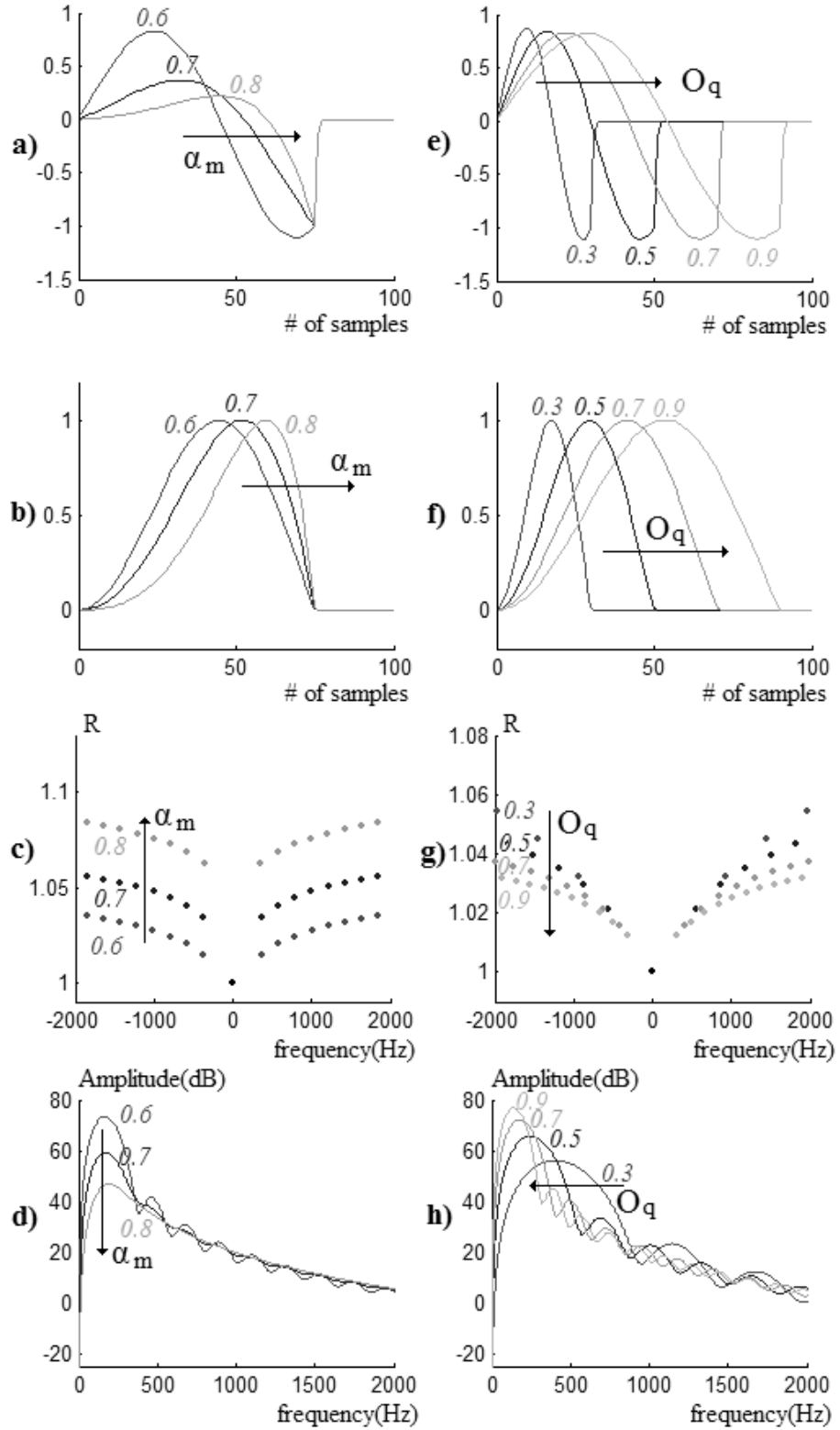


Fig. 9: Effect of parameter variation on the ZZT of glottal flow signals.  
 Asymmetry coefficient variation: a) differential glottal flow signals, b) glottal flow signals (equalized), c) ZZT representations, d) amplitude spectra.  
 Open quotient variation: e) differential glottal flow signals, f) glottal flow signals (equalized), g) ZZT representations, h) amplitude spectra.

### Contribution of the return phase to the ZZT of the LF model glottal flow signal

The return phase exponential component ( $Ta$  indicated on Fig. 2) of the differential LF function contributes to the ZZT representation by a group of zeros inside the unit circle, aligned in parallel to the unit circle and the distance of these lined zeros to the unit circle is proportional to the exponential decay coefficient. Again, there exists a gap on the real axis (Fig. 8c, Fig. 10c). Its effect on the magnitude spectrum is a slope, *spectral tilt*, change for the high frequency part of the magnitude spectrum.

In Fig. 10, the effect of varying the return phase coefficient on the ZZT pattern of differential glottal flow is shown. Three glottal flow signals are created by keeping all time parameters of the LF model constant but only varying  $Ta$  (three values are assigned: 0.02, 0.06, and 0.1). The differential glottal flow and glottal flow signals obtained are presented in Fig. 10a and Fig. 10b. ZZT and amplitude spectra are presented in Fig. 10c and Fig. 10d. A variation in the return phase causes a modification in the first phase of the signal since the integral of a one period length differential glottal flow signal should equal to zero not to have a DC component in the glottal flow signal. The LF model is designed with such consideration. If we only consider the change introduced in the return phase, we observe that an increase in  $Ta$  (which results in slower closing of the vocal folds) corresponds to shifting of the ZZT of the return phase closer to the unit circle which corresponds to more suppression of the high frequency components. For this reason, the glottal flow signal (labeled with  $Ta=0.1$ ), which corresponds to a high  $Ta$  value has its return phase ZZT closest to the unit circle and the spectral tilt is higher (i.e. the energy of high frequency components are lower).

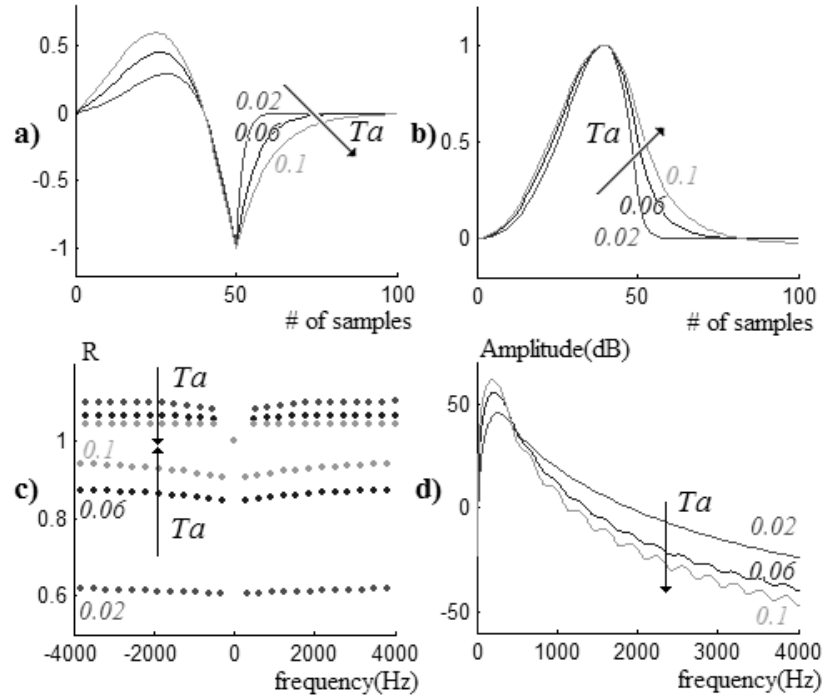


Fig. 10: Effect of spectral tilt variation to ZZT of glottal flow signals.  
a) Glottal flow derivative, b) glottal flow, c) ZZT representation, d) magnitude spectrum.

### III.3.3. ZZT representation and source-filter model of speech

In Fig. 11, we present the ZZT patterns for the source filter model of speech [Fant, 1960] for voiced speech. For simplicity, the lip radiation component is included in the source signal as a derivation, resulting in differential glottal flow signal on the second column: speech signals can be expressed as  $s(t)=d(U_g(t)*v(t))/dt$  where  $U_g(t)$  stands for the glottal flow signal,  $v(t)$  stands for the

vocal tract filter impulse response and the lip radiation component is approximated by a derivation operation. Equivalently, we can re-write this as:  $s(t)=(d(U_g(t))/dt)*v(t)$  including the derivative operation in the glottal flow part.

In Fig. 11, each row presents the model in one domain: the 1st row is in the time domain, the 2nd row is in the  $z$ -domain (ZZT representation) and the 3rd row is in the log-magnitude spectrum. The operators are: convolution (\*), union (U) and addition (+). We now discuss in detail the second row, the ZZT-representations. Since we have discussed the ZZT patterns for differential glottal flow signals in the previous section, there remains two components to be discussed: impulse train (the first column in Fig. 11) and vocal tract filter (the third column in Fig. 11).

The ZZT pattern for an impulse train (2nd row 1st column in Fig. 11) is such that, zeros are equally spaced on the unit circle with the exception that there exist gaps at all harmonics of the fundamental frequency, which create the harmonic peaks on the magnitude spectrum (3rd row 1st column in Fig. 11). The location of zeros for an impulse train (Eq. 3.14) with period  $P$  can be analytically found by finding the roots of its  $z$ -transform (Eq. 3.15). The roots of the denominator in (Eq. 3.15) are expressed in (Eq. 3.16) and the roots of the numerator are expressed in (Eq. 3.17).  $P$  roots of the denominator cancels  $P$  roots of the numerator resulting in  $P(M-1)$  zeros for the impulse train  $z$ -transform, located on the same circle and  $P$  zero gaps exist on the zero-circle at multiples of the fundamental frequency (i.e. the harmonic frequencies),  $2\pi m/P$ .

$$x(n) = \sum_{k=0}^{M-1} \delta(n-kP), n = 0, 1 \dots N-1, N \geq P(M-1) \quad (3.14)$$

$$X(z) = \sum_{n=0}^{N-1} \sum_{k=0}^{M-1} \delta(n-kP) z^{-n} = \sum_{k=0}^{M-1} z^{-kP} = \frac{1 - \left(\frac{1}{z}\right)^{PM}}{1 - \left(\frac{1}{z}\right)^P} \quad (3.15)$$

$$Z_d = e^{j2\pi m/P}, m = 1, 2 \dots P-1 \quad (3.16)$$

$$Z_n = e^{j2\pi m/PM}, m = 1, 2 \dots PM-1 \quad (3.17)$$

The zeros of the vocal tract filter response are mainly inside the unit circle due to the decreasing exponential character and there exists gaps for the formant locations, which create formant spectral peaks. Again, we observe the wing-like character for the ZZT pattern of vocal tract response depending on the location of the truncation point for the time-domain response. The reader is invited to watch the movie "causalResponseZeros1.avi" available on <http://www.tcts.fpms.ac.be/demos/zzt/index.html> for a demonstration of the effects of the truncation point to the ZZT pattern of a causal all-pole filter response.

**It is interesting to note here that the set of ZZT of speech is just the union of ZZT sets of the three components. This is due to the fact that the convolution operation in time-domain corresponds to multiplication of the  $z$ -transform polynomials in  $z$ -domain. What is interesting is that the ZZT of each component appear at a different area on the  $z$ -plane and have effect on the magnitude spectrum relative to their distance to the unit circle. The closest zeros to the unit circle are the impulse train zeros and they cause the spectral dips on the magnitude spectrum, which give rise to harmonic peaks. Vocal tract zeros are the second closest set and the zero-gaps due to formants contribute to the magnitude spectrum with formant peaks on the spectral envelope. Glottal flow ZZT are further away from the unit circle and their contribution on the magnitude spectrum is rather vague and distributed along the frequency axis.**

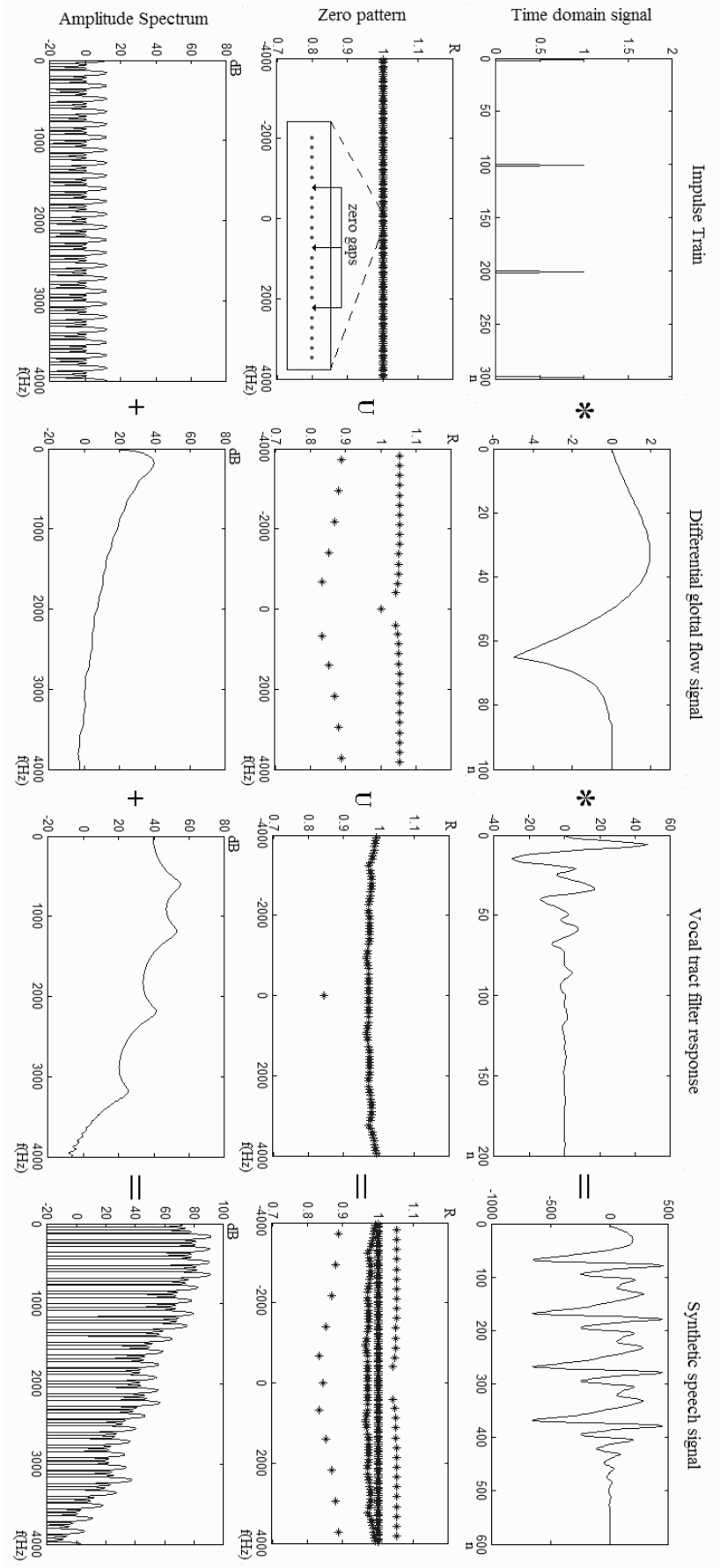


Fig. 11: ZZT and source-filter model  
(note that the amplitude spectra added are in dBs)



### III.3.4. ZYT of windowed synthetic speech signals

In real-life applications, the windowing operation is essential especially for the spectral analysis of signals. The effect of windowing is very drastic on ZYT patterns. The size, location and function of the window play an important role on the resulting ZYT patterns for the windowed signal. It is very difficult to analytically study the effect of windowing on the ZYT patterns since we face the difficult problem mentioned previously: the windowing operation is a term-wise multiplication operation in the time-domain, and there is no known methodology for estimating roots of the resulting  $z$ -transform polynomial directly from the roots of the two polynomials that are subject to term-wise coefficient multiplication. For this reason, our discussions in this section are again based only on experimental observations of ZYT patterns on the  $z$ -plane.

In Fig. 12, we present the effect of windowing on a truncated frame with single excitation. The glottal flow derivative, the synthetic speech signal obtained by all-pole filtering (with resonances at 600Hz, 1200Hz, 2200Hz and 3200Hz) of the glottal flow derivative and the windowed synthetic speech signal are presented respectively in Fig. 12a, Fig. 12b and Fig. 12c together with their ZYT representation and amplitude spectra. The ZYT patterns in Fig. 12, compared to the ZYT patterns for the source-filter model of speech in Fig. 11, do not contain the zero pattern of an impulse train on the unit circle since there is only a single excitation. On the ZYT representations in Fig. 12, the corresponding glottal formant zero-gap and the vocal tract zero-gaps are indicated. As discussed previously, the glottal formant zero-gap is located outside of the unit circle and the vocal tract zero-gaps are located inside the unit circle. The ZYT of the windowed synthetic speech signal in Fig. 12c has an important property: a zero-free region exists around the unit circle and for this reason a noise-free group delay function with clear formant peaks is obtained (Fig. 12c). We show in the following chapters that having a zero-free region around the unit circle is very advantageous for phase (or group delay) spectrum processing. In addition, it is also the basis for our source-tract decomposition algorithm (called the ZYT-decomposition), since the zeros (and the zero-gaps) for the vocal tract and the glottal flow fall on two different sides of the unit circle.

The ZYT representation in Fig. 12c includes two lines of zeros (with some ripples): one outside the unit circle and one inside the unit circle with gaps creating formant peaks on the spectrum. The reason for this alignment is as follows: once the window is placed such that the increasing exponential part of a single speech frame (due to the first phase (Eq. 3.11) of the glottal flow signal) is multiplied with the first half of the window, which is also increasing, and the decreasing exponential part (due to the vocal tract filter response and to the return phase of the glottal flow when it exists) is multiplied with the second half of the window, which is also decreasing, the ZYT of the resulting windowed speech has a pattern close to that of the glottal flow (with additional patterns inside the unit circle due to the vocal tract filter). For the cases where there exists a non-zero return phase of the glottal flow signal, its zeros are combined with those of the vocal tract resulting in a single line of zeros inside the unit circle. When the window is not centered on the increasing-decreasing function turning point, the ZYT-pattern is destroyed, and zeros do not group on the two sides of the unit circle. Therefore, GCI-synchronous windowing is necessary to obtain separate ZYT patterns for glottal flow and vocal tract contributions, which provides the opportunity to perform decomposition.

The windowing operation shifts the two zero-groups (lines) (Fig. 12b) away from the unit circle (Fig. 12c). The zero-gaps are also shifted out of the unit circle but the frequency location stays almost unchanged. The glottal formant zero-gap creates a negative peak on the group delay function, which is labeled as a glottal formant peak on the third plot in Fig 12c since it is outside the unit circle. All the other peaks on group delay are positive since the zero-gaps creating those peaks are located inside the unit circle. This is an interesting observation since it sheds light into some phase related auditory phenomena, which is reported in some studies but no explanation/theory has been provided. For example in [Sun, 1997], it has been shown (without any theoretical background but mainly with observations) that introducing negative group delay (via some randomization) the perceptual quality of speech coding increases. In addition, this observation leads us to question the minimum-phase property assigned to speech signals in many applications. In the following chapters, we discuss a mixed-phase model for speech signals based on anti-causality of the glottal formant.

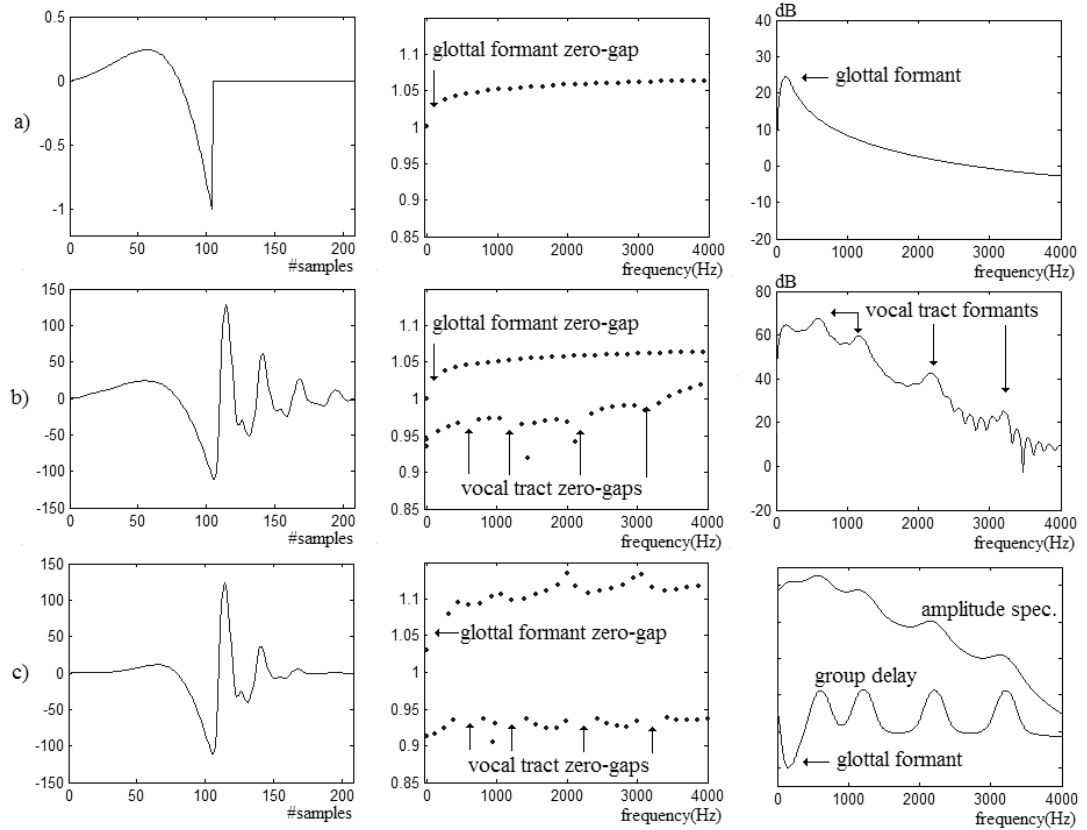


Fig. 12: ZTZ patterns for a single excitation speech frame.

a) Time-domain glottal flow derivative signal, corresponding ZTZ representation and magnitude spectrum, b) time-domain synthetic speech signal, corresponding ZTZ representation and magnitude spectrum, c) time-domain Blackman windowed synthetic speech signal, corresponding ZTZ representation, magnitude spectrum and group delay function scaled and plotted together.

As the next step we start testing if similar patterns can be obtained when real speech frames are windowed. We study the effects of window location, window size and window function on ZTZ patterns and study the criteria for optimum windowing.

### Effect of window location on ZTZ patterns

In Fig. 13, we demonstrate the windowing location effect on ZTZ patterns. A Blackman window of two pitch period size is slid in six steps within a pitch period and the resulting ZTZ are presented (please refer to the Appendix for window function definitions). In addition, a movie demonstration is available on <http://www.tcts.fpms.ac.be/demos/ztz/index.html> : real2T0Blackman.avi.

The six ZTZ representations in Fig. 13 show that the influence of window location on ZTZ patterns is indeed very important. Among the six possibilities presented, the fourth window, centered on the glottal closure instant (GCI) matches the theoretical expectations: there exists a zero-free region around unit circle in the [0-5000Hz] frequency region as discussed in the previous section. In Fig. 14, we present a zoomed plot (of the shaded area on Fig. 13) of this complicated picture and show the windowed signal waveform, ZTZ representation and the corresponding magnitude spectrum and group delay function.

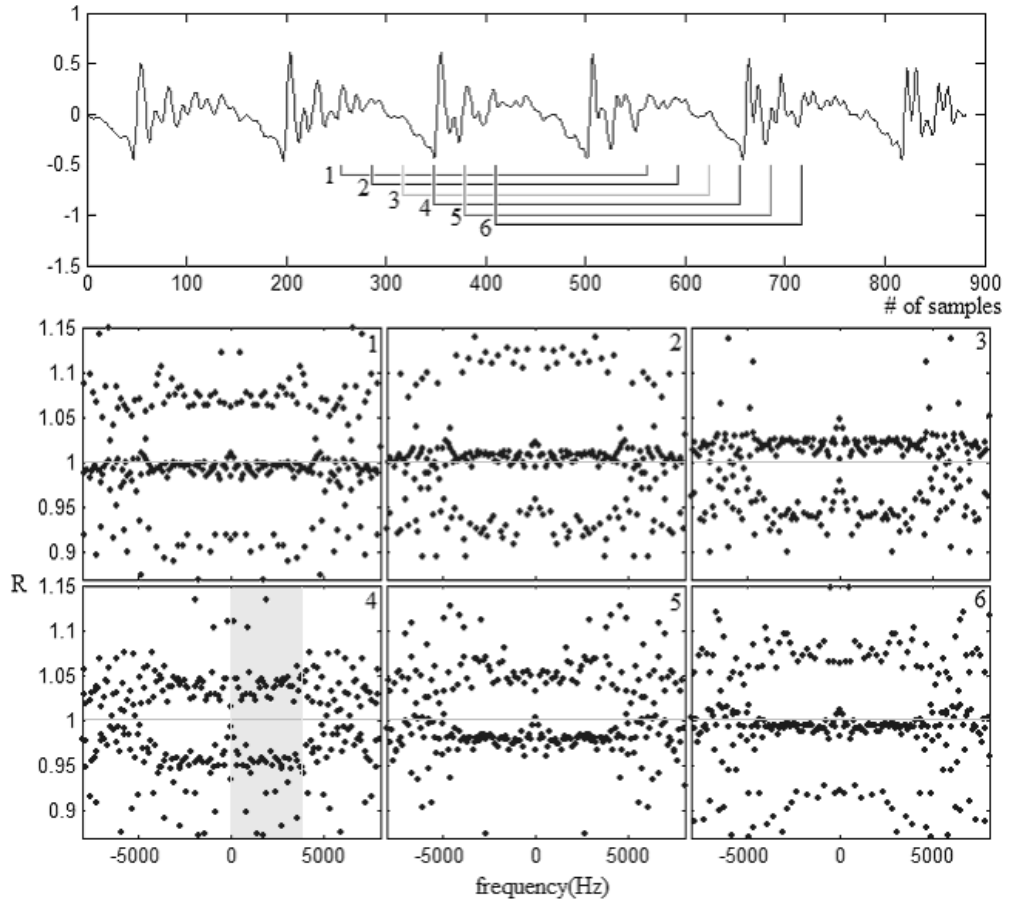


Fig. 13: Effect of windowing location on ZZT of a real speech signal. Each (Blackman) window position is indicated on the signal on the top figure with reference numbers. The ZZT representation of the resulting windowed data for each window is presented with the window index indicated on the right-top corner of the figure.

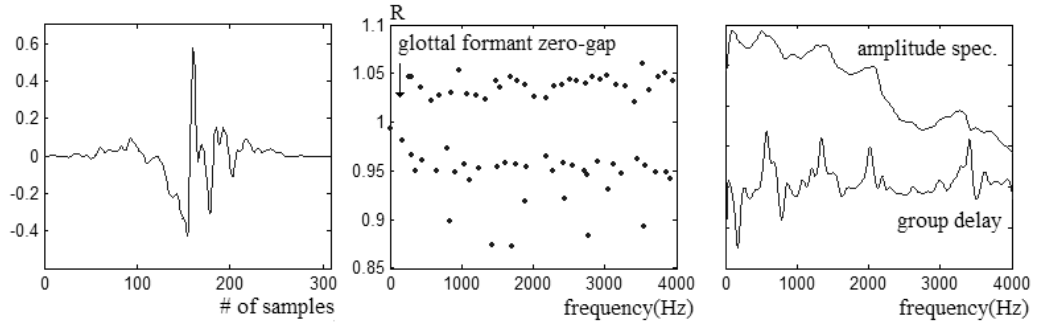


Fig. 14: GCI synchronous windowing. Time-domain waveform of the windowed signal, corresponding ZZT representation, magnitude spectrum and group delay scaled and plotted together.

The group delay contains peaks, which corresponds to the formant peaks observed on the magnitude spectrum. In addition, for the lowest frequency peak in magnitude spectrum, the group delay peak has a negative direction since the corresponding glottal formant zero-gap is outside the unit circle. This is further discussed in Section IV.4.

We conclude that GCI synchronous windowing is necessary to obtain windowed signals with ZZT patterns that match the theory presented.

### Effect of window function on ZZT patterns

Window function also has important influences on the ZZT patterns. In Fig. 15, we present the ZZT patterns obtained by windowing (with six different windowing functions) the real speech data in Fig. 14 by a two pitch period size window centered at GCI.

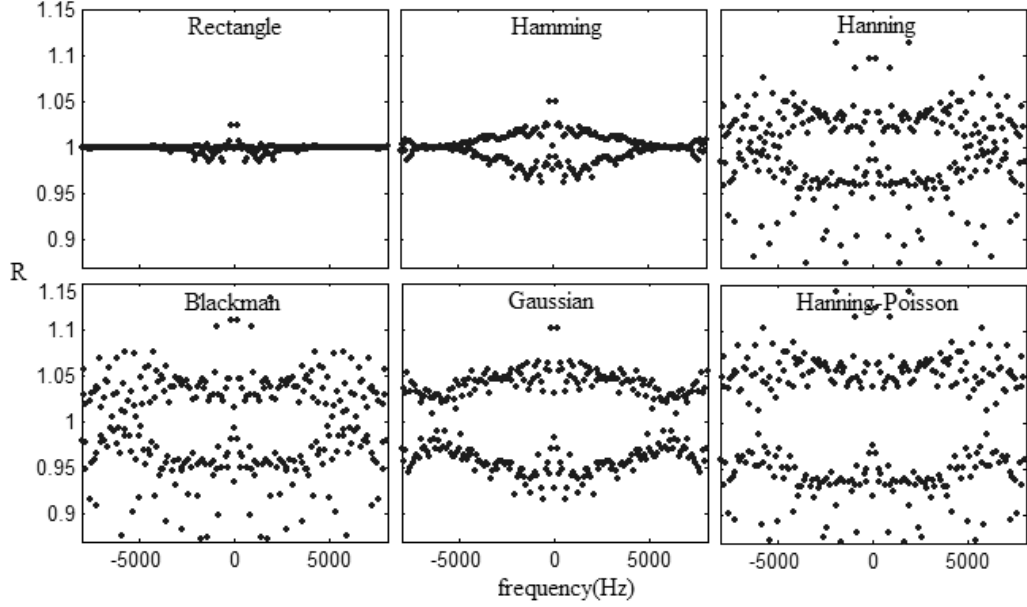


Fig. 15: Effect of window function to ZZT patterns.

GCI synchronous windowing using several window functions: rectangle, Hamming, Hanning, Blackman, Gaussian and Hanning-Poisson

The resulting ZZT patterns are quite different for the different window functions used. Optimality of a windowing function depends on the particular task and the ZZT properties desired. For example, for obtaining noise-free group delay functions, it is important to have a zero-free region around unit circle and given this criterion Blackman, Gaussian and Hanning-Poisson windowing functions are advantageous.

When GCI synchronous windowing is concerned, it is possible to consider the window being composed of two parts, an increasing first half boosting the glottal flow first phase component of the signal and a decreasing second half boosting the vocal tract response and return phase of glottal flow. This is mainly due to the specific localization of the events in the speech signal, and it only holds for speech signals for which phase spectrum is not modified by some filtering operation. We can consider two parts independently and adjust the contribution of each half which leads to “asymmetric windowing” of the speech frame. In Fig. 16 we present the effect of two different asymmetric windows on the ZZT representation and spectrum of speech frame used in Fig. 14.

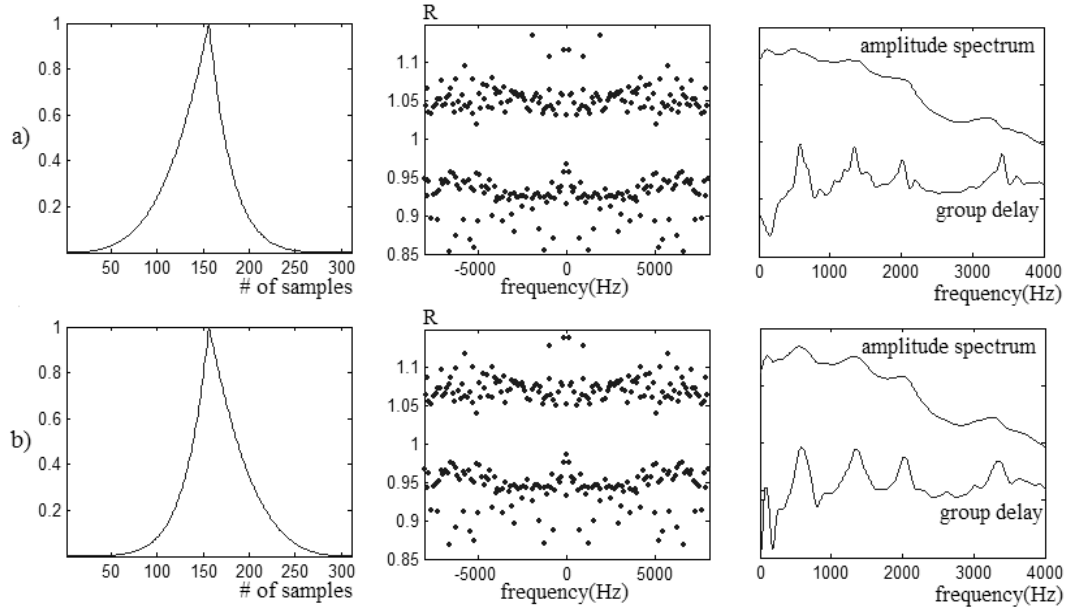


Fig. 16: GCI synchronous Hanning-Poisson asymmetric windowing. Each row includes the window function used and ZYT representation, group delay and magnitude spectrum of the windowed signal.

The window in Fig. 16a has a second half decaying coefficient higher than first half decaying coefficient (and vice versa for the window in Fig. 16b). We have previously mentioned that multiplication with an exponential shifts the ZYT in the radial direction. When we compare the two ZYT plots, we observe that for the first window, ZYT inside the unit circle due to vocal tract response is shifted further away from the unit circle therefore the formant peaks are less prominent. Equally, for the second window glottal flow ZYT are shifted away from the unit circle resulting in glottal formant peak becoming less prominent. This example shows that asymmetric windowing can be applied to adjust the level of contribution of the glottal flow or the vocal tract in the resultant magnitude spectrum once the window is centered at GCI and the window size is less than two pitch periods.

### Effect of window size on ZYT patterns

The window size determines the number of zeros in the ZYT representation (given that the first sample is non-zero).

In Fig. 17, we present the ZYT patterns obtained by GCI synchronous windowing with different window sizes (one, two, three and four pitch periods respectively). The first row provides a larger view and the second row a zoomed view of the ZYT representations.

When window size gets larger than two pitch periods some zeros close to the unit circle are observed. As discussed previously, the existence of a zero-free region around unit circle is important for group delay processing and for source-tract separation by ZYT-decomposition. Therefore in most of the cases in our analysis tools, we use a maximum window size of two pitch periods. However, one can consider other applications where the existence of these zeros is advantageous for analysis, for example in  $f_0$  estimation.

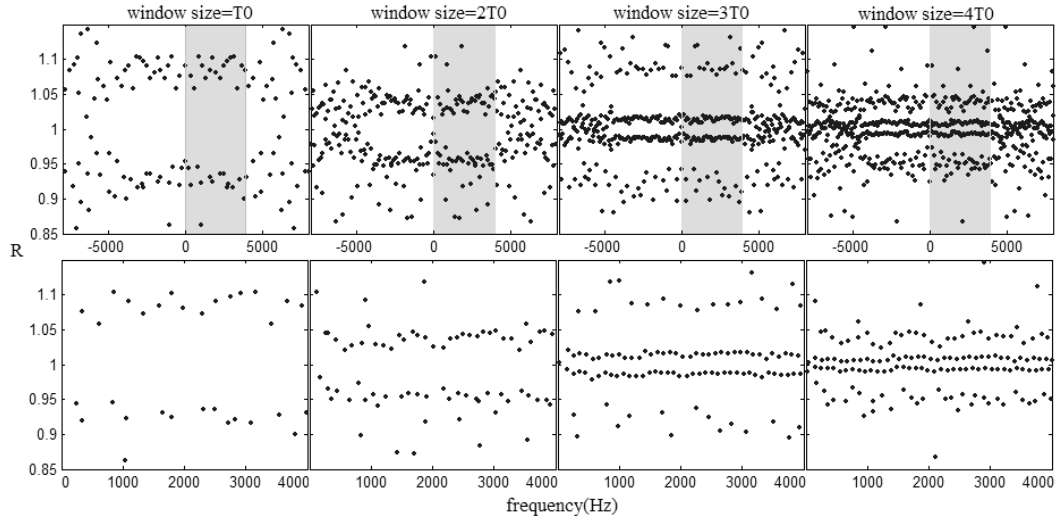


Fig. 17: Effect of window size on ZZT of windowed speech signals. First row: large view of the ZZT representation, second row: zoomed view of the shaded part of the ZZT representation. Blackman window is used.

### III.3.5. ZZT of aperiodic components in speech

It is a common practice in speech processing to consider speech being composed of two types of components: periodic and aperiodic. The periodic part is considered to be due to the periodic excitation created at vocal folds which result in voiced speech. For all parts of speech, whether voiced or unvoiced, some aperiodic contribution exists in varying levels due to various sources like air flow through the vocal folds containing some random fluctuations or noise.

By intuition, we assume that the roots of a polynomial with random coefficients are located randomly on the  $z$ -plane. Observations on a few examples support that assumption. On an example, we study the contribution of the random component on the ZZT of speech signals.

In Fig. 18, we present time domain signals and the ZZT of windowed noise (Fig. 18a), glottal flow derivative (Fig. 18b), excitation obtained by summing these two components (Fig. 18c) and the result of vocal tract filtering of these three signals (in Fig. 18d, 18e and 18f respectively). The ZZT of noise is quite unorganized compared to that of the ZZT of glottal flow derivative and the speech signal obtained by filtering the glottal flow derivative. No patterns are observed for noise or the noise contribution in speech; the contribution can rather be considered as disturbances in the ZZT patterns. For this reason, the ZZT of unvoiced speech is not studied in detail in this thesis (but we provide noise robustness test results when ZZT is used in a proposed algorithm). In addition, the same properties are observed for additive noise, so they are also considered as unorganized disturbances to the ZZT patterns and the degree of disturbance is related to the relative amplitude of the noise component to the periodic speech components.

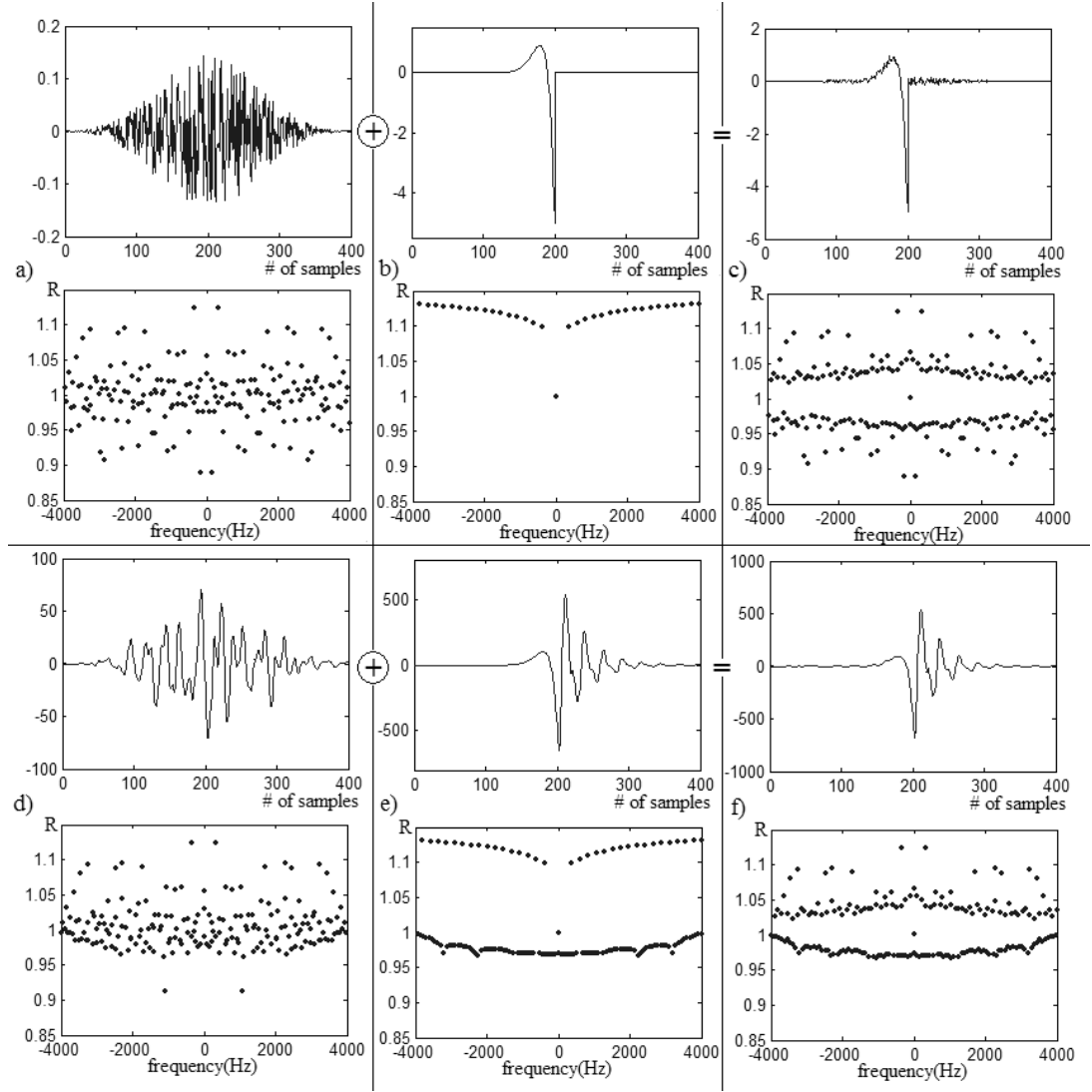


Fig. 18: Aperiodic component contribution to ZZZ of speech.

Each couple of plots includes the time domain signal and the ZZZ representation. a) windowed noise with uniform distribution, b) glottal flow derivative, c) excitation signal obtained by summing a and b, d) vocal tract filtered a, e) vocal tract filtered b, f) vocal tract filtered c. Blackman window is used.

### III.3.6. Conclusion

In this chapter we have presented the ZZZ representation and discussed ZZZ patterns for discrete time speech signals. We have shown that the glottal flow component of ZZZ occurs outside the unit circle and the speech signals are mixed-phase due to this component.

The theory and discussions presented in this chapter will further lead us to a source-tract decomposition algorithm, which will be presented in the applications part of the thesis (section V.1).

A study of windowed speech ZZZ was presented, which is important for the spectral study of speech signals in real-life applications. The usefulness of the ZZZ representation becomes clearer in the next chapter where group delay functions are studied together with zero locations.

---

# Chapter IV:

## Chirp group delay processing of signals

---

### IV.1. Introduction

As we have discussed in the previous sections, speech signals exhibit mixed-phase signal characteristics due to anti-causality of the glottal flow signals. The mixed-phase characteristics can only be observed on the phase component of the FT spectrum but not on the magnitude spectrum. For this reason, group delay processing is especially important for studying characteristics of the glottal flow and the vocal tract contributions separately. However the phase spectrum is often considered to be difficult to process to extract useful information. By nature, the phase component of the FT spectrum is in a wrapped form (see next section for definitions) and the first derivative of the unwrapped phase spectrum (the group delay function) is much more easy to study both for numerical and observation based analysis.

B.Yegnanarayana and H.A. Murthy have published plenty of papers in a period of 20 years on processing the group delay function. In this chapter, we first present a review of their theoretical discussions and methods, which provides the basic terminology and the current state-of-the-art in group delay processing of speech signals.

Then we present a mixed-phase speech model and discuss the phase/group delay spectrum characteristics of speech signals theoretically. We show that, when minimum-phase assumption for speech signals is used or a conversion to minimum-phase version is applied in group delay processing (a common step in many algorithms), the glottal flow characteristics are mixed with that of vocal tract characteristics in the phase/group delay spectrum domain.

Our second focus point is on the reliable estimation of a smooth phase spectrum with well-preserved resonance structure for speech signals. Such a study is of ultimate need for most of the phase processing applications listed in the introduction chapters. The main difficulties in correct phase spectrum estimation and unwrapping are mostly related with the ZZT close to the unit circle, which cause spikes on the derivative of the phase spectrum (group delay) [Yegnanarayana & Murthy, 1992]. We show through ZZT representations that windowing plays a very important role in determining the phase characteristics of the resulting signal and in the reliable estimation of phase information since the location of ZZT (or the existence of zeros close to the unit circle) is very much affected by the windowing operation.

As the final step we propose the chirp group delay as a spectral representation for signals. Chirp group delay (CGD) is potentially useful for tracking resonances of signals and also for spectral feature extraction for speech recognition. **“By both manipulating the ZZT and adjusting the analysis circle radius for CGD computation, we can guarantee certain distance of zeros to the analysis circle and obtain spike-free functions revealing clear formant peaks. This is one of the basic ideas proposed and used through out this thesis.”** In this part, we only discuss the basic theoretical issues; the applications of chirp group delay processing are presented in the next part of the thesis.



---

## IV.2. Methods proposed by Yegnanarayana and Murthy for group delay processing

B. Yegnanarayana and H.A Murthy are the two researchers who most contributed to the rather small literature of group delay processing. They have discussed both theoretical aspects of the issue and applications in formant tracking [Murthy *et al*, 1989 a, Murthy & Yegnanarayana, 1991 a] and glottal instant marking [Smits & Yegnanarayana, 1995]. In the group delay processing literature, their methods are closest to our own methods. Therefore we find it necessary to provide a review of their studies below.

### Terminology

The following list of definitions of terminology is provided in their early papers [Yegnanarayana *et al*, 1984, Murthy & Yegnanarayana, 1991 b].

For a discrete time digital signal  $\{x(n)\}$ ,  $n=0,1,2,\dots,N-1$ , the  $z$ -transform is expressed as:

$$X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} \quad (4.1)$$

or in a rational form as:

$$X(z) = Az^{-n_0} \frac{N(z)}{D(z)} \quad (4.2)$$

where  $A$  is a real constant,  $n_0$  is an integer,  $D(z)$  is the denominator polynomial and  $N(z)$  is the nominator polynomial. The roots of  $D(z)$  are the poles and roots of  $N(z)$  are the zeros of the  $z$ -transform.

**Causality:** A signal  $\{x(n)\}$  is said to be causal if  $x(n)=0$  for all negative values of  $n$ .

**Minimum-phase signal:** If  $n_0=0$  and if all poles and zeros of the  $z$ -transform are inside the unit circle, the signal is said to be minimum-phase.

**Maximum-phase signal:** If all poles and zeros of the  $z$ -transform are outside the unit circle, the signal is said to be maximum-phase.

**Mixed-phase signal:** If all poles and zeros of the  $z$ -transform lie both inside and outside the unit circle, the signal is said to be mixed-phase.

The Fourier Transform (FT) is expressed as

$$X(\omega) = X(z) \Big|_{z=e^{j\omega}} \quad (4.3)$$

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)} \quad (4.4)$$

where  $|X(\omega)|$  is the magnitude and  $\theta(\omega)$  is the phase component.  $\theta(\omega)$  is a wrapped phase information:  $-\pi \leq \theta(\omega) \leq \pi$ . The FT can also be expressed as

$$X(\omega) = |X(\omega)|e^{j[\theta(\omega)+2\pi\lambda(\omega)]} \quad (4.5)$$

where  $\lambda(\omega)$  is an integer such that  $[\theta(\omega)+2\pi\lambda(\omega)]$  is a continuous function of  $\omega$ . Then  $\theta(\omega)$  is called the principle phase value or the wrapped phase function. The unwrapped phase function is:

$$\theta_u(\omega) = \theta(\omega) + 2\pi\lambda(\omega) \quad (4.6)$$

The group delay function is defined as the negative derivative of the unwrapped phase function.

$$\tau_p(\omega) = \frac{-d\theta_u(\omega)}{d\omega} \quad (4.7)$$

This is called the group delay function derived from the FT phase function, therefore denoted as  $\tau_p(\omega)$ . Eq. 4.7 can also be expressed as:

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \quad (4.8)$$

where  $Y(\omega)$  is the FT of  $nx(n)$  and  $R$  and  $I$  refer to real and imaginary parts. The advantage of this representation is that phase unwrapping, which is a problematic task, is not necessary.

The authors also find it necessary to define a version of group delay derived from the magnitude spectrum,  $|X(\omega)|$ , denoted as  $\tau_m(\omega)$ . Their definition is based on the cepstral relation between group delay and magnitude spectrum for a minimum-phase signal.

Given the Fourier transform  $V(w)$  of a minimum-phase signal  $v(n)$ ,

$$V(\omega) = |V(\omega)|e^{j\theta_v(\omega)} \quad (4.9)$$

it can be shown that

$$\ln|V(\omega)| = c(0)/2 + \sum_{n=1}^{\infty} c(n) \cos(n\omega) \quad (4.10)$$

and the unwrapped phase function

$$\theta(\omega) = -\sum_{n=1}^{\infty} c(n) \sin(n\omega) \quad (4.11)$$

where  $c(n)$  are the cepstral coefficients. Then the group delay function obtained by finding the derivative with respect to  $w$  can be expressed as

$$\tau(\omega) = \sum_{n=1}^{\infty} nc(n) \cos(n\omega) \quad (4.12)$$

Comparing the two equations Eq. 4.10 and Eq. 4.12, for a minimum-phase signal, we see that the natural logarithm of the amplitude function and the group delay function are related through cepstrum (this is further discussed in the next sub-section).

This property does not hold for mixed-phase signals, therefore the authors follow an alternative path and define two sets of cepstral coefficients,  $c_1(n)$  and  $c_2(n)$  for magnitude function and phase function:

$$\ln|X(\omega)| = c_1(0)/2 + \sum_{n=1}^{\infty} c_1(n) \cos(n\omega) \quad (4.13)$$

$$\theta(\omega) = -\sum_{n=1}^{\infty} c_2(n) \sin(n\omega) \quad (4.14)$$

They define two group delay functions, one derived from magnitude spectrum:

$$\tau_m(\omega) = \sum_{n=1}^{\infty} nc_1(n) \cos(n\omega) \quad (4.15)$$

and one derived from the phase spectrum:

$$\tau_p(\omega) = \sum_{n=1}^{\infty} nc_2(n) \cos(n\omega) \quad (4.16)$$

which are equivalent when the signal is a minimum-phase signal. However for maximum-phase signals  $\tau_m = -\tau_p$  and for mixed-phase signals there is no direct formula.

***The relation between phase and amplitude of frequency response of minimum-phase systems***

As a matter of fact, it is well known that for minimum-phase systems, the phase function can be computed directly from the amplitude function. Such a relation has been shown by Bode [Bode, 1945] that for a stable minimum-phase transfer function  $H(\omega)$ , the phase of the system at  $\omega_0$  is given by:

$$\angle H(\omega_0) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{d(\ln|H(\omega)|)}{d\nu} \ln \coth \frac{|\nu|}{2} d\nu \quad (4.17)$$

where  $\nu = \ln(\omega/\omega_0)$ . For the computation of this integral transform in the discrete time, the following method can be used [Cizek, 1970, Damara-Venkata *et al*, 2000]: Given a sampled magnitude spectrum  $|H[i]|$  for  $i=0, \dots, M-1$  where  $M$  is the DFT length, the sampled phase spectrum can be computed by:

$$\theta = -jDFT\{s \bullet IDFT\{a\}\} \quad (4.18)$$

where  $\bullet$  corresponds to term-wise vector multiplication and

$$s = [\text{sgn}[0], \text{sgn}[1], \dots, \text{sgn}[M-1]] \quad (4.19)$$

$$a = [\log|H(0)|, \log|H(1)|, \dots, \log|H(M-1)|]$$

such that

$$\text{sgn}[i] = \begin{cases} 0, & i = 0, \frac{M}{2} \\ 1, & 0 < i < \frac{M}{2} \\ -1, & \frac{M}{2} < i < M \end{cases} \quad (4.20)$$

Similarly, such a relation is expressed through Hilbert transform in [Papoulis, 1962] as follows:

If the causal and minimum-phase-shift function  $h(t)$  contains no singularities at the origin, then with:

$$H(\omega) = e^{-(\alpha(\omega) + j\theta(\omega))} \quad (4.21)$$

the phase and the amplitude functions,  $\alpha(\omega)$  and  $\theta(\omega)$  satisfy these equations:

$$\theta(\omega_0) = \frac{\omega_0}{\pi} \int_{-\infty}^{\infty} \frac{\alpha(\omega)}{\omega^2 - \omega_0^2} d\omega \quad (4.22)$$

$$\alpha(\omega_0) = \alpha(0) - \frac{\omega_0^2}{\pi} \int_{-\infty}^{\infty} \frac{\theta(\omega)}{\omega(\omega^2 - \omega_0^2)} d\omega \quad (4.23)$$

$\theta(\omega)$  can be uniquely determined from  $\alpha(\omega)$ , and for the determination of  $\alpha(\omega)$  one needs not only  $\theta(\omega)$  but also the constant  $\alpha(0)$ .

The numerical evaluation of the two integrals is in general complicated. By a change in the independent variable, a simpler set of equations, known as Wiener-Lee transforms are derived. Introducing the variable  $\delta$  defined by

$$\omega = -\tan \frac{\delta}{2} \quad (4.24)$$

If variable exchange is applied to phase and magnitude components we get (please refer to [Papoulis, 1962] for the complete derivation/proof):

$$\alpha(\omega) + j\theta(\omega) = \alpha(\tan \frac{\delta}{2}) - j\theta(\tan \frac{\delta}{2}) = \tilde{\alpha}(\delta) - \tilde{\theta}(\delta) \quad (4.25)$$

$$\tilde{\alpha}(\delta) = d_0 + d_1 \cos \delta + \dots + d_n \cos n\delta + \dots \quad (4.26)$$

$$\tilde{\theta}(\delta) = e_0 + e_1 \sin \delta + \dots + e_n \sin n\delta + \dots \quad (4.27)$$

and the coefficients are related by:

$$d_n = -e_n \quad (4.28)$$

if  $h(t)$  is causal and minimum-phase-shift.

## Difficulties in group delay processing

At the frequency values where one or more zeros exist very close to the unit circle, spikes are observed on the group delay function. One explanation for this phenomenon is that the term  $|X(\omega)|^2$  gets very small in Eq. 4.8 at those frequency bins [Hegde *et al*, 2004 a].

We can also add a geometric explanation at this point. The FT can be expressed through ZZT representation as

$$X(\omega) = x(0)e^{(j\omega)(-N+1)} \prod_{m=1}^{N-1} (e^{j\omega} - Z_m) \quad (4.29)$$

$$\text{Ang}(X(\omega)) = \theta(\omega) = (-N+1)\omega + \sum_{m=1}^{N-1} \text{Ang}(e^{j\omega} - Z_m) \quad (4.30)$$

where  $Z_m$  are the zeros. Each factor in Eq. 4.29 corresponds, in the  $z$ -plane, to a vector starting at  $Z_m$  and ending at  $e^{j\omega}$  and rate of change in the phase component (the group delay) is very high at frequency bins very close to the zero (Fig. 19).

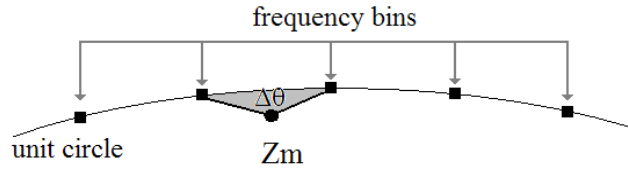


Fig. 19: Geometric interpretation for spikes in the group delay function at frequency locations close to a zero

For actual speech signals, many zeros appear to be very close to the unit circle. The effect of zeros close to the unit circle can easily be observed both on amplitude spectra and group delay functions. In Fig. 20, we present a windowed real speech frame with its ZZT representation and spectrum plots. The zeros close to the unit circle are shown in between dashed lines in Fig. 20b and they are also superimposed on spectrum plots to draw attention to their relation with the spectral dips in the magnitude spectrum (Fig. 20c) and with spikes on the group delay function (Fig. 20d).

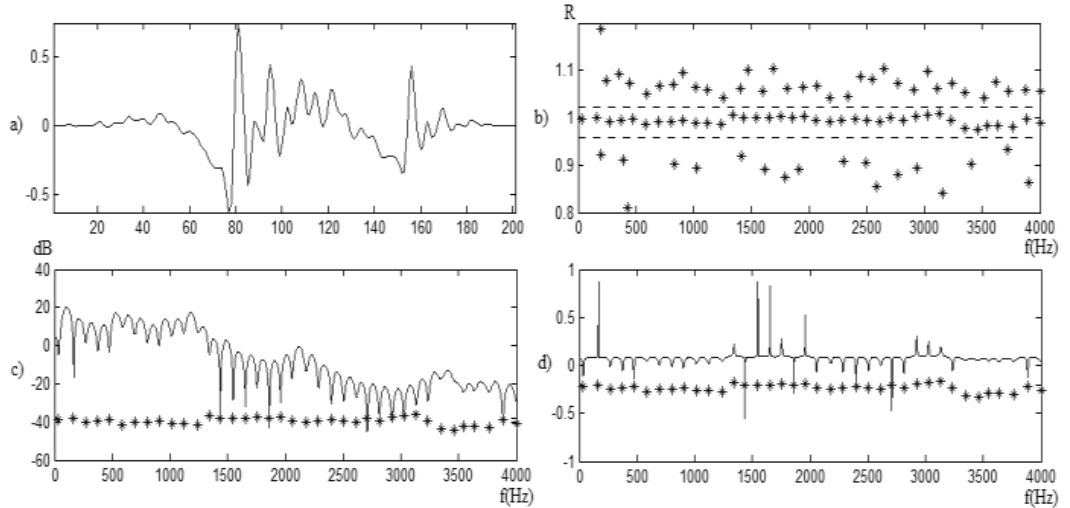


Fig. 20: Effects of zeros to the spectrum of a signal.

a) Hanning windowed real speech frame taken from a natural utterance (phoneme /a/ in the word “party”), b) ZZT representation in polar coordinates (zeros close to the unit circle are indicated by labeling a region in-between dashed lines), c) magnitude spectrum, d) group delay function. The zeros close to the unit circle are superimposed on c) and d) to show their effect.

The general shape of the group delay function is mainly dictated by the zeros close to the unit circle

(Fig. 20d waveform is like a DC function with spikes due to zeros close to the unit circle and no other spectral information is observed), and this domination of spikes hide the formant structure in the group delay function. For the magnitude spectrum, the spiky effect of zeros close to the unit circle is much reduced and we can still observe formants.

In the papers of Yegnanarayana and Murthy, the authors mention the existence of spikes on the group delay due to the zeros close to the unit circle and propose methods to remove those effects to obtain smooth group delay function. In our study, we aim at understanding the sources of zeros close to unit circle (by studying ZZT representations of speech signals) and developing methods that guarantee certain distance of zeros from the frequency points where FT (or the chirp  $z$ -transform (see Section IV.6)) is computed. We have shown in Section III.3.4 that the existence of zeros close to the unit circle is highly dependent on how windowing is performed. We show through ZZT representations that by appropriate windowing, group delay functions free of spikes can be obtained without further de-noising of the group delay on voice speech frames. We also propose computation of group delay on circles other than the unit circle (which results in what we name the chirp group delay) as an efficient way of avoiding spikes on the final spectral representation. Before presenting our methods, we present the methods for spike removal proposed by Yegnanarayana and Murthy.

### Processing group delay of the minimum-phase version of a signal

For getting rid of spikes due to the zeros, the authors propose several versions of group delay function estimation from minimum-phase version of a given signal [Yegnanarayana *et al*, 1988, Murthy *et al*, 1989 a, Murthy & Yegnanarayana, 1991 a]. The common steps involved in their methods are:

1. Obtaining the magnitude spectrum for a short-time windowed speech frame
2. Smoothing the magnitude spectrum via cepstral smoothing and
3. Computing smooth minimum-phase group delay from this representation through cepstrum.

Such processing removes the spikes from the group delay function computed. However, the resulting spectrum is a form of smoothed magnitude spectrum representation and the conversion to minimum-phase destroys the phase information available in the signal. The advantage of this representation compared to the magnitude spectrum is that the formant peaks appear with better resolution [Murthy *et al*, 1989 a]. The authors propose formant tracking algorithms by picking peaks on this new representation.

### Modified group delay function

Recently, Murthy and her colleagues proposed another group delay function, which does not include the “conversion to minimum-phase” step. The basic idea behind the calculation of this new representation called the Modified Group Delay (MODGD) function is: smoothing the  $|X(\omega)|^2$  component (in Eq. 4.8) (through cepstrum) which is considered as the main source of spikes. However, this modification alone cannot remove all spikes (see Fig. 21 for an example). The resulting spectrum still includes spikes but some formant peaks can be observed. To further reduce spikes, two new parameters are introduced:  $\gamma$  and  $\alpha$  which need to be fine-tuned according to the environment. The modified group delay function is defined as:

$$\tau_{\text{mod}}(\omega) = \left( \frac{\tau_p(\omega)}{|\tau_p(\omega)|} \right) \left( |\tau_p(\omega)| \right)^\alpha \quad (4.31)$$

where

$$\tau_p(\omega) = \left( \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{S(\omega)^{2\gamma}} \right) \quad (4.32)$$

where  $X(\omega)$  and  $Y(\omega)$  refer to the FT of  $x(n)$  and  $nx(n)$  respectively,  $R$  and  $I$  refer to real and

imaginary parts and  $S(\omega)$  is the cepstrally smoothed version of  $|X(\omega)|$ .

The authors propose to use this new representation in feature extraction for speech recognition and report improvement when combined with the power spectrum based feature construction [Hegde *et al*, 2004 b]. However, the results have been falsified in another paper [Zhu & Paliwal, 2004]. In the applications chapter of this thesis dissertation, we provide comparative tests in speech recognition for feature construction using the MODGD, Zhu and Paliwal's alternative: the product spectrum, the magnitude spectrum, and the group delay functions we specifically propose in this thesis.

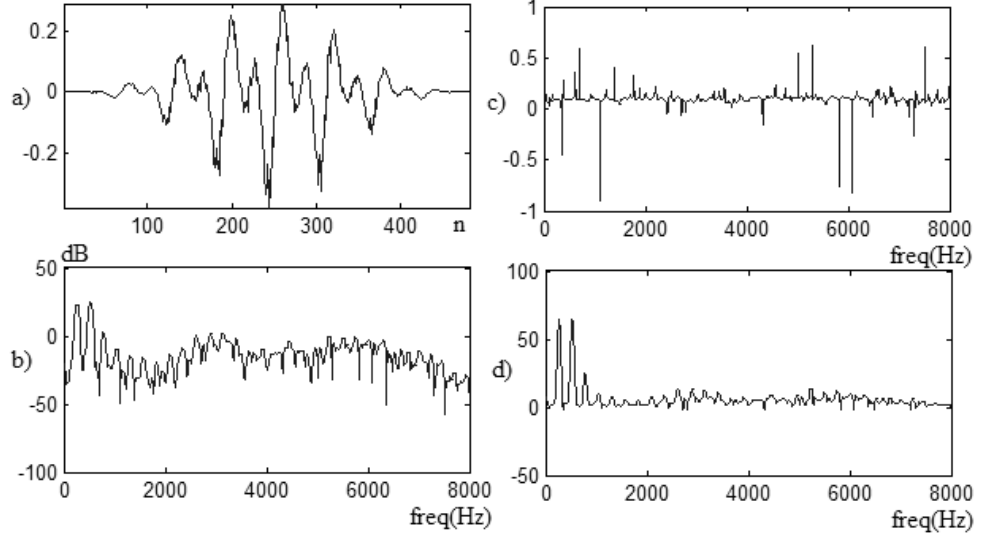


Fig. 21: Modified group delay function.

a) Time-domain speech signal, b)magnitude spectrum, c)group delay function, d)modified group delay function

The group delay functions proposed by Yegnanarayana and Murthy (which appear to contain either amplitude-only information or a mixture of amplitude and phase information) seem to be potentially useful in some spectral processing applications that process some global spectral characteristics of the speech signals (like ASR systems). However, such phase representations cannot provide real phase information that can be attributed to glottal flow or vocal tract components or improve our understanding of phase characteristics of speech signals. To enhance our knowledge about the phase content of speech signals we need to study phase characteristics in the view of the speech production theory and phase properties of finite length discrete time signals. Such a study leads to a better understanding of the phase component properties as we show in the next section.

### IV.3. Phase processing of mixed-phase signals

As defined in the previous section, the term “mixed-phase” (as defined for signals) refers to signals which have zeros on both sides (outside/inside) of the unit circle. There is some confusion when we consider the signal as an output of a system and attribute similar characteristics to the system (especially if the system is infinite impulse response (IIR)). When we analyze real-life signals, windowing is unavoidable, as we do not have the facilities to analyze infinite length signals. As we have shown previously, windowing changes the characteristics of signals to a great extent. Below, we demonstrate how window size may change the phase characteristics of a signal. A signal is created by exciting a minimum-phase all-pole system (with poles at  $[0.9432 + 0.2264i \ 0.9432 - 0.2264i]$ , and sampling frequency at 16000Hz) by a single impulse. Two windowed versions are obtained using different sized rectangular windows, one having minimum-phase characteristics and one having mixed-phase characteristics.

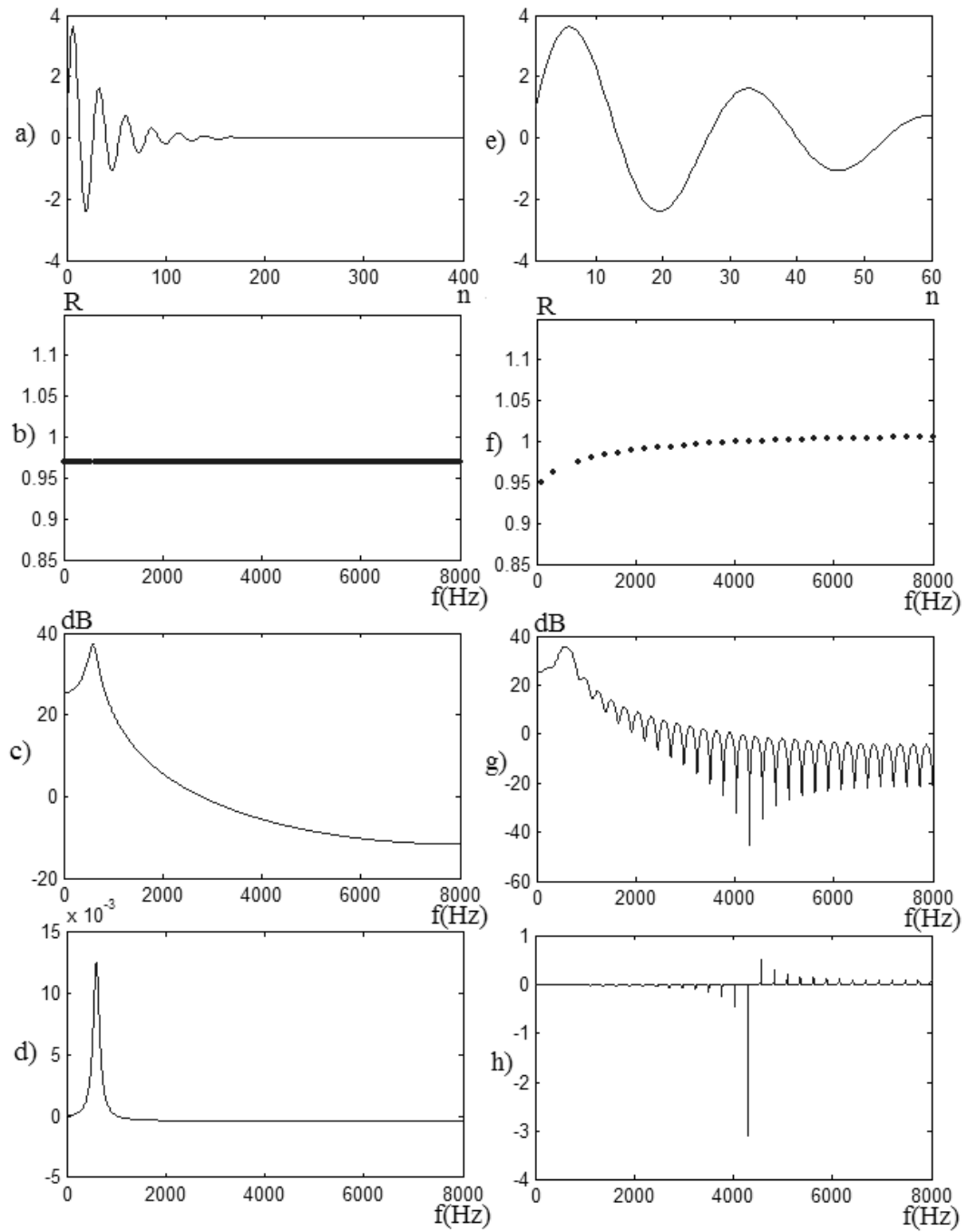


Fig. 22: Effect of windowing to phase characteristics of a signal.

Minimum-phase system output truncated at a zero-crossings: a) time domain signal, b) ZTZ, c) magnitude spectrum, d) group delay function. Minimum-phase system output truncated at a non zero-crossing right-hand boundary: e) time domain signal, f) ZTZ, g) magnitude spectrum, h) group delay function.

Considering the zero locations in Fig. 22b, the signal in Fig. 22a is a minimum-phase signal since all zeros occur inside the unit circle. The spectra for this signal are smooth and contain a peak corresponding to the single resonance due to the pole-pair. A second version of the windowed signal is obtained by taking the first 60 points of the same discrete time signal (Fig. 22e). This time the signal is mixed-phase since some of the zeros fall outside the unit circle (Fig. 22f). Therefore, the phase characteristics of a truncated signal cannot be directly attributed to the system of which the signal is considered to be a response.

Mixed-phase characteristics have been observed for speech signals starting from the early times of speech processing. However, very few studies address these observations and develop methods accordingly. A study of mixed-phase characteristics of speech signals is especially necessary in analysis/modification/synthesis methods since synchronization of excitation instants is very important in such systems. Introducing distortions in phase/synchronization leads to hoarseness and roughness in re-constructed speech.

One of the early uses of mixed-phase speech processing is [Quatieri, 1979]. Quatieri's study aims at reducing the dispersion of the signal in the time-domain caused by minimum-phase (or zero-phase) reconstruction. Such dispersion introduces hoarseness in re-constructed speech. Quatieri proposes a mixed-phase homomorphic system that uses complex cepstrum to retain the mixed-phase characteristics of the analyzed signal. He reports that using mixed-phase reconstruction provides generally higher quality speech than its minimum-phase counterpart.

In such approaches, the mixed-phase characteristics appear as an observed phenomenon and the general approach in processing is: "keep it as is not to introduce distortions". The sources of mixed-phase characteristics and the extent to which the components of the speech production mechanism contribute to these characteristics is not a largely studied issue. Below, we first show that speech signals are mixed-phase as a result of the production model. Then we discuss the effects of windowing discrete time speech signals and the contribution of windowing on the final speech data that is subject to analysis in our speech technology systems.

---

## IV.4. Mixed-phase speech model

Being the output of a physical system, the speech signal is assumed to be stable. Together with the causality feature, this assumption draws important guidelines for speech analysis. Once it is also assumed that the speech signal is causal, we end up with the minimum-phase speech model: all the poles of a signal that is causal and stable must lie inside the unit circle on the  $z$ -plane.

Here, we discuss a mixed-phase model of speech, where we assume that speech is obtained by convolving an anti-causal and stable source signal with a causal and stable vocal tract filter. In this model, some resonances of the signal correspond to poles outside the unit-circle on the  $z$ -plane but these poles are anti-causal, and therefore still stable. These anti-causal poles correspond to resonances of the glottal source signal, while the causal-stable poles (inside the unit circle on  $z$ -plane) correspond to the vocal tract resonances. The speech signal is a mixed-phase signal obtained by exciting a minimum-phase system (vocal tract system) by a maximum-phase signal (glottal source signal). It should be noted that the return phase component of the glottal source signal is included in the vocal tract component since it also has minimum-phase characteristics.

This assumption is based on the characteristics of glottal flow models (LF [Fant, 1995], KLGLOTT88 [Klatt & Klatt, 1990]) and was already discussed to some extent in a few studies. Citing from [Jackson, 1989]: "Since the human vocal tract is obviously causal, the implication of the non-causal impulse response resulting from cepstral analysis is simply that the standard causal source-filter approach with impulse train source is inadequate to represent the actual speech waveform, even though it can model the power spectrum... The author conjectures that the causal portion of the impulse response corresponds primarily to the closed-glottis state because the primary vocal tract excitation occurs at the open-to-closed glottis transition. Assuming this to be true, the anti-causal portion of the impulse response then corresponds primarily to the open-glottis state." Gardner [Gardner, 1994] has shown that mixed-phase models are appropriate for modeling voiced speech due to the maximum-phase nature of the glottal excitation. He shows that use of an anti-causal all-pole filter for the glottal pulse is necessary to resolve magnitude and phase information correctly. Anti-causality of the glottal flow signal has also been discussed within the context of spectrum of glottal waveform models in [Doval & d'Alessandro, 1997] and the authors point the similarity of the phase spectrum of KLGLOTT88 signal to an anti-causal filter phase spectrum. The authors compare the impulse response of an anti-causal all-pole system with KLGLOTT88 synthesized glottal flow signal and the main difference is reported to be the oscillations due to truncation.

An intuitive method to present this property is to compare time-domain signals of glottal flow excitation and causal and anti-causal filter responses. In Fig. 23, we present such an example: the



glottal flow signal looks like a time-reversed causal filter response (anti-causal filter response). (similar plots are available in [Gardner & Rao, 1997])

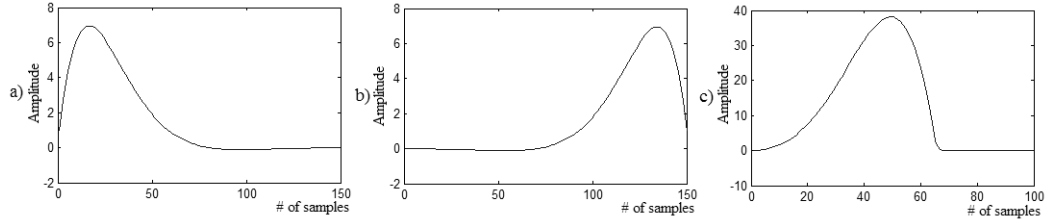


Fig. 23: Effect of anti-causality on the time-domain waveform.

a) a causal filter response, b) an anti-causal filter response, c) LF glottal flow signal

For the stability of an anti-causal all-pole system, all of the poles have to be out of the unit circle and therefore the system has to be maximum-phase. The mixed-phase model assumes that speech signals have two types of resonances; anti-causal resonances of the glottal source signal and causal resonances of the vocal tract filter.

In Fig. 24, we present the mixed-phase model for voiced speech impulse response. The lip radiation component is included in the glottal excitation component resulting in glottal flow derivative and the return phase component of the glottal flow is included in the vocal tract response which provides us finally with two components: a maximum-phase component and a minimum-phase component. The excitation of the resulting mixed-phase impulse response with a period impulse train results in realistic voiced speech signals.

Several spectral representations for the two components and the resulting mixed-phase speech signals are presented. The amplitude spectra appear as presented in most of the textbooks. The group delay functions exhibit interesting properties, which have not been discussed in the literature. Due to the anti-causality of the glottal flow, the corresponding group delay function includes a negative peak, which also appears in the speech group delay function. It is obvious by comparing magnitude spectrum and group delay functions of speech that mixed-phase characteristics (spectral components from maximum and minimum-phase parts) can only be observed on the group delay but not on the magnitude spectrum. This shows that studying phase/group delay information is especially important if we want to capture the characteristics of glottal flow and of the vocal tract components separately. It is important to note here that the operation of “conversion to minimum-phase” used in studies of Yegnanarayana and Murthy merges all components in a minimum-phase version signal. Therefore, phase/group delay information extracted from such minimum-phase version signals carries superimposed information from glottal flow and vocal tract.

The all-pole representation simply presents the locations of poles and regions of convergence (ROC) (indicated as shaded areas on the  $z$ -plane). For stability, the unit circle should be included in the ROC and it is the case for the three representations. A pole pair is attributed to the glottal flow derivative component. For computational details about the location of these poles and the corresponding glottal flow parameters, the reader is referred to [Doval *et al.*, 2003]. Estimating these poles from speech signals by LP analysis is an alternative way of glottal excitation analysis we have recently tested. We present our algorithm in the applications section of this thesis dissertation (section V.4). The two similar methods we could find in the literature are [Jackson, 1989, Gardner & Rao, 1997].

The ZZT representations are also provided. Comparing this all-zero representation with the all-pole representation, we see that poles correspond to zero-gaps in the ZZT representation. The previously discussed pattern of grouping of zeros outside and inside the unit circle is in accordance with the model presented here.

In this section we have presented the mixed-phase model for the voiced speech impulse response. However, the observation of the discussed mixed-phase characteristics is not trivial on short-time speech signals. In the previous section, we have demonstrated that the phase characteristics of a windowed speech signal is not only dependent on the system that produces speech and that windowing plays an important role in the final phase characteristics of the actual discrete time speech signals. In the next section, we study this second part of the phase analysis problem.

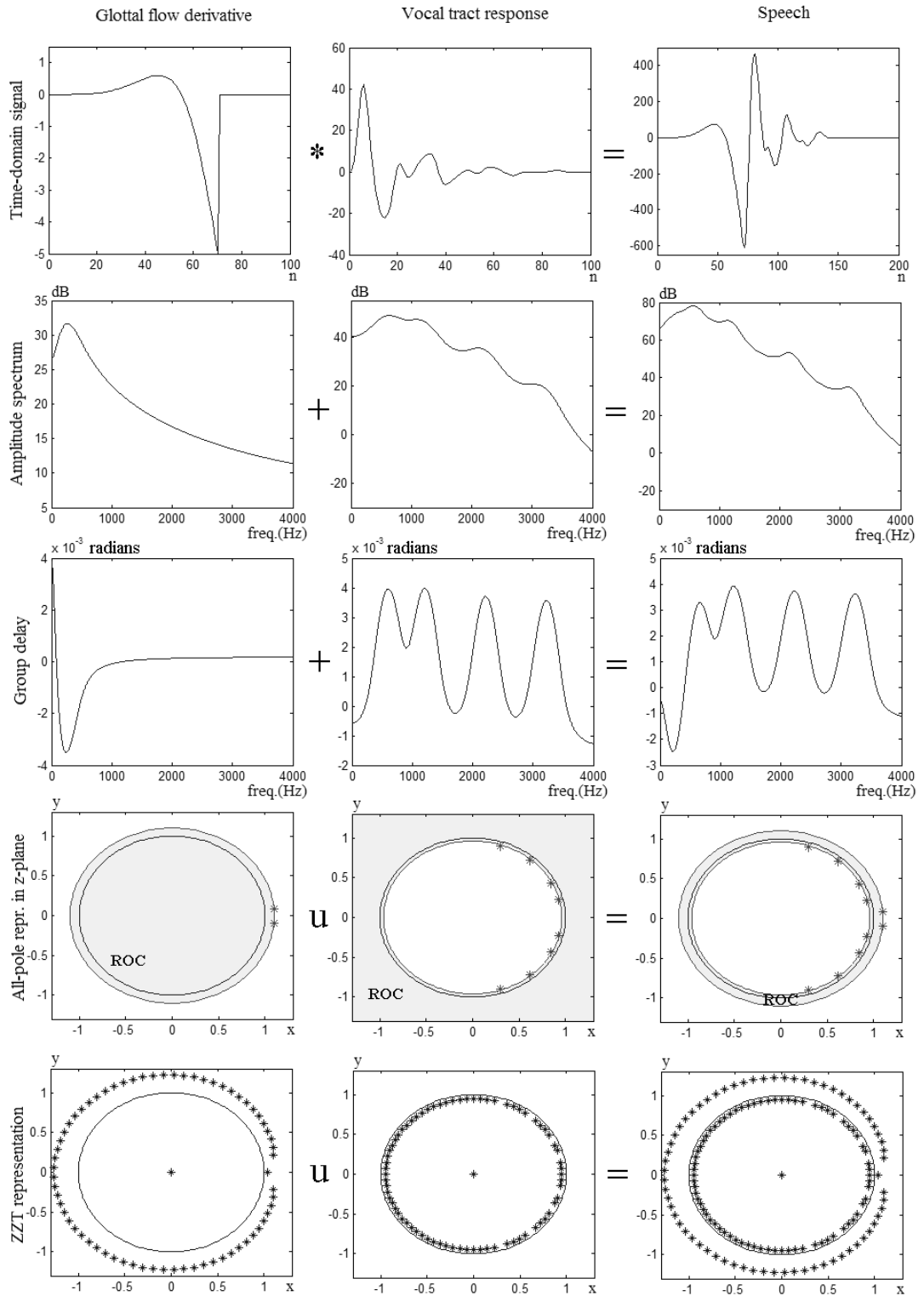


Fig. 24: The mixed-phase speech model

---

## IV.5. Effects of windowing on group delay functions

The effects of windowing on the phase spectrum estimation have been addressed in a few recent studies [Zolfaghari *et al*, 2003, Alsteris & Paliwal, 2004]. Especially the study of Alsteris and Paliwal [Alsteris *et al*, 2004] is a good example of the importance of windowing in phase estimation: the authors show that Liu's extensive study [Liu *et al*, 1997] on phase contribution to speech intelligibility can provide quite different results when the window function and window shift is modified in the procedure. They show that Liu's choice of window function, Hamming, needs to be replaced by Rectangular and such modification completely changes the results leading to different conclusions. Still the studies mentioning the importance of the windowing operation lack background explanation and the preferences are often based only on 'trial and error' methodologies. The ZYT is a very appropriate representation to study effects of windowing on the phase spectrum and we use it as the basic tool in this chapter.

The effect of windowing in zero-patterns is drastic as shown in Section III.3.4. In addition, we have discussed the difficulty introduced by zeros on the unit circle to phase computation. The study of effects of windowing on ZYT is therefore very important to understand the actual phase characteristics of the speech signal. However, there is an important lack of mathematical theory for studying roots of high degree polynomials since it is too complicated. For studying windowing, the main difficulty stems from the fact that term-wise multiplication of two discrete time signals in time-domain corresponds to term wise multiplication of  $z$ -transform polynomial coefficients and how roots of the polynomial are displaced after this operation is an issue very hard to predict analytically (This is impossible in the general case. However for very special cases like multiplication with an exponential, it is possible as shown in Eq. 3.8). For this reason, we studied the zero-patterns of windowed data and the group delay functions by observations rather than by mathematical analysis, on various examples. One of the best ways of studying such variations is to create movies and observe changes due to variations in: windowing size, location and function. We have created movies by shifting windows on signals and observing ZYT and group delay function using:

- Synthetic and real speech signals,
- Window sizes with  $T_0$ ,  $2T_0$  and  $3T_0$
- Window functions: Rectangular, Hamming, Hanning, Blackman, Gaussian, Hanning-Poisson

Candidate window functions are chosen according to their popularity and their spectral characteristics. In Appendix A, the definitions and plots for these functions are presented. Some of the movies created are available on <http://www.tcts.fpms.ac.be/demos/zyt/index.html>.

Due to space limitations, here we present only selected plots. The windowing effects to ZYT have already been presented in Section III.3.4. To avoid duplication we present only group delay plots in this section and refer to the plots in Section III.3.4 when necessary. There are actually three dimensions to study: window location, size and function effects.

### Effects of window location on group delay functions

We have previously shown in Fig. 22 that truncation points (window boundaries) play an important role in the final phase characteristics of the windowed signal. A discontinuity at these boundaries are likely to lead to mixed-phase characteristics even though the signal is a minimum or a maximum-phase system response. One other important point about location of the window is the synchronization of the window center with certain instants like GCI instant for speech signals. We have shown in Section III.3 that ZYT patterns are close to ZYT of speech production system impulse response patterns when window is centered at the GCI instant. Once the window is placed such that the increasing exponential part is multiplied with the first half of the window, which is also increasing, and the decreasing exponential part is multiplied with the second half of the window, which is also decreasing, the ZYT outside the unit circle are kept outside and ZYT inside

the unit circle are kept inside and further pushed away from the unit circle. This results in a zero-gap on the unit circle and smooth group delay functions with characteristics matching the mixed-phase model we have presented. Below in Fig. 25 (equivalent ZZT plots presented in Fig. 12 in section 3), we present the group delay functions obtained from a real speech signal for six different locations of the window.

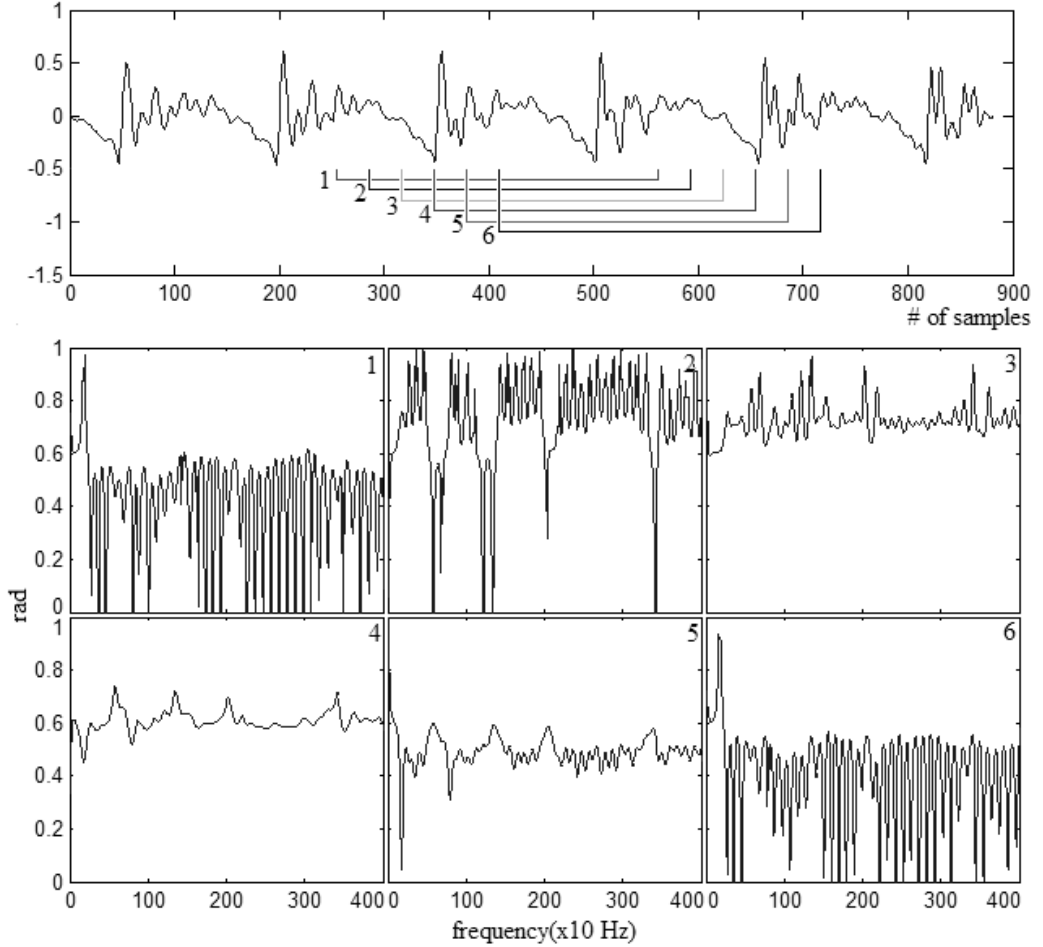


Fig. 25: Effect of windowing location to group delay of a real speech signal. Each (Blackman) window position is indicated on the signal on the top figure with reference numbers. The group delay function of the resulting windowed data for each window is presented with the window index indicated on the right-top corner of the figure.

Therefore two criteria are derived from these observations for reliable phase/group delay function estimation: window center should be synchronized with GCI instants and the boundaries should correspond to zero-crossings of the signal. The second condition may result in asymmetric windowing when the distance from zero-crossing on two sides of the GCI are not the same. Actually, this does not appear to be an important problem in the examples we have studied. Using a smooth window function with zero boundaries removes such discontinuities. Matching with zero-crossing is necessary only for windows with non-zero boundaries and asymmetric windows can be used in that case, i.e. two sides of the window may have different lengths.

### Effects of window size on group delay functions

Window size is also important. There is especially a big difference in group delay functions obtained with a window size smaller than two pitch periods and a window size bigger than two pitch periods. For windows larger than two pitch periods, the signal contains several periods, which means an impulse train component can be considered to be included. This results in ZZT of

impulse train to appear close to the unit circle introducing spikes in the group delay function. This is demonstrated in Fig. 26 where window center is at GCI and we only vary the window size. A window size in the  $T_0$ - $2T_0$  range appears to be a good choice for group delay processing.

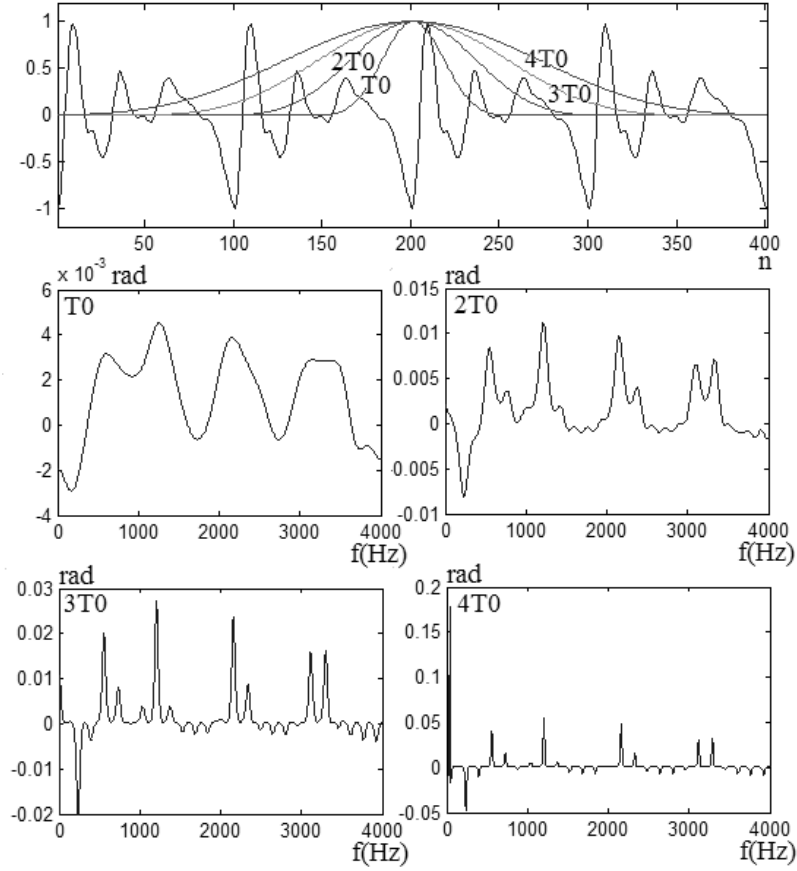


Fig. 26: Effect of windowing size to group delay of a synthetic speech signal. Each (Blackman ) window size is indicated on the window waveform on the top figure. The group delay function of the resulting windowed data for each window is presented with the window size indicated on the left-top corner of the figure.

### Effects of window function on group delay functions

Windowing function is also important but comparatively less important than window size and location once we limit ourselves with commonly used window functions listed above (see Fig. 27 for group delay functions obtained on the same data frame using different window functions). We observed that three types of windowing functions provide best group delay functions: Blackman, Gaussian and Hanning-Poisson.

The Hanning-Poisson windows provide the smoothest group delay functions since the Poisson contribution of the window is composed of exponential functions. Windowing with a Hanning-Poisson results in multiplication of exponentials of the Poisson function and the speech signal, thus addition of decay coefficients of the window and the glottal flow and vocal tract responses (Eq. 3.11 and Eq. 3.12). This shifts zeros further away from the unit circle. For this reason, Hanning-Poisson window is preferable in group delay based analysis methods. Hanning-Poisson and Gaussian are the functions for which the smoothness of the representation can be adjusted to some level with the decay coefficient (these two window functions have an independent user controlled parameter for adjusting decay coefficients).

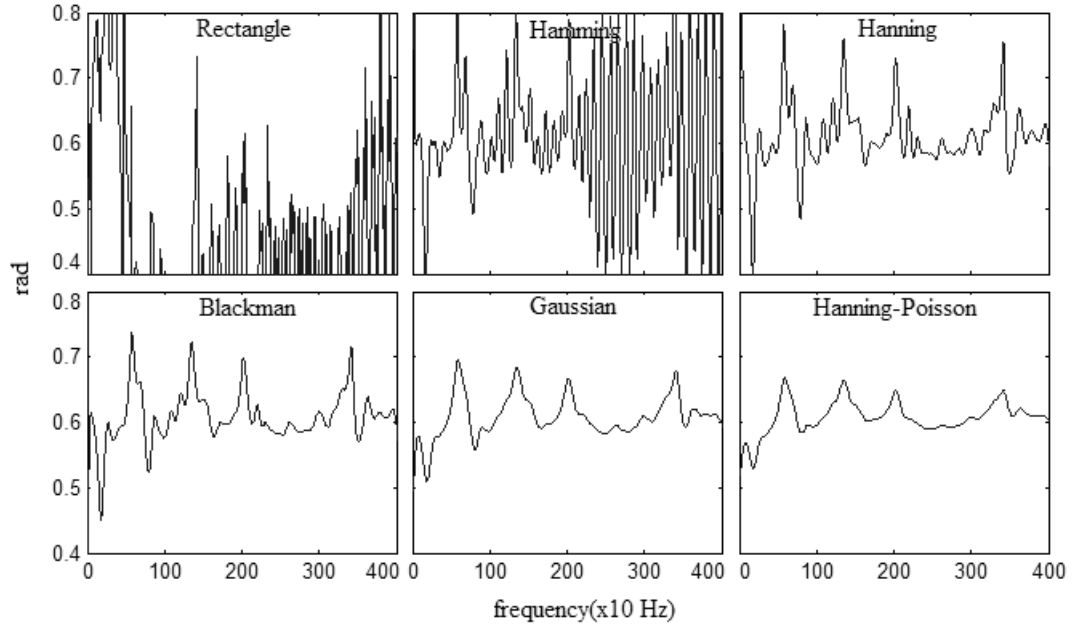


Fig. 27: Effect of windowing function to group delay

### Group delay spectrogram

Finally, we show that group delay functions, computed on data with proper windowing (two pitch period size Hanning-Poisson windowing centered at GCI instants), provide the formant structure, on a real speech example (BrianNormal3.wav from Voqual 03 database [www-Voqual03], for which the uttered sentence is “*she has left for a great party today*” with modal phonation). We obtained spectrogram-like plots from the positive part of the group delay functions computed for voiced frames of the complete speech data. Fig. 28 shows the spectrogram obtained by group delay functions and amplitude spectra and their correlation is obvious for the formant tracks. This figure shows that the group delay functions indeed carry resonance information of the signal once windowing is properly performed.

### Conclusion

In this section we have shown that windowing plays a very important role in reliable group delay estimation. This point is very important for phase processing since even very recent studies concerning phase information do not take it into consideration: not respecting the criteria listed above results in unreliable phase estimation, which leads to loss of research time/effort.

Until this point we have discussed how to get rid of masking spikes due to inappropriate windowing that hide the speech characteristics (like formants) in the actual group delay functions. Apart from windowing, there are still sources of spikes like the noise component in speech. As we have discussed in Section III.3, the ZZT of noise components may appear on the unit circle and introduce spikes. In addition, GCI and pitch detection are not always very robust which may result in inappropriate size and location of the window. These factors reduce the robustness of group delay computation. In the next section, we further target new group delay functions that are easy to process to obtain resonance information and which are more robust to compute. We introduce chirp group delay processing methods as alternative ways of obtaining phase related spectral information. Such representations are potentially useful in various speech applications like automatic speech recognition (ASR). They also facilitate studying minimum-phase and maximum-phase contributions in the signals separately.

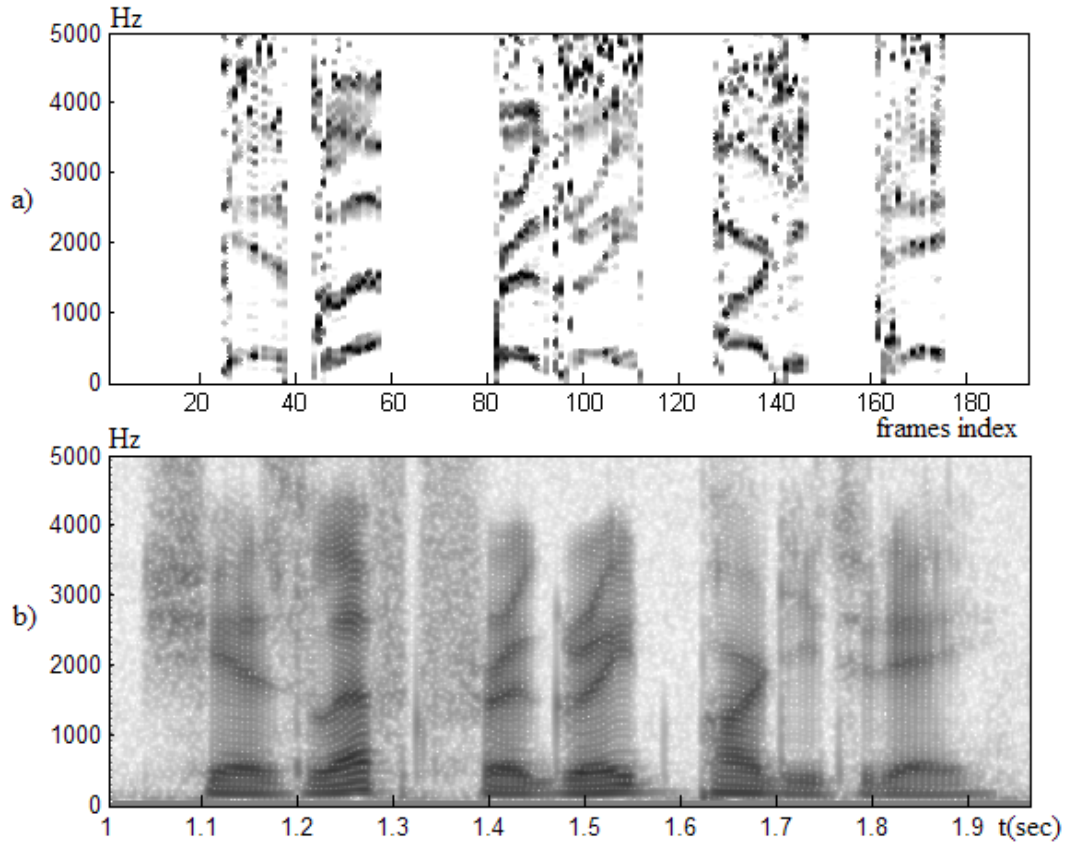


Fig. 28: Group delay spectrogram.

(a) Group delay and (b) amplitude spectrogram for the sentence “she has left for a great party today”

## IV.6. Chirp group delay processing of speech

We define the term chirp group delay<sup>3</sup> as the negative derivative of the phase spectrum (the group delay function) computed from chirp  $z$ -transform [Rabiner *et al*, 1969], that is  $z$ -transform computed on a circle/spiral other than the unit circle. Given the chirp- $z$  transform  $CZT(\omega)$ , the chirp group delay,  $CGD(\omega)$ , is defined by Eq. 4.35:

$$CZT(\omega) = X(z)|_{z=\rho e^{j\omega}} = \sum_{n=0}^{N-1} x(n) (\rho e^{j\omega})^{-n} = a(\omega) + jb(\omega) \quad (4.33)$$

$$\mathcal{G}(\omega) = \arctan\left(\frac{b(\omega)}{a(\omega)}\right) \quad (4.34)$$

$$CGD(\omega) = -\frac{d(\mathcal{G}(\omega))}{d\omega} \quad (4.35)$$

where  $\rho$  is the radius of the analysis circle. It is interesting to note that an existing fast Fourier Transform (FFT) implementation can be used to compute  $CZT(\omega)$  by re-writing the equation as:

<sup>3</sup> Initially the term “differential phase spectrum” has been used in our early papers. After a suggestion by Kuldeep K. Paliwal, we have decided to use “chirp group delay” instead.

$$CZT(\omega) = \sum_{n=0}^{N-1} (x(n)\rho^{-n})(e^{j\omega})^{-n} \quad (4.36)$$

$$CZT(\omega) = \tilde{X}(\omega) = \tilde{X}(z) \Big|_{z=e^{j\omega}} \quad (4.37)$$

$$\tilde{x}(n) = x(n)\rho^{-n}, n = 0, 1, 2, \dots, N-1 \quad (4.38)$$

Therefore, for computation of the CGD from a given signal, it is sufficient to term-wise multiply the data array with an exponential array and compute the group delay with direct formula Eq. 4.8.

Equivalently, given a ZZT representation for a signal, we can compute the chirp  $z$ -transform using the equation:

$$CZT(\omega) = x(0)(\rho e^{j\omega})^{(-N+1)} \prod_{m=1}^{N-1} (\rho e^{j\omega} - Z_m) \quad (4.39)$$

The main motivation for processing CGD computed on circles other than the unit circle is to get rid of spikes created by zeros of the  $z$ -transform (ZZT) which mask formant peaks on group delay functions.

**“By both manipulating the ZZT and adjusting the analysis circle radius for CGD computation, we can guarantee certain distance of zeros to the analysis circle. This is one of the basic ideas proposed and used through out this thesis.”**

We have discussed the difficulties involved in group delay processing and the link to windowing. We have proposed criteria for appropriate windowing. Although the group delay we obtain, when the criteria for windowing is taken into consideration, is incomparably smoother and spike-free, there are still some problematic issues for some cases like: errors in GCI detection, presence of additive noise in speech, errors in pitch period estimation leading to including more than two pitch periods in the window frame (which introduce extra zeros close to the unit circle). In Fig. 29, we present one such example of real speech. The formant structure is observed on group delay and most of spikes are avoided. But still we cannot guarantee that no zero will be close to the unit circle and the group delay contains some noise and two sharp spikes. For most of the speech applications this is undesirable.

For applications like formant tracking and speech recognition, we propose to use CGD with some rough control on zero locations so that group delay is computed on a zero-free region. Two new representations are proposed here for this purpose: a GCI synchronous and an asynchronous version. The asynchronous version is more advantageous than the synchronous method in terms of: i) computational efficiency, ii) independency from GCI synchronization, iii) robustness to noise. However the actual phase information is destroyed in the asynchronous version, since it contains only the information available in the magnitude spectrum. We briefly present the two methods below.



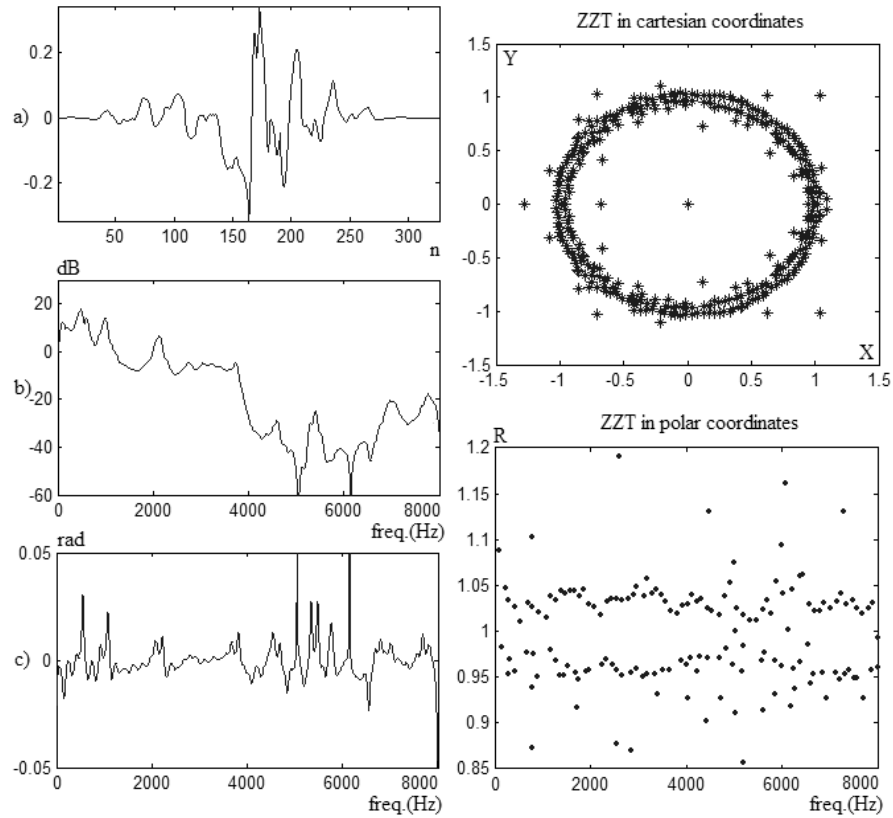


Fig. 29: Remaining problems for group delay of GCI synchronously windowed data.  
a) GCI synchronously windowed speech data, b) magnitude spectrum, c) chirp group delay

#### ***Chirp Group Delay of GCI-Synchronously Windowed Speech (CGDGCI)***

Two steps are necessary in the computation of the CGDGCI representation: suppression of the ZZT outside the unit circle on GCI synchronously windowed data and then computing the CGD outside the unit circle from zeros inside the unit circle. This representation contains only the phase information of the minimum-phase component of the data. For GCI synchronously windowed speech signals, the minimum-phase component is due to vocal tract and return phase of the glottal flow, therefore such a representation can be used for formant tracking and speech recognition applications successfully.

In Fig. 30, we present the CGDGCI computed on the example in Fig. 29. The ZZT after suppression of ZZT outside the unit circle and the analysis circle is presented on the left figure of Fig. 30. The CGDGCI obtained is much smoother than the group delay in Fig. 29c.

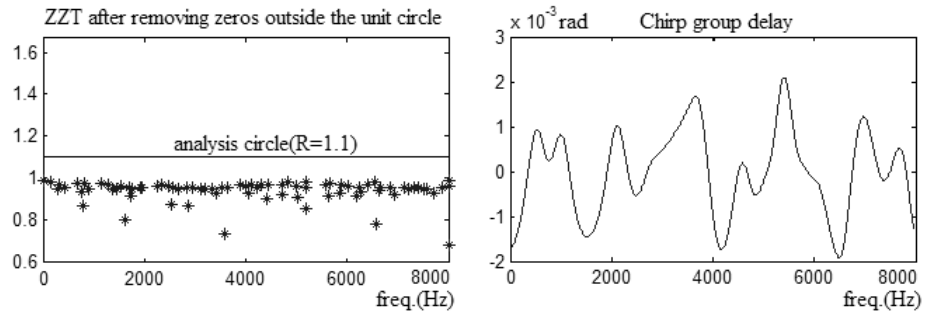


Fig. 30: First method of chirp group delay processing.  
a) ZZT after suppression of ZZT outside the unit circle, b) chirp group delay

We want to stress once more that it is extremely useful to consider CGD and ZZT together as in

CGDGCI: without any control on ZZT, CGD are likely to suffer from random peaks/spikes (as regular group delay functions). An example is presented in Fig. 31. Although most of the ZZT appear at a certain distance from the analysis circle, a single zero close introduces a large peak hiding formant information.

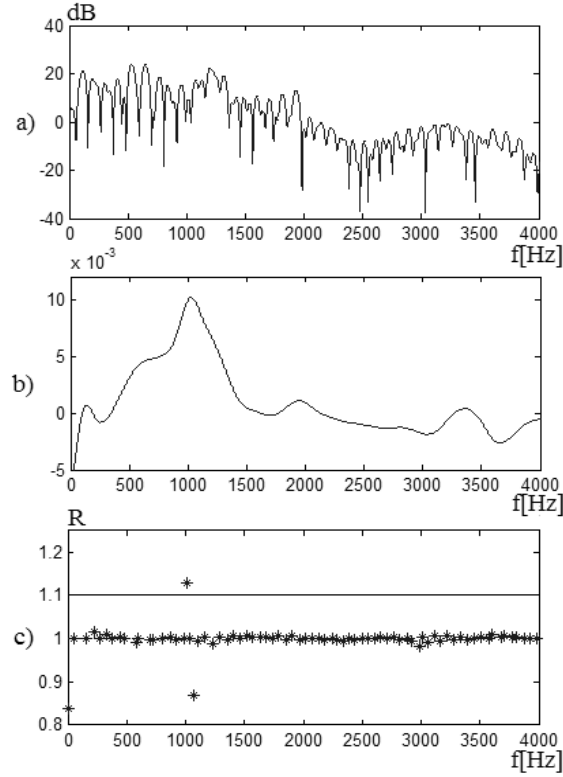


Fig. 31: Effect of ZZT on chirp group delay

a) magnitude spectrum of a real speech frame, b) chirp group delay computed on  $R=1.1$ , c) ZZT representation on  $z$ -plane (analysis circle indicated by a line at  $R=1.1$ )

A formant tracking algorithm will be presented in the applications part of the thesis dissertation, which simply picks the peaks on CGDGCI [Bozkurt *et al*, 2004c]. With comparative tests both on synthetic and real speech, we show that our formant tracker is high quality. However it has two drawbacks: GCI detection is necessary and it is computationally heavy due to calculation of zeros. We developed the second representation to get rid of these difficulties.

#### ***Chirp Group Delay of The Zero-Phase Version (CGDZP)***

Again the procedure contains two steps for the computation of CGDZP: computation of the zero-phase version of the signal (inverse FT of  $|X(\omega)|$ ) and computation of the CGD on a circle outside the unit circle ( $\rho=1.12$  appears to be a good choice by experience) using the chirp  $z$ -transform. Conversion to zero-phase guarantees that all of the zeros occur very close to the unit circle therefore the resulting chirp group delay function is very smooth with well-resolved formant peaks. However, the phase information is destroyed for this case, therefore the representation contains only the information available in the magnitude spectrum but formant peak resolutions appear with higher resolution.

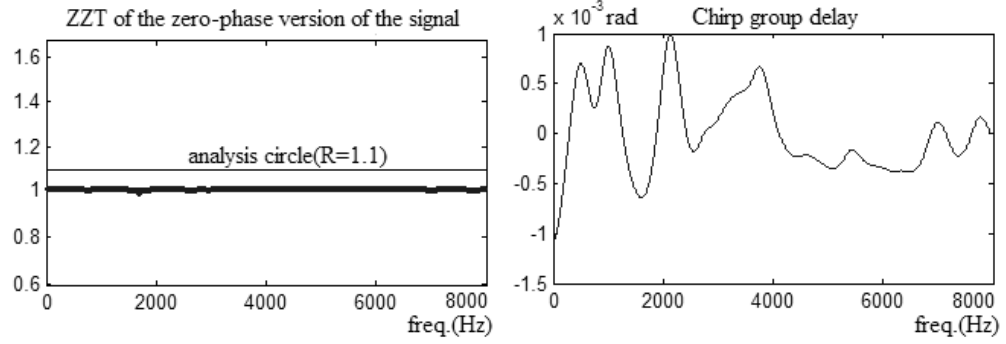


Fig. 32: Second method of chirp group delay processing.

a) ZZT after the zero-phasing operation, b) chirp group delay

We provide an example of the chirp group delay obtained by this second method in Fig. 32. In preliminary tests, we have observed that this method is more robust to noise than the first method. The other advantages are: GCI synchronization is not necessary and there is no need for computation of zeros (therefore the computational load is much lower).

---

## IV.7. Conclusion

In this section, we have first reviewed the group delay processing theory and discussed the difficulties involved. We have shown that windowing plays an important role and reliable/clean group delay functions revealing resonance information can be obtained from speech signals if windowing is properly performed. This is an important step in speech analysis since phase characteristics of speech signals are known to be important for perception but their analysis have always been reported to be very difficult. In some studies concerning speech coding and synthesis, the obscurity of phase information is reported and for improving naturalness of re-constructed speech, phase randomization techniques are tried (often by trial and error methodologies).

Windowing is a secondary issue (in importance) even in most of the speech analysis studies. Many phase studies use rectangular or Hamming windows, which are not appropriate from our point of view. Here we showed that for phase analysis, windowing is indeed one of the key issues. We expect that the outcomes of this study will be useful for improving effectiveness of speech analysis algorithms where phase information is valuable.

We have also presented a mixed-phase model for speech. In that model, the glottal flow first phase (“the active part”) has maximum phase characteristics and the remaining components have minimum phase characteristics and therefore the combination, the speech signal, has mixed-phase characteristics. When actually discrete-time speech signals are concerned, the truncation/windowing operation contribution needs to be taken into consideration and a good way to study this is to study the locations of the ZZT of the resulting truncated/windowed signal. As we have shown in Fig. 22, a truncated minimum phase filter impulse response may have minimum phase or mixed phase characteristics depending on the truncation boundaries.

In addition, two new chirp group delay functions are presented (CGDGCI and CGDZP). Among the two representations, CGDZP appears to be especially useful in practical applications like speech recognition since: the load for computation of it is low and it is more robust to noise. However it mainly contains magnitude information. CGDGCI computation is heavy however it provides us the opportunity to estimate minimum phase and maximum phase components’ phase information separately.

*Second Part*

# APPLICATIONS OF ZZT AND CHIRP GROUP DELAY PROCESSING IN SPEECH ANALYSIS

**T**he second part of this thesis is dedicated to the use of the two representations presented in part one in various speech analysis and parameter estimation problems. In the first chapter, we describe a source-filter decomposition algorithm using the ZZT representation. The algorithm classifies zeros of glottal flow component and zeros of vocal tract components and constructs FT spectra of these two components. The second application presented is a glottal flow parameter estimation algorithm using the ZZT-decomposition. The proposed algorithm estimates the frequency of glottal formant, which is potentially useful in studying some voice quality variations. Third application is in formant tracking and both the ZZT representation and chirp group delay processing theory are utilized. Three formant trackers are presented with varying complexity and robustness. The formant trackers are tested both on synthetic and real speech signals and shown to be effective by comparing to three state-of-the-art formant trackers: that of Praat, WinSnoori and Wavesurfer. The fourth algorithm presented is based on linear predictive (LP) modeling of mixed-phase speech signals. This algorithm does not include ZZT or chirp group delay processing but the mixed-phase speech model discussed in the second part and the well-known LP-covariance approach to modeling/analysis from literature. Finally the fifth application is in speech recognition. In this part we show that chirp group delay function carries equivalent or complementary information to power spectrum, which can improve speech recognition performance.

---

# Chapter V:

## Applications of ZZT and Chirp Group Delay Processing in Speech Analysis

---

### V.1. ZZT-decomposition for source-filter separation of speech

In chapter III, we have shown that separate patterns for the glottal flow first phase and the vocal tract (plus the return phase) contributions exist on the ZZT of speech signals. In this section, we propose an easy to implement and high quality source-tract decomposition algorithm using this property.

In section V.1.1 we present the ZZT-decomposition algorithm and section V.1.2 is dedicated to the evaluation of the algorithm. A similar decomposition can theoretically be performed by cepstral deconvolution. Therefore, for completeness we discuss the link between ZZT decomposition and the mixed-phase decomposition by complex cepstrum in section V.1.3. Finally, the conclusion is presented in section V.1.4.

#### V.1.1. The ZZT-decomposition algorithm

The decomposition algorithm we propose is based on the patterns of GCI synchronously windowed speech signals: the ZZT outside the unit circle (UC) is mainly due to the glottal flow first phase and the ZZT inside the unit circle is mainly due to the vocal tract filter and the glottal flow return phase. Grouping zeros into two sets by their location on the  $z$ -plane, the signal can be decomposed into those two parts. In Fig. 33, we present our ZZT-decomposition algorithm for source-tract separation based on the characteristics of the ZZT of GCI synchronously windowed data.

The decomposition starts with a pitch detection algorithm (PDA) and a voiced/unvoiced decision<sup>4</sup> (ZZT decomposition can be performed only for voiced frames). Given a first estimate of the pitch mark locations, GCI detection is performed with the technique defined in [Kawahara *et al*, 2000], which is based on processing of the evolution signal of center of gravity of windowed speech signals. A Blackman window with a size of two pitch periods and centered at GCIs is observed to be a good choice for the GCI synchronous windowing operation (as discussed in Section IV.5). Zeros are separated into two subsets based on their radius. Computing DFTs for each group is straightforward using Eq. 5.1. No  $z$ -transform polynomial re-computation is performed from given roots since some overflow problems are observed. Direct DFT computation from the roots appears to be safer.

---

<sup>4</sup> The pitch detection and voiced/unvoiced classification algorithm used in this study based on MBE [Griffin & Lim, 1988] analysis of speech signals. Most of the tools used for such analysis are taken directly from the database processing tools of the Mbrola project [www-Mbrola]. In addition to the available tools, a pitch marker is implemented based on processing the phase of the first harmonic. Details of these tools are not included in this thesis since they are rather out of topic and the theory presented here is independent of the particular pitch tracker. For further details, the reader is referred to [Bozkurt *et al*, 2004 f].

$$X(e^{j\varphi}) = Ge^{(j\varphi)(-N+1)} \prod_{m=1}^{N-1} (e^{j\varphi} - Z_m) \quad (5.1)$$

The most important detail for ZZT-decomposition algorithm is the existence of a single zero on the real axis due to the anti-causal portion of the signal, which in some cases falls inside the unit circle. This zero can be observed on figures Fig. 8c and 2nd row 2nd column of Fig. 11. At the moment no governing rule has been found for the classification of this zero (inside or outside?) and an heuristic approach is used for this problem; if no zero has been found on the real axis in the range  $R=[1 \ 1.1]$ , then the closest zero on the real axis to the point ( $R=1$ ,  $\varphi=0$ ) is removed from the set of zeros inside the unit circle set and put in the set of zeros outside the unit circle.

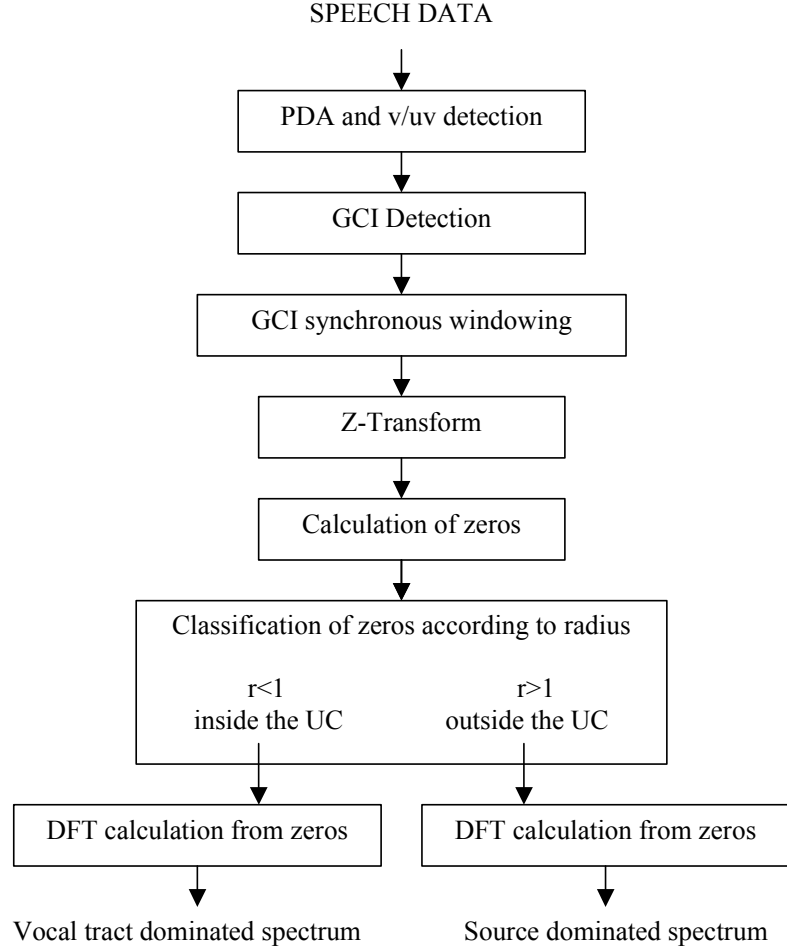


Fig. 33: The ZZT-decomposition algorithm

### V.1.2. Examples and evaluation of the decomposition algorithm

For demonstrating the efficiency of the algorithm, we first present example signal plots for decomposition of synthetic and real speech signals. For a more complete evaluation of the ZZT decomposition we present tests in a parameter estimation scheme: that of glottal formant frequency estimation (section V.2).

#### Synthetic speech example

We first present the effectiveness of the decomposition method on a synthetic speech example in

Fig. 34 and Fig. 35. Speech is synthesized by filtering the periodic glottal flow excitation (LF signal without a return phase) in Fig. 34a, by a four pair all-pole vocal tract filter. The resulting signal is presented in Fig. 34b. Our choice of window location for the ZZT-decomposition is presented on the signals in Fig. 34a and Fig. 34b and the windowed speech signal is presented in Fig 34c with the ZZT representation in Fig 34d. ZZT-decomposition is performed from the ZZT representation in Fig. 34d.

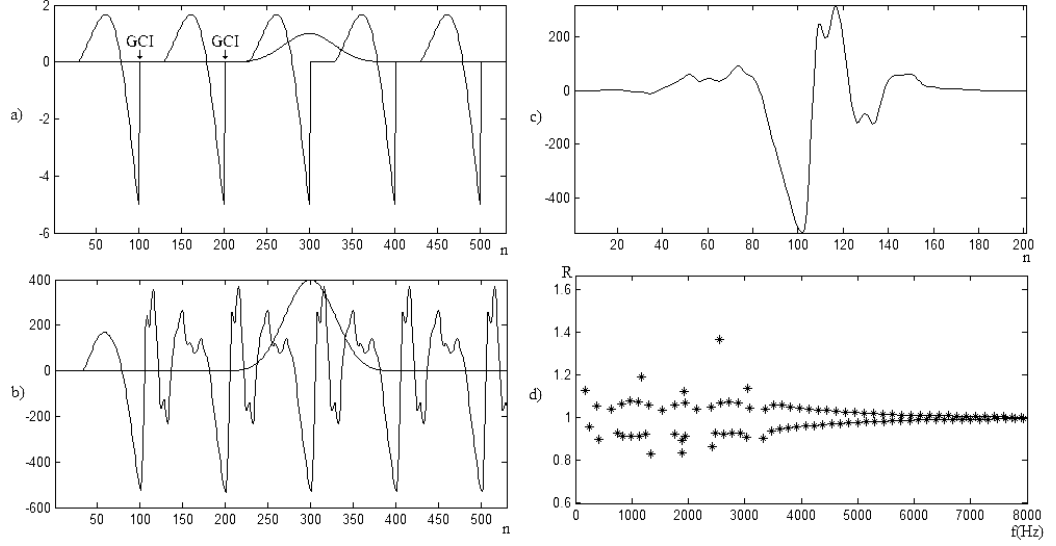


Fig. 34: Synthetic speech signal for testing ZZT-decomposition.

a) periodic glottal flow signal, b) speech synthesized by filtering the excitation signal in a, c) windowed synthetic speech, d) ZZT of the windowed speech frame

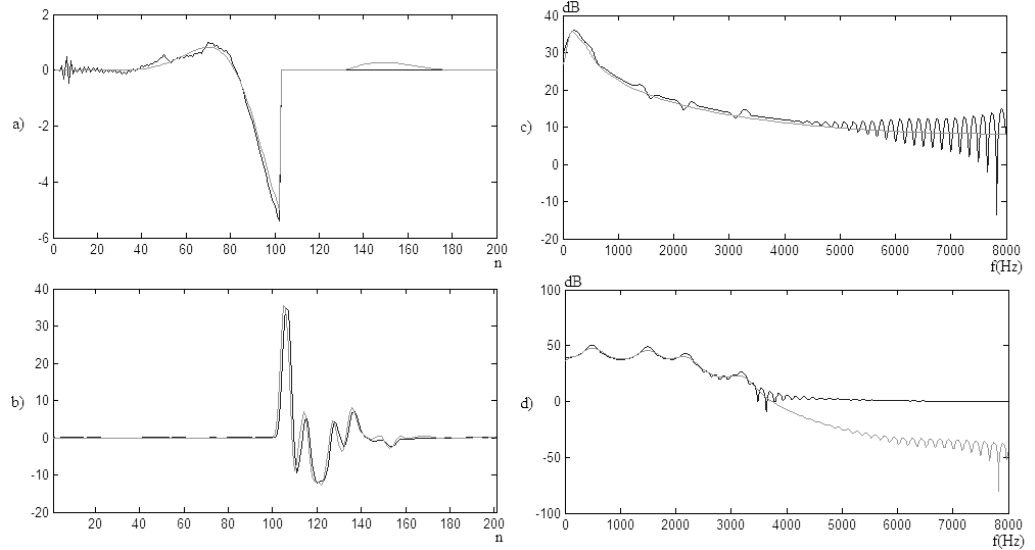


Fig. 35: ZZT-decomposition results of synthetic speech.

a) time-domain signal obtained by inverse DFT from spectrum calculated from zeros outside the unit circle, b) time-domain signal obtained by inverse DFT from spectrum calculated from zeros inside the unit circle, c) magnitude spectrum for the zeros outside the unit circle, d) magnitude spectrum for the zeros inside the unit circle, in a and b, estimated signals are presented together with the original signals (in gray) used for synthesis after the same windowing operation.

The amplitude spectra of the decomposition results are Fig. 35c and Fig. 35d. Also, the time-domain signals obtained by inverse FT of the two spectra are presented in Fig. 35a and Fig. 35b.

For comparison, all estimated signals are plotted together with the actual windowed excitation signal and windowed vocal tract impulse response used for synthesis. The original and estimated signals are very close and ZZT-decomposition is indeed capable of separating source and vocal tract signals to a high extent (not completely though, small variations due to vocal tract formants are observable both on the time domain glottal flow signal and its magnitude spectrum) and parameter estimation can be performed effectively on the resulting spectra and time domain signals.

Next, we present a similar example where both results obtained by the ZZT-decomposition and by the well-known PSIAIF [Alku, 1992 a] algorithm for source-tract decomposition are plotted together with the original excitation signal (Fig. 36). The PSIAIF algorithm is an iterative algorithm, which decomposes glottal flow and vocal tract components by LP analysis. The user sets fixed numbers of poles for the glottal flow and the vocal tract components. Then a sequence of inverse filtering analysis is applied to the speech signal to estimate each component.

Speech is synthesized with the following parameters: open quotient=0.7, asymmetry coefficient=0.65, vocal tract formant frequencies =[600 1200 2200 3200] Hz. We present the glottal flow signals and their amplitude and group delay spectra. The color codes are: PSIAIF (red), ZZT-decomposition (green) and the original (blue).

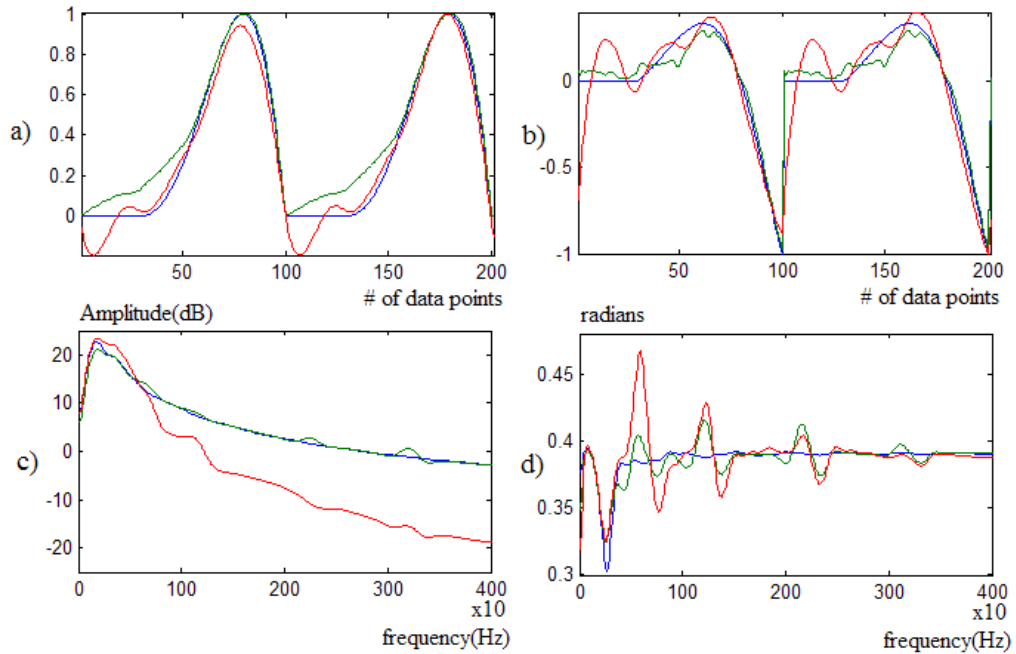


Fig. 36: ZZT-decomposition example on a synthetic speech frame (comparison of ZZT-decomposition and PSIAIF).

Three superimposed signals: original glottal flow used for synthesis (blue), glottal flow estimated by PSIAIF (red), glottal flow estimated by ZZT-decomposition (green). a) time domain glottal flow signal, b) time domain glottal flow derivative signal, c) magnitude spectrum of the glottal flow derivative, d) group delay of the glottal flow derivative signal.

It is hard to see from the time domain signals which method performs better but in the spectral domain the differences are clearer. The magnitude spectrum of the ZZT-decomposition estimate is quite better than that of PSIAIF. The same holds for the group delay function except the third and fourth formant peaks where errors in the ZZT-decomposition estimate is slightly worse. In this example, the ZZT-decomposition better estimates the glottal flow contribution compared to the PSIAIF method, which is known to be the reference method currently used in many studies. A reliable and complete comparison of the two methods would need that we perform tests on a large variety of real speech signals and this is not possible at the moment due to unavailability of reference data. Therefore we only presented a single example for visual comparison of the decomposition results. The disadvantages of ZZT-decomposition compared to PSIAIF are the need for GCI estimation and computational complexity.



## Real speech example

Here we present a real speech decomposition example in Fig. 37 and Fig. 38. The speech frame is taken from vowel /a/, from the word “party” (The file named BrianNormal2.wav in the Voqual 2003 [www-Voqual03] database). An example of vowel /a/ is presented since it is a rather an easy type of signal for visual inspection of formant locations. The displayed frequency range is 0-4000Hz for better viewing and for the reason that ZZT-decomposition is more robust in this frequency range (since the zero gap around the unit circle is prominent in this region). The actual magnitude spectrum of the windowed speech frame is presented in Fig. 37c and the amplitude spectra of glottal flow and vocal tract contributions obtained by ZZT-decomposition are presented in Fig. 38c and Fig. 38d.

The zero locations in Fig 37b show that indeed a zero gap exists around the unit circle for the windowed data and for this reason the corresponding magnitude spectrum (Fig. 37c) is smooth with apparent formant peaks. The ZZT-decomposition procedure results in separating the first peak as the glottal formant peak (Fig. 38c) and the rest of the formant peaks are included in the vocal tract contribution part (Fig. 38d). This fulfills our expectation for the decomposition of this signal since theoretical values of the formant frequencies for vowel /a/ ( $F1 \sim 600\text{Hz}$ ,  $F2 \sim 1200\text{Hz}$ ) are in agreement with the formant peaks observed in Fig. 38d. For sounds with low  $F1$  frequency (for example /i/), mid-low open quotient and high pitch, glottal formant ( $F_g$ ) and  $F1$  peak share the same frequency region making visual inspections very difficult. Our decomposition for such examples gives  $F_g$  and  $F1$  to be very close but due to lack of reference data it is difficult to check the reliability of the estimate. In the following sections, we present our tests based on the analysis of synthetic signals in which one part of the tests is especially set for cases where  $F_g$  is very close to or higher than  $F1$  frequency (high pitch female speech).

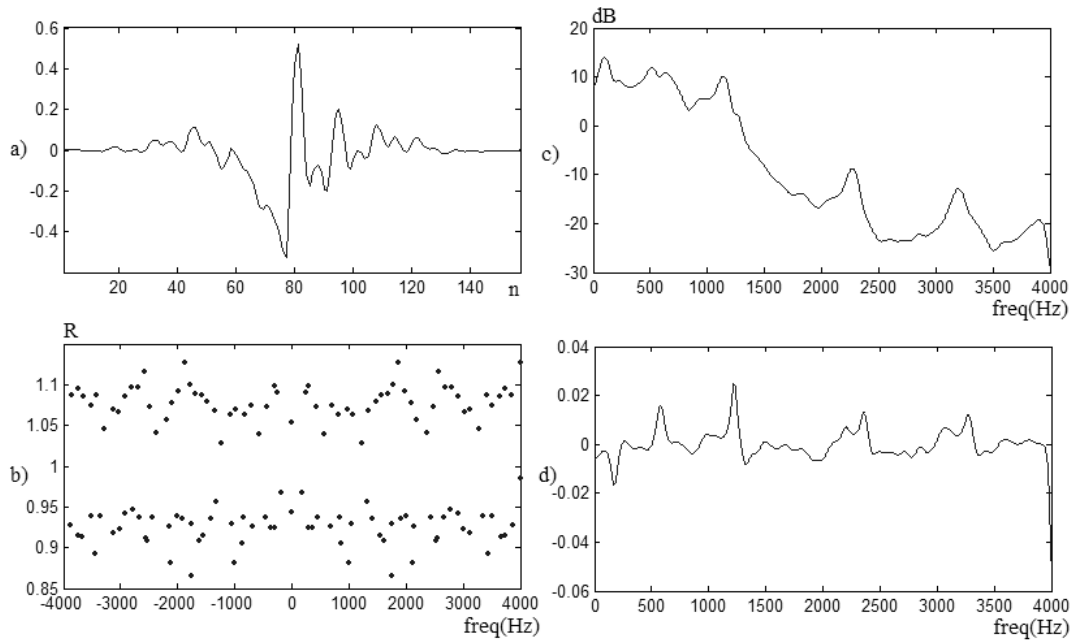


Fig. 37: ZZT of the real speech frame (/a/) windowed synchronously with GCI.  
a) time domain windowed signal, b) ZZT representation, c) magnitude spectrum, d) group delay

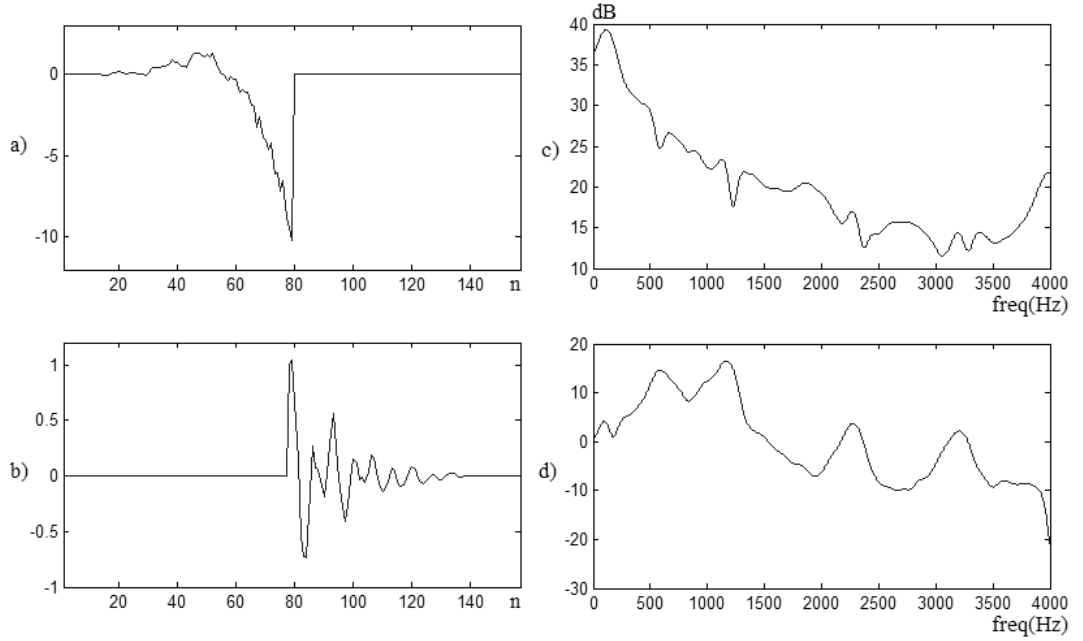


Fig. 38: ZT-decomposition result for a real speech frame.

a) time-domain source signal estimate, b) time domain vocal tract response estimate, c) magnitude spectrum of the source estimate, d) magnitude spectrum of the vocal tract estimate

## Robustness tests

Investigating the robustness of the decomposition method to GCI detection errors, to *FI* variations, to additive noise and to return phase variations is necessary.

### Robustness to GCI detection errors

Robustness tests for GCI detection errors is handled by introducing error in the GCI estimate sample by sample on a synthetic speech and a real speech example, and checking the glottal formant peak location in the decomposition result (the estimate for the source spectrum). In Fig. 39, we present the amplitude spectra of estimated glottal flow derivative signals when the error is systematically introduced in the GCI estimate for the synthetic signal (left column) and the real speech signal (right column).

The left column figures show that the decomposition results are very similar in the  $\pm 23\%$  error range and no longer reliable beyond these limits (glottal formant peak disappears and formant peaks appear in the estimated source spectrum). Our main concern is the location of the glottal formant peak and it stays almost unaltered in this range. Similar results are obtained for the real speech example (right column figures); however the error range tolerated is smaller:  $-17\%$  to  $+13\%$ . Though results may vary from signal to signal, we think that these tests show that GCI sensitivity is rather low: most of the state-of-the-art techniques for GCI estimation provide estimates with enough precision to perform ZT-decomposition reliably. This is of course an important result.

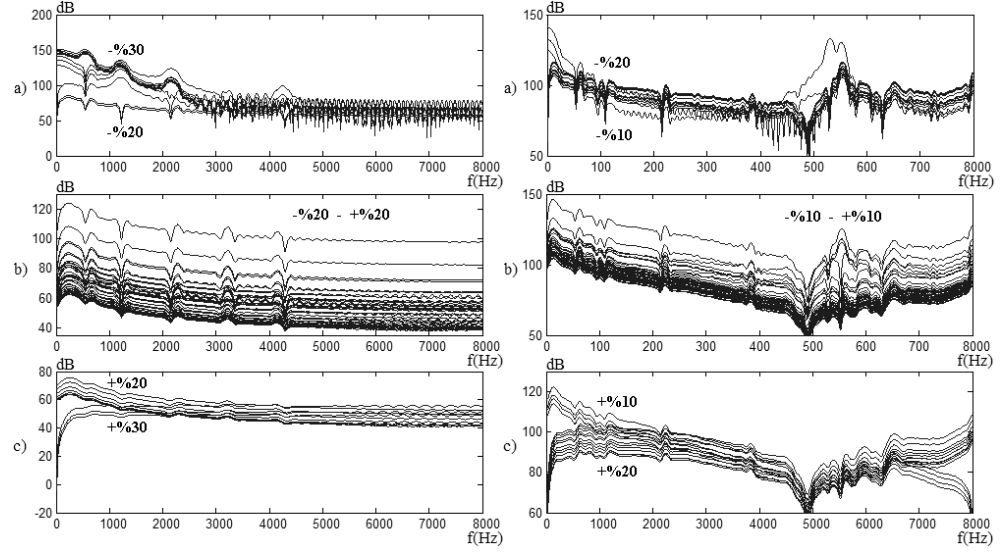


Fig. 39: Tests for robustness to GCI estimation errors.

Magnitude spectrum of estimated glottal flow derivative for synthetic speech signal(left column): a) error in GCI in range  $-\%30T_0$  to  $-\%20T_0$ , b) error in GCI in range  $-\%20T_0$  to  $+\%20T_0$ , c) error in GCI in range  $+\%20T_0$  to  $+\%30T_0$ .

Magnitude spectrum of estimated glottal flow derivative for real speech signal (right column): a) error in GCI in range  $-\%20T_0$  to  $-\%10T_0$ , b) error in GCI in range  $-\%10T_0$  to  $+\%10T_0$ , c) error in GCI in range  $+\%10T_0$  to  $+\%20T_0$ .

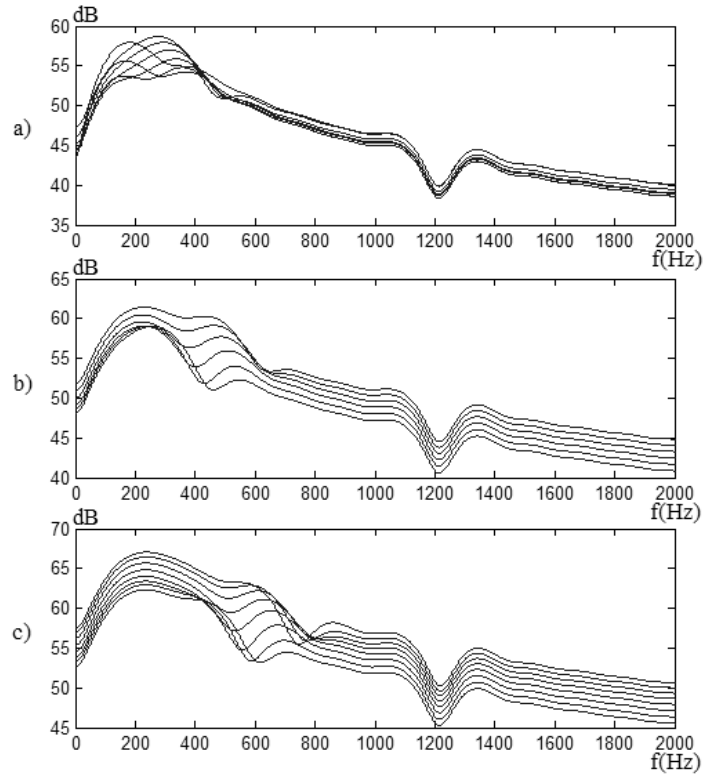


Fig. 40: Tests for robustness to  $F1$  variations.

Magnitude spectrum of estimated glottal flow derivative for synthetic speech signal: a)  $F1$  variation in range 200Hz to 350Hz, b)  $F1$  variation in range 375Hz to 500Hz, c)  $F1$  variation in range 525Hz to 700Hz.

### Robustness to $F1$ variations

As the next step, we studied effects of  $F1$  variations. Speech signals are synthesized by varying only the  $F1$  frequency and keeping all other parameters constant. In Fig. 40, the amplitude spectra of estimated glottal flow derivative signals are presented.

We observe in Fig. 40 that glottal formant peak is resolved reliably after  $F1$  exceeds 375Hz. This shows that ZZT-decomposition is not reliable when  $F1$  and glottal formant share the same frequency band. This is an important drawback. However this sensitivity exists for all source-tract decomposition methods and further improvement is generally necessary.

### Robustness to additive noise and return phase variations

Finally we studied robustness to additive noise and return phase variations on a synthetic speech signal (Fig. 41). For additive noise tests, random white noise is created and mixed with the synthetic speech signals. The signal-to-noise-ratio (SNR) is varied from  $\infty$  (no noise) to 25dB. Left column figures present the final noisy speech signal and estimation results for various mixing levels. Right column figures present the excitation signal with variation on the return phase and the corresponding estimation results.

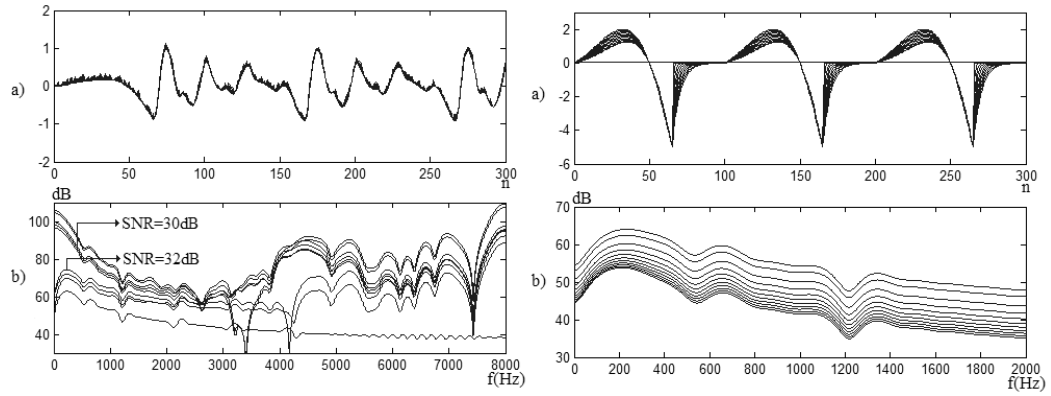


Fig. 41: Tests for robustness to additive noise and return phase variations.

Left column: Estimation results when mixing level of additive noise is varied (in range  $\infty$  (no noise) to 25dB): a) time domain speech signals with various levels of additive noise, b) amplitude spectra of estimated glottal flow derivative signals.

Right column: Estimation results when return phase coefficient is varied: a) time domain glottal flow derivative signals used for synthesis with various return phase coefficients, b) amplitude spectra of estimated glottal flow derivative signals.

In Fig. 41, we observe that robustness to noise is low: at an SNR level as high as 30dB, the glottal formant peak is no more available on the estimated spectrum. It seems that the most important problem introduced by noise addition is the variation introduced on the location of the zero on the real axis. Such variations cause, for some cases, the exclusion of the zero on the real axis due to glottal flow first phase from the set of glottal flow ZZT. A systematic study is difficult since we need to study changes in the root locations of high degree polynomials when some noise is added to the polynomials coefficients. This effect does not seem to be a linear function of SNR so we think there is no reason to study this systematically by varying SNR levels for all types of phonation and formant settings.

The second test presented in Fig. 41 is return phase coefficient variation. The robustness is high for these variations; the glottal formant peak location stays almost unaltered.

Summarizing the results: ZZT-decomposition suffers mainly from  $F1$  variations when  $F1$  is close to the glottal formant peak in the spectrum and is not robust to additive noise. ZZT-decomposition is robust to GCI errors in  $\pm 10\%$  range and variations in the return phase coefficient of the glottal flow component.

### V.1.3. Mixed-phase decomposition using complex cepstrum

ZZT decomposition corresponds to decomposing a mixed-phase signal into its maximum-phase and minimum-phase components (except the zero on the real axis, which affects the resulting spectra to a large extent). Theoretically, a similar operation can be performed using the complex cepstrum. Therefore for completeness, we discuss cepstral decomposition in this section. As we shall see, computational difficulties exist for complex cepstrum (due to phase unwrapping problems) and direct cepstral decomposition does not lead to source-tract decomposition as we show by examples. Developing a cepstrum based method needs further research.

#### Links between ZZT and complex cepstrum

A  $z$ -transform polynomial for a given signal  $x(n)$  can be expressed in the ZZT form as:

$$X(z) = Az^{-r} \prod_{k=1}^{Mi} (1 - a_k z^{-1}) \prod_{k=1}^{Mo} (1 - b_k z) \prod_{k=1}^{Mz} (1 - c_k z^{-1}) \quad (5.2)$$

$$X(z) = Az^{-r} X_{\min}(z) X_{\max}(z) X_{\text{zero}}(z) \quad (5.3)$$

where  $(1 - a_k z^{-1})$  correspond to zeros inside the unit circle,  $(1 - b_k z)$  correspond to zeros outside the unit circle and  $(1 - c_k z^{-1})$  correspond to zeros on the unit circle.  $X_{\min}(z)$  is the minimum-phase component,  $X_{\max}(z)$  is the maximum-phase component and  $X_{\text{zero}}(z)$  is the zero-phase component, and where  $A$  stands for the gain coefficient and  $z^{-r}$  is the time-shift factor.

Given a mixed-phase signal, one can decompose it into the minimum-phase, the maximum-phase components and the zero-phase components. The direct way to perform this operation is to compute the ZZT and classify them according to their distance to the origin of the  $z$ -plane. An alternative way, which is computationally more efficient, is to use the complex cepstrum for this operation (if no zero occurs on the unit circle).

The complex cepstrum,  $\hat{x}[n]$ , is defined as ([Quatieri, 2002])

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[X(\omega)] e^{j\omega n} d\omega \quad (5.4)$$

and the even component of the complex cepstrum, denoted as  $c[n]$ , is referred to as the real cepstrum.

$$c[n] = \left( \hat{x}[n] + \hat{x}[-n] \right) / 2 \quad (5.5)$$

which is also defined through magnitude spectrum only as:

$$\hat{c}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log|X(\omega)| e^{j\omega n} d\omega \quad (5.6)$$

By definition, the logarithm of the complex FT spectrum is;

$$\log[X(\omega)] = \log|X(\omega)| + j \arg\{X(\omega)\} \quad (5.7)$$

In order that  $\log[X(\omega)]$  be unique,  $\arg\{X(\omega)\}$ , the phase spectrum, should be unique. However there is an ambiguity in the definition of the phase: it can be defined as  $\arg\{X(\omega)\} = PV(\arg\{X(\omega)\}) + 2\pi k$  where  $PV$  stands for the principle value (in interval  $[-\pi, \pi]$ ) and  $k$  is any integer value. This ambiguity necessitates the unwrapping operation and appears as a computational difficulty in complex cepstrum computation.

If we discard the time shift component, the gain factor of the rational form of  $X(z)$ , and assume no roots occur on the unit circle, the complex logarithm is:

$$\log(X(z)) = \log\left(\prod_{k=1}^{M_i} (1 - a_k z^{-1}) \prod_{k=1}^{M_o} (1 - b_k z)\right) \quad (5.8)$$

and referring to the power series expansion formulas:

$$\log(1 - az^{-1}) = -\sum_{n=1}^{\infty} \frac{a^n}{n} z^{-n}, |az^{-1}| < 1 \quad (5.9)$$

$$\log(1 - bz) = -\sum_{n=1}^{\infty} \frac{b^n}{n} z^n, |bz| < 1 \quad (5.10)$$

where Eq. 5.9 corresponds to a right-sided sequence and Eq. 5.10 corresponds to a left-sided sequence according to z-transform properties. We can re-write the complex cepstrum as:

$$\hat{x}[n] = -\sum_{k=1}^{M_i} \frac{a_k^n}{n} u[n-1] + \sum_{k=1}^{M_o} \frac{b_k^{-n}}{n} u[-n+1] \quad (5.11)$$

where  $u[n]$  is the unit step function. This suggests the complex cepstrum is a two sided sequence which is combination of a right-sided sequence due to the zeros outside the unit circle and a left-sided sequence due to the zeros inside the unit circle [Quatieri, 2002]. Therefore a given mixed-phase signal can be decomposed into minimum-phase and maximum-phase components as defined in [Oppenheim *et al*, 1976]:

$$\hat{x}[n] = \hat{x}_{\min}[n] + \hat{x}_{\max}[n] \quad (5.12)$$

$$\hat{x}_{\min}[n] = \begin{cases} 0, n < 0 \\ \frac{1}{2} \hat{x}[0], n = 0 \\ \hat{x}[n], n > 0 \end{cases} \quad (5.13)$$

$$\hat{x}_{\max}[n] = \begin{cases} \hat{x}[n], n < 0 \\ \frac{1}{2} \hat{x}[0], n = 0 \\ 0, n > 0 \end{cases} \quad (5.14)$$

The question to raise at this point is: does this type of decomposition lead to the same result as the ZZT-decomposition? To answer this question, first we present an example of decomposition results for a GCI-synchronously windowed synthetic speech frame using ZZT and complex cepstrum.

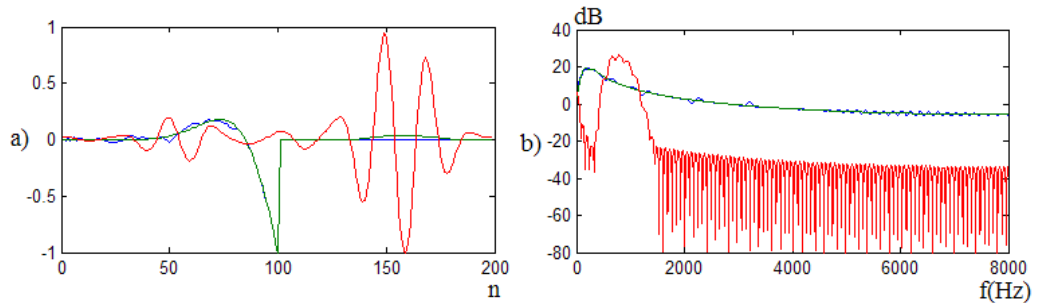


Fig. 42: Comparison of complex cepstrum and ZZT decomposition results  
Red: complex cepstrum decomposition result, blue: ZZT decomposition result, green: original glottal flow derivative signal

Fig. 42 shows that complex cepstrum decomposition into maximum-phase and minimum-phase components does not directly lead to source-tract decomposition in the example the ZZT

decomposition provides high quality decomposition. It seems that the main difficulty is in the phase unwrapping operation in calculation of the complex cepstrum.

In the contrary, in the homomorphic decomposition literature we could find two figures where successful source decomposition is demonstrated: Fig.5 in [Oppenheim & Schaffer, 1968] and Fig. 6.18 in [Quatieri, 2002]. However, referring to the following statements in [Oppenheim & Schaffer, 1968]: “Since no clear statements can be made about the relative importance of the maximum and minimum-phase components of the glottal pulse, the notion of recovering the maximum-phase component has no obvious implications. However, if we retain values in the complex cepstrum for positive and negative values of  $n$ , then combined vocal tract and glottal pulse information can be recovered with appropriate phase relations”, we think that the authors could not get reliable decomposition in all examples and that lead to some confusion. In [Quatieri, 2002], the figure caption stresses: “...for this particular example...” which again makes us think that the provided good result cannot be generalized. We have conducted a detailed search in literature and could find no method proposed using this potential of complex cepstrum for source-tract decomposition. The fact that there exist a few demonstrations but no methods suggests that again some factors like windowing plays an important role (since phase unwrapping seems to be one of the main problems in cepstral decomposition) and researchers could not observe reliable outputs in their trials to use this property. Further research is needed to check if similar results can be obtained using the complex cepstrum.

---

#### V.1.4. Conclusions

A ZZT-decomposition method based on classifying roots of the  $z$ -transform of windowed speech data was presented. The decomposition is of high quality though not complete. The contribution of the vocal tract in the glottal-flow-dominated spectrum is observed as ripples of low amplitude, while the contribution of glottal flow in the vocal tract dominated spectrum is hardly observed. The proposed algorithm is very easy to implement but computationally heavy due to the need of finding roots of high degree polynomials. For this reason, it is more appropriate for off-line database processing.

In the robustness tests, we have observed that two factors have important effects to efficiency: additive noise and low  $FI$  values when  $Fg$  is high. Further research is necessary to improve robustness of the method for these variations.

---

## V.2. Application to glottal flow parameter estimation

Apparently, source and vocal tract parameter estimation can be performed on the resulting two components of the ZZT-decomposition.  $Fg$  (glottal formant frequency) is one such parameter we are mainly interested in. Since time-domain glottal flow parameter estimation methods are sensitive to noise, we find spectral parameter estimation methods to be more robust. Tracking the maximum valued peak location of the magnitude spectrum of zeros outside the unit circle, we can easily get an estimate for  $Fg$ .

The mathematical expression for  $Fg$  is derived from glottal flow models in [Doval *et al*, 2003] as:

$$F_g = \frac{f_g(\alpha_m)}{O_q T_0} \quad (5.15)$$

where  $\alpha_m$  is the asymmetry coefficient,  $O_q$  is the open quotient,  $T_0$  is the pitch period (these parameters are indicated on Fig. 2) and  $f_g$  refers to the numerator which is only a function of  $\alpha_m$ . The authors mention (and as presented in Fig. 9) that  $\alpha_m$  mainly controls the quality factor of the glottal formant and does not play an important role on the  $Fg$  value. This suggests that given the pitch period,  $Fg$  estimate can be useful to detect open quotient variations in speech, which is considered to be one of the important dimensions of voice quality variations.

### V.2.1. Testing the $Fg$ estimation algorithm

#### Tests with synthetic speech

We first tested our algorithm with synthetic speech signals. As test signals, two periodic excitation (glottal flow derivative) signals are synthesized with the LF model [Fant, 1995] at constant pitch frequencies: 100Hz and 200Hz. All the parameters of the glottal flow have been kept constant except the open quotient, which is varied linearly in the range (0.3-0.98). No spectral tilt component is included for simplicity ( $Ta$ , the return phase decaying exponential time coefficient is set to zero) and the asymmetry coefficient is set to 0.7. These two excitation signals are then passed through three second order resonant filters for the vowel formants; /a/ ( $F1=600\text{Hz}$ ,  $F2=1200\text{Hz}$ ,  $F3=2200\text{Hz}$ ,  $F4=3200\text{Hz}$ ), /l/ ( $F1=300\text{Hz}$ ,  $F2=1800\text{Hz}$ ,  $F3=2200\text{Hz}$ ,  $F4=3200\text{Hz}$ ) and /u/ ( $F1=300\text{Hz}$ ,  $F2=800\text{Hz}$ ,  $F3=2200\text{Hz}$ ,  $F4=3200\text{Hz}$ ), thereby obtaining six synthetic speech signals. Then  $Fg$  is estimated using the proposed algorithm on all the signals (including pure excitation signals for reference). Since  $Fg$  is independent from formant variations, we expect to obtain the same  $Fg$  estimates for all synthetic vowels created and the glottal flow signal itself (as reference). The results are presented in Fig 43.



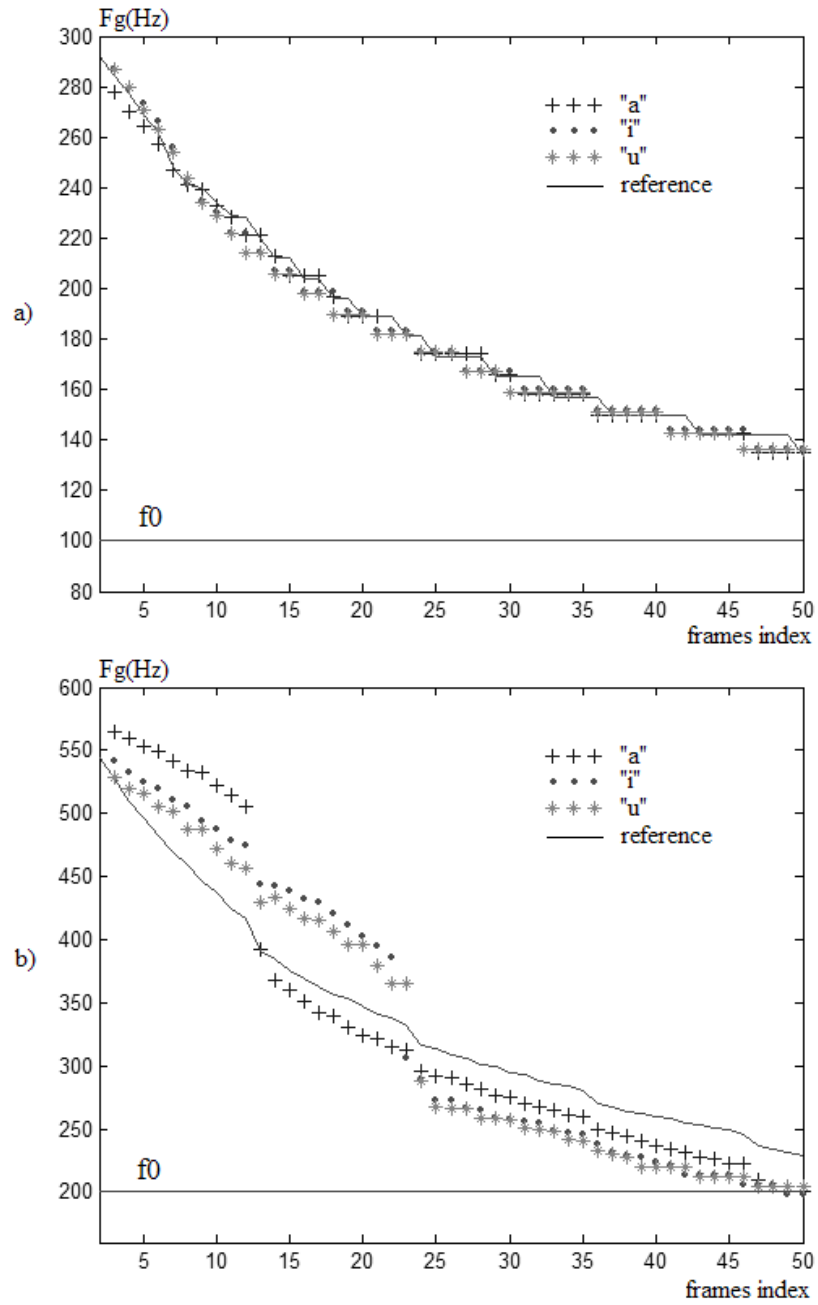


Fig. 43:  $F_g$  estimation results for two excitation signals.  
a)  $f_0=100\text{Hz}$ , b)  $f_0=200\text{Hz}$

As expected, the  $F_g$  estimate plots have the basic form  $y=1/x$  since open quotient is linearly varied and  $F_g$  is inversely proportional to open quotient. The robustness of the estimation depends on the relative location of glottal formant to the first formant of the vocal tract ( $F1$ ). For the frames where  $F_g$  is lower than 300Hz, the estimates for the vowels compared to the reference estimate are very close. For higher  $F_g$  values, the maximum peak location of the magnitude spectrum corresponding to zeros outside the unit circle is more affected by  $F1$ , and peak picking is sensitive to this effect. The ZZT-decomposition result for the worst estimation in the test is shown in Fig 44.

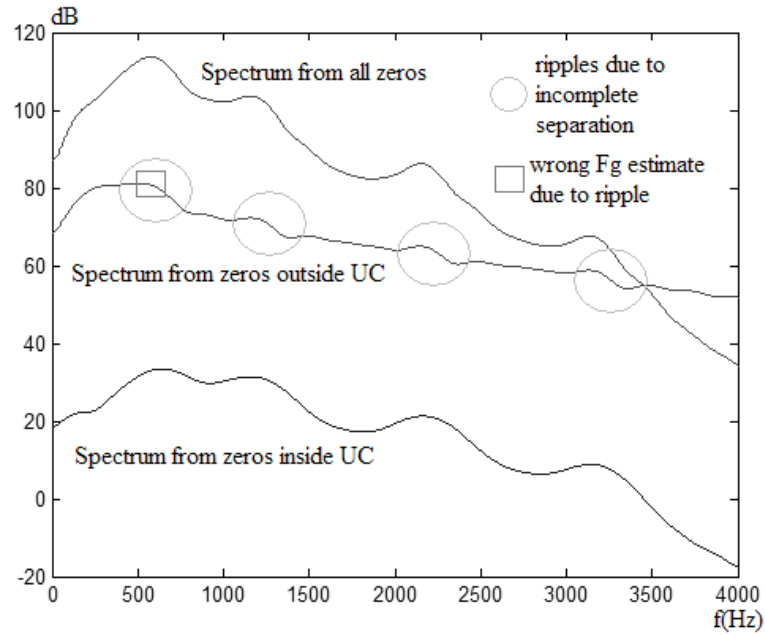


Fig. 44: Magnitude spectrums of ZZT-decomposition for the worst glottal formant frequency estimation with peak picking.  
12th frame from the  $f_0=200\text{Hz}$  test where the vowel is “a”, Fig. 42b

The ripples due to incomplete separation of vocal tract from glottal flow are marked with circles on the glottal flow dominated spectrum. The ripple due to  $F_1$  causes the maximum valued point of the spectrum to appear at a higher frequency (marked with a rectangle) introducing an error in the  $F_g$  estimate. To improve robustness of our algorithm, we plan to apply a curve fitting method instead of peak picking in our future studies.

### Tests with real speech

The  $F_g$  estimation method has also been tested on a real speech signal for which we could obtain a reliable open quotient estimate. A sustained vowel /a/ with flat pitch and decreasing open quotient has been uttered and EGG signals were recorded in parallel. Using the method described in [Henrich *et al.*, 2000], open quotient (Oq) estimate was obtained<sup>5</sup>. As previously mentioned and also demonstrated in Section III.3.2, the effect of asymmetry coefficient variation to  $F_g$  variation is rather minor. For this reason, for our speech example with almost constant pitch, we expect the  $F_g$  estimate to be highly correlated with the inverse of the open quotient estimate (Eq. 5.15). In Fig. 45, we present the  $F_g$  estimate plotted together with  $f_0$  and with the inverse of the open quotient estimate scaled by a constant.

<sup>5</sup> Thanks to Nathalie Henrich for providing the speech data and the EGG based open quotient estimates.

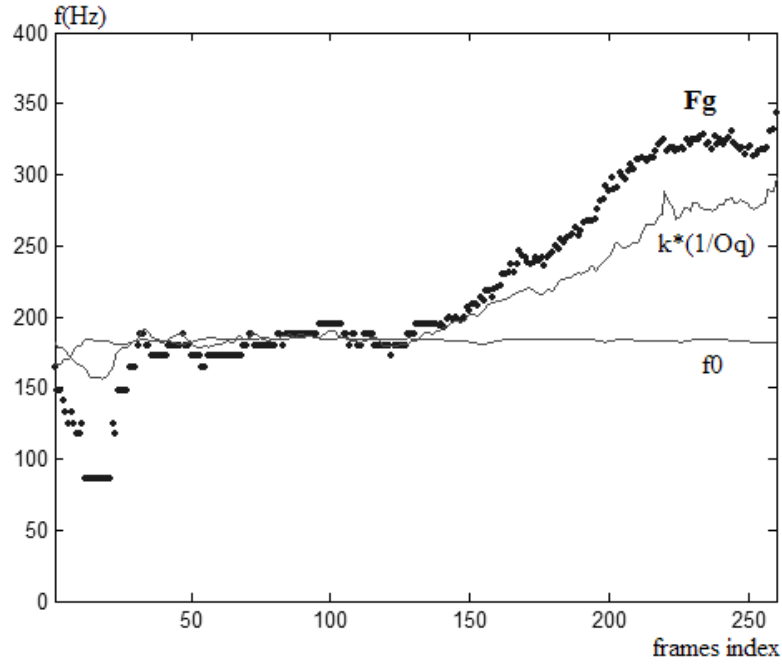


Fig. 45: Comparing glottal formant frequency estimate with inverse of open quotient estimate and  $f_0$  estimate ( $k=115$ )

The glottal formant frequency estimate and inverse of open quotient estimate plotted in Fig. 45 have high correlation which indicates that our algorithm is effective not only on synthetic signals but can track glottal flow variations of real speech signals.

### V.2.2. Conclusions

We have presented a method for estimating the glottal formant frequency. It is mainly composed of picking the maximum valued peak on the magnitude spectrum of the glottal flow obtained by ZZT-decomposition from GCI synchronously windowed speech signals.

The proposed algorithm was tested on synthetic speech and the results show that the ripples on glottal flow spectrum due to incomplete separation introduces errors when  $F_g$  and  $F_1$  values are close. It is also the sensitivity of peak picking to small ripples which causes the wrong estimates. For our future studies we target improving the quality of the  $F_g$  estimation method by replacing peak picking with a more robust method.

Due to lack of reference methods and data, tests on real speech were limited. On a single real speech example, we have shown that open quotient variations can be tracked with the  $F_g$  estimation algorithm. Robustness tests on larger real speech databases are necessary.

---

## V.3. Application to formant tracking

One of the potential applications of chirp group delay processing is formant tracking. During this thesis study, three versions of a formant tracker have been developed (the first two being presented in [Bozkurt & Dutoit, 2003, Bozkurt *et al*, 2004a]). All the versions are based on peak picking on the chirp group delay computed outside the unit circle. At each update of the formant tracker, quality and robustness are improved by solving a problem. Below we present all of the three versions.

### V.3.1. Formant tracker – first version

The first method is simply based on the fact that most of the ZZT of a speech frame is located around the unit circle and the spikes on the group delay due to these zeros can be avoided by simply calculating the chirp group delay outside the unit circle. Once the radius of the analysis circle is appropriately set, peaks due to formants of speech signals can be observed on chirp group delay functions even with better resolution than on the corresponding magnitude spectrum. This first version of the formant tracker was actually developed before ZZT of windowed speech was studied. The method presented was considered to be theoretically correct but the tests have shown that it lacks robustness for some unknown reason.

The source of problem became apparent after the study of ZZT of windowed speech. We conducted experiments comparing the chirp group spectrum and ZZT of signals. We observed that zeros can appear nearly everywhere on the  $z$ -plane though most of them are closely located to the unit circle and a single zero close to the analysis circle where chirp group delay is computed introduces important errors in the formant estimation result (this issue is discussed in detail in Section IV.6 with example figures). Therefore, the first version of the formant estimation algorithm was very sensitive to such unexpected zeros close to the analysis circle and we had to find a means to guarantee certain distance between zeros and the analysis circle. A second version of the formant tracker was developed to avoid this problem.

### V.3.2. Formant tracker – second version (DPPT)

The second method we have developed [Bozkurt *et al*, 2004c], differential phase peak tracking (DPPT<sup>6</sup>), guarantees some distance between the analysis circle and the zeros by removing some of the zeros from the set of ZZT. Actually an optimum way to achieve this correctly (without destroying the vocal tract information available in ZZT) is to apply the ZZT-decomposition on GCI synchronous windowed data as apriori step. Then the chirp group delay is computed outside the unit circle from zeros inside the unit circle. This refers to the CGDGCI representation defined in section IV.6. The flow chart of the method is presented in Fig. 46.

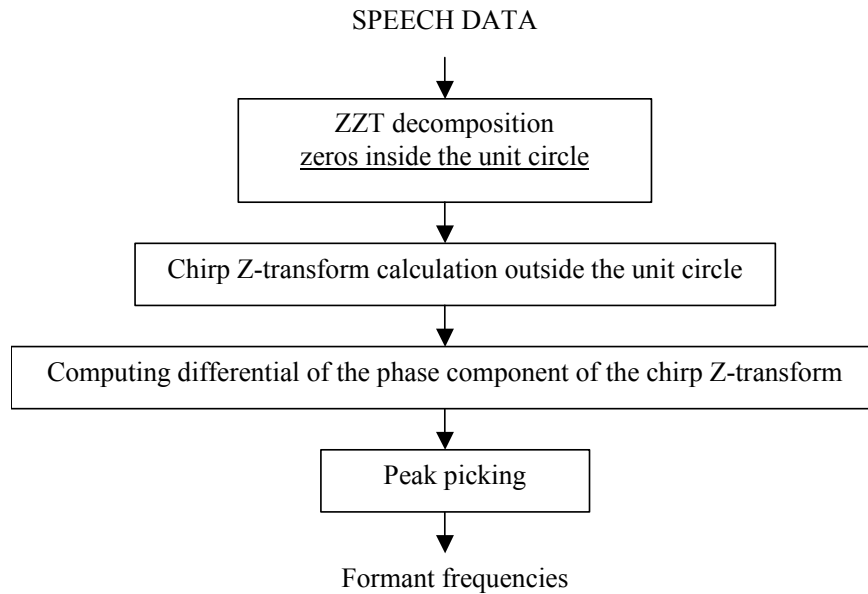


Fig. 46: The Differential-Phase Peak Track (DPPT) algorithm

---

<sup>6</sup> Initially the term “differential phase spectrum” was used instead of chirp group delay. Therefore the formant tracker was named as DPPT.

Once the  $z$ -transform is calculated on a circle outside the unit circle with radius  $R=1+\delta R$ , the closest zero is at least  $\delta R$  away from the analysis circle since all zeros are inside the unit circle. As  $\delta R$  increases, the differential phase spectrum gets smoother. Defining an optimum value is a compromise between smoothness and high resolution and  $\delta R=0.06$  is empirically found to be an appropriate value. The formant tracking algorithm simply picks the peaks on the chirp group delay function.

We have conducted tests for comparing the efficiency of the second version of the formant tracker to that of two publicly available formant trackers: that of Praat [www-Praat] and WinSnoori [www-WinSnoori]. Both synthetic speech and real speech examples were used for testing. The results of the tests are reported below.

## Tests

### Stimuli

For the tests with synthetic speech, a single synthetic speech chunk with pitch frequency and formant frequency variations so as to uniformly sample the  $f_0$ - $F1$ - $F2$ - $F3$ - $F4$  parameter space (see Fig. 47) was designed. Speech was synthesized by all-pole filtering a periodic excitation signal. The excitation signal was created by using the LF model [Fant, 1995] with fixed open quotient (0.65) and asymmetry coefficient (0.7) and  $f_0$  was varied from 200 Hz to 100 Hz by a sinusoidal function.

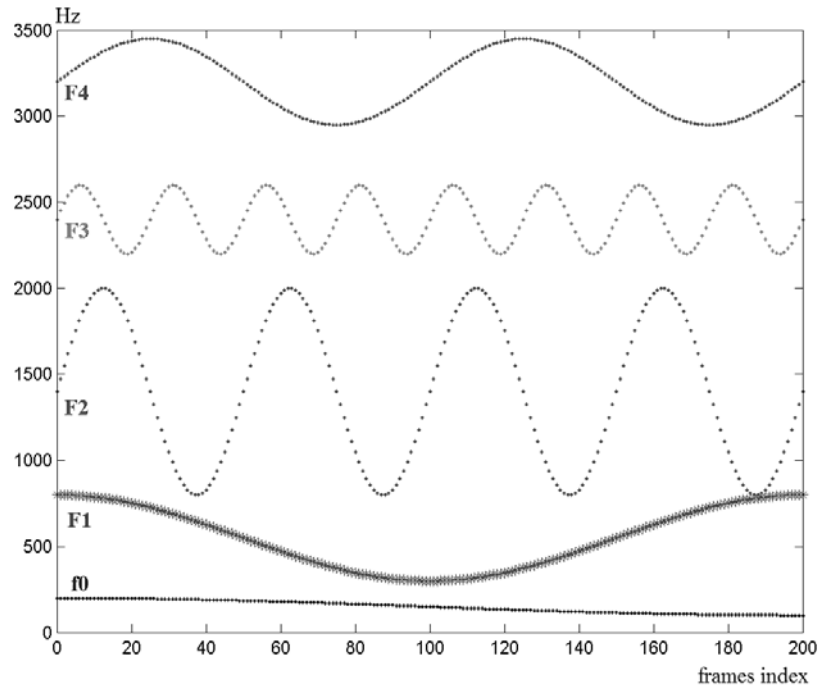


Fig. 47: Parameter space of the synthetic speech stimuli

For the tests with real speech, two female and two male examples with large formant movements were used.

### Results

The formant tracks obtained for the synthetic speech signal by DPPT is presented in Fig. 48 together with the actual formant tracks used in synthesis. In addition, average percentage error rates and formant missing rate (percentage of frames where no estimate for that particular formant is available) for the outputs of the three systems are provided in Table 1. Average percentage error rates are calculated only for the formants, which are not missed (i.e. a formant missed does not contribute to the error rate by 100% but is simply not included in calculations). Plots for outputs

from two other formant trackers are available on the CD-ROM: “\formantTrackingTests\formantEstimates\plots\synth” directory.

**Table 1:** Formant tracking error rates for DPPT, Winsnoori and Praat

	Average percentage error				Formant missing rate			
	F1	F2	F3	F4	F1	F2	F3	F4
DPPT	6.8	1.8	1	0.8	0	17.1	3.5	0
Win-Snoori	2.8	1.9	0.6	-	0	0	0	-
Praat	3.8	3.8	4.7	13.8	0	0	0	24.4

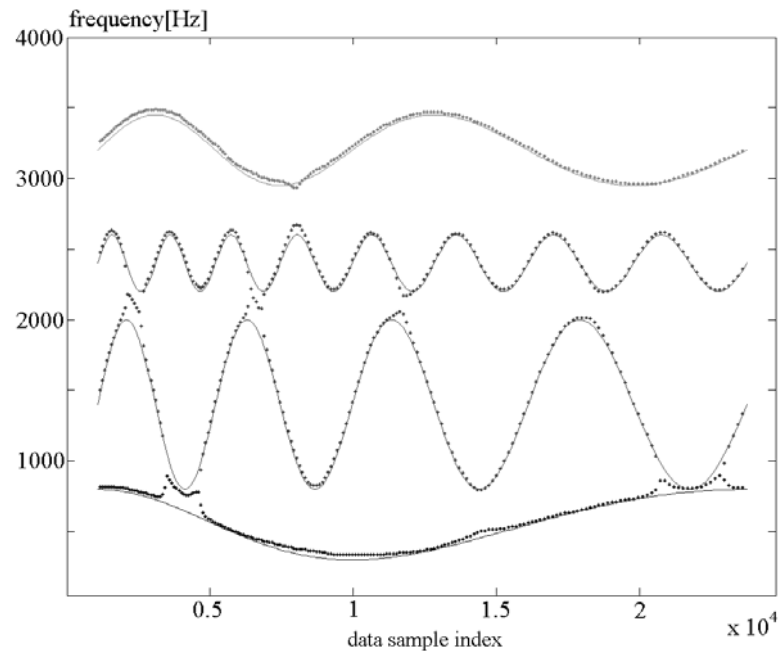


Fig. 48: DPPT formant tracks (dots) and formant synthesis parameters (solid lines)

For demonstrating the results of tests with real speech, we provide only plots but no error rates due to unavailability of reference data for formant frequencies of real speech chunks. Here, we provide the DPPT peaks picked for one of the real speech examples (Fig. 49), for which the formant tracks on spectrogram are obvious. All plots obtained from three formant trackers on four real speech examples are presented in the CD-ROM.

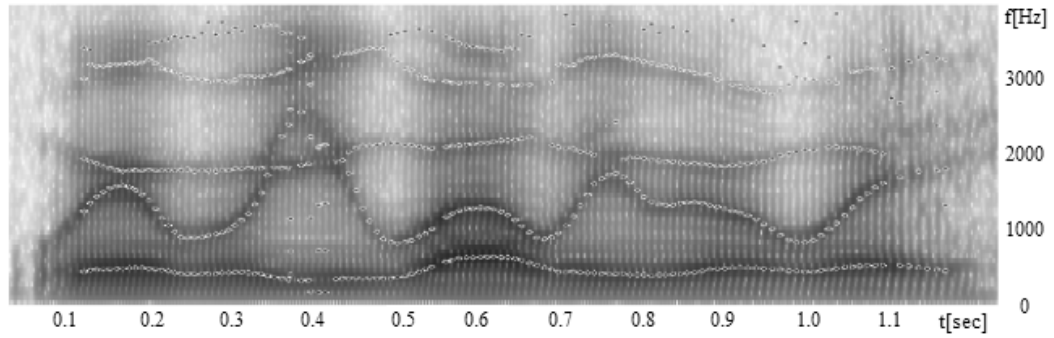


Fig. 49: Differential phase spectrum peaks indicated on the spectrogram of a speech signal. For the phrase: "where were you while we were away?" from The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc CD1-1.1, DR1/msjs1/sx9.wav

### Discussion

It is interesting to observe that the robustness of methods for tracking formants on real speech and synthetic speech are quite different. In tests with synthetic speech, Winsnoori results are best for the first three formants. But the fourth formant track is not available so we cannot compare the results for  $F4$ . DPPT's quality is very close to Winsnoori's except in frames where formants get very close to each other (can be easily observed on Fig. 48) and it provides the  $F4$  track with high precision. For frames where formant frequencies are very close, DPPT tracks a single peak instead of two peaks, which causes a high formant miss rate for  $F2$ . This is one of the important drawbacks of the DPPT algorithm. Praat's robustness on analysis of synthetic speech is lowest except for the  $F1$  track.

However, the robustness of the three methods for analyzing real speech are quite different than that for synthetic speech. For three of the four real speech examples, DPPT is either the best or among the best two. But it gives the worst results among the three techniques for the fourth example. Praat's quality seems to be more constant than the other two methods when all four examples are considered. Winsnoori is effective mainly for tracking  $F1$ . It has moderate quality in tracking  $F2$  and fails to provide reliable  $F3$  estimates for most of the frames (and cannot provide  $F4$  track).

The results show that DPPT has similar quality as the two state-of-the-art methods and is an effective method for formant tracking. Its main advantage is in tracking high order formants and the main reason for this is the spectral tilt-free property of the differential phase spectra. The main drawbacks are: need for GCI marking and high computational load due to need for calculation of roots of high degree polynomials. We developed the third method to get rid of these difficulties.

### V.3.3. Formant tracker – third version (Fast-DPPT)

The third method guarantees the absence of zeros outside the unit circle by computing a zero-phase version of the signal. A zero-phase version of a signal can be computed by applying inverse FT to the magnitude spectrum of the signal. Again the chirp group delay is computed outside the unit circle (named the CGDZP in section IV.6) and peak picking is performed.

The third version of the formant tracking algorithm was developed (presented in Fig. 50) including one additional step of optimizing the radius ( $R$ ) of the analysis circle by an iterative procedure. As discussed before, the smoothness and resolution of peaks of the chirp group delay varies with the radius of the analysis circle. Given a fixed number of formants to track (usually five), the iterative procedure optimizes the  $R$  value by decrementing/incrementing it with steps of 0.01: if the number of peaks picked is higher than the number of formants to be tracked,  $R$  is incremented and vice versa. For tracking of five formants on a 30 msec frame-size analysis, the optimum value found for most of the examples is  $R=1.12$ , so this value is taken as the optimum initial value for the iterative procedure.

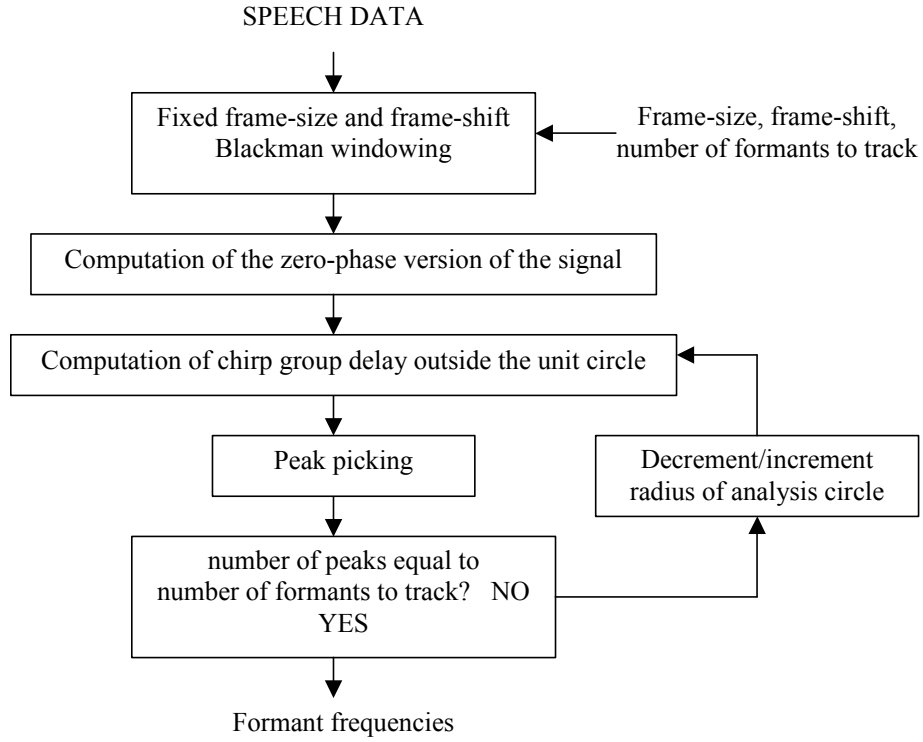


Fig. 50: The final version of the formant tracking algorithm using peak picking on chirp group delay computed outside the unit circle.

In Fig. 51, we present the histogram of results for iteration of  $R$  on 2865 speech frames (1453 male, 1412 female). The initial value is set to  $R=1.12$  and the limits of the iteration are  $[1.05 \ 1.25]$  where the number of formants is fixed to five. This figure shows that  $R=1.12$  is an appropriate value for tracking five formant peaks. To reduce computational time of the algorithm one can remove the iteration block and set  $R=1.12$  for frame size of 30 msec. with 16000Hz sampling rate.

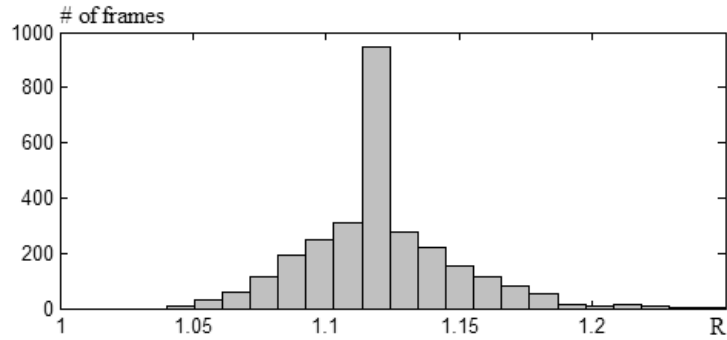


Fig. 51: Histogram of  $R$  values obtained by the iterative procedure.

The main advantages of this last version of formant tracker to the previous version are: there is no need for GCI instant synchronous analysis; there is no need for finding roots of a large order polynomial (therefore it is named as Fast-DPPT) and it is more robust.

## Tests

### *Procedure and Stimuli*

The final version of the formant tracker is compared to formant tracker of Praat and the formant tracker of Wavesurfer [www-WaveSurfer]. In this test Wavesurfer was preferred to WinSnoori



because WinSnoori had the lowest quality in the previous tests. A preliminary test was conducted with few real speech examples to compare WaveSurfer and WinSnoori, and WaveSurfer was more effective. So, the final tests include a comparison between our final algorithm and two best formant trackers we could access from the internet. In addition, only real speech examples were used in this second test (but the set of examples are enlarged) since tests on synthetic speech was not found to be very useful in terms of measuring the robustness of the formant trackers in the first tests. The first tests have shown that the robustness of the methods varies a lot from synthetic speech to real speech examples.

The set of real speech examples were thus enlarged in the second test: it included five female and five male examples with large formant movements. The female example set contains: one sentence in Japanese (from [www-Voqual03]), one sentence in French (from [www-Voqual03]), one sentence in English (from [www-DarpaDBA]) and two sentences in Danish (from the DES database [Hansen et al, 1996]). The male example set contains: three sentences in English (from [www-DarpaDBA]) and two sentences in Danish (from the DES database [Hansen et al, 1996]). The files are available in the CD-ROM, in the “\wave\” directory.

## Results

In Fig. 52, we present one of the figures for formant tracking results. The rest of the results, formant tracks of nine other real speech examples using three formant trackers, are presented in Appendix C (all figures are also available in the CD-ROM: ‘formantTrackTesting\formantEstimates\plots’ directory).

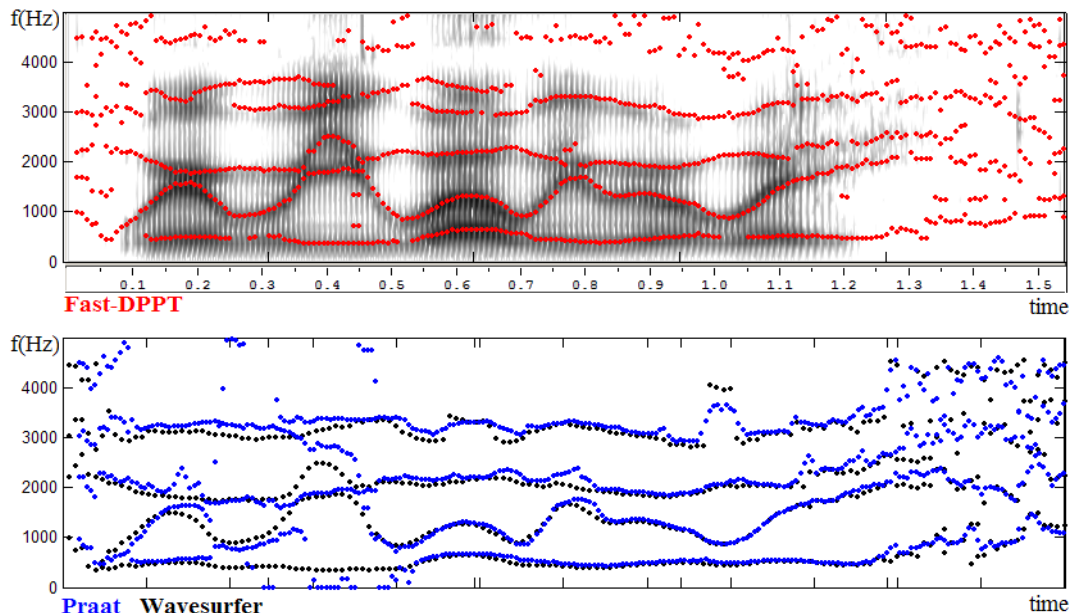


Fig. 52: Formant tracking example 1  
Male speech example, 1SX9.wav from [www-DarpaDBA]: “where were you while we were away”

The results of the three systems are comparable; they all provide high quality formant tracks. For some examples, the formant trackers of Praat and WaveSurfer provide more continuous plots since a smoothest path finding algorithm is included.

The fact that high quality formant tracking can be performed on chirp group delay with a simple algorithm, as presented here, shows that chirp group delay is an effective spectral representation and can potentially be used in various other applications.

---

## V.4. A Linear Prediction (LP) algorithm to estimate the glottal flow component from speech signals<sup>7</sup>

In Section IV.4, the mixed-phase characteristic of speech was explored. In this section, we define an algorithm with low complexity, which performs a new use of the linear prediction analysis (covariance method [Makhoul, 1975]) to retrieve the maximum-phase component of speech signals. When an all-pole model is studied for such a mixed-phase signal, and when the covariance method is used, some of the poles fall outside the unit circle. In speech processing, poles outside the unit circle are most of the time avoided/reflected due to the minimum-phase assumption. In this study, we follow the inverse path: we try to find outside poles for estimating glottal flow characteristics.

In the literature we could find two studies, which try to estimate poles outside the unit circle for such a purpose. In [Jackson, 1989], Jackson obtains maximum-phase and minimum-phase components of the speech signals by complex cepstral decomposition and analyses these components by LP. However, we have shown/discussed in section V.1.3 that complex cepstrum decomposition is not reliable due to necessity of phase unwrapping. Therefore, this method is likely to suffer from phase unwrapping problems (as a matter of fact no testing was presented for the proposed algorithm). Gardner and Rao use linear and non-linear least squares estimation methods to estimate the parameters of a non-causal model [Gardner & Rao, 1997]. In this approach, harmonic model parameters (magnitude and phase) are estimated and recursive algorithms are proposed to minimize error functions.

We propose to use LP-covariance analysis to estimate a pole pair outside the unit circle corresponding to the anti-causal poles of the source signal component in the mixed-phase speech model. Given the pair of anti-causal poles, a procedure to resynthesize the anti-causal part of the glottal flow, and then an open quotient estimation method, are proposed. Evaluations show that the method is high quality for analyzing synthetic speech but lacks robustness in analysis of natural speech.

### V.4.1. The MixLP algorithm

The proposed Mixed-Phase Linear Prediction (MixLP) algorithm, for detecting a pole pair outside the unit circle corresponding to the contribution of the maximum-phase glottal flow signal, is presented in Fig. 53a. First, a glottal closure instant (GCI) synchronous windowing is applied to the speech signal and a single pitch period length signal in-between two consecutive GCI marks is extracted. The resulting speech frame is integrated, to remove the lip radiation contribution. LP-covariance analysis [Makhoul, 1975] is applied to this signal, which is expected to result in a pole pair outside the unit circle (due to the glottal flow contribution as presented in the causal/anti-causal model by Doval *et al* [Doval *et al*, 2003]) and several other pole pairs inside the unit circle. This is a particular property of the LP covariance analysis, usually considered as a sign of instability of the estimation algorithm [Makhoul, 1977].

---

<sup>7</sup> This part of the study was mostly implemented and tested by Francois Severin within the STOP project of TCTS Lab.

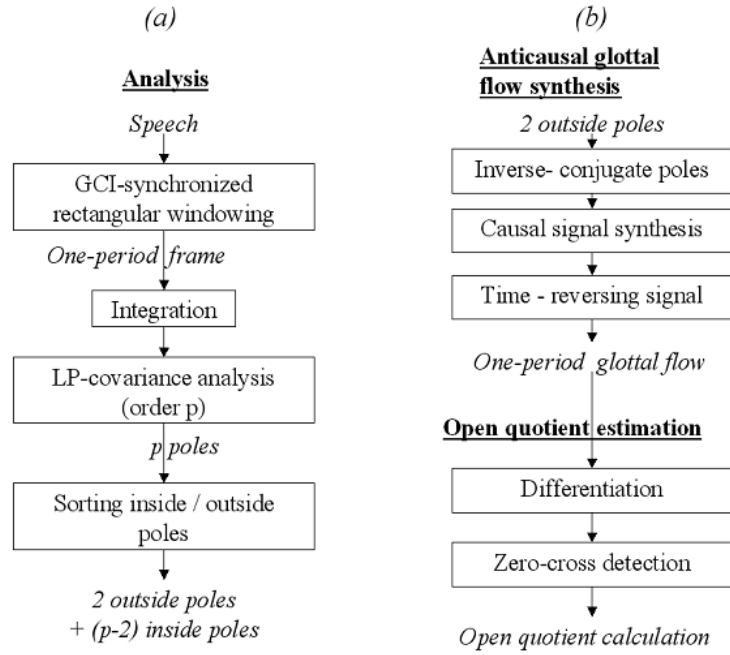


Fig. 53: MixLP algorithm flow diagram

It is interesting to mention some of the investigations performed during the design of the algorithm. Our first investigation was to find optimum windowing since the existence of poles outside the unit circle heavily depends on the applied windowing. Although we have shown through ZZT representation that mixed-phase characteristics can be observed on the Fourier transform of the windowed speech signal when the window is centered at GCI, such windowing is not appropriate for estimation of poles outside the unit circle with LP-covariance. By applying sliding window analysis and checking the correctness of estimates on synthetic speech signals, we have observed that the end of the window must be synchronized with the GCI (a few samples before the GCI is a good choice for safety) and including even a few data samples after the GCI results in no poles outside the unit circle most of the time.

A second investigation was on the order of the LP analysis, tested in the range [2-32] for 16000Hz synthetic speech signals (for which the LF model was used to synthesize glottal flow excitation and filtered by a four pole-pair all-pole vocal tract filter). The LP degree that provided best estimates is 14 or higher.

## Tests

In order to test the MixLP algorithm we have designed a method to estimate the open quotient (Fig. 53b) from poles outside the unit circle. This includes the re-synthesis of the glottal flow from the poles, which is achieved by: synthesis of a causal signal by computing the impulse response of a two-pole filter with the inverse-conjugate poles, and time reversal of this signal. A differentiation provides the differentiated glottal flow. In Fig. 54 we present an example of a glottal flow estimate using the MixLP method, together with a glottal flow estimate using the inverse filtering algorithm PSIAIF [Alku, 1992].

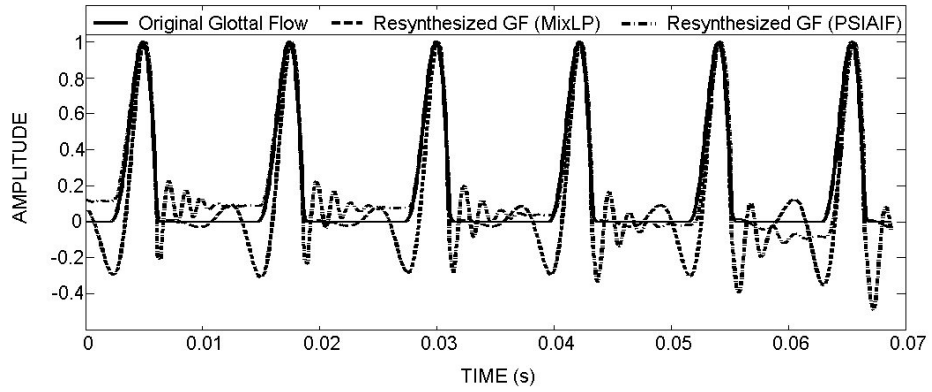


Fig. 54: Re-synthesized glottal flow signals obtained with the MixLP and PSIAIF algorithms

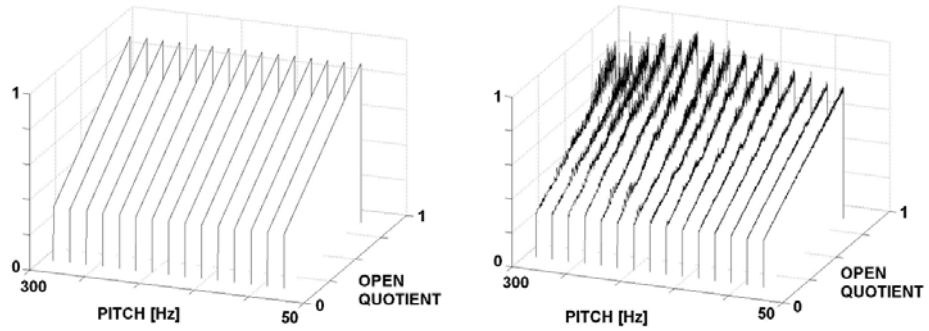


Fig. 55: MixLP open quotient estimation results

Left figure: stimuli, right figure: estimated open quotient

Test signals are sustained vowels with constant first formant and return phase for several values of the pitch period and open quotient is varied linearly

For evaluation of the open quotient estimation method, tests were conducted on synthetic speech signals, in which several parameters (pitch, spectral tilt, first formant frequency and open quotient) were varied systematically and higher formant frequencies are kept constant. Here we only present the output of our test for checking the robustness of estimation to pitch variations in Fig. 55. Some conclusions are: the error is small when the open quotient is higher than 0.7, and otherwise it is negligible. Moreover, the open quotient is better estimated if the return phase is short, and especially if the pitch is high. This open quotient estimation method was also compared to a well-known algorithm ([Henrich *et al*, 1999]). Both methods provide similar results but the MixLP estimation method is more effective when the first formant frequency is small. A comparison with [Gardner & Rao, 1997] would be useful, however the complexity of their algorithm is very high and considering the time limitation of the thesis study, it was not feasible.

The open quotient estimation on natural speech was also tested. As a reference for the open quotient estimation tests, we used open quotient estimates obtained from differential electro-glottograph (EGG) signals by using a threshold method. Observations on a few natural utterances showed that the MixLP estimation method is not robust as the estimation error depends on the phonetic context. The development of the MixLP algorithm is stopped at this stage due to time constraints. It is not obvious to us at this point why robustness is very low for real speech analysis and to further study the problem, considerable amount of research time is needed.

### Conclusion

A linear prediction based method was presented for estimating the maximum-phase glottal flow signal and the open quotient. Tests showed that open quotient estimation can be successfully performed on synthetic signals with LP-covariance analysis but the method lacks robustness when real speech signals are analyzed.

---

## V.5. Application to speech recognition

The Automatic Speech Recognition (ASR) process aims at transcribing the text content of a given speech utterance.

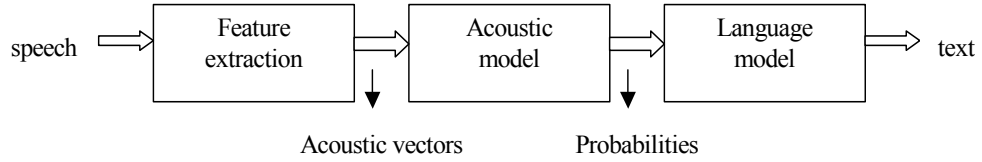


Fig. 56: ASR process

Our main concern is the first block of an ASR process (Fig. 56): the feature extraction where spectral computation is necessary. Therefore we only consider the feature extraction part and do not discuss further details of the ASR technology.

The feature extraction block aims at effectively reducing the amount of data to be processed by extracting  $k$ -sized acoustic feature vectors  $x_n = \{x_{n1}, \dots, x_{nk}\}$  for each  $N$ -size sample window. Most of the techniques use spectral envelope processing and amplitude/power spectrum has been the preferred component of the Fourier Transform (FT) for feature extraction. In this section, we investigate the possibilities of using chirp group delay function for feature extraction.

Two recent studies address this problem and propose two group delay based features: the modified group delay function (MODGDF) [Hegde *et al*, 2004 b] and the product spectrum (PS) [Zhu & Paliwal, 2004]. In [Hegde *et al*, 2004 b] Hedge *et al*, shows that the MODGDF representation captures complementary information to that of the power spectrum and ASR performance can be improved by combining MODGDF features and MFCC. However the same results could not be obtained in [Zhu & Paliwal, 2004] and the authors propose a new representation as alternative, the product spectrum (PS).

We have proposed three group delay based representations in chapter IV. In this section, we list the five group delay based representations used for feature extraction and we compare all five representations (plus the power spectrum for reference) in an ASR experiment<sup>8</sup>. The results show that two of the representations that we propose provide good results (outperforming the other group delay functions) and contain equivalent or complementary information to the power spectrum that is potentially useful for improving ASR performance.

### V.5.1. Group delay based features

The five group delay based representations we have used are:

- 1) The Modified Group Delay Function (**MODGDF**) [Hegde *et al*, 2004 b] explained in section IV.2. Its computation includes spectral smoothing of magnitude spectrum, which is further used for group delay computation.
- 2) The Product Spectrum (**PS**) defined as the product of the power spectrum and the group delay function in [Zhu & Paliwal, 2004]:

$$PS = |X(\omega)|^2 \tau_p(\omega) \quad (5.16)$$

- 3) **GDGCI** defined as the group delay function computed on GCI-synchronously windowed speech data (as discussed in Section IV.5).

---

<sup>8</sup> ASR experiments were performed by Laurent Couvreur, TCTS Lab., We thank him for his collaboration.

4) **CGDGCI** is defined as the chirp group delay computed on GCI-synchronously windowed data after ZZT-decomposition in Section IV.6 (the chirp group delay is computed on the circle with radius  $\rho=1.12$ ).

5) **CGDZP** is defined as the chirp group delay computed on zero-phase version of the signal in Section IV.6 (the chirp group delay is computed on the circle with radius  $\rho=1.12$ ).

#### *Comparison of Proposed Methods via Spectral Plots*

Fig. 57 presents a typical time-domain speech signal and its group delay function.

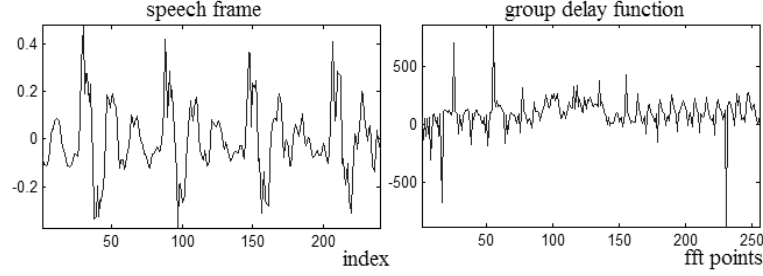


Fig. 57: Time-domain signal of a 30 ms speech frame and its group delay function. The frame example is extracted from the noise-free utterance “mah\_4625” of the test set A of the AURORA-2 [Hirsch & Pearce, 2000] and corresponds to vowel /i/ in word “six”.

As expected, the group delay function computed directly on the speech frame contains mainly spikes and resonance information cannot be observed. In Fig. 58, we present the five group delay based representations together with the power spectrum for this speech frame.

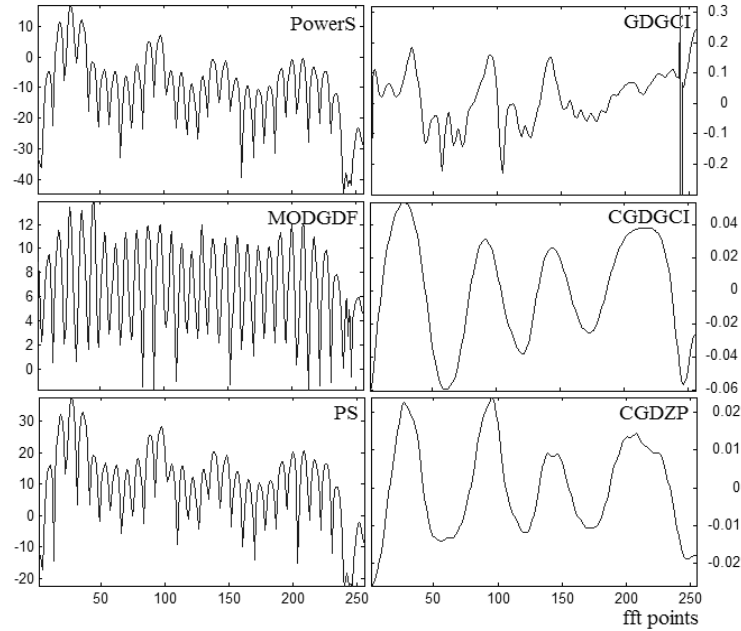


Fig. 58: Power spectrum (PowerS) and group delay based functions for the speech signal frame in Fig. 2.

The formant peaks appear with high resolution in GDGCI, CGDGCI and CGDZP where in MODGDF the spectral envelope appears to be blurred and PS is actually very similar to the power spectrum (as it is the case in [Zhu & Paliwal, 2004]). GDGCI includes a spike at high frequencies due to a zero, which cannot be avoided by only GCI-synchronous windowing. As more noise is added to signals, such spikes occur more frequently, therefore the robustness of GDGCI to noise is rather low. Thanks to zero removal techniques and zero-phasing, CGDGCI and CGDZP are more

robust to noise. In Fig. 59, we also present spectrogram plots obtained using the described group delay functions as well as the classical power spectrum.

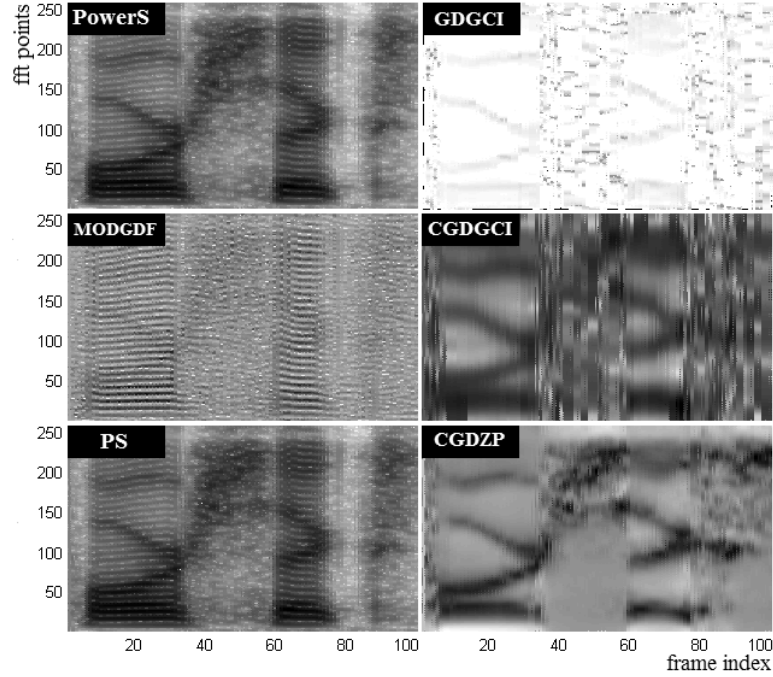


Fig. 59: Spectrogram plots of a noise-free utterance. Only the first half of the file, “mah\_4625a”, that contains the digit utterance “46” is presented.

The formant tracks can be well observed on all of the spectrograms except for MODGDF, and PS is very close to PowerS as already shown in Fig. 1 of [Zhu & Paliwal, 2004] and in Fig. 58 and Fig. 59 above. GDGCI representation is vague to some level. This is mainly due to the fact that unvoiced frames include spikes with large amplitudes that force a low contrast on the plots. Actually, the group delay functions computed on unvoiced frames mostly do not contain resonance information but random spikes. GDGCI and CGDGCI are actually the two representations that really suffer from this problem.

These observations suggest that the representations have some potential in an ASR framework. The main concern is if they can provide complementary information to the power spectrum and improve performance.

### Computation of features for ASR

The most common feature extraction for ASR systems consists of computing power-based Mel-frequency cepstral coefficients (MFCC) [Huang *et al*, 2001], that is, a Mel filterbank is applied to the power spectrum and an inverse discrete cosine transform (IDCT) is computed on the logarithm of its outputs. The main reason for such processing is to capture the essential shape of the power spectrum with a few coefficients well conditioned for pattern recognition. A similar scheme can be applied to the group delay functions in order to derive phase-based feature extractions for ASR systems. The simplest approach consists in replacing the power spectrum in the MFCC algorithm by the group delay function computed via one of the analysis techniques described in the previous section.

In this work, we use a Mel filterbank with 24 triangular filters and 12 IDCT coefficients are computed for 30 ms frames shifted by 10 ms. Note that the logarithm is not applied on the outputs of the filterbank when fed with a phase spectrum. These coefficients are augmented with the frame log-energy and their (delta-) delta coefficients. We finally come up with six feature extractions: MFCC as a reference and five group delay based methods.

## V.5.2. ASR experiments

### *ASR system*

The ASR system that is considered in this work relies on the STRUT toolkit [www-Strut]. It merely consists of three blocks. First, the feature extraction chops the discrete speech signal into overlapping frames and computes for each frame a set of acoustic coefficients using one of the algorithms described in the previous sections. Next, the acoustic coefficient vectors are fed into the acoustic model that is here based on the Multi Layer Perceptron (MLP) / Hidden Markov Models (HMM) paradigm [Bouclard & Morgan, 1994]. In this framework, the phonemes of the language under consideration are modeled by HMM's whose observation state probabilities are estimated as the outputs of a MLP. Such an acoustic model is trained beforehand in a supervised fashion on a large speech database containing a few hours of phonetically segmented speech material. Finally, the word decoder searches for the most likely word sequence given the sequence of probability vectors for all the frames. Here, the search is constrained by a phonetic lexicon and a word grammar, which together define all the authorized sequences of phonemes. Here, the search is performed as a one-pass frame-synchronous Viterbi algorithm [Huang *et al*, 2001] without any pruning constraints.

### *Speech Database*

The AURORA-2 database [Hirsch & Pearce, 2000] was used in this work. It consists of connected English digit utterances sampled at 8kHz. More exactly, we used the clean training set, which contains 8440 noise-free utterances spoken by 110 male and female speakers, for building our acoustic models. These models were evaluated on the test set A. It has 4004 different noise-free utterances spoken by 104 other speakers. It also contains the same utterances corrupted by four types of real-world noises (subway, babble, car, exhibition hall) at various signal-to-noise ratios (SNR) ranging from 20dB to -5dB. During the recognition experiments, the decoder was constrained by a lexicon reduced to the English digits and no grammar was applied.

**Table 2:** ASR performances for various feature extraction on the AURORA-2 task. Results are given in terms of word error rate (WER) in percent.

Feature Extraction	SNR(dB)						
	$\infty$	20	15	10	5	0	-5
MFCC	1.9	6.7	18.6	45.2	75.1	88.8	91.5
MODGDF	3.2	19.0	41.7	68.7	86.1	91.0	92.3
PS	2	6.7	19.4	45.3	75.5	89	92.2
GDGCI	8.8	32.8	49.4	69	88.3	98.6	100
CGDGCI	3.2	12.3	25.6	50.8	80.8	97	99.8
CGDZP	1.8	5.8	12.2	29.4	62.6	88.7	97.6

**Table 3:** ASR performances for features combined with MFCC on the AURORA-2 task. Results are given in terms of word error rate (WER) in percent.

Feature Extraction	SNR(dB)						
	$\infty$	20	15	10	5	0	-5
MODGDF	2.1	8.5	23.9	52.7	79.5	89.5	91.5
PS	1.9	6.7	18.6	44.4	74.6	88.5	91.6
GDGCI	2.1	7.8	16.8	36	64.4	88	96.1
CGDGCI	1.8	5.8	12.2	29.1	58	83.8	93.8
CGDZP	1.7	5	10.4	24.8	52.7	82.3	91.1

### *Experimental Results*

Tab. 2 gives the word error rates (WER) for the ASR system tested with the feature extractions described. Errors are counted in terms of word substitutions, deletions and insertions, and error rates are averaged over all noise types. In Tab. 3, the results are also provided when combining



MFCC feature extraction with the others. The combination is simply performed by taking a weighted geometric average of the probability outputs of the combined acoustic models:

$$p_{12} = p_1^\lambda \cdot p_2^{1-\lambda} \quad (5.17)$$

where  $p_{12}$ ,  $p_1$  and  $p_2$  denote the combined probability and the probability provided by the two combined acoustic model, respectively. The combination parameter  $\lambda$  takes its value in the range (0,1) and is optimized for every combination.

### V.5.3. Discussion and conclusion

Our main target in this study was to test if a phase/group delay function carries equivalent or complementary information to that of the power spectrum in the framework of feature extraction for ASR systems. The results presented in Tab. 3 shows that the group delay functions CGDGCI and CGDZP have this potential: the values in the last two rows of Tab. 3 compared to the MFCC-only results (first row in Tab. 2) are in all cases lower except for the extreme noise setting SNR=-5dB.

In our in-detailed analysis, we have observed that the GDGCI, which is the pure group delay function computed on GCI-synchronous data without further processing, mainly suffers from window size problems (including several pitch periods result in zeros on the unit circle). In addition, GDGCI and CGDGCI do not carry reliable information for unvoiced frames.

The AURORA-2 task was chosen for its simplicity and ease of comparison to the already available results in [Zhu & Paliwal, 2004]. Further experiments will be performed on other tasks in order to confirm the present results about the usefulness of phase information for ASR systems.

---

# Chapter VI:

## Conclusion and Future Works

---

### VI.1. Conclusions

This study proposed two new spectral representations and demonstrated their use in various speech analysis problems: source-tract separation, glottal flow parameter estimation, reliable phase spectrum estimation, formant tracking and feature extraction in speech recognition. The two representations (ZZT and chirp group delay) are strongly linked. For most of the cases they need to be considered simultaneously, even in algorithms where only one is processed. The combination of the ZZT representation with the chirp group delay processing algorithms provides an effective framework for the study of the resonance characteristics of the source and filter components of speech.

#### *The ZZT representation and its applications*

The first spectral representation we proposed was: the zeros of the z-transform (ZZT) representation, which we defined as the set of roots of the Z-transform polynomial for a discrete time signal. There are mainly two useful points of the ZZT representation for speech signals: i) it sheds light into many difficulties involved in phase spectrum processing and for this reason provides us with the opportunity to design better methodologies, ii) patterns exist in the ZZT of speech signals which make it possible to design a new spectral decomposition method for source-tract separation.

Being a form of z-transform representation, ZZT representation is especially useful for studying some properties of the Discrete Fourier Transform (DFT) of a signal, especially the phase component of it. In the theory part of the thesis, we have studied ZZT representation of: some elementary signals, the LF glottal flow model, source-filter model of speech and windowed speech signals. Through a systematic study of ZZT of windowed speech (with various windowing functions, size and location (with respect to important instants in speech signals) on various speech signals), we showed that windowing lies at the very heart of the problem of spikes in the derivative of phase spectrum (the group delay function) due to zeros close to the unit circle. The spikes in the group delay function appear as an important obstacle in speech processing: the often cited and unsolved “phase unwrapping problem” is mainly due to these spikes on the group delay function. We showed that avoiding these spikes is possible by performing the windowing appropriately: glottal closure instant synchronous windowing with a size of two pitch periods and with one of the three windowing functions: Blackman, Gaussian or Hanning-Poisson.

The fact that we can obtain spike-free group delay functions is an important step for phase processing. There are actually plenty of signal processing applications, which can benefit from the results of this study. In many signal processing studies, phase estimation is considered to be a difficult problem and discarded. However, for some applications, the phase information is essential or at least is an important factor of the efficiency of the algorithms. The methods defined in this thesis provide hopefully a potential to remove some of the obstacles in the phase estimation problem.

The systematic study of the ZZT of windowed speech signals has one more important output: separate patterns for the glottal flow and vocal tract contributions can be observed. The ZZT representation of a GCI synchronously windowed speech frame includes two lines/groups of zeros: one outside the unit circle and one inside the unit circle with gaps creating formant peaks on the spectrum. These observations have led us to design a spectral source-tract separation algorithm based on ZZT-decomposition. Our methodology involves no modeling but direct separation in the spectral domain. In addition, such an observation both supports the studies discussing anti-causality of the glottal flow component in literature and suggests a mixed-phase model for speech signals. For completeness of the

theoretical background, we also discussed the mixed-phase model for speech signals through ZZT patterns of source-filter model of speech.

The ZZT-decomposition method was first demonstrated by synthetic speech and real speech examples. The decomposition is of high quality though not complete. The contribution of the vocal tract in the glottal-flow-dominated spectrum was observed as ripples of low amplitude, while the contribution of glottal flow in the vocal tract dominated spectrum was hardly observed. Then some robustness tests were presented using synthetic speech. We have observed that ZZT-decomposition is robust to GCI errors in  $\pm 10\%$  range and variations in the return phase coefficient of the glottal flow component. However, it suffers from F1 variations (when F1 is close to the glottal formant peak in the spectrum) and is not robust to additive noise. Further research is necessary to improve robustness of the method for these variations. The proposed algorithm is very easy to implement but computationally heavy due to the need of finding roots of high degree polynomials. For this reason, it is more appropriate for off-line database processing.

The ZZT-decomposition was also tested in a glottal flow parameter estimation scheme: an algorithm for glottal formant ( $F_g$ ) tracking was proposed. It is mainly composed of picking the maximum valued peak on the magnitude spectrum of the glottal flow obtained by ZZT-decomposition from GCI synchronously windowed speech signals. The frequency of glottal formant is potentially useful in studying voice quality variations of speech.

The proposed algorithm was tested on synthetic speech and the results showed that the ripples on glottal flow spectrum due to incomplete separation introduces errors when  $F_g$  and  $F_1$  values are close. In addition, the sensitivity of peak picking to small ripples reduces robustness of the method. For our future studies we target improving the quality of the  $F_g$  estimation method by replacing peak picking with a more robust method.

Due to lack of reference methods and data, tests on real speech were limited. On a single real speech example, we have shown that open quotient variations can be tracked with the  $F_g$  estimation algorithm. Further validation with real speech decomposition tests is necessary.

#### ***The chirp group delay (CGD) representation***

Another representation proposed in this thesis was the ‘chirp group delay’. The chirp group delay simply corresponds to the group delay function computed on a circle in z-plane other than the unit circle. It is the negative derivative of the phase component of the chirp z-transform.

The necessity of such a representation is due to the difficulties involved in group delay processing for some applications like formant tracking. Although the group delay one can obtain after appropriate windowing reveals formant peaks, it is not possible to guarantee absence of zeros close to the unit circle for noisy speech even when windowing is appropriately performed. Therefore we have developed chirp group delay processing as an alternative, for which spike-freeness can be guaranteed. Robust spectral processing can be achieved using this representation. It also facilitates studying minimum-phase and maximum-phase contributions in the signals separately.

#### ***Applications of ZZT and CGD***

One of the potential applications of chirp group delay processing is formant tracking. Three formant trackers were presented with varying complexity and robustness. At each update of the formant tracker, quality and robustness were improved by solving a problem. All the versions are based on peak picking on the chirp group delay computed outside the unit circle and two versions include some ZZT manipulation of a given signal.

The formant trackers were tested both on synthetic and real speech signals. The first version of the formant tracker was observed to have robustness problems and the second version was developed by solving one important problem involved. The results for the second method showed that it has similar quality as the two state-of-the-art methods (that of Praat and WinSnoori) and is an effective method for formant tracking. Its main advantage is in tracking high order formants and main drawbacks are: need for GCI marking and high computational load due to need for calculation of roots of high degree polynomials. The third version of the formant tracker was developed to get rid of these difficulties. This final version was further compared with formant trackers of Praat and Wavesurfer by performing tests on real speech. The results of the three systems are comparable; they all provide high quality formant tracks. The fact that high quality formant tracking can be performed on chirp group delay with a simple algorithm shows that chirp group delay is an effective spectral representation and can potentially be used in various other applications.

Another application was in speech recognition, more specifically in acoustic feature extraction. For this application, we tested if chirp group delay function carries equivalent or complementary information to power spectrum, which can improve speech recognition performance.

We proposed three chirp group delay based features and compared them to two recently proposed group delay based features from the literature and the MFCC. The AURORA-2 task (isolated digit recognition testing) was chosen for its simplicity and ease of comparison to the already available results in literature. The results showed that two of the group delay functions we proposed (CGDGC and CGDZP) improve recognition rates when combined with MFCC. We conclude that chirp group delay function carries equivalent or complementary information to power spectrum, which can improve speech recognition performance. Between the two representations, CGDZP is more applicable in speech recognition since the computational load is low and it is more robust to noise. Further experiments need to be performed on other tasks in order to confirm the present results about the usefulness of phase information for ASR systems.

#### ***Other applications studied***

We have also studied linear prediction analysis of the mixed-phase speech signals. We have developed the MixLP algorithm that estimates a pole pair outside the unit circle corresponding to the anti-causal poles of the source signal component in the mixed-phase speech model. Given the pair of anti-causal poles, a procedure to resynthesize the anti-causal part of the glottal flow, and an open quotient estimation method, were proposed. Tests performed showed that the method is high quality for analyzing synthetic speech but lacks robustness in analysis of natural speech.

Due to time constraints the application areas were kept limited to these topics. However we believe that the two representations can be further used in many other speech analysis applications.

## **VI.2. Future works**

We have shown that the developed algorithms are high quality. However, some problems exist and there is still room for development. In addition, further testing is necessary for some of the algorithms. The thesis has been concluded at this point since testing every possible path for using the two representations is practically not possible in a single thesis research period.

The speech recognition tests we have handled are rather limited though sufficient to demonstrate the potential; we have only tested baseline systems in limited test settings. The results are very promising. Therefore we find it interesting to further study this issue to integrate the proposed algorithms in a complete continuous speech recognition system and perform large tests. Such a study would need considerable amount of time since most speech recognition algorithms include fine-tuning heuristics rules and many trial-errors are necessary for finding those rules.

Unfortunately the ZZT-decomposition for source-tract separation lacks robustness for two factors: additive noise and closeness of  $F1$  and glottal formant. Since the algorithm uses the roots of large degree polynomials, studying the effect of noise is not easy. We think that  $F1$  and glottal formant interaction can be further studied in detail using the ZZT representation. It is not easy to foresee a good methodology to study these problems at this point. Hopefully, visual study of the two phenomena for various examples should result in better comprehension of the problems and lead to solutions as it did during this thesis work.

ZZT-decomposition needs to be further tested on continuous speech with various voice qualities. We have discussed the difficulties of finding reference data for such tests. We plan to further test the algorithm comparing glottal formant frequency variations with open quotient estimates from EGG signals recorded in parallel to speech signals. We cannot conclude at this point that ZZT-decomposition can be used for studying glottal flow variations in different phonation types before performing these tests. We could only show its effectiveness in limited test settings. Further tests and improvement also necessitate considerable amount of research time.

One of the potentially important outputs of this thesis is that it shows that formant information exists on the group delay functions computed directly from the phase spectra without further processing. We think that some of the concatenative speech synthesis problems using the sinusoidal/harmonic model can be re-studied based on this observation. For example spectral smoothing is usually performed at concatenation points for the magnitude spectrum; for the phase spectrum some form of continuity of phases are targeted

and algorithms are proposed for this operation. Spectral smoothing on the magnitude spectrum to some extent corresponds to imposing smooth formant transitions at concatenation points. Guaranteeing smooth formant transition for the phase component is also important to reduce discontinuities. The only concatenation algorithm we could find in literature using the group delay function at some levels is the STRAIGHT system [Kawahara et al, 2001]: for keeping a fine temporal structure of the signal, storing/coding the phase spectrum, etc. We believe that concatenation systems can be further improved if formant information in phase is further studied.

One other research path to follow is spectral parameter estimation for voice quality analysis using (chirp) group delay. Compared to the magnitude spectrum, events are more localized in chirp group delay function, for example formant peaks occur with smaller bandwidths and sharper amplitudes. Spectral tilt is one of the most important parameters for voice quality analysis and it is difficult to estimate it from the magnitude spectrum since the effect is somehow distributed on a large frequency band. Studying effect of spectral tilt variations on chirp group delay functions can potentially result in developing effective algorithms for spectral tilt estimation.

Actually, phase processing is not only necessary for speech processing. We hope that the discussions presented in this thesis will be also accessible and useful to researchers from different fields like: radar signal processing, medical imaging, sound source localization, optics, solid state physics, geophysics, holography, etc.

# Appendix A: Window functions

This appendix includes the definitions for the window functions mentioned in this thesis manuscript. For a large study of window function spectra please refer to [Harris, 1978].

Notes:

$N$ : window size,  $n=0,1,2,\dots,N-1$ .  $\alpha$  is a user defined variable for which default values is specified as 2.5 in MATLAB. The Hanning-Poisson function is obtained by simply term-wise multiplication of Hanning and Poisson windows.

Window function	Definition	Matlab Function
Rectangular	1	Rectwin
Hanning	$w(n) = \frac{1}{2} \left[ 1 - \cos\left(\frac{2\pi n}{N-1}\right) \right]$	Hanning
Hamming	$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$	Hamming
Blackman	$w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cos\left(\frac{4\pi n}{N-1}\right)$	Blackman
Gaussian	$w(n) = \exp\left(-0.5 \left( \frac{\alpha \left(n - \frac{N-1}{2}\right)^2}{N/2} \right)\right)$	Gausswin
Poisson	$w(n) = \exp\left(-\frac{\alpha \left n - \frac{N-1}{2}\right }{N/2}\right)$	-

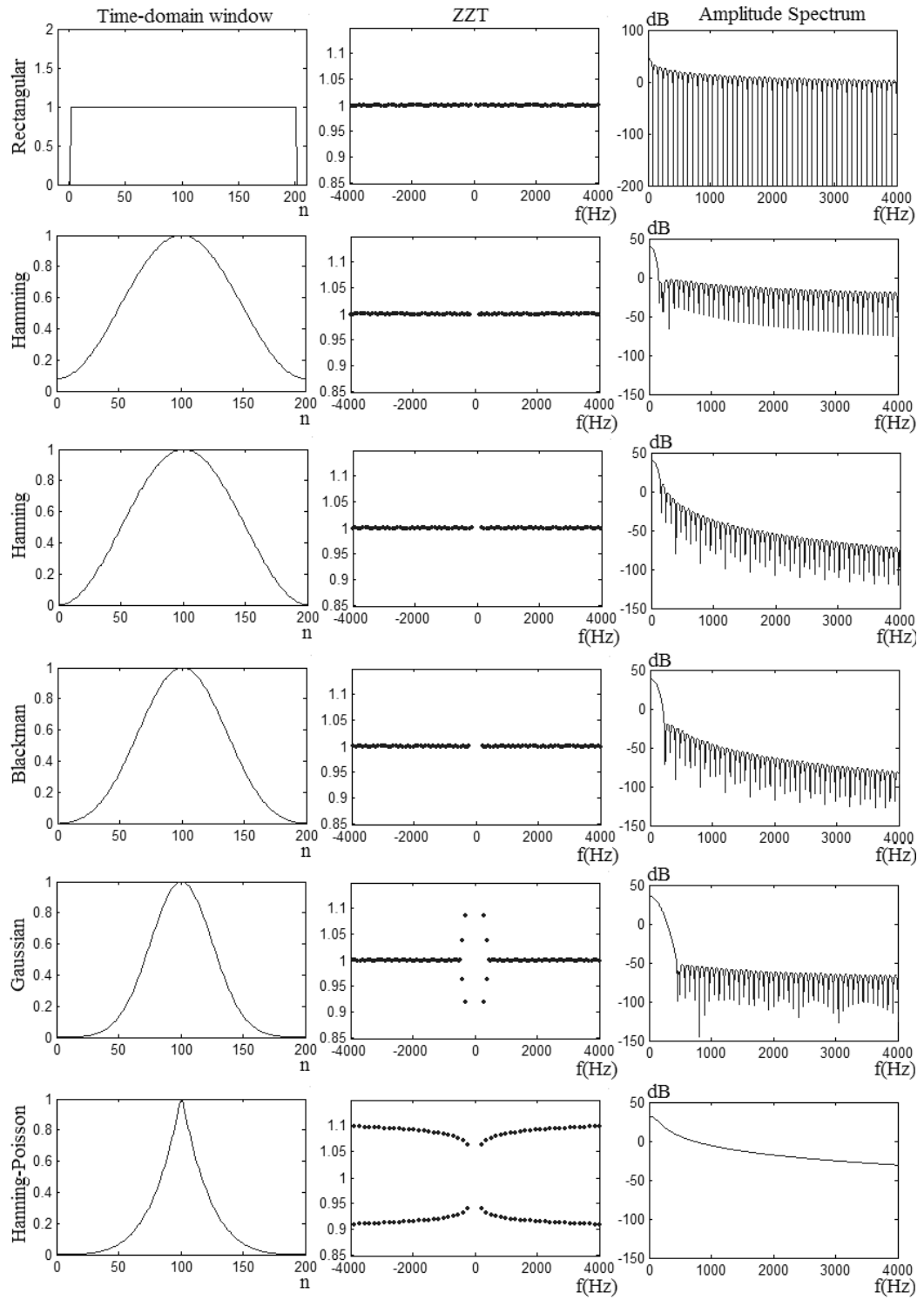


Fig. 60: Window functions  
Time domain function, ZFT representation and magnitude spectrum

# Appendix B: Relation between poles and spectral peaks of an all-pole filter

The basic idea in the linear predictive (LP) analysis is that a speech sample can be approximated as a linear combination of past speech samples [Rabiner & Schafer, 1978]. Here we present only the very basics to explain the links between the poles of an all-pole system and its resonance frequencies. For a detailed description of LP analysis, the reader is referred to [Makhoul, 1977]. In the LP model, a discrete-time sequence  $s[n]$  is expressed in terms of a weighed sum of the past samples of  $s[n]$  :

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

where  $\tilde{s}[n]$  is the predicted sequence and  $e[n]$  refers to the prediction error. This is also referred as the auto-regressive (AR) model.

The predicted sequence is considered to be an output of a linear predictor with coefficients  $\alpha_k$

$$\tilde{s}[n] = \sum_{k=1}^p \alpha_k s[n-k]$$

The prediction error,  $e[n]$ , is defined as :

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

The goal in LP analysis is to estimate the filter coefficients  $a_k$  for a specific order  $p$  that will minimize the mean-squared prediction error over short segment of the speech waveform. Then the resulting system function for the AR model can be expressed as :

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}}$$

or equivalently in an all-pole form :

$$H(z) = \frac{A}{\prod_{k=1}^p (1 - d_k z)(1 - d_k^* z)}$$

where  $d_k$  are the poles of the transfer function (and  $*$  is the complex conjugate operator).

Considering the source-filter model of speech, the complete transfer function of the speech production system corresponds to:

$$H(z) = AG(z)V(z)R(z)$$

where  $G(z)$  is the glottal flow transfer function,  $V(z)$  is the vocal tract transfer function and  $R(z)$  is the radiation load transfer function. Separating the three contributions is a difficult problem and various algorithms have been proposed in literature (e.g. [Alku, 1992 a]). In fact, the glottal flow function is considered to include a maximum-phase pole pair and a zero [Doval *et al*, 2003] and the radiation load is considered to include a zero [Quatieri, 2002]. Although any zero can be expressed by infinite number of poles, the AR model representation is practically more convenient to express the vocal tract transfer function which is often approximated by an all-pole minimum-phase filter.

$$V(z) = \frac{1}{\prod_{k=1}^N (1 - c_k z)(1 - c_k^* z)}$$

where  $N$  is smaller than  $p$ , and  $c_k$  are the poles of the vocal tract transfer function. Each resonance of the vocal tract filter corresponds to a pole pair  $(c_k, c_k^*)$  in this representation and contributes to the speech spectrum by local spectral peaks. In the figure below, a simple three pole-pair system is presented (the



frequency axis of the two spectra are labeled in Hz and the sampling frequency is selected as 16000Hz. The main aim of this choice is facilitating any comparison to existing speech processing literature where formants are often mentioned with their actual frequency in Hz.). Each pole pair and the corresponding peaks on the log-magnitude spectrum and the group delay function are connected with lines. In the pole-plot in polar coordinates, the shaded area on the pole-plot in cartesian coordinates is presented.

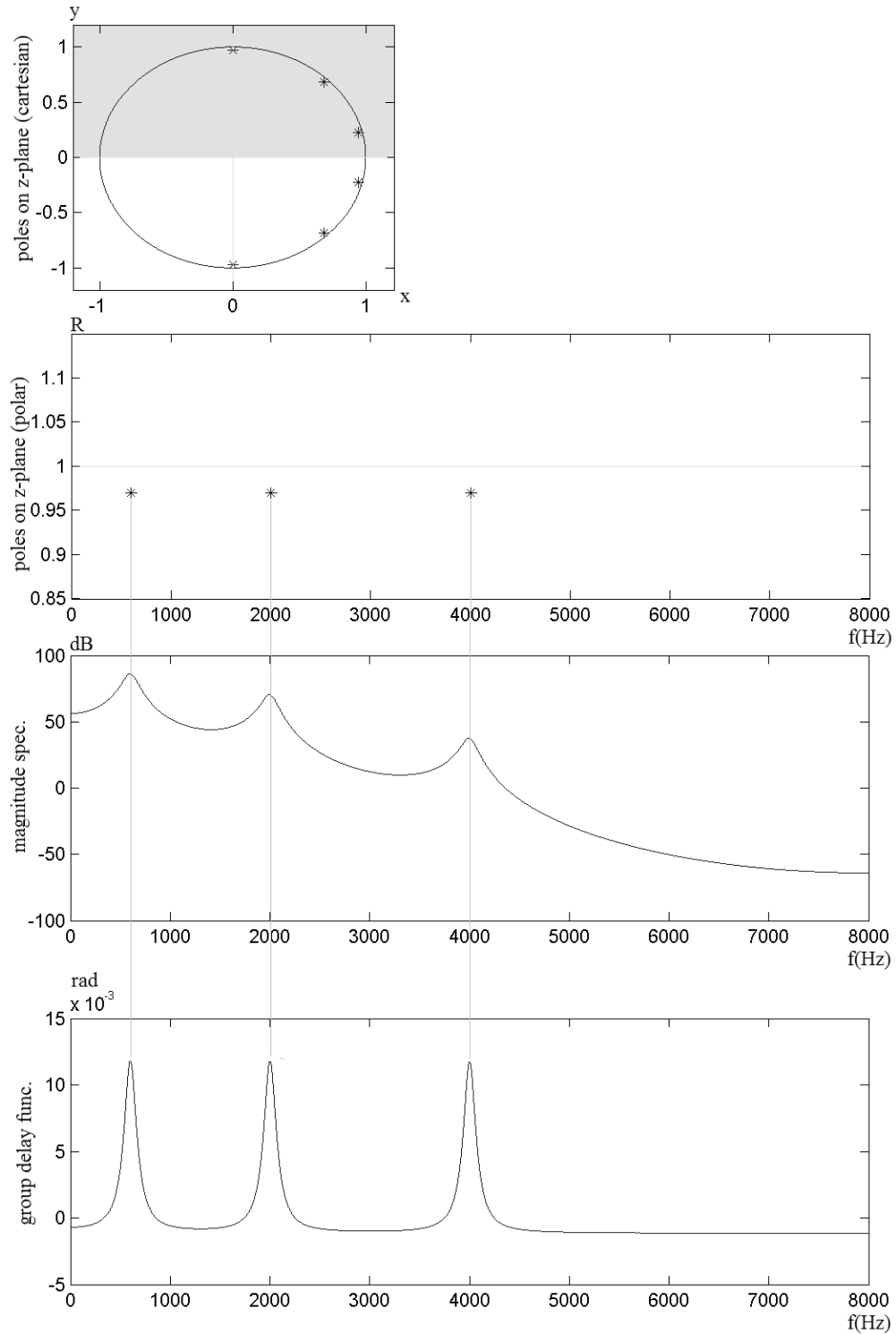


Figure Appendix B: Link between pole pairs and spectral peaks

# Appendix C: Formant tracking examples

In this part, we present plots of formant tracking results from three formant trackers: Fast-DPPT, Praat and WaveSurfer. The list of wave files used is:

Gender of the speaker	File name	Language	Source
Male	BrianLou5.wav	English	Voqual03
Male	JackBrown.wav	English	Private
Male	WNEU_SE9.HO.wav	Danish	DES
Male	WNEU_SE9.JZB.wav	Danish	DES
Female	f24cb1_6.wav	French	Voqual03
Female	01_08-338.34.wav	Japanese	Voqual03
Female	4SX9.wav	English	DARPA
Female	WNEU_SE9.DHC.wav	Danish	DES
Female	WNEU_SE9.KLA.wav	Danish	DES

Voqual03: Database made available for the workshop Voqual03: Voice quality: Functions, analysis and synthesis, Geneva, August 2003. [www-Voqual03]

DARPA: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc CD1-1.1 [www-DarpaDBA]

DES: Danish Emotional Speech Database [Hansen & Engberg, 1996]

Each figure contains two sub-plots: the upper subplot is the spectrogram superimposed to the formant tracks obtained by the Fast-DPPT formant tracker (indicated in red); the lower subplot contains the formant tracks obtained by the formant trackers of Praat (indicated in blue) and Wavesurfer (indicated in black). The spectrograms are obtained using the WaveSurfer software.

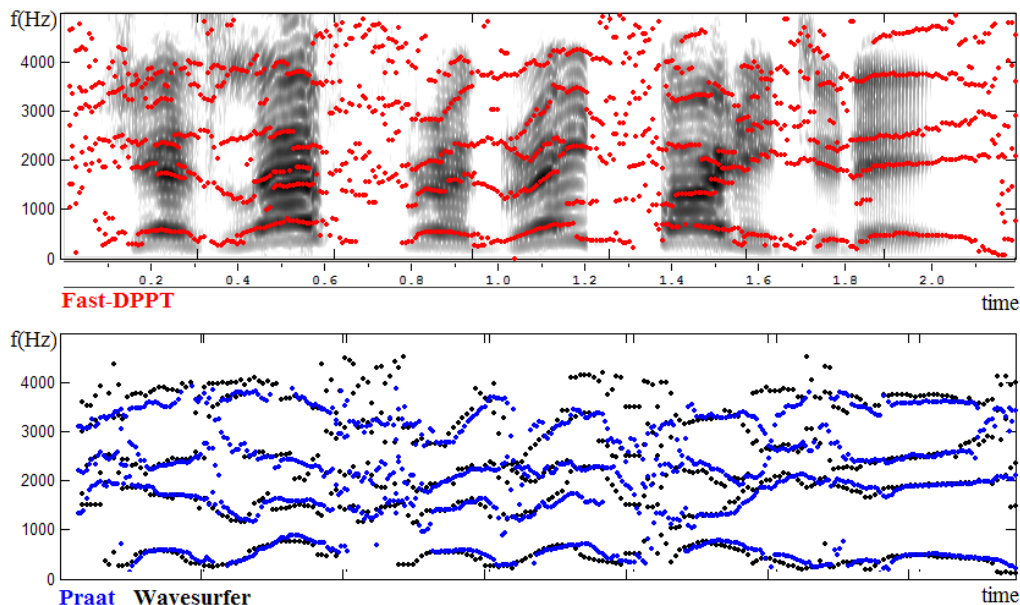


Fig. 61: Formant tracking example 2  
BrianLou5.wav

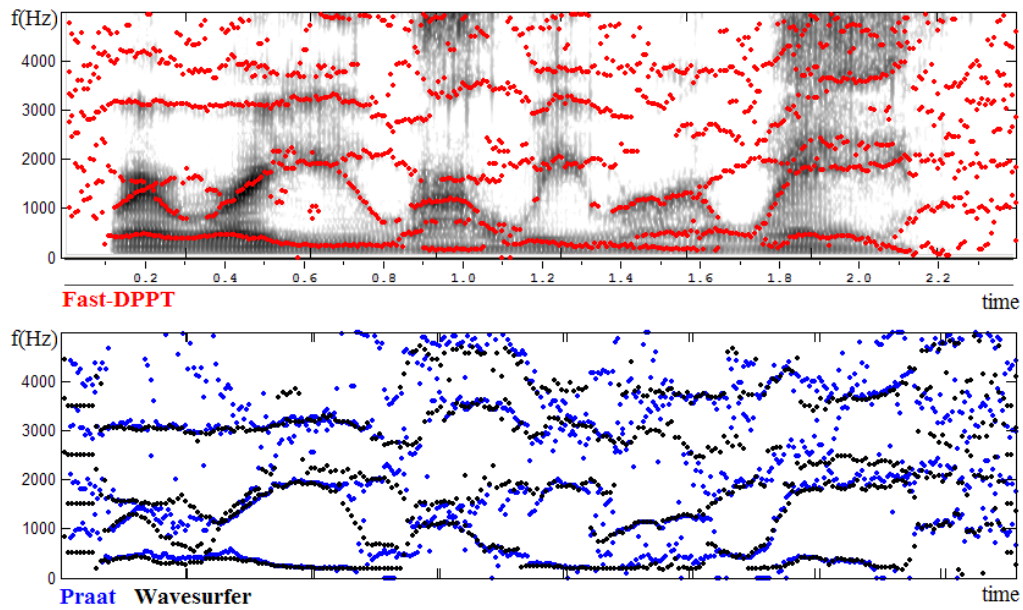


Fig. 62: Formant tracking example 3  
jackBrown.wav

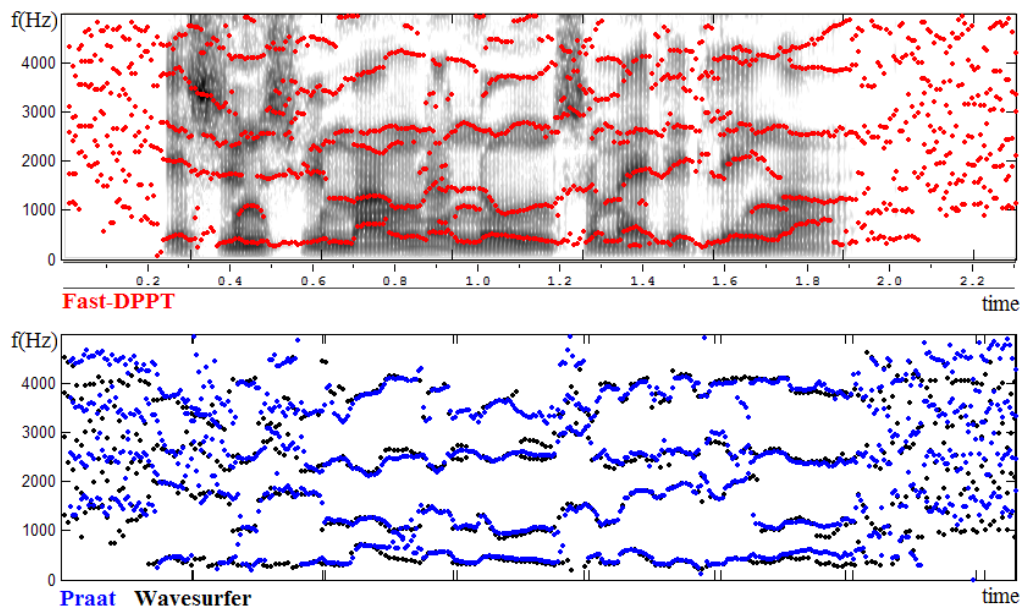


Fig. 63: Formant tracking example 4  
WNEU\_SE9.HO.wav

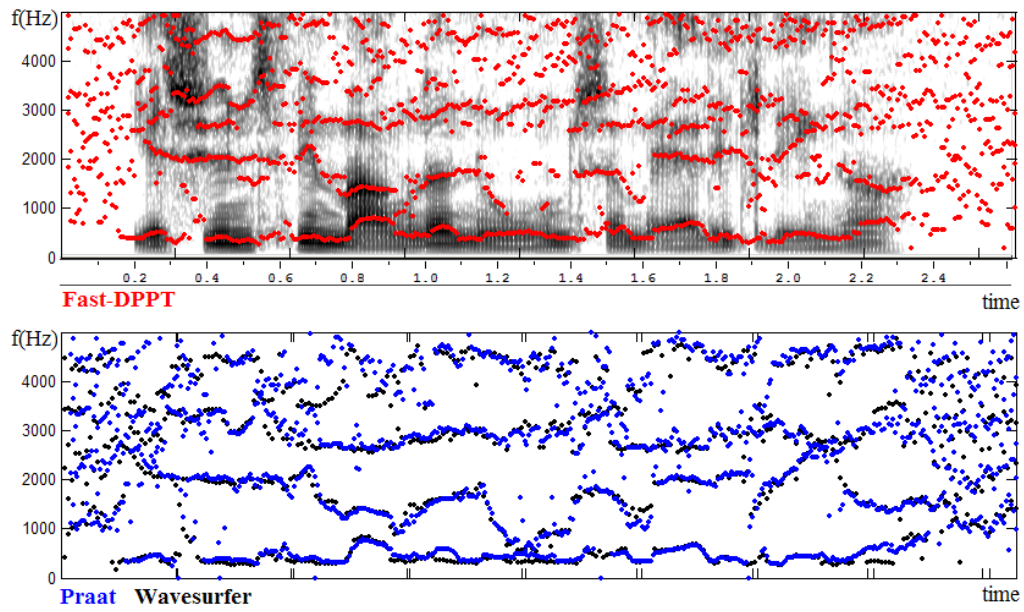


Fig. 64: Formant tracking example 5  
WNEU\_SE9.JZB.wav

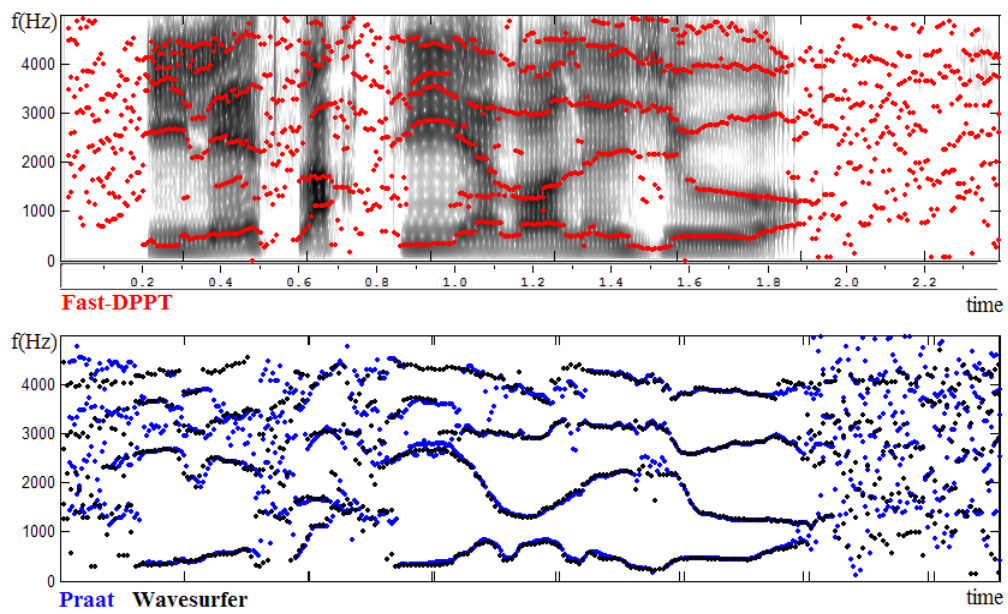


Fig. 65: Formant tracking example 6  
f24cb1\_6.wav

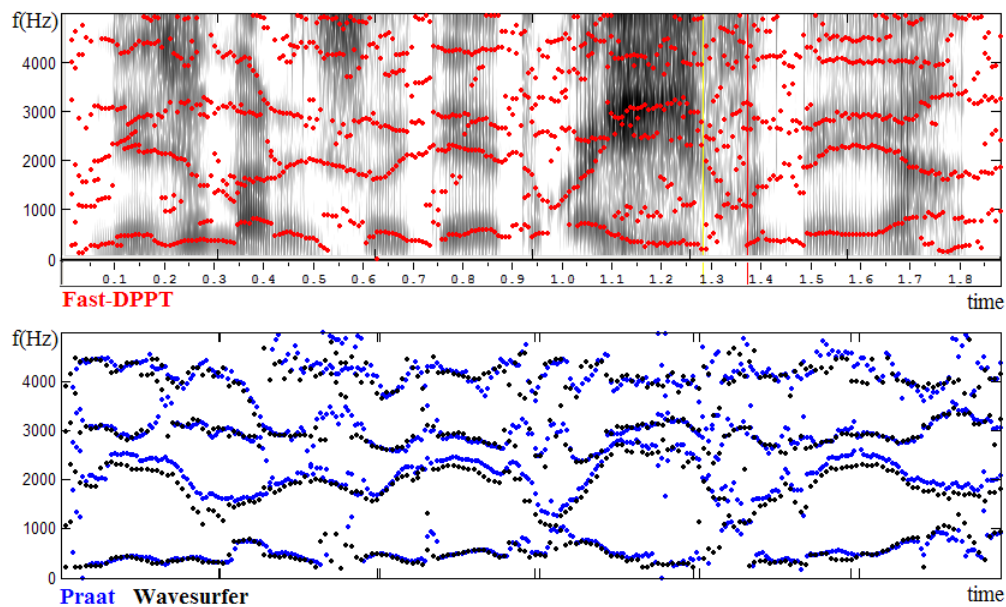


Fig. 66: Formant tracking example 7  
01\_08-338.34.wav

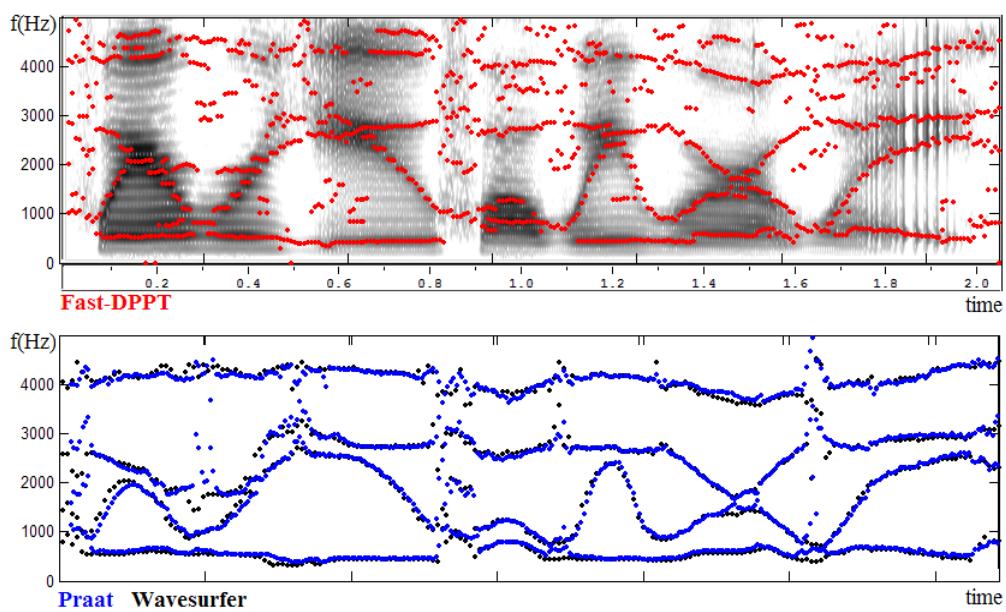


Fig. 67: Formant tracking example 8  
4sx9.wav



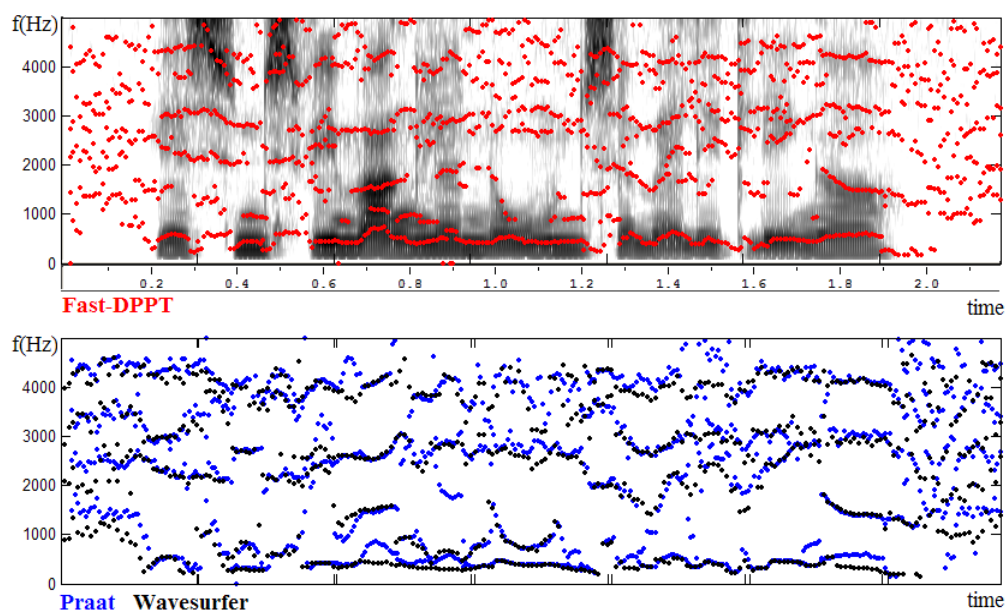


Fig. 68: Formant tracking example 9  
WNEU\_SE9.DHC.wav

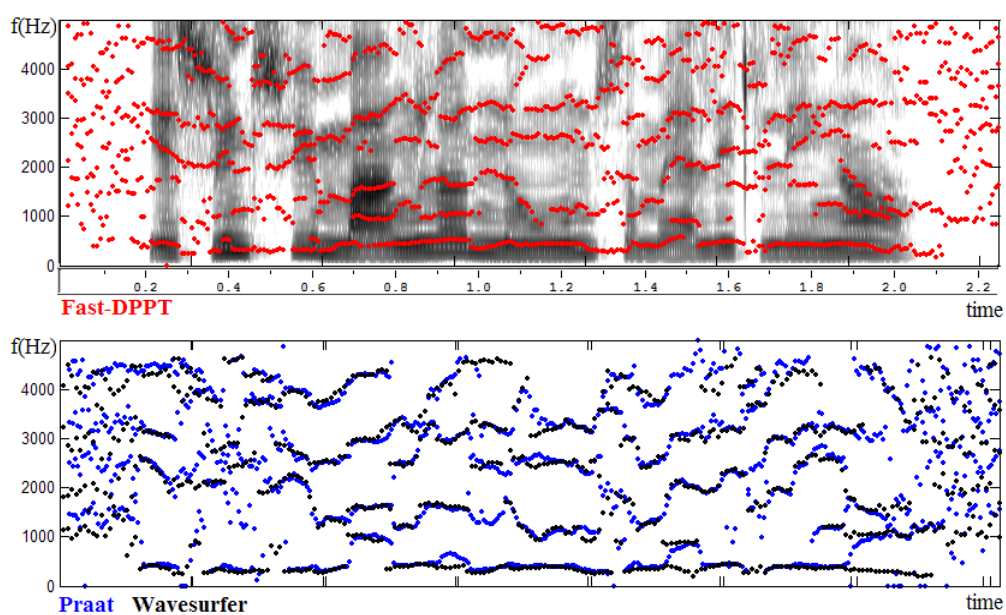


Fig. 69: Formant tracking example 10  
WNEU\_SE9.KLA.wav

# Appendix D: Publications not referred in the thesis manuscript

Publications/studies of Baris Bozkurt during the thesis research period that are not referred in the thesis manuscript:

- F.Severin, B.Bozkurt, T.Dutoit, ‘HNR extraction in voiced speech, oriented towards voice quality analysis’, *Proc. EUSIPCO*, Antalya (Turkey), 2005.
- Z.Hammal, B.Bozkurt, L.Couvreux, D.Unay, A.Caplier, T.Dutoit, ‘Passive versus active: vocal classification system’, *Proc. EUSIPCO*, Antalya (Turkey), 2005.
- N.D’Allesandro, R.Sebbe, B.Bozkurt, T.Dutoit, ‘MAXMBROLA: A Max/MSP Mbrola-based tool for real-time voice synthesis’, *Proc. EUSIPCO*, Antalya (Turkey), 2005.
- O.Turk, M.Schroeder, B.Bozkurt, L.M.Arslan, ‘Voice Quality Interpolation for Emotional Text-to-Speech Synthesis’, *Proc. Interspeech*, Lisbon (Portugal), 2005.
- T.Dutoit, B.Bozkurt, ‘Speech synthesis’, chapter to appear in “*Handbook on Signal Processing in Acoustics*”, Springer Verlag (in press), 2005.
- B.Bozkurt, M.Ozkan, ‘Gorme ozurluler icin okuma yardimcisi’ *Biyomedikal Muhendisligi Enstitusu Bulteni, Bogazici Univ. Yayinlari*, Istanbul, sayi:3, 2004.
- O.Ozturk, B.Bozkurt, T.Ciloglu, T.Dutoit, ‘Acgozlu algoritma kullanilarak turkce metin veritabani hazirlanmasi’, *Proc. SIU, Sinyal Isleme Uygulamalari*, Istanbul, 2003.
- B.Bozkurt, T.Dutoit, O.Ozturk, ‘Text Design For TTS Speech Corpus Building Using A Modified Greedy Selection’, *Proc. EUROSPEECH*, Geneva, pp 277-280, 2003.
- B.Bozkurt, T.Dutoit, V.Pagel, ‘Synthèse vocale par sélection d’unité: une méthode pour la redéfinition de la courbe intonative’, *Proc. JEP, Journées d’Etude sur la Parole*, pp 121-124, Nancy, 2002.
- B.Bozkurt, R.Prudon, T.Dutoit, C.D’Alessandro, V.Pagel, ‘Improving Quality of MBROLA Synthesis for Non-Uniform Units Synthesis’, *Proc. of the IEEE TTS 2002 Workshop*, Santa Monica, September 2002.
- B.Bozkurt, T.Dutoit, V.Pagel, ‘Re-defining intonation from selected units for non-uniform units based speech synthesis’, *Proc. of SPS-2002 IEEE Benelux Signal Processing Symposium*, Leuven, 2002.
- B.Bozkurt, M.Bagein, T.Dutoit, ‘From MBROLA to NU-MBROLA’, *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Blair Atholl, Scotland, 2001.
- B.Bozkurt, T.Dutoit, ‘An implementation and evaluation of two diphone-based synthesizers for Turkish’, *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Blair Atholl, Scotland, 2001.

# References

- [Abel, 1826] N. H. Abel, 'Beweis der Unmöglichkeit, algebraische Gleichungen von höheren Graden als dem vierten allgemein aufzulösen.' *J. reine angew. Math.* 1, 65, 1826. Reprinted in Abel, N. H. OE (Ed. L. Sylow and S. Lie). Christiania [Oslo], Norway, 1881. Reprinted in New York: Johnson Reprint Corp., 1988, pp. 66-87.
- [d'Alessandro & Doval, 1997] C. d'Alessandro and B. Doval, 'Spectral representation and modeling of glottal flow signals' *Proc. ESCA Workshop Larynx*, Marseille, June 1997, pp. 87-90.
- [d'Alessandro & Doval, 1998] C. D'Alessandro, B. Doval, 'Voice quality modification using periodic-aperiodic decomposition and spectral processing of the voice source signal.' *Proc. 3rd International Workshop on Speech Synthesis*, Jenolan Caves, Australia, November, 1998.
- [d'Alessandro & Doval, 2003] C. d'Alessandro, B. Doval, 'Voice quality modification for emotional speech synthesis', *Proc. Eurospeech*, Genève, Suisse, September 2003, pp. 1653-1656.
- [Alku, 1992 a] P. Alku, 'Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering.' *Speech Communication*, vol. 11, no. 2-3, 1992, pp. 109-117.
- [Alku, 1992 b] P. Alku, 'An automatic method to estimate the time-based parameters of the glottal pulse form.' *Proc. ICASSP*, vol. 2, San Francisco, 1992, pp. 29-32.
- [Alku et al, 1997] P. Alku, H. Strik, Vilkman, 'Parabolic spectral parameter- A new method for quantification of the glottal flow.' *Speech Communication*, vol. 22, 1997, pp. 67-79.
- [Alku et al, 2000] P. Alku, J. Svec, E. Vilkman, F. Sram, 'Analysis of voice production in breathy, normal and pressed phonation by comparing inverse filtering and videokymography.' *Proc. ICSLP*, Beijing 2000, pp. 885-888
- [Alsteris & Paliwal, 2004] L. Alsteris, K.K. Paliwal, 'Importance of window shape for phase only reconstruction of speech.' *Proc. ICASSP*, 2004, pp. 573-576.
- [Andersen & Jensen, 2001] T.H. Andersen, K. Jensen, 'On the importance of phase information in additive analysis/synthesis of binaural sounds.' *Proc. of the International Computer Music Conference*, Havana, Cuba, 2001.
- [Banno et al, 2001] H. Banno, K. Takeda, F. Itakura 'A study on perceptual distance measure for phase spectrum of stimuli.' *Proc. of ICASSP, Speech, and Signal Processing*. vol.5, 2001, pp.3297-3300
- [Bäckström et al, 2002] T. Bäckström, P. Alku, E. Vilkman, 'Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range.' *IEEE Transactions on Speech and Audio Processing*, vol. 10, March 2002, no. 3, pp. 186-192.
- [Bode, 1945] H.W. Bode, 'Network analysis and feedback amplifier design.' New York, Van Nostrand, 1945.
- [Bourlard & Morgan, 1994] H. Bourlard, N. Morgan, 'Connectionist Speech Recognition: A Hybrid Approach.' Kluwer Academic Publisher, 1994.
- [Bozkurt & Dutoit, 2003] B. Bozkurt, T. Dutoit, 'Mixed-phase speech modeling and formant estimation, using differential phase spectrums.' *Proc. ISCA ITRW VOQUAL*. Geneva, 2003, pp. 21-24.
- [Bozkurt et al, 2004 a] B. Bozkurt, B. Doval, C. D'Alessandro, T. Dutoit, 'Appropriate windowing for group delay analysis and roots of Z-transform of speech signals.' *Proc. Eusipco*, Vienna, 2004.
- [Bozkurt et al, 2004 b] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, 'A method for glottal formant frequency estimation.' *Proc. Interspeech-ICSLP*. Korea, 2004.



- [Bozkurt *et al*, 2004 c] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, 'Improved differential phase spectrum processing for formant tracking.' *Proc. Interspeech-ICSLP*, Korea, 2004.
- [Bozkurt *et al*, 2004 d] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, 'Zeros of z-transform (ZZT) decomposition of speech for source-tract separation.' *Proc. Interspeech-ICSLP*, Korea, 2004.
- [Bozkurt *et al*, 2004 e] B. Bozkurt, F. Severin, and T. Dutoit, 'An algorithm to estimate anti-causal glottal flow component from speech signals.' Accepted for publication in *'Nonlinear Speech Processing: Algorithms and Analysis'* edited by G. Chollet, A. Esposito, M. Faundez, and M. Marinaro, Springer Verlag.
- [Bozkurt *et al*, 2004 f] B. Bozkurt, T. Dutoit, R. Prudon, C. D'alessandro, V. Pagel, 'Reducing discontinuities at synthesis time for corpus-based speech synthesis.' In *'Text To Speech Synthesis: New Paradigms and Advances'* by S. Narayanan, A. Alwan, eds., Prentice Hall, chap. 1, 2004.
- [Bozkurt *et al*, 2005] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech", *IEEE Signal Processing Letters*, vol. 12, no. 4, pp.344-347, 2005.
- [Bozkurt & Couvreur, 2005] B. Bozkurt, L. Couvreur, 2005, "On the use of phase information for speech recognition", *Proc. EUSIPCO*, Antalya (Turkey), 2005.
- [Bruce *et al*, 2002] I. C. Bruce, N. V. Karkhanis, E. D. Young, and M. B. Sachs, 'Robust formant tracking in noise.' *Proc. ICASSP*, vol. 1, Florida, 2002, pp. 281–284.
- [Campbell & Marumoto, 2000] N. Campbell, T. Marumoto, 'Automatic labeling of voice-quality in speech databases for synthesis.' *Proc. ICSLP*, Beijing, China, 2000, pp.468-471.
- [Charpentier, 1986] F.J. Charpentier, 'Pitch detection using the short-time phase spectrum.' *Proc. ICASSP*, Tokyo, 1986, pp. 113-116.
- [Charpentier, 1988] F.J. Charpentier, '*Traitement de la parole par analyse-synthese de Fourier application a la synthese par diphones.*' PhD dissertation, Ecole Nationale-Superieure des Telecommunications, 1988.
- [Chatwal & Constantinides, 1987] H. S. Chatwal, A. G. Constantinides, 'Speech spectral segmentation for spectral estimation and formant modeling.' *Proc. ICASSP*, Dallas, 1987, pp. 316–319.
- [Chavez *et al*, 2002] S. Chavez, QS Xiang, L. An, 'Understanding phase maps in MRI: a new outline phase unwrapping method.' *IEEE Transactions on Medical Imaging*, vol. 21, issue: 8, 2002, pp. 966 – 977.
- [Chen & Loizou, 2004] B. Chen , P. C. Loizou, 'Formant frequency estimation in noise.' *Proc. ICASSP*, Montreal, 2004, pp. 581-584.
- [Chen & Zebker, 2002] C. W. Chen, H. A. Zebker, 'Phase unwrapping for large SAR interferograms: statistical segmentation and generalized network models.' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, 2002, pp. 1709-1719.
- [Childers, 1995] D. Childers, 'Modeling the glottal volume velocity for three voice types.' *Journal of the Acoustical Society of America(JASA)*, vol. 97(1), 1995, pp. 505-519.
- [Cizek, 1970] V. Cizek, 'Discrete Hilbert Transform.' *IEEE Trans. Audio Electroacoustics*, vol. 18, no. 4, Dec. 1970, pp. 340-343.
- [Costantini *et al*, 1999] M. Costantini, A. Farina, F. Zirilli, 'A fast phase unwrapping algorithm for SAR interferometry.' *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, Issue 1, 1999 , pp. 452 – 460.
- [Cummings & Clements, 1995] K. E. Cummings and M. A. Clements, 'Glottal Models for Digital Speech Processing: A Historical Review and New Results.' *Digital Signal Processing: A Review Journal*, vol. 5, No. 1, January 1995, pp. 21-42.
- [Damara-Venkata *et al*, 2000] N. Damara-Venkata, B. L. Evans, S. R. McCaslin, 'Design of Optimal Minimum-phase Digital FIR Filters Using Discrete Hilbert Transforms.' *IEEE Transactions on Signal Processing*, vol. 48, no. 5, May 2000, pp. 1491-1495.

- [Deller *et al*, 1999] J. Deller, John H. L. Hansen, John G. Proakis, '*Discrete-time Processing of Speech Signals*.' Wiley-IEEE Press, 1999, ISBN: 0-7803-5386-2.
- [Deng & Sun, 1994] L. Deng, Don X. Sun, 'A statistical framework for automatic speech recognition using the atomic units constructed from overlapping articulatory features.' *Journal of the Acoustical Society of America(JASA)*. vol. 95, no. 5, 1994, pp. 2702-2719.
- [Ding & Kasuya, 1996] K. Ding, H. Kasuya, 'A novel approach to the estimation of voice source and vocal tract parameters from speech signals' *Proc. ICSLP*, Philadelphia, 1996. pp. 1257-1260.
- [Doval & d'Alessandro, 1997] B. Doval, C. d'Alessandro, 'Spectral correlates of glottal waveform models: an analytic study.' *Proc. ICASSP*, Munich 1997, pp. 446-452.
- [Doval & d'Alessandro, 1999] B. Doval, C. d'Alessandro, 'The spectrum of glottal flow models.' *Limsi Documentation*, no:99-07, 1999.
- [Doval *et al*, 2003] B. Doval, C. d'Alessandro, and N. Henrich, 'The voice source as a causal/anti-causal linear filter.' *Proc. ISCA ITRW VOQUAL*, Geneva 2003, pp. 15-19.
- [Dutoit & Leich, 1993] T. Dutoit, H. Leich, 'MBR-PSOLA : Text-to-speech synthesis based on an MBE re-synthesis of the segments database.' *Speech Communication*, v. 13, no 3-4, 1993, pp.435-440.
- [Dutoit & Gosselin, 1996] T. Dutoit, B. Gosselin, 'On the use of a hybrid harmonic/stochastic model for TTS synthesis-by-concatenation.' *Speech Communication*, v. 19, 1996, pp.119-1443.
- [Edelman & Murakami, 1995] A. Edelman, H. Murakami, 'Polynomial roots from companion matrix eigenvalues.' *Mathematics of Computation*. vol. 64 , Issue 210, 1995, pp. 763 – 776.
- [Fant, 1960] G. Fant, '*Acoustic Theory of Speech Production*.' Mouton and Co. Netherlands, 1960.
- [Fant, 1995] G. Fant, 'The LF-model revisited. Transformation and frequency domain analysis.' *Speech Trans. Lab.Q.Rep., Royal Inst. of Tech.* vol. 2-3, Stockholm, 1995, pp 121-156.
- [Fant, 1997] G. Fant, 'The voice source in connected speech.' *Speech Communication*, vol. 22, 1997, pp.125-139.
- [Friedman, 1985] D. H. Friedman, 'Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech.' *Proc. ICASSP*, 1985, pp. 1121 – 1124.
- [Frolova & Taxt, 1996] G. V. Frolova, T. Taxt, 'Homomorphic deconvolution of medical ultrasound images using a Bayesian model for phase unwrapping.' *Proc. of Ultrasonics Symposium*, vol.2, San Antonio, TX, 1996, pp. 1371 – 1376.
- [Gardner, 1994] W. R. Gardner, '*Modeling and Quantization Techniques for Speech Compression Systems*.' PhD dissertation, University of California, San Diego, 1994.
- [Gardner & Rao, 1997] W. R. Gardner, B. D. Rao, 'Noncausal all-pole modeling of voiced speech.' *IEEE Trans. Speech and Audio Processing*, vol.5, no.1, 1997, pp. 1-10.
- [Garofolo *et al*, 1986] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. Pallett, , and N. L. Dahlgren, 'The darpa timit acoustic-phonetic continuous speech corpus.' *Tech. Rep. Speech Disc CD1-1.1*, NIST, Gaithersburg, MD, 1986.
- [Gauffin & Sundberg, 1989] J. Gauffin, J. Sundberg, 'Spectral correlates of glottal voice source waveform characteristics.' *Journal of Speech and Hearing Research*, vol. 32, 1989, pp 556-565.
- [George & Smith, 1997] E. B. George, M. J. T. Smith, 'Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model.' *IEEE Trans. on Speech and Audio Processing*, vol. 5, September 1997, pp. 389-406.
- [Gobl & Chasaide, 2001] C. Gobl, A. Ní Chasaide, 'Dynamics of the glottal source signal: implications for naturalness in speech synthesis.' *Improvements in Speech Synthesis*, Chapter 27. Wiley & Sons, Chichester, UK. 2001.
- [Gobl, 2003] C. Gobl, '*The voice source in speech communication*.' PhD Thesis, KTH Speech Music and Hearing, Stockholm, 2003.
- [Golub & Loan, 1996] G. H. Golub, C. F. Van Loan, '*Matrix Computations*.' Johns Hopkins University Press, 3rd Edition, 1996.

- [Griffin & Lim, 1988] D.W. Griffin, J.S. Lim, 'Multi-band excitation vocoder.' *IEEE Trans. Acoustics, Speech, and Signal Processing*, 36(8), 1988, pp. 1223–1235.
- [Hansen & Engberg, 1996] A.V. Hansen, I.S. Engberg, 'Documentation of the Danish emotional speech database', 1996, Aalborg University.
- [Hanson *et al*, 1994] H. M. Hanson, P. Maragos, and A. Potamianos, 'A system for finding speech formants and modulations via energy separation.' *IEEE Transactions on Speech and Audio Processing*, vol. 2, 1994, pp. 436-443.
- [Hanson, 1995] H.M Hanson, 'Glottal Characteristics of Female Speakers.' Ph.D. Thesis, Harvard University, 1995.
- [Hanson & Chuang, 1999] H. M. Hanson, E. S. Chuang, 'Individual variations in glottal characteristics of female speakers.' *Journal of the Acoustical Society of America(JASA)*, vol. 106, no. 2, 1999, pp. 1064-1077.
- [Harris, 1978] F. J. Harris, 'On the use of windows for harmonic analysis with the Discrete Fourier Transform.' *Proc. of the IEEE*, vol. 66, no. 1, 1978, pp. 51-83.
- [Hedelin, 1988] P. Hedelin, 'Phase compensation in all-pole speech analysis.' *Proc. ICASSP*, 1988, pp. 339-342.
- [Hegde *et al*, 2004 a] R. M. Hegde, H. A. Murthy and V. R. Gadde, 'The modified group delay feature: A new spectral representation of speech.' *Proc. of the Interspeech-ICSLP*, Korea 2004.
- [Hegde *et al*, 2004 b] R. M. Hegde, H. A. Murthy and V. R. Gadde, 'Continuous speech recognition using joint features derived from the modified group delay function and MFCC.' *Proc. of the Interspeech-ICSLP*, Korea 2004.
- [Hegde *et al*, 2004 c] R. M.Hegde, H. A.Murthy and V. R. Gadde, 'Application of the modified group delay function to speaker identification and discrimination.' *Proc. ICASSP*, vol. 1, 2004, pp. 517-520.
- [Helmholtz, 1875] H.L.F von Helmholtz, 'On the Sensations of Tone.' 1875, English translation by A.J. Ellis, Longmans, London, 1912.
- [Henrich *et al*, 1999] N. Henrich, B. Doval, C. d'Alessandro, 'Glottal open quotient estimation using linear prediction.' *Proc. Inter. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, 1999.
- [Henrich *et al*, 2000] N. Henrich, B. Doval, C. d'Alessandro and M. Castellengo, 'Open quotient measurements on EGG, speech and singing signals.' *Proc. 4th Int. Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research*, Jena, 2000.
- [Henrich, 2001] N. Henrich, 'Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception.' PhD thesis, Université Paris 6, 2001.
- [Hirsch & Pearce, 2000] H. G. Hirsch, D. Pearce, 'The AURORA experimental framework for the performance evaluation of speech recognition Systems under noisy conditions.' *Proc. ASR 2000*, Paris, Sep. 2000.
- [Huang *et al*, 2001] X. Huang, A. Acero and H.W. Hon, 'Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.' Prentice Hall, 2001.
- [Jackson, 1989] L.B.Jackson, 'Noncausal ARMA modeling of voiced speech.' *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. 37, no.10, 1989, pp. 1606-1608.
- [Kawahara, 1997] H.Kawahara, 'Speech representation and transformation using adaptive interpolation of weighted spectrum: VOCODER revisited.' *Proc. ICASSP*, 1997, pp. 1303-1306.
- [Kawahara *et al*, 2000] H. Kawahara, Y. Atake, and P. Zolfaghari, 'Accurate vocal event detection method based on a fixed-point to weighted average group delay.' *Proc. ICSLP*, Beijing, 2000, pp. 664–667.
- [Kawahara *et al*, 2001] H. Kawahara, J. Estill, O. Fujimura, 'Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT.' *Proc. MAVEBA*, Firenze Italy, September 2001.

- [Kawai & Tsuzaki, 2004] H. Kawai and M. Tsuzaki, 'Voice quality variation in a long-term recording of a single speaker speech corpus.' In *Text to Speech Synthesis: New Paradigms and Advances* by Narayanan, S. and Alwan, A. (eds.). Prentice Hall, 2004.
- [Kopec *et al*, 1977] G. Kopec, A.V. Oppenheim and J. Tribolet. 'Speech analysis by homomorphic prediction.' *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 1, 1977, pp. 40-49.
- [Klabbers, 2000] E. Klabbers, '*Segmental and prosodic improvements to speech generation.*' PhD Thesis, Eindhoven University of Technology (TUE), 2000.
- [Klatt & Klatt, 1990] D.H. Klatt, L.C. Klatt, 'Analysis, synthesis, and perception of voice quality variations among female and male talkers.' *Journal of the Acoustical Society of America*, 87(2), February 1990, pp.820-57.
- [Kopec, 1986] G. E. Kopec, 'Formant tracking using hidden Markov models and vector quantization.' *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, 1986, pp. 709-729.
- [Laprie & Berger, 1996] Y. Laprie, M.O. Berger. 'Cooperation of regularization and speech heuristics to control automatic formant tracking.' *Speech Communication*, vol. 19, no. 4, 1996, pp. 255–270.
- [Laprie, 2004] Y. Laprie, 'A concurrent curve strategy for formant tracking.' *Proc. Interspeech-ICSLP*, Korea, 2004.
- [Li & Levinson, 2002] D. Li, S.E. Levinson, 'A linear phase unwrapping method for binaural sound source localization on a robot.' *Proc. of ICRA '02. IEEE International Conference on Robotics and Automation*, vol.1, 2002, pp. 19 – 23.
- [Liu *et al*, 1997] L. Liu, J. He, G. Palm, 'Effects of phase on the perception of intervolic stop consonants.' *Speech Communication*, vol. 22, 1997, pp. 403-417.
- [Lu & Smith, 1999] H.L. Lu and J. O. Smith, 'Joint estimation of vocal tract filter and glottal source waveform via convex optimization.' *Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics(ICASSP)*, New Paltz, New York, 1999.
- [Lu, 2002] H.L. Lu, '*Toward a high-quality singing synthesizer with vocal texture control.*' PhD Thesis, Stanford University, 2002.
- [Macon, 1996] M. W. Macon, '*Speech and voice synthesis based on sinusoidal modeling.*' PhD Thesis, Oregon Graduate Institute, 1996.
- [Makhoul, 1975] J. Makhoul, 'Linear prediction: A tutorial review.' *Proc. IEEE*, vol. 63, 1975, pp 561-580.
- [Makhoul, 1977] J. Makhoul, 'Lattice methods for linear prediction.' *IEEE Trans. On Acoustics, Speech, And Signal Processing*. Vol.25, 1977, pp. 423-428.
- [Marques, 1989] J. S. Marques, '*Sinusoidal modeling of speech: Application to medium to low bit rate coding.*' , PhD Thesis, Technical University of Lisbon, 1989.
- [Marques *et al*, 1990] J. S. Marques, L. B. Almeida and J. M. Tribolet, 'Harmonic coding at 4.8 KP/S.' *Proc. ICASSP*, 1990, pp.17-20.
- [McAulay & Quatieri, 1986] R.J. McAulay, T.F. Quatieri, 'Speech analysis/synthesis based on a sinusoidal representation.' *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP*-34(4), pp. 744-754.
- [McAulay & Quatieri, 1987] R. J. McAulay, T. F. Quatieri, 'Multirate sinusoidal transform coding at rates from 2.4 KBPS to 8 KBPS.' *Proc. ICASSP*, 1987, pp. 38.7.1-38.7.4.
- [McAulay & Quatieri, 1991] R. J. McAulay and T. F. Quatieri, 'Sinusoidal coding.' In *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), ch. 4, Marcel Dekker, 1991, pp. 165-172.
- [McCandless, 1974] S. McCandless, 'An algorithm for automatic formant extraction using linear prediction spectra.' *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 22, 1974, pp. 135–141.
- [Metz *et al*, 1991] S.W. Metz, J.A. Heinen, R.J. Niederjohn, T.V. Sreenivas, 'Auditory modeling applied to formant tracking of noise-corrupted speech.' *Proc. of the International Conference on Industrial Electronics, Control and Instrumentation*, vol. 3, 1991, pp. 2120–2124.

- [Milenkovic, 1986] P. Milenkovic, 'Glottal inverse filtering by joint estimation of an AR system with a linear input model.' *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 31, 1986, pp. 28-42.
- [Mobijs, 2000] B. Mobijs, 'Corpus-based speech synthesis: methods and challenges.' In '*Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung*', vol. 6, no. 4, 2000, pp. 87-116.
- [Moriya & Honda, 1986] T. Moriya, M. Honda, 'Speech coder using phase equalization and vector quantization.' *Proc. ICASSP*, 1986, pp. 1701-1704.
- [Murthy *et al*, 1989 a] H. A. Murthy, K. V. Murthy, and B. Yegnanarayana, 'Formant extraction from phase using weighted group delay function.' *Electronics Letters*, vol.25, no.23, 1989, pp. 1609-1611.
- [Murthy & Yegnanarayana, 1989 b] K. V. M. Murthy, B. Yegnanarayana, 'Effectiveness of representation of signals through group delay functions.' *Signal Processing*, vol.17, no.2, 1989, pp. 141-150.
- [Murthy & Yegnanarayana, 1991 a] H A. Murthy, B. Yegnanarayana, 'Formant extraction from group delay function.' *Speech Communication*, vol.10, no.3, 1991, pp. 209-221.
- [Murthy & Yegnanarayana, 1991 b] H. A. Murthy, B. Yegnanarayana, 'Speech processing using group delay functions.' *Signal Processing*, vol. 22, no. 3, 1991, pp. 259-267.
- [Murthy & Yegnanarayana, 1999] P.S. Murthy, B. Yegnanarayana, 'Robustness of group delay based method for extraction of significant instants of excitation in speech using group delay function.' *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, 1999, pp. 609-619.
- [Mustafa, 2003] K. Mustafa, '*Robust Formant Tracking for Continuous Speech with Speaker Variability*.' MS Thesis. McMaster Univ. December 2003.
- [Oliveira, 1993] C. Oliveira, 'Estimation of source parameters by frequency analysis.' *Proc. of the Eurospeech*, Berlin, 1993, pp.99-102.
- [Oppenheim & Schafer, 1968] A. V. Oppenheim, R. W. Schafer, 'Homomorphic analysis of speech.' *IEEE Transactions on Audio and Electroacoustics*, vol. 16, 1968, pp. 221-226.
- [Oppenheim, 1969] A. V. Oppenheim, 'A speech analysis-synthesis system based on homomorphic filtering,' *Journal of the Acoustical Society of America (JASA)*, vol. 45, no. 2, February 1969, pp. 458-465.
- [Oppenheim *et al*, 1976] A. V. Oppenheim, G. Kopec and J. Tribolet, 'Signal analysis by homomorphic prediction.' *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, 1976, pp. 327-332.
- [Oppenheim *et al*, 1999] A. V. Oppenheim and R. W. Schafer, with J.R. Buck. '*Discrete-Time Signal Processing*.' Second Edition. Prentice-Hall, Inc.: Upper Saddle River, NJ, 1999.
- [Paliwal & Alsteris, 2003] K. K. Paliwal, L. Alsteris, 'Usefulness of phase spectrum in human speech perception.' *Proc. Eurospeech*, Geneva, 2003, pp. 2117-2120.
- [Papoulis, 1962] A. Papoulis, '*The Fourier Integral and Its Applications*.' New York: McGraw-Hill, 1962, pp. 198-203.
- [Patterson, 1987] R.D. Patterson, 'A pulse ribbon model of monaural phase perception.' *Journal of the Acoustical Society of America*, vol. 82, no. 5, 1987, pp. 1560-1586.
- [Pescic, 2003] P. Pescic, '*Abel's Proof: An Essay on the Sources and Meaning of Mathematical Unsolvability*'. MIT Press, ISBN 0262162164.
- [Plumpe & Quatieri, 1999] M. D. Plumpe, T. F. Quatieri, 'Modelling of the glottal flow derivative waveform with application to speaker identification.' *IEEE Trans. on Speech and Audio*, vol. 7, no. 5, Sept 1999, pp. 569-585.
- [Pobloth & Kleijn, 1999] H. Pobloth, W. B. Kleijn, 'On phase perception in speech.' *Proc. ICASSP*, 1999, pp. 29-32.

- [Pollard, 1997] M. P. Pollard, B. M. G. Cheetham, M. D. Edgington, 'Shape invariant pitch and time-scale modification of speech by variable order phase interpolation.' *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 919-922.
- [Potamianos & Maragos, 1996] A. Potamianos, P. Maragos, 'Speech formant frequency and bandwidth tracking using multiband energy demodulation.' *Journal of the Acoustical Society of America (JASA)*, vol. 99, 1996, pp. 3795-3806.
- [Potamianos & Maragos, 1999] A. Potamianos P. Maragos, 'Speech analysis and synthesis using an AM-FM modulation model.' *Speech Communication*, vol. 28, 1999, pp. 195-209.
- [Quatieri, 1979] T. F. Quatieri, 'Minimum and mixed-phase speech analysis-synthesis by adaptive homomorphic deconvolution.' *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no.4, 1979. pp. 328-335.
- [Quatieri & McAulay, 1989] T. F. Quatieri, R. J. McAulay, 'Phase coherence in speech reconstruction for enhancement and coding applications.' *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Glasgow, Scotland, May 1989
- [Quatieri, 2002] T. F. Quatieri, '*Discrete-time speech signal processing: Principles and Practice*.' Uper Saddle River, NJ: Prentice-Hall, 2002.
- [Rabiner *et al*, 1969] L.R.Rabiner, R.W.Schafer and C.M.Rader, 'The chirp z-transform algorithm and its application.' In *Bell System Tech. J.* vol. 48, 1969, pp. 1249-1292.
- [Rabiner & Schafer, 1978] L. R. Rabiner, R. W. Schafer, '*Digital processing of speech signals*.' Englewood Cliffs, NJ: Prentice-Hall, 1978, pp. 365-372.
- [Rao & Kumaresan, 2000] A. Rao, R. Kumaresan, 'On decomposing speech into modulated components.' *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, 2000, pp. 240-254.
- [Riegelsberger *et al*, 1993] E. L. Riegelsberger, A. K. Krishnamurthy, 'Glottal source estimation: Methods of applying the LF model to inverse filtering.' *Proc. ICASSP*. Minneapolis, 1993, pp. 542-545.
- [Rothenberg, 1973] M. Rothenberg, 'A new inverse filtering technique for deriving the glottal airflow during voicing.' *Journal of the Acoustical Society of America (JASA)*, vol. 53, , 1973, pp. 1632-1645.
- [Rothenberg, 1981] M. Rothenberg, 'Acoustic interaction between the glottal source and the vocal tract' In *Vocal Fold Physiology* edited by Stevens, K. N. and Hirano, M., Tokyo: Univ. of Tokyo Press, pp. 305-328.
- [Schafer & Rabiner, 1970] R.W.Schafer, L.R. Rabiner, 'System for automatic formant analysis of voiced speech.' *Journal of Acoustical Society America*, vol. 47, no. 2, 1970 pp. 634-648.
- [Schroeder, 1959] M.R. Schroeder, 'New results concerning monaural phase sensitivity.' *Journal of the Acoustical Society of America*, 31, 1959, pp.1597.
- [Schroeder & Strube, 1986] M. R. Schroeder, H.W. Strube, 'Flat-spectrum speech.' *Journal of the Acoustical Society of America*. vol. 79, no. 5, 1986, pp.1580-1583.
- [Sitton *et al*, 2003] G.A. Sitton, C.S. Burrus, J.W. Fox, S. Treitel, 'Factoring very-high-degree polynomials.' *Signal Processing Magazine*, IEEE. vol. 20, Iss. 6, 2003, pp. 27 – 42.
- [Smits & Yegnanarayana, 1995] R. Smits, B. Yegnanarayana, 'Determination of instants of significant excitation using group delay function.' *IEEE Trans. Speech Audio Processing*, vol. 3, 1995, pp. 325-333.
- [Snell & Milinazzo, 1993] R. C. Snell, F. Milinazzo, 'Formant location from LPC analysis data.' *IEEE Trans. Speech Audio Processing*, vol. 1, 1993 pp. 129-134.
- [Spanias, 1994] A.S. Spanias, 'Speech coding: a tutorial review.' *Proc. of the IEEE*, 1994, pp: 1541-1582.
- [Sproat & Olive, 1995] R. Sproat, J. Olive, 'An approach to text-to-speech synthesis.' In *Speech Coding and Synthesis*, by Kleijn *et al*. Elsevier 1995, pp. 611-633.

- [Strik, 1998] H. Strik, 'Automatic parameterization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses.' *Journal of the Acoustical Society of America (JASA)*, vol. 103, no. 5, 1998, pp. 2659-2669.
- [Stylianou, 1996 a] Y. Stylianou, 'Decomposition of speech signals into a deterministic and a stochastic part.' *Proc. of ICSLP '96*, vol. 2, Philadelphia, 1996, pp. 1213-1216.
- [Stylianou, 1996 b] Y. Stylianou, '*Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis. Ecole Nationale Supérieure des Télécommunications, 1996.
- [Stylianou, 1998 a] Y. Stylianou, 'Concatenative speech synthesis using a harmonic plus noise model.' *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.
- [Stylianou, 1998 b] Y. Stylianou, 'Removing phase mismatches in concatenative speech synthesis.' *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998, pp. 267-272.
- [Stylianou, 1999] Y. Stylianou, 'Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis.' *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, Arizona, 1999, pp.377-380.
- [Stylianou, 2001] Y. Stylianou, 'Removing linear phase mismatches in concatenative speech synthesis.' *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, March 2001, pp.232-239.
- [Sun, 1995] D. X. Sun, 'Robust estimation of spectral center of gravity trajectories using mixture spline models.' *Proc. of the 4th European Conference on Speech Communication and Technology*. Madrid, 1995, pp. 749-752.
- [Sun, 1997] X. Sun, 'Phase modeling of speech excitation for low bit-rate sinusoidal transform coding.' *Proc. ICASSP*, 1997, pp. 1691-1694.
- [Talkin, 1987] D. Talkin, 'Speech formant trajectory estimation using dynamic programming with modulated transition costs.' *Journal of the Acoustical Society of America (JASA)*, S1, March 1987, pp. S55.
- [Thomson, 1988] D. L. Thomson. 'Parametric models of the magnitude/phase spectrum for harmonic speech coding.' *Proc. ICASSP*, 1988, pp.378-381.
- [Traunmüller & Eriksson, 1997] H. Traunmüller, A. Eriksson, 'A method of measuring formant frequencies at high fundamental frequencies.' *Proc. EuroSpeech'97*, vol.1, pp. 477 – 480.
- [Vyacheslav & Zhu, 2003] V. Vyacheslav, Y. Zhu, 'Deterministic phase unwrapping in the presence of noise.' *Optics Letters*, vol. 28, no. 22, November, 2003, pp. 2156-2158.
- [Watanabe, 2001] A. Watanabe, 'Formant estimation method using inverse-filter control.' *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, 2001, pp. 317-326.
- [Welling & Ney, 1998] L. Welling, H. Ney, 'Formant estimation for speech recognition.' *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, January 1998, pp. 36-48.
- [Wong et al, 1979] D. Y. Wong, J. D. Markel and J. Augustine H. Gray, 'Least squares glottal inverse filtering from the acoustic waveform.' *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, August 1979, pp. 350-355.
- [Xia & Epsy-Wilson, 2000] K. Xia, C. Epsy-Wilson, 'A new strategy of formant tracking based on dynamic programming.' *Proc. ICSLP*, Beijing, 2000.
- [Yan et al, 2004] Q. Yan, E. Zavarehei, S. Vaseghi, D. Rentzos, 'A formant tracking LP model for speech processing in car/train noise.' *Proc. Interspeech-ICSLP*, Korea. 2004.
- [Yegnanarayana et al, 1984] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, 'Significance of group delay functions in signal reconstruction from spectral magnitude or phase.' *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.32, no.3, June 1984, pp. 610-623.
- [Yegnanarayana et al, 1988] B. Yegnanarayana, G. Duncan, H. A. Murthy, 'Improving formant extraction from speech using minimum-phase group delay spectra.' *Proc. EUSIPCO*, Grenoble, France, 1988, pp. 447-450.

- [Yegnanarayana & Murthy, 1992] B. Yegnanarayana, H. A. Murthy, 'Significance of group delay functions in spectrum estimation.' *IEEE Trans. on Signal Processing*, vol. 40, no. 9, Sept. 1992, pp. 2281-2289.
- [Yegnanarayana *et al*, 1998] B. Yegnanarayana, C. d'Alessandro, V. Darsinos, 'An iterative algorithm for decomposition of speech signals into periodic and aperiodic components.' *IEEE Transaction on Speech and Audio Processing*, vol. 6 (1):1-11, 1998.
- [You, 2004] H. You, 'Application of long-term filtering to formant estimation.' *Proc. Interspeech-ICSLP*, Korea, 2004.
- [Zheng & Hasegawa-Johnson, 2003] Y. Zheng, M. Hasegawa-Johnson, 'Particle filtering approach to Bayesian formant tracking.' *Proc. IEEE Workshop on Statistical Signal Processing*, 2003.
- [Zhu & Paliwal, 2004] D. Zhu, K. K. Paliwal, 'Product of power spectrum and group delay function for speech recognition.' *Proc. ICASSP*, Montreal, 2004, pp. 125-128.
- [Zolfaghari *et al*, 2003] P. Zolfaghari, T. Nakatani, T. Irino, H. Kawahara, and F. Itakura, 'Glottal closure instant synchronous sinusoidal model for high quality speech analysis/synthesis.' *Proc. of Eurospeech*, Geneva, 2003. pp. 2441-2444.
- [Zolfaghari & Robinson, 1996] P. Zolfaghari, T. Robinson, 'Formant analysis using mixtures of gaussians.' *Proc. ICSLP*, Philadelphia, 1996, pp. 1229-1232.
- [www-DarpaDBA] [http://www ldc.upenn.edu/readme\\_files/timit.readme.html](http://www ldc.upenn.edu/readme_files/timit.readme.html)
- [www-Ellis] D. Ellis, Lecture notes on speech production : <http://www.ee.columbia.edu/~dpwe/e6820/>
- [www-Mbrola] <http://tcts.fpms.ac.be/synthesis/mbrola.html>
- [www-Strut] J.-M. Boite, L. Couvreur, S. Dupont and C. Ris, Speech Training and Recognition Unified Tool (STRUT), <http://tcts.fpms.ac.be/asr/project/strut>.
- [www-Praat] <http://www.praat.org>
- [www-Voqual03] <http://www.limsi.fr/VOQUAL>
- [www-WinSnoori] <http://www.loria.fr/~laprie/WinSnoori/>



