Technological University Dublin

ARROW@TU Dublin

Doctoral

Engineering

2012-2

Phase-Distortion-Robust Voice-Source Analysis

Alan O'Cinneide

Technological University Dublin

Follow this and additional works at: https://arrow.tudublin.ie/engdoc

Part of the Engineering Commons

Recommended Citation

O Cinneide, A. (2012) *Phase-Distortion-Robust-Voice-Source Analysis.* Doctoral Thesis. Dublin, Technological University Dublin. doi:10.21427/D7FW3B

This Theses, Ph.D is brought to you for free and open access by the Engineering at ARROW@TU Dublin. It has been accepted for inclusion in Doctoral by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, vera.kilshaw@tudublin.ie.

Phase-Distortion-Robust Voice-Source Analysis

Alan Ó Cinnéide, M.A., B.A.

February 2012

Thesis submitted to the Dublin Institute of Technology

for the degree of Doctor of Philosophy

Supervisors:	Dr. David Dorran
	Dr. Mikel Gainza
	Prof. Eugene Coyle

Abstract

This work concerns itself with the analysis of voiced speech signals, in particular the analysis of the glottal source signal. Following the source-filter theory of speech, the glottal signal is produced by the vibratory behaviour of the vocal folds and is modulated by the resonances of the vocal tract and radiation characteristic of the lips to form the speech signal. As it is thought that the glottal source signal contributes much of the non-linguistic and prosodical information to speech, it is useful to develop techniques which can estimate and parameterise this signal accurately.

Because of vocal tract modulation, estimating the glottal source waveform from the speech signal is a blind deconvolution problem which necessarily makes assumptions about the characteristics of both the glottal source and vocal tract. A common assumption is that the glottal signal and/or vocal tract can be approximated by a parametric model. Other assumptions include the causality of the speech signal: the vocal tract is assumed to be a minimum phase system while the glottal source is assumed to exhibit mixed phase characteristics. However, as the literature review within this thesis will show, the error criteria utilised to determine the parameters are not robust to the conditions under which the speech signal is recorded, and are particularly degraded in the common scenario where low frequency phase distortion is introduced. Those that are robust to this type of distortion are not well suited to the analysis of real-world signals.

This research proposes a voice-source estimation and parameterisation technique, called the Power-spectrum-based determination of the R_d parameter (PowRd) method. Illustrated by theory and demonstrated by experiment, the new technique is robust to the time placement of the analysis frame and phase issues that are generally encountered during recording. The method assumes that the derivative glottal flow signal is approximated by the transformed Liljencrants-Fant model and that the vocal tract can be represented by an all-pole filter. Unlike many existing glottal source estimation methods, the PowRd method employs a new error criterion to optimise the parameters which is also suitable to determine the optimal vocal-tract filter order.

In addition to the issue of glottal source parameterisation, nonlinear phase recording conditions can also adversely affect the results of other speech processing tasks such as the estimation of the instant of glottal closure. In this thesis, a new glottal closing instant estimation algorithm is proposed which incorporates elements from the state-of-the-art techniques and is specifically designed for operation upon speech recorded under nonlinear phase conditions. The new method, called the Fundamental RESidual Search or FRESS algorithm, is shown to estimate the glottal closing instant of voiced speech with superior precision and comparable accuracy as other existing methods over a large database of real speech signals under real and simulated recording conditions.

An application of the proposed glottal source parameterisation method and glottal closing instant detection algorithm is a system which can analyse and re-synthesise voiced speech signals. This thesis describes perceptual experiments which show that, under linear and nonlinear recording conditions, the system produces synthetic speech which is generally preferred to speech synthesised based upon a state-of-the-art time-domain-based parameterisation technique.

In sum, this work represents a movement towards flexible and robust voice-source analysis, with potential for a wide range of applications including speech analysis, modification and synthesis.

Declaration

I certify that this thesis, which I now submit for examination for the award of PhD, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Dublin Institute of Technology (DIT) and has not been submitted in whole or in part for another award in any other third level institution.

The work reported on in this thesis conforms to the principles and requirements of the DIT's guidelines for ethics in research. DIT has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature:

Date:

Acknowledgements

Firstly, I would like to thank my supervisors, Drs. David Dorran and Mikel Gainza, for their guidance throughout my time at DIT. The research contained within this document, and indeed the quality of my work generally, has been greatly improved from their technical, structural and stylistic insights. I'd also like to extend my thanks to Prof. Eugene Coyle, who offered support from afar.

Special thanks goes to my colleagues at the Audio Research Group: Dan Barry, Derry FitzGerald, Martin Gallagher, Rajesh Jaiswal and Cillian Kelly. The group provided a stimulating working atmosphere and amusing conversation, whether discussing the details of Fourier analysis or the minutiae of music theory. Indeed, I must acknowledge colleagues from the research community generally, here in Dublin and abroad, who have been a source of inspiration for me.

I'm also very grateful for my wonderful family who have supported me through my research career and my entire life generally. I'm very proud to be a part of it.

Finally, I owe a huge debt of gratitude to my *schatje*, Chantal de Kleijn. She sustained me throughout the whole ordeal of postgraduate research, and in many ways this work is also hers. Thank you, my love - without you, I would not have been able.

For my Dad, Breanndán Ó Cinnéide

Contents

List of Abbreviations

ϕ LS	Least-squares phase
AEVT	Adaptive Estimation of the Vocal Tract transfer function
ARMAX	AutoRegressive-Moving Average eXogenous
ARX	AutoRegressive eXogenous
ARX-LF	ARX model of speech with LF model of the derivative glottal pulse
CALM	Causal-Anti-causal Linear Model of glottal flow
CCD	Complex Cepstrum Decomposition
CPIF	Closed-Phase Inverse Filtering
DAP	Discrete All-Pole
DEGG	Derivative ElectroGlottoGraph
DFT	Discrete Fourier Transform
DIT	Dublin Institute of Technology
DSP	Digital Signal Processing

DTFT	Discrete Time Fourier Transform
DYPSA	DYnamic programming projected Phase Slope Algorithm
EGG	ElectroGlottoGraph
EGG	Electroglottograph
FAR	False Alarm Rate
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FRESS	Fundamental RESidual Search
GCI	Glottal Closing Instant
GELP	Glottal Excited Linear Prediction
GOI	Glottal Opening Instant
GSBIF	Glottal-Spectrum-Based Inverse Filtering
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
I-S	Itakura-Saito distance measure
IAIF	Iterative Adaptive Inverse Filtering
IIR	Infinite Impulse Response

IPA	International Phonetic Alphabet
IR	Identification Rate
LF	Liljencrants-Fant
LPC	Linear Predictive Coding
MA	Moving Average
MOS	Mean Opinion Score
MR	Miss Rate
MSP	Mean Squared Phase
PowARXLF	Power-Spectrum-based ARX-LF parameterisation of voiced speech
PowRd	Power spectrum determination of the R_d parameter
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SIGMA	Singularity in EGG by Multiscale Analysis
SIM	Simultaneous Inverse filtering and Model matching
SWIPE'	Sawtooth Waveform Inspired Pitch Estimator using prime harmonics
SWLP	Stabilised Weighted Linear Prediction
SWT	Stationary Wavelet Transform
TD-LS	Time-Domain Least-Squares

YAGA	Yet Another GCI/GOI Algorithm
ZFR	Zero Frequency Resonator
ZZT	Zeros of the Z-Transform

List of Tables

List of Figures

List of Publications

The following is a list of publications relating to and arising from the work within this thesis.

- A Frequency-Domain Approach to ARX-LF Voiced Speech Parameterisation and Synthesis, Ó Cinnéide, Alan; Dorran, David; Gainza, Mikel; Coyle, Eugene, *Interspeech 2011*, Florence, Italy, 2011.

- Glottal Inverse Filtering with Automatic Filter Order Selection, Ó Cinnéide, Alan; Dorran, David; Gainza, Mikel; Coyle, Eugene, *ISSC 2011*, Trinity College Dublin, Ireland, 2011.

- Exploiting Glottal Formant Parameters for Glottal Inverse Filtering and Parameterisation, Ó Cinnéide, Alan; Dorran, David; Gainza, Mikel; Coyle, Eugene, *Interspeech 2010*, Chiba, Japan, 2010.

- Towards a Method to Determine the Glottal Formant Parameters of Voiced Speech without Time-Domain References, Ó Cinnéide, Alan; Dorran, David; Gainza, Mikel; Coyle, Eugene, *ISSC 2010*, University College Cork, Ireland, 2010.

- On the Appearance of a Positive Real Pole in the Results of Glottal Closed-Phase Linear Prediction, Ó Cinnéide, Alan; Dorran, David; Gainza, Mikel; Coyle, Eugene, *EUSIPCO 2010*, Aalborg, Denmark, 2010.

Chapter 1

Introduction

Speech is a complex, information-dense acoustic pressure wave. In addition to the basic lexical contents of the message, the speech signal also contains information beyond the linguistic level. This paralinguistic information conveys details about the physical and emotional state of the speaker in a manner which is intertwined with the spoken words. Following analog-to-digital conversion, the samples of the speech signal, and indirectly the linguistic and paralinguistic information represented by those samples, can be processed and analysed using digital signal processing (DSP) techniques.

This thesis describes DSP techniques which estimate and analyse the voice-source signal. According to the source-filter theory of speech production (Fant, 1970), voiced speech may be separated into a source signal resulting from the periodic vibrations of the vocal folds and the response of the vocal-tract filter. The voice-source signal, often referred to as the glottal source signal as it results from the opening and closing of the space between the vocal folds called the glottis, provides the acoustic energy source for speech. This signal is thought to be the main conveyer of the paralinguistic information within speech (Gobl, 2003).

Estimation and parameterisation of the voice signal is useful for a wide range speech processing applications. Below is an list of various applications:

Speech Synthesis As was found in (Rosenberg, 1970), the inclusion of more accurate glottal source waveform shape increased the naturalness of synthetic vowel signals produced by a speech formant synthesiser. Recently, this technique is witnessing a resurgence of interest with the advent of HMM-based speech synthesis (Cabral, 2010; Raitio *et al.*, 2011).

Voice Modification Parameterisation of the speech signal allows for straightforward modification of those parameters which can produce various physically related transformations (Childers, 1995; Lu, 2002; Vincent, 2007; Degottex, 2010).

Voice Quality Characterisation (Childers and Lee, 1991) identified factors of glottal flow signal could be used to characterise voice quality.

Speaker Identification In (Plumpe *et al.*, 1997) it was demonstrated that parameterisation of the voice-source signal could be used to supplement vocal tract information in order to increase the accuracy of speaker identification systems.

Voice Conversion Voice conversion is the application of DSP techniques to speech signals for the purposes of converting the characteristics of a source speaker to those of a target speaker (Sündermann, 2008). Researchers have found that the inclusion of glottal source information improves the performance of a voice conversion system in terms of voice quality (del Pozo, 2008; Pérez and Bonafonte,

2011) and the retention of speaker identity (Pérez and Bonafonte, 2011).

Speech Coding Speech coding utilises signal processing techniques in order to reduce the necessary bandwidth for the transmission and storage of speech signals (Spanias, 1994). Low bit rate speech coding is possible using the Glottal Excited Linear Prediction (GELP) speech coder (Hu and Wu, 2000), where the system was shown to be superior to a similar coding scheme in a Mean Opinion Score (MOS) test.

For these applications and others, tools which enable the analysis and parameterisation of the voice-source signal are of great interest to the speech research community.

Intertwined with the problem of estimation and analysis of the voice-source signal, glottal closing instant (GCI) estimation is also a critical issue for many voice-source analyses. Knowledge of the time interval of glottal closure is an important speech processing task, particularly for voice-source estimation and parameterisation (Wong *et al.*, 1979; Fujisaki and Ljungqvist, 1986; Vincent *et al.*, 2005). GCI estimation is useful for other purposes including speech synthesis (Stylianou, 2001) and prosodical modifications (Moulines and Laroche, 1995), as GCIs indicate the relative positions of the glottal pulses. Because of its importance for voice-source analysis and related issues, this thesis will also investigate accurate GCI estimation.

1.1 Thesis Aims and Scope

This thesis proposes DSP methods for the analysis of the voice-source signal which are more robust than existing technologies. In particular, this work focuses upon developing techniques which are insensitive to the presence of low frequency phase distortion commonly imparted by electro-acoustic equipment. This phenomenon disturbs the phase relationship of the components of the speech signal and can change the time-domain signal shape, with little or no perceptual effect. As it is often necessary that the speech signal exhibit a certain preconceived time-domain shape, speech processing algorithms which are robust to this common distortion are of great benefit to the speech research community.

Speech which has been recorded using phase linear equipment is exceptional; indeed, it has been claimed that most electro-acoustic equipment imparts some degree of phase distortion (Doval and d'Alessandro, 2006). Researchers avoid this phenomenon by utilising specialised recording equipment (Lehto *et al.*, 2007). Alternatively, the distortion is corrected by inversely applying a transfer function which approximates the phase response the recording instrumentation (Holmes, 1975; Berouti *et al.*, 1977; Hedelin, 1986; Brookes and Chan, 1994). However, specialised recording equipment is often unavailable and, as will be discussed, correcting the distortion is often unfeasible. For these reasons, speech processing methods which exhibit robustness to the phase spectrum of a signal are potentially very useful.

Many voice-source estimation methods are particularly sensitive to the phenomenon of phase distortion. The analysis of the acoustic voice source is already a difficult problem because of the intrinsic hidden nature of the waveform. The ill-posed question of voice-source estimation then has the consequence that circular logic is necessary to estimate it: before determining the voice source, one must make assumptions regarding some or all of its characteristics. A common assumption is that the glottal signal can be approximated by a time-domain parametric model or exhibit certain time-domain characteristics (*e.g.* a closed phase). As mentioned above, phase distortion makes assumptions based upon the time-domain shape of the signal unreliable. Like many existing voice-source estimation methods, this thesis proposes a method which also utilises parametric voice-source and vocal-tract models. However, rather than adopting a time-domain-based approach, the method proposed in this work operates upon the power spectrum. The method is presented in Chapter 6.

Another common speech processing task which can also be adversely affected by phase distortion is glottal closing instant estimation. Unlike voice-source estimation, these methods do not necessarily rely upon strict assumptions regarding signal shape, however the auxiliary signals which they employed to locate the GCIs occasionally do. A GCI estimation method which is explicitly robust to phase distortion is presented in Chapter 7.

Finally, corroborating results of voice-source analysis algorithms is difficult, owing to the lack of an appropriate benchmark. This work will follow the path taken by other researchers and compare the algorithms with synthetic speech signals whose parameters are known, but also exploit ElectroGlottoGraph (EGG) signals where appropriate. Additionally, the subjective preference of a group of listeners in a perceptual experiment is also utilised to validate the approach presented in this work. As low frequency phase distortion is the focus of this work, the transfer function used to represent this phenomenon have been measured from a professional studio. Additional examples of low-frequency-phase-distorted transfer functions are taken from those described by other researchers.

1.2 Thesis Structure

This thesis contains eight main chapters, which can be roughly segmented into three parts: Background (Chapters 2 to 5), Investigation (Chapters 6 to 8), and Conclusions (Chapter 9). In addition to these sections, there are also four appendices which describes some technical details which were inappropriate for the main thesis body.

Background The Background section informs the reader of the general area of investigation and reviews the state-of-the-art technology.

Chapter 2 discusses the speech production system and introduces the models which are used to engineers to conceptualise it, including the models utilised within this work. Emphasis is given to the source-filter theory of speech which lays the foundation for glottal source estimation. The models which feature in this study are also introduced in this chapter.

A literature review was undertaken of voice-source estimation and parameterisation techniques in Chapter 3. There it is shown that the efficacy of many voice-source parameterisation techniques may be seriously degraded in the presence of low frequency phase distortion. Even in the case of ideal recording conditions, certain techniques may give inaccurate results due to the position of the analysis frame. Additionally, existing power-domain-based approaches are not well designed for real speech signals.

A review of glottal closing instant techniques is given in Chapter 4. Though more robust to phase distortion than voice-source parameterisation techniques, glottal closing instant algorithms may also be degraded. Without certain modifications, the performance of the state-of-the-art can be seriously compromised.

The issues with the state-of-the-art methods for voice-source estimation and GCI estimation are collated in Chapter 5, which then serves as the departure point for the contributions of this study.

Investigation The Investigation section proposes new techniques for voice-source analysis which solve the issues identified in the previous Background section.

A novel power-spectrum-based approach for voice-source estimation and parameterisation is proposed in Chapter 6, and called the PowRd method. By transforming the data, unreliable phase information may be separated from the signal. Similar approaches have exploited this transformation, but have assumed that the filter order necessary for parameterisation was known. Additionally, these methods make no attempt to avoid the high frequency noise which is often present in real speech signals. The PowRd method utilises a novel error criterion which is suitable for the identification of the optimal filter order, in addition to the optimal voice-source parameters. Experiments with real and synthetic speech validate the approach.

A new glottal closing instant estimation method is proposed in Chapter 7. Drawing from the various state-of-the-art algorithms, the new FRESS algorithm estimates the glottal closing instant of voiced speech with superior precision and comparable accuracy as other existing methods over a large database of real speech signals under real and simulated recording conditions. Unlike other algorithms, the method is explicitly robust to any phase disturbances which may have been imparted upon the signal.

Amongst the applications for the proposed voice-source analysis algorithms is speech synthesis. Chapter 8 describes a perceptual experiment where the synthetic speech synthesised with parameters extracted by a power-spectrum-based approach similar to the PowRd method is compared with a time-domain parameterisation approach for a variety of phase conditions, real and simulated. The synthetic speech produced using parameters based upon the power spectrum approach are found to be generally preferred by a group of listeners, further validating and showing the potential of the method for robust speech analysis/synthesis.

Conclusions Finally, Chapter 9 summaries the findings of this research. The main conclusions of the work are drawn and the drawbacks and limitations of the developed techniques indicate possible directions for future work.

1.2.1 Summary of Contributions

Contribution 1: Robust Voice Source Parameterisation Algorithm The first contribution of this work is a glottal source parameterisation method which is robust to phase distortion and which also chooses the optimal filter order. The method is based upon the power spectrum of the speech signal, and is thus not sensitive to the phase spectrum of the speech signal and any distortions that may have been imparted to it. The technique has the considerable advantage that it is robust to the time position of the analysis frame, therefore not requiring accurate timing information regarding pulse location. Additionally, the proposed method attempts to avoid high-frequency noise in the signal by adopting a harmonic plus noise type signal model.

The approach, called the PowRd method, is described in Chapter 6.

In order to determine the filter order, the use of a novel error criterion, the Relative Itakura-Saito error function, is proposed. The usual Itakura-Saito error generally decreases with increasing filter order; this new function does not have this property and can be used to obtain a robust estimate of the order of the vocal tract all-pole filter. This parameter is usually fixed by other vocal-tract filter estimation methods. Experiments demonstrate that the new function lends the PowRd method increased robustness over existing state-of-the-art methods in the typical situation where the filter order is unknown.

Contribution 2: Robust Glottal Closing Instant Estimation Algorithm Another contribution of this work is the FRESS algorithm for glottal closure instant estimation, described in Chapter 7. This approach is an extension of existing GCI detection methods with certain modifications and extensions to improve accuracy of estimation of the epochs of the speech signal which have been recorded under non-ideal conditions. The method uses a low order Infinite Impulse Response (IIR) filter to determine the fundamental sinusoidal signal which oscillates with the fundamental frequency. Landmarks are then extracted from this simple signal, which are then aligned with the peaks of the normalised energy contour signal. The realigned landmarks indicate likely regions of glottal closure. Epoch candidates are extracted from a search for maxima of the low pass Linear Predictive Coding (LPC) residual signal. The most likely sequence of glottal epochs is then determined by a dynamic programming algorithm.

A comparative experiment shows that the FRESS algorithm offers similar accuracy and higher precision than other approaches in both linear and nonlinear phase conditions.

Contribution 3: Speech Analysis/Synthesis System The third contribution of this work is a speech analysis/synthesis system, introduced in Chapter 8. The system smooths the obtained speech parameters of a method similar to the PowRd approach. The synthetic speech is produced using an overlap-add scheme similar to existing approaches.

The system is robust to phase distortion of the analysed speech signal and is therefore suitable for the analysis/synthesis of recorded speech, regardless of the phase characteristics of the recording equipment. A comparative perceptual experiment with 50 listeners demonstrate that the new system is capable of synthesising speech which generally preferred to a similar method based on a time-domain speech parameterisation scheme.

Minor Contribution For the voice-source estimation/parameterization portion of this study, the determination of the frequency-domain information of many LF model pulses is required. This is a computationally demanding operation, owing to the numerous correlation operations required. As a minor contribution, this study describes two methods which substantially improves the speed of these calculations in an informal comparison test. These methods and the experiment are described in Appendix D.

Chapter 2

Speech Anatomy and Models

This chapter presents the fundamental of speech production, its anatomy and models. In order to make the complex operation of speech production both comprehensible and amenable to mathematical analysis, engineers and linguists have developed a model based upon a basic understanding of the physical speech process - the acoustic theory of speech production (Fant, 1970). The model is often referred to as the source-filter model of speech as it broadly parallels the conceptualisation of the speech production system as *phonation* and *articulation*, modeling speech as a phonating source shaped by a filter representing the articulators. Though an acknowledged simplification, this model has experienced success across many areas of speech processing including synthesis, recognition and modification.

Source-filter theory is also the theoretical foundation upon which voice-source analysis rests. The theory supports the voice-source estimation algorithms reviewed in Chapters 3 and 4 in addition to the proposed techniques in 6 and 7.

This chapter will discuss the anatomy of the human speech production apparatus, and detail the process of speech production following the functions of phonation and articulation. Following the introduction of the anatomical structure and function of human speech production, the source-filter model of speech is discussed and some prevalent models for both source and filter are described.

2.1 Anatomy of the Speech Production System

The production of speech is an elaborate process resulting from the motor coordination of many constituent parts. In addition to producing speech, the organs of the speech production system have multiple functions within the body including alimentation and respiration. For the purposes of speech production, these organs can be roughly divided into two categories: those of phonation and those of articulation (Honda, 2007). The speech organs of phonation include the lungs and larynx, while the organs of articulation include the various cavities above the larynx in addition to the tongue, teeth and lips. A sagittal plane representation of these apparatus is shown in Figure 2.1. This section briefly discusses both processes of phonation and articulation and the associated anatomical organs.

2.1.1 Phonation

The word *phonation* derives from the ancient Greek $\varphi\omega\nu\eta$ (pronounced “foni”) meaning “voice”; indeed, the phonatory organs generate the acoustic energy source from which the articulators form spoken speech. The largest organs of phonation are the lungs. The lungs are the primary organ of respiration, but for phonatory purposes the lungs can be considered air reservoirs which through the contraction of the diaphragm

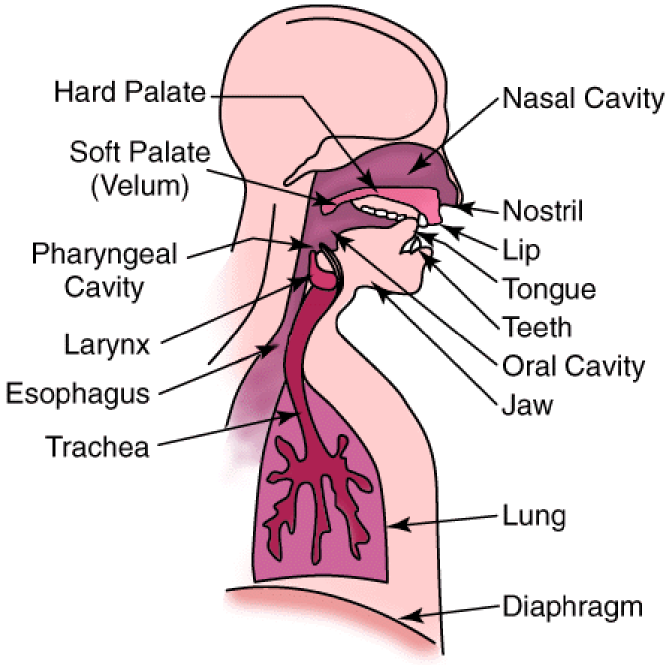


Figure 2.1: A sagittal perspective of the human speech production system, from (Flanagan, 1972).

have the capability to force air through the trachea to the larynx.

The second organ of phonation is the larynx, a diagram of which is given in Figure 2.2. The larynx is an organ in the neck, composed of soft tissue and encased in cartilage. The laryngeal cartilage may protrude under the skin of the neck to form the laryngeal prominence or “Adam’s apple”. Housed within the larynx are the vocal folds, which can be held open or sealed together by muscular coordination. The area between the vocal folds is referred to as the glottis.

The primary purpose of the larynx is to form a protective closure above the respiratory system during swallowing. The sealed larynx can also be used to increase abdominal pressure during certain human functions such as heavy lifting. During

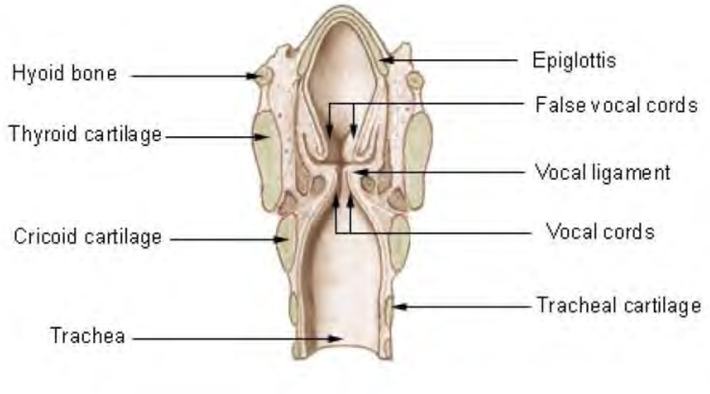


Figure 2.2: Coronal plane perspective of the human larynx, from (National Cancer Institute, Retrieved September 12th, 2009).

speech production however, the larynx is the conduit through which air flows from the trachea to the pharynx where a quasi-periodic phonatory source can be generated due to the behaviour of the glottis.

Speech phonation is divided into two broad categories: voiced and unvoiced. During unvoiced phonation, air from the lungs is expelled through the trachea into the vocal tract, unhindered through the open glottis. This results in a noisy signal which is used to generate many consonant sounds including fricatives (*e.g.* /s/, /f/), and plosives (*e.g.* /p/). During voiced phonation however, the laryngeal muscles tighten the vocal folds, resulting in their quasi-periodic oscillation between open and closed states, exciting the vocal tract with bursts of air. This excitation source is used for many vowels and sonorant type sounds.

The myoelastic-aerodynamic theory (Van den Berg, 1958) hypothesises that the vibration is a result of the interplay between two forces and is illustrated in Figure 2.3. The vocal folds are held shut by muscular tension (top left in Figure 2.3), which are then forced open as a result of increased air pressure from the lungs (top right). Note that the vocal folds may not fully close along their length for certain speakers/voice qualities *etc.* The subsequent movement of air through the space between the vocal folds causes a pressure drop between the folds which produces a suction effect, forcing the folds back together (bottom). This phenomenon is known as Bernoulli's principle. The process then repeats for the duration of the tension placed upon the vocal folds. The cycle of vibration produces an acoustic signal which is often referred to as the glottal flow signal.

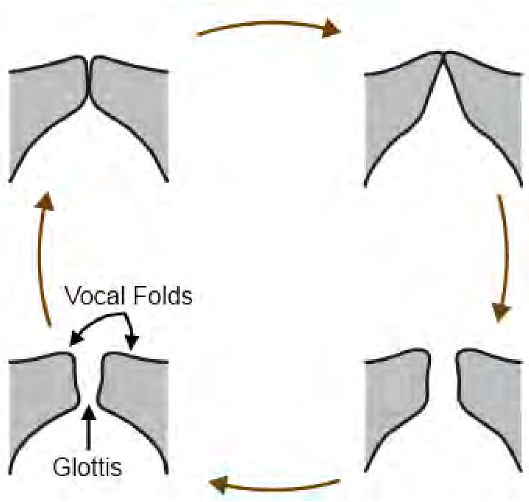


Figure 2.3: A schematic view of the vocal fold vibratory cycle, showing the opening and closing of the vocal folds (in grey) during voiced phonation. Adapted from (Honda, 2007).

Constrictions within the vocal tract can introduce a second turbulent noise source into the speech. When a constricted vocal tract is articulating a voiced source from the larynx, this can result in a mixed excitations combining periodic and aperiodic sound sources. Voiced fricatives such as /z/ and /v/ are such examples of phones which are produced in this manner.

The characteristics of the phonation source are generally thought to contain many perceptual indicators of voice quality (Klatt and Klatt, 1990; Childers and Lee, 1991), where voice quality is defined on a breathy to pressed scale. It also contains most of the prosodical information of an utterance, and is responsible for the fundamental frequency of the speech signal (related to the glottal vibratory cycle) and the duration of a utterance.

2.1.2 Articulation

The organs of articulation are generally thought to be the major contributor to the intelligibility of speech. During speech, the articulators, *i.e.* the tongue, teeth and lips, move between various geometrical configurations in coordination with the behaviour of the organs of phonation. Different configurations exhibit different resonant characteristics which are imprinted upon the laryngeal excitation signal. These spectral peaks are called formants[1].

For many sounds, the vocal tract forms a single multi-chambered tube from the top of the larynx to the lips during which the velum or soft palate at the top of the pharynx is closed. However, during nasalised sounds, *e.g.* /m/ and /n/, the velum is lowered and mouth cavity sealed, which couples vocal tract with the nasal cavities, producing a more complex geometry. A schematic diagram of the speech production system is given in Figure 2.4.

[1] The word formant comes from the Latin verb *formāre* meaning “to shape”.

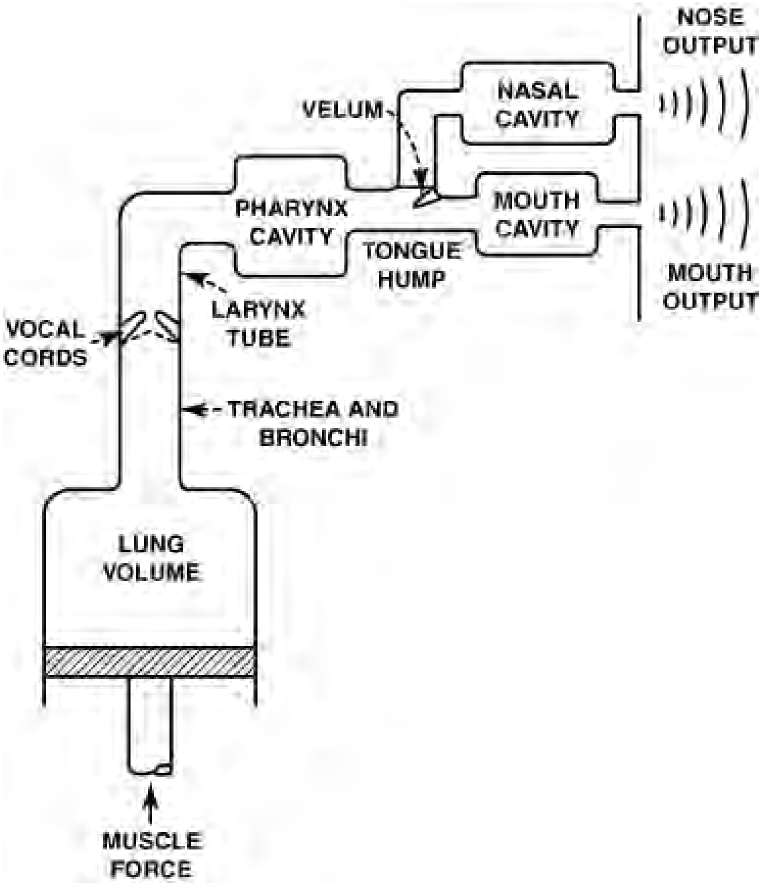


Figure 2.4: A schematic view of the speech production system, showing the various chambers of the vocal tract. From (Flanagan, 1972).

2.2 Source-Filter Theory of Speech

The division of the speech production process into two separate functions - phonation and articulation - also serves as the starting point from which speech processing interprets and analyses speech as it gives rise to the idea that speech can be interpreted as a source signal exciting a filter, the so-called source-filter theory of speech (Fant, 1970). The basic idea is that the acoustic energy source of speech - quasi-periodic puffs of air or a turbulent air stream - is shaped by the resonances and anti-resonances of the vocal tract and then radiated at the lips or nostrils.

As the articulators move relatively slowly, the speech signal can be assumed to be time-invariant over small time intervals - generally taken to be on the order of 25ms. Additionally, if the relationship between the source and filter is assumed to be linear, the source-filter theory of speech can be stated in the Z-domain:

$$S(z) = G(z)V(z)L(z) \tag{2.1}$$

where $S(z)$ is the Z-transform of the speech signal, comprising the multiplication of three components: $G(z)$ the glottal flow, $V(z)$ the vocal-tract filter and $L(z)$ the lip radiation.

Through the simplifying assumption that speech production can be approximated as a linear time-invariant system, the source-filter theory forms the basis for high quality rule-based speech synthesis systems (Klatt, 1987), efficient speech coding algorithms (Rabiner and Schafer, 1978), inverse filtering techniques (Wong *et al.*, 1979) and many other applications.

2.2.1 Source-Filter Model Limitations

The core assumption of the linear source-filter theory is that the operation of the source and filter are independent of each other. In actuality, they interact in a complex non-linear fashion that has yet to be satisfactorily described (Plumpe *et al.*, 1997). This interaction violates the linear, time-invariant assumption of source-filter theory. Instead, there is a nonlinear relationship between the activity of the source and the filter transfer function, which causes the filter to vary in time. In (Ananthapadmanabha and Fant, 1982), an electrical circuit was derived to model the glottis which demonstrated that the subglottal coupling during the open phase of the glottal signal tended to modulate the formant frequencies and bandwidths of the vocal-tract filter, most noticeably for the first formant. If the vocal-tract filter is assumed to be unchanging during the open phase of the laryngeal cycle, the modulation of the vocal tract resulting from the interaction appears as a ripple component superimposed upon the glottal signal. Interaction effects may also produce asymmetric glottal pulses. Figure 2.5 shows a synthetic glottal pulse with super-imposed ripple component imparted by increasing first formant bandwidth and center frequency during the open phase.

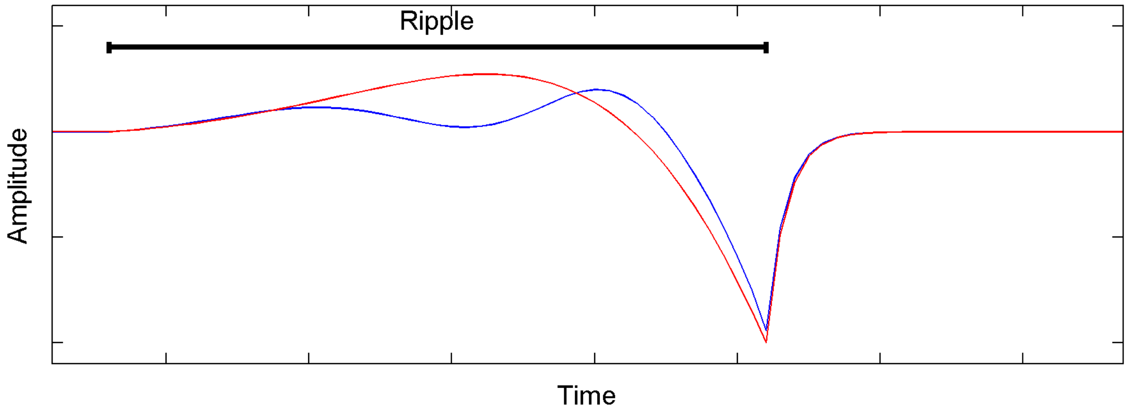


Figure 2.5: A synthetic glottal pulse with ripple component superimposed over the open phase, resulting from first formant modulation.

However, it has been claimed that the interaction is weak and that it is common to ignore it (Rabiner and Schafer, 1978). Additionally, it is thought not to contribute largely to speech perception (Nord *et al.*, 1984; Gobl, 2003).

2.3 Source-Filter Models

Following from the previous section, the linear source-filter theory of speech production can be decomposed into three elements:

- a glottal flow signal $G(z)$,
- a vocal-tract filter $V(z)$, and
- the radiation characteristics of the lips, $L(z)$.

This section will discuss the prevalent models of these three components.

2.3.1 Lip Radiation

Though more closely modeled as a source upon a sphere, it is reasonable to approximate the radiating surface of the lips as set in an infinite plane baffle, which avoids the calculation of complex diffraction effects (Rabiner and Schafer, 1978). In the digital domain, this characteristic $L(z)$ attenuates low frequencies, and is modeled as a first order differentiating filter (Rabiner and Schafer, 1978):

$$L(z) = 1 - z^{-1} \tag{2.2}$$

The amplitude response of L(z) is given in Figure 2.6.

More complex models of lip radiation have also been proposed (Laine, 1982), yet the approximation of Equation 2.2 has seen wide adoption due to its simplicity.

2.3.2 Vocal Tract

The vocal tract is defined as the various cavities which exist between the larynx and the lips through which a sound wave may pass during speech and includes the

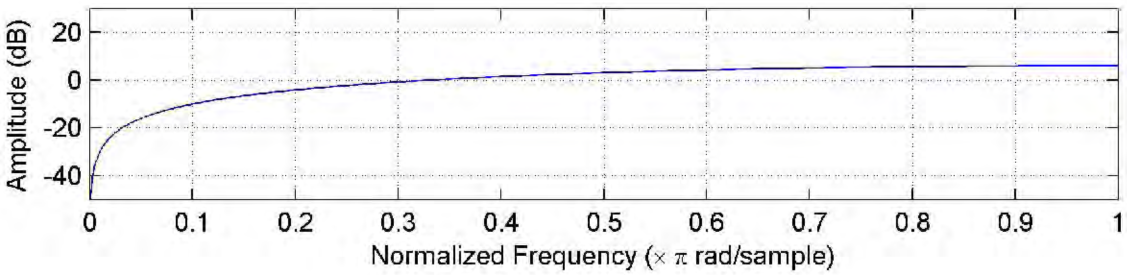


Figure 2.6: The amplitude response of the digital lip radiation characteristic $L(z)$, following Equation 2.2.

pharynx, mouth and nasal cavities. Physical models of the vocal tract begin with the approximation of the structure as a simple tube, to which wave propagation models can be applied in order to determine its spectral characteristics. The simplest approximation is a single dimensional tube model which is explained in detail here, but higher dimensionality models have also been proposed (Mullen, 2006; Birkholz *et al.*, 2006). However, the spectral model which derives from the single dimensional case (*i.e.* the all-pole vocal-tract filter model) has witnessed a wide deployment across a number of different applications including voice-source analysis and is thus focused upon in this section.

As a initial approximation, the tract is modeled as a uniform lossless acoustic tube which is open at one end. Additionally, it is useful to assume that acoustic waves from the larynx travel through the vocal tract along a single dimension as a plane wave (Rabiner and Schafer, 1978). Planar wave propagation assumes that the acoustic pressure wave moves in a direction perpendicular to the walls of the vocal tract otherwise complicated scattering may occur. This assumption is reasonable for frequencies whose wavelength is large in comparison with the diameter of the vocal tract. As the vocal tract varies along its length during articulation between 0.01 to 0.03 m in diameter, planar wave propagation is assumed valid for frequencies less than 5kHz.

By solving the wave equations defined by this system, the acoustic behaviour of the tube can be shown to have an infinite number of poles equally spaced along the imaginary axis in the Laplace domain (Rabiner and Schafer, 1978). Each complex conjugate pole pair comprises a formant. The analog transfer function of the acoustic tube $V_a(s)$ can then be described by the equation (Rabiner and Schafer, 1978):

$$V_a(s) = \frac{2e^{\frac{-sL}{c}}}{1 + e^{\frac{-s2L}{c}}} \tag{2.3}$$

where L represents the length of the tube and c represents the speed of sound. The poles s_n of $V_a(s)$ are therefore given by:

$$s_n = \pm j\frac{(2n+1)\pi c}{2L}, \quad n = 0, \pm 1, \pm 2, \cdots \tag{2.4}$$

The magnitude response of the above equations for an acoustic tube of the length of a average male vocal tract ($17cm$) with the speed of sound at sea level ($340ms^{-1}$) is given in Figure 2.7.

Note that this first approximation of the vocal tract as an idealised tube lacks the ability to model a varying shape. A more sophisticated vocal tract model is that of a non-uniform acoustic tube, where the cross-sectional area of the tube changes along its length, which can be approximated as a set of concatenated tubes of different dimensions (Atal and Hanauer, 1971; Rabiner and Schafer, 1978; Markel and Gray, 1982), as shown in Figure 2.8. It can be shown that, although the transfer function can still be represented using only poles, the formants of the cavity are no longer

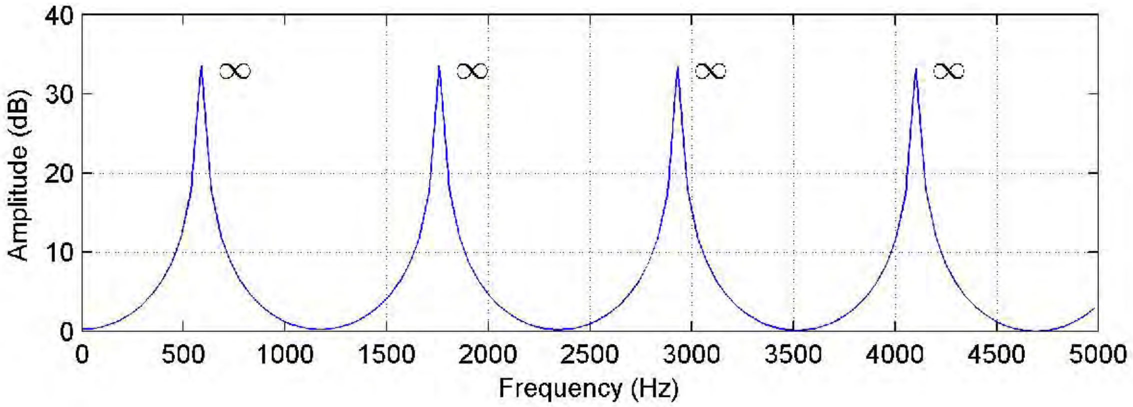


Figure 2.7: A detail of the frequency response of an idealised lossless acoustic tube of length similar to male vocal tract. ∞ implies a infinitesimally small bandwidth.

equally spaced as with the uniform model and are a function of the dimensions of the tube (Rabiner and Schafer, 1978).

While acoustic tube models are useful for giving an approximation of the vocal tract transfer function, they differ from the actual vocal tract in a number of respects excluding the effects of, for example, vocal tract wall vibration, viscous friction, thermal conduction and the curvature of the vocal tract. The mathematical derivations of these effects result in intractable frequency-dependent equations of motion, though by implementing a number of simplifying assumptions regarding this complex behaviour (Rabiner and Schafer, 1978), numerical solutions to the equations of motion can be determined. The net effect of these components is to broaden the bandwidths of the vocal tract formants and shift them slightly in frequency. The reader is referred to (Rabiner and Schafer, 1978) (pp. 66-70) for the detailed mathematics of this approach.

In the digital domain, an all-pole filter such as a nonuniform acoustic tube rep-

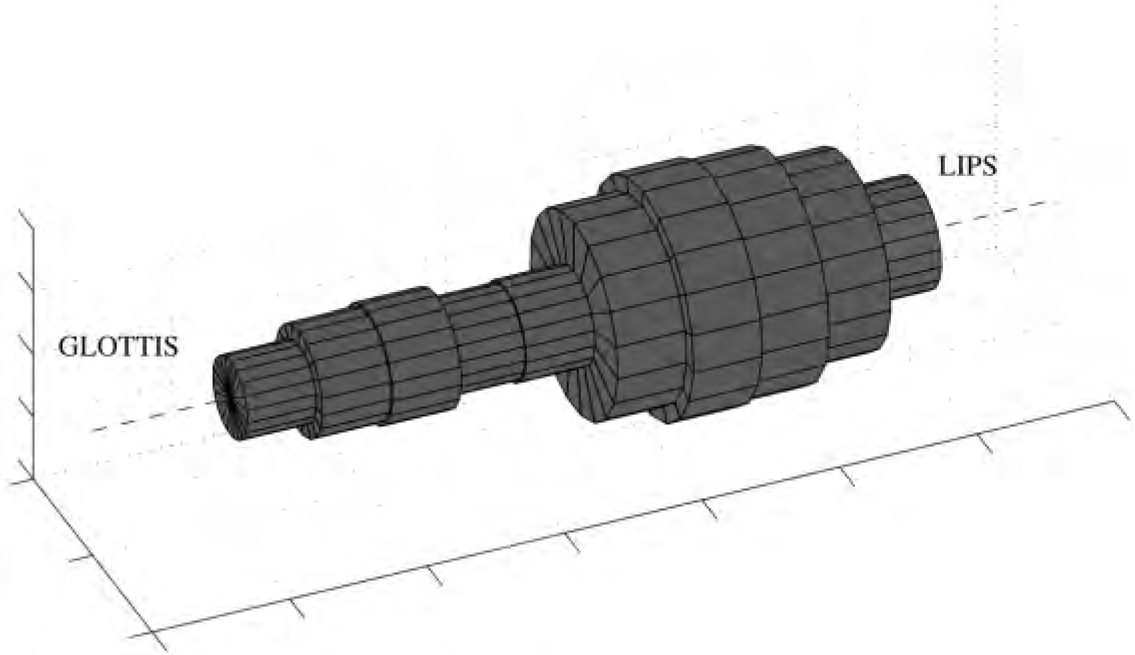


Figure 2.8: Above gives a representation of a non-uniform acoustic tube comprising of concatenated cylinders of differing diameters, often used to model to vocal tract. Taken from (Bäckström, 2004).

resenting the vocal tract can be described in terms of its formants. The transfer function $V(z)$ of such a system with p complex conjugate poles[2] can be expressed:

$$V(z) = \frac{1}{\prod_{k=1}^{p/2}(1 - c_k z^{-1})(1 - c_k^* z^{-1})} \tag{2.5}$$

where c_k and c_k^* are the complex conjugate pairs describing the k^{th} formant. c_k is of the form:

$$c_k = e^{\frac{-B_k \pi}{f_s}} \left(\cos \frac{2\pi F_k}{f_s} + i \sin \frac{2\pi F_k}{f_s} \right) \tag{2.6}$$

[2] In the case that the transfer function has real poles, $V(z)$ assumes a similar slightly different form to the one expressed by Equation 2.5.

where F_k and B_k represent the center frequency and bandwidth in hertz of the k^{th} formant, respectively, and f_s the sampling frequency.

The vocal-tract filter order p is dependent on the length of the vocal tract according to the equation:

$$p = \frac{2Lf_s}{c} \tag{2.7}$$

where L is the length of the vocal tract and c the speed of sound. As a typical male vocal tract is 17cm in length and the speed of sound in air is approximate 340ms^{-1}, the usual "rule of thumb" for choosing the filter order is given by the equation:

$$\begin{aligned} p &= \frac{2 \times 0.17\, f_s}{340} \qquad (2.8) \\ &= \frac{f_s}{1000}(+\gamma) \qquad (2.9) \end{aligned}$$

Thus, for male speech, the vocal tract is characterised by a two-pole formant for every 1kHz of signal bandwidth (Markel and Gray, 1982). Often a "fudge factor" γ is introduced to supply an extra flexibility to the analysis with a small number of additional poles (1 or 2) (Markel and Gray, 1982).

The parameters and amplitude responses of a variety of vocal-tract filters are given in Appendix A.

Though the all-pole model of the vocal tract is appropriate for many speech sounds when the vocal tract approximates a nonuniform acoustic tube, there exists certain sounds where zeros may be introduced into the frequency response. For example, during certain phones the nasal cavity may be coupled with the pharynx deviating from the nonuniform tube model, being better described as a branched tube. The sealed oral cavity (*e.g.* at the lips for /m/, the alveolar ridge and tongue for /n/, *etc.*) can then trap frequencies, creating zeros in the spectrum. Additionally, zeros may occur in other non-continuant type phones. However, as noted in (Atal and Hanauer, 1971), poles are more salient than zeros and spectral zeros can be approximated by multiple poles.

2.3.3 Glottal Source

As previously mentioned, phonation can be loosely categorised into two different states: voiced and unvoiced. During the voiced state of phonation, quasi-periodic pulses of acoustic energy, resulting from the opening and closing of the glottis, are introduced into the vocal tract. This section discusses this signal, known as the glottal flow waveform, and its prevalent models.

Although not a focus of this work, the presence of noise in the glottal waveform is also very important for both voiced and unvoiced phonatory states. This section also briefly discusses the phenomenon.

Physical Models Like the vocal tract, the complex biological system of the vocal folds is simplified in order to model its physical behaviour: mechanically-coupled masses are the typical approach, *e.g.* the two-mass model (Ishizaka and Flanagan, 1972), the one-mass model (Drioli and Avanzini, 2000) and the three-mass model (Story, 2003). These systems can be described by equations of motion, which, when coupled with a similar vocal tract simulation, can be used to calculate the synthetic speech waveforms numerically (Ishizaka and Flanagan, 1972).

The main drawback of physical glottal models is the number of parameters which are required to produce the glottal pulses (*e.g.* the two-mass model (Ishizaka and Flanagan, 1972) has 19 parameters), and determining these parameters has proved difficult and intricate. Additionally, the complex relationship between glottal physiology and the parameters of the models make control of these models difficult (Sciamarella and d'Alessandro, 2004)

Acoustic Models Unlike the vocal tract, it is common to model the characteristics of the acoustic glottal waveform directly, without explicit reference to the physical process which created it. Instead of the glottal flow signal, it is usually the derivative glottal flow signal which is modeled. This follows from a rearrangement of the source-filter representation of speech. Because of the linearity of the speech production system, the order of operations can be changed. Thus, the differentiator representing the lip radiation characteristic $L(z)$ is often applied directly to the glottal flow signal $G(z)$ to give the derivative glottal flow signal $G'(z)$:

$$G'(z) = G(z)L(z) \tag{2.10}$$

Given this combination, linear speech production is equivalent to the derivative glottal flow exciting the vocal tract:

$$S(z) = G'(z)V(z) \tag{2.11}$$

Throughout this work, the term *voice source* refers to representations of $G'(z)$, *i.e.* the voice source is the acoustic signal which results be inverse filtering the speech signal by the vocal tract.

Though the models differ in their formulation, derivative glottal flow models exhibit broadly similar characteristics as discussed in (Doval and d'Alessandro, 2006). When discussing time-domain glottal models, it is useful to divide the flow into three different stages which correspond to the phases of the glottal voicing cycle: an open

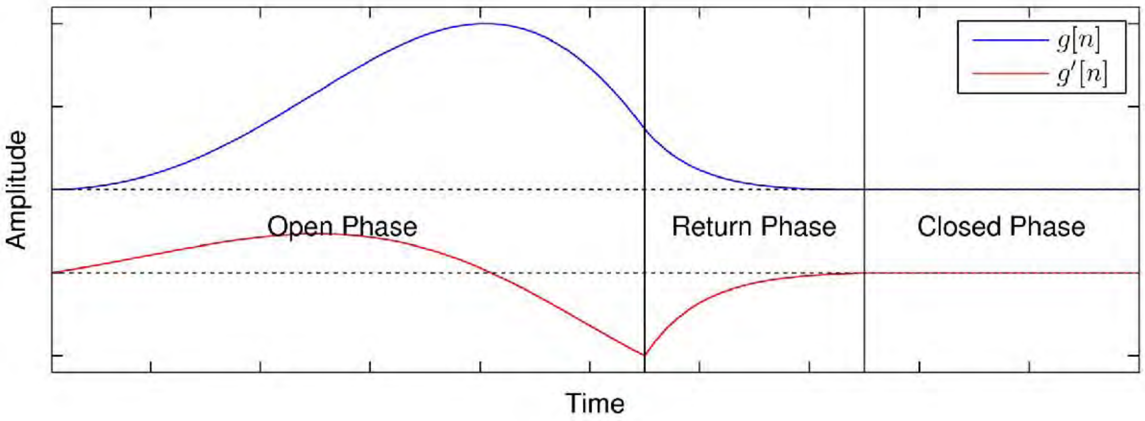


Figure 2.9: The figure above gives the time-domain representations of synthetic glottal flow $g[n]$ and derivative glottal flow $g'[n]$ pulses, in addition to the time intervals of the various phases of the glottal cycle.

phase corresponding to when the vocal folds are open, a return phase corresponding to when the folds are rapidly closing and a closed phase corresponding to when the folds are closed. Figure 2.9 shows a synthetic time-domain waveform of a both a glottal pulse and a derivative glottal pulse waveform.

In the frequency domain, the main features of the glottal source signal are a spectral maximum, sometimes termed the “glottal formant”, and spectral tilt. Figure 2.10 illustrates these two characteristics. The glottal formant is a peak of spectral energy in the region of the signal's fundamental frequency. The term glottal formant is in fact a misnomer as there is no resonance effect like in the vocal tract: rather, the position of the spectral energy boost is dependent mostly upon the shape and duration of the signal's opening phase as well as the fundamental frequency.

The other salient characteristic of derivative glottal flow models is its decrease in spectral energy with increasing frequency. The signal's spectral tilt has been shown to be of the utmost importance for the perception of voice quality (Klatt and Klatt, 1990). Contrary to the glottal formant, the spectral tilt is a result mainly due to the change of the time-domain waveform from its opening phase into its return phase. The most abrupt change, *i.e.* no return phase, imparts the least spectral tilt and therefore introduces the most high frequency energy into the spectrum while a more gradual change introduces less.

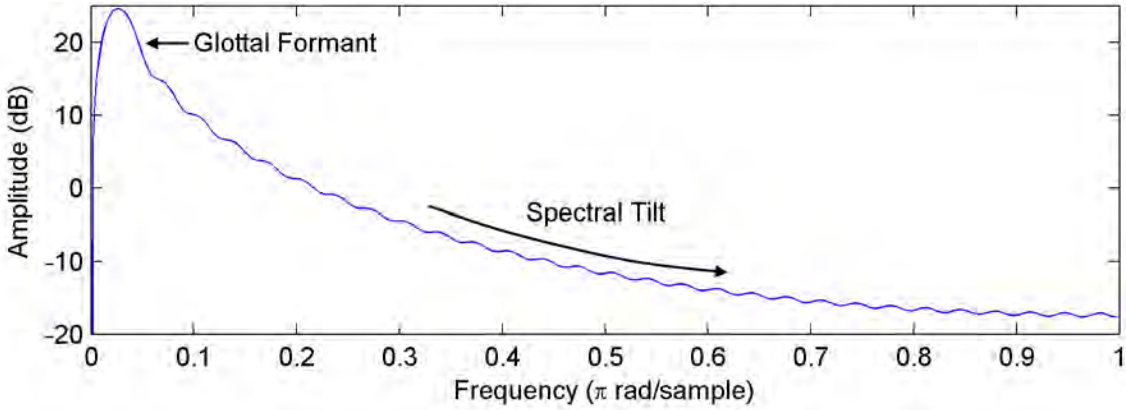


Figure 2.10: The above figure give the magnitude spectrum of a synthetic derivative glottal flow pulse, illustrating the low frequency energy peak known as the glottal formant, and the frequency roll-off referred to as the pulse's spectral tilt.

2.3.3.1 Glottal Flow Models

Two Pole Glottal Flow Model Traditionally, the glottal flow was modeled in the Z-domain as a two-pole anti-causal filter of the form:

$$G(z) = \frac{1}{(1 - e^{\alpha} z^{-1})^2} \tag{2.12}$$

where $\alpha \simeq 0$. In the time domain, the truncated impulse response of $G(z)$ is similar to Figure 2.11. This model has experienced some success for early linear predictive voice models (Atal and Hanauer, 1971; Markel and Gray, 1982).

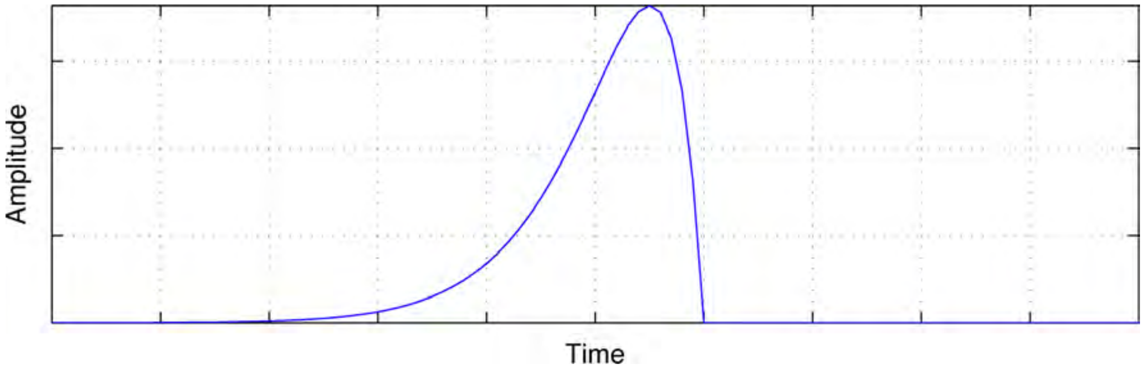


Figure 2.11: The above figure gives the truncated impulse response of the two pole model of $G(z)$, as given by Equation 2.12.

As α is close to zero, the single zero of the lip radiation differentiator is often thought to approximately cancel one of the glottal poles during speech production. Thus, the derivative glottal flow $G'(z)$ related to the two-pole glottal flow model can often expressed:

$$G'(z) = \frac{1}{1 - e^{\alpha} z^{-1}} \tag{2.13}$$

This model exhibits the -6 dB/oct roll off generally attributed to the spectrum of the derivative glottal source, though the glottal formant is not represented.

Rosenberg and KLGLOTT88 Models In experiments exploring the effect of shape of the glottal pulse on the perception of synthetic speech (Rosenberg, 1970), a simple polynomial model of the glottal flow was proposed. It describes derivative glottal flow in a piecewise fashion: the open phase as an inverted parabola, with no flow during the closed phase. The shape of the model are given by the pitch period T_0 and open quotient O_q, defined as the ratio of the opening time of the pulse to the pitch period, while a third parameter b defines its scale. The time-domain shape of the model is given by the equation:

$$u_{RO}(n) = \begin{cases} b(2\,O_qT_0\,n - 3\,n^2) & 0 < n \leq O_qT_0 \\ 0 & O_qT_0 < n \leq T_0 \end{cases} \tag{2.14}$$

The KLGLOTT88 voicing source is a model of the glottal flow originally incorporated into Klatt's formant synthesiser KLSYN88 (Klatt and Klatt, 1990) which extends the Rosenberg polynomial model.

The piecewise formulation of Rosenberg model offers no other possibility than an immediate abrupt closure with a fixed spectral tilt; in order to allow for a more gradual closure, the KLGLOTT88 model (Klatt and Klatt, 1990) extends the basic Rosenberg shape by applying a spectral tilt filter to the waveform. This spectral tilt filter $TL(z)$ is first order IIR filter, parameterised by a single pole a distance μ away from the origin in the Z-plane. The transfer function of $TL(z)$ can therefore be written:

$$TL(z) = \frac{1}{1 - \mu z^{-1}} \tag{2.15}$$

In addition to imparting a more gradual return phase upon the pulse, the filtering operation also affects the open phase of the pulse in a manner that is difficult to predict. The time-domain shapes of the basic Rosenberg and KLGLOTT88 pulse shapes are given in Figure 2.12.

In their analytical study of the spectra of glottal models, (Doval and d'Alessandro,

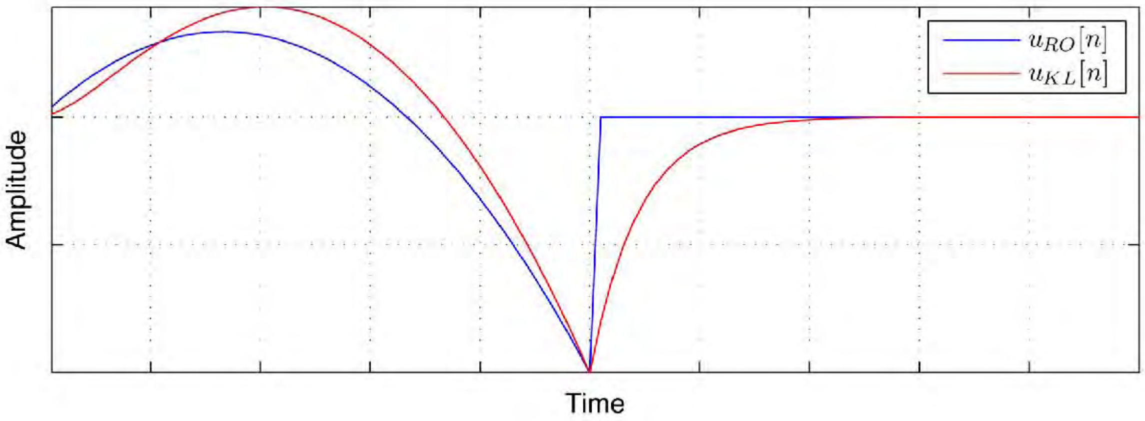


Figure 2.12: Above are the time-domain shapes of the Rosenberg and KLGLOTT88 models of derivative glottal flow. The models have been normalised in amplitude to facilitate comparison.

2006) notes the relative simplicity of the KLGLOTT88 model compared with other models. The quadratic polynomial that forms the open phase implies a fixed asymmetry coefficient for the model. A consequence of this is that this position of the glottal formant is dependent upon only the open quotient.

Liljencrants-Fant Model The LF model, proposed in (Fant *et al.*, 1985), represents the general flow shape of the derivative glottal flow over one glottal cycle. Like the Rosenberg model, the LF model is described in the time domain by a piece-wise function, given as follows:

$$u_{LF}(n) = \begin{cases} E_0 e^{\alpha n} \sin \omega_g n & 0 \leq n < T_e \\ \frac{-E_e}{\epsilon T_a}(e^{-\epsilon(n-T_e)} - e^{-\epsilon(T_c - T_e)}) & T_e \leq n \leq T_c \\ 0 & T_c \leq n < T_0 \end{cases} \tag{2.16}$$

The first segment is an exponentially increasing sinusoid of angular frequency

ω_g, bandwidth α, and scaled by E_0. This portion of the waveform characterises the derivative glottal flow from the instant of glottal opening at 0, through the time axis at T_p, to the maximum negative extreme E_e at instant T_e.

At this point the return phase of the LF model begins. This portion models the glottal closure as a modified exponential function which returns to zero at a rate determined by the steepness of the slope of the tangent to the function at T_e. The distance of this tangent's time axis intercept from T_e is called T_a, and is referred to as the effective duration of the return phase. The parameter ϵ is the decay constant of the exponential. The total number of samples in the pulse is the pitch period, T_0. An example LF model pulse is given in Figure 2.13.

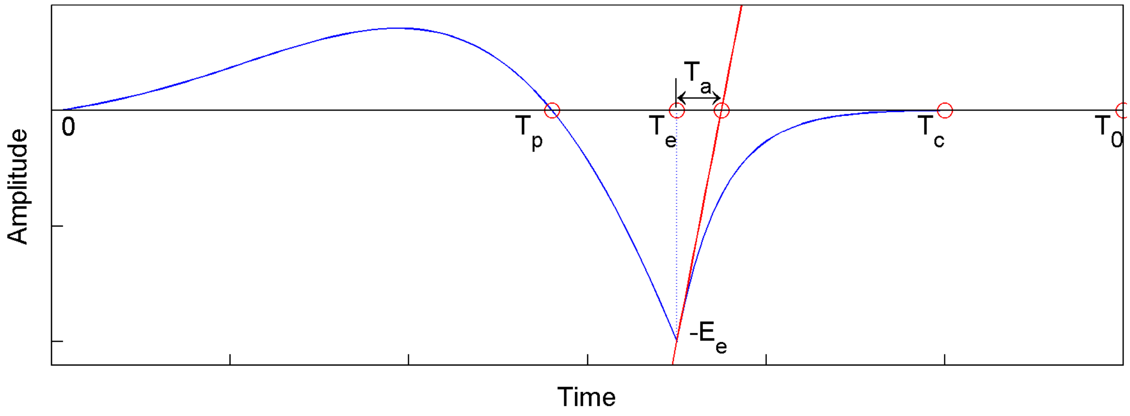


Figure 2.13: This figure gives a representation of a synthetic glottal pulse generated according to the LF model, as given by Equation 2.16.

Another often used parameter set of the LF model are the R parameters, $\{R_a, R_g, R_k\}$, which are calculated:

$$R_a = \frac{T_a}{T_0} \tag{2.17}$$

$$R_k = \frac{T_e - T_p}{T_p} \tag{2.18}$$

$$R_g = \frac{T_0}{2T_p} \tag{2.19}$$

Because of the natural covariation of the LF model parameters observed in real speech, researchers have attempted to reduce the degrees of freedom of the LF model and describe the entire glottal shape using a single parameter. The so-called basic waveshape parameter R_d, proposed in (Fant, 1995), is calculated:

$$R_d = \frac{f_{ac}}{d_{peak}} \frac{1}{T_0 0.11} \tag{2.20}$$

where f_{ac} and d_{peak} are given in Figure 2.14.

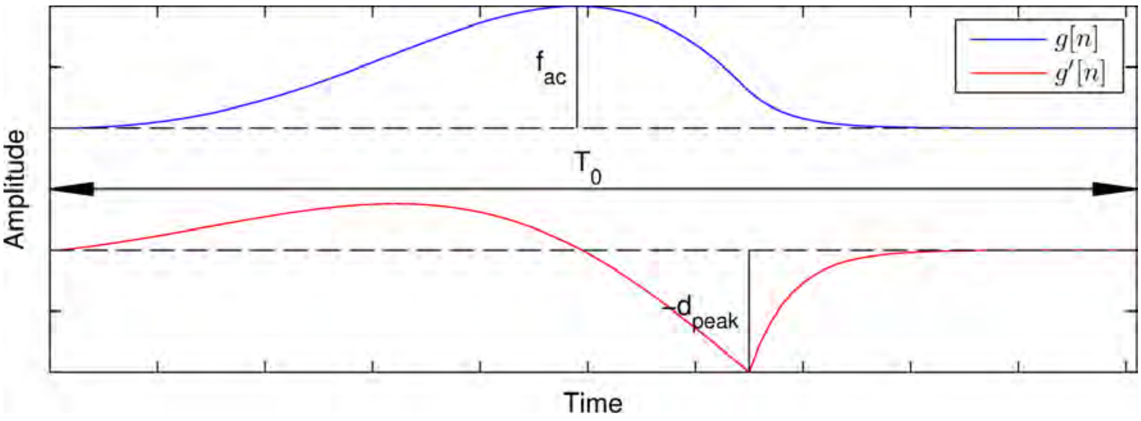


Figure 2.14: The above figure illustrates the amplitude and time-interval measurements necessary to calculate the R_d parameter of a glottal pulse, according to Equation 2.20.

Following a statistical analysis of vowels and voiced consonants (Fant, 1995), the following statistical relationships were devised to predict the R_a and R_k parameters from the R_d parameter:

$$R_a = \frac{-1 + 4.8R_d}{100} \tag{2.21}$$

$$R_k = \frac{22.4 + 11.8R_d}{100} \tag{2.22}$$

The R_g parameter is not predicted using statistical analysis, rather it is found by solving the following relationship to ensure that pulse which conforms to the LF model is produced:

$$R_g = \frac{R_d R_k}{\frac{4}{0.11}(0.5 + 1.2R_k) - 4R_a R_d} \tag{2.23}$$

Thus, given an amplitude scale and fundamental frequency, a limited range of R_d values ($0.25 \leq R_d \leq 3$) can be utilised to generate LF model pulses. At low values, the R_d parameter generates to LF model pulse with small open and return phases, corresponding to a pressed phonation type, while at high values, both open and return phases are longer, corresponding to breathy phonation. Figure 2.15 gives three LF model pulses generated from three R_d different values.

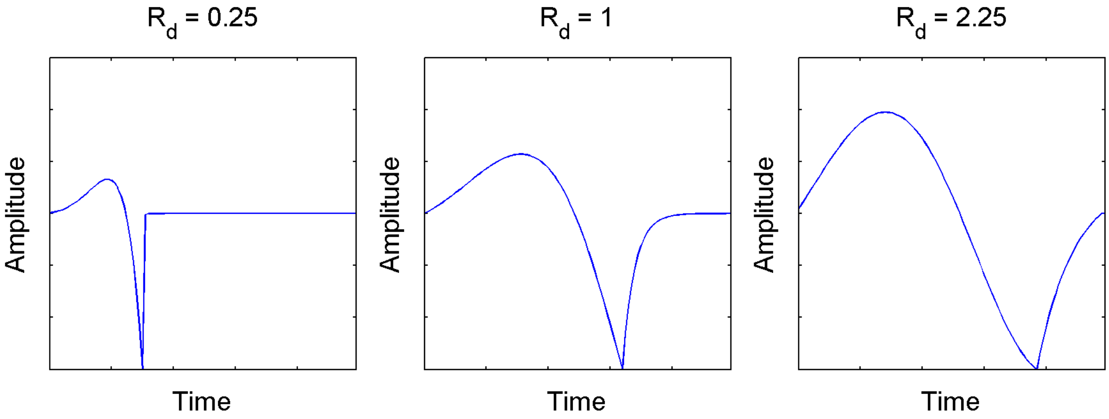


Figure 2.15: The figure above gives example LF model pulses generated using different R_d values.

Fujisaki-Ljungqvist Model The six parameter Fujisaki-Ljungqvist model, proposed in (Fujisaki and Ljungqvist, 1986), represents the derivative glottal flow pulse with four segments represented by polynomial functions. The formula is given by:

$$u_{FL}(n) = \begin{cases} A - \frac{2A+R\alpha}{R}n + \frac{A+R\alpha}{R}n^2, & 0 < n \leq R \\ \alpha(n-R) - \frac{3B-2F\alpha}{F^2}(n-R)^2 + \frac{2B-F\alpha}{F^3}(n-R)^3, & R < n \leq W \\ C - \frac{2(C-\beta)}{D}(n-W) + \frac{C-\beta}{D^2}(n-W)^2, & W < n \leq W+D \\ \beta, & W+D \leq n < T_0 \end{cases} \tag{2.24}$$

where

$$\alpha = \frac{4AR - 6FB}{F^2 - 2R^2} \tag{2.25}$$

$$\beta = \frac{CD}{D - 3(T_0 - W)} \tag{2.26}$$

Sudden or gradual discontinuities are permitted at glottal closure in addition to glottal opening from the parameter A. The pulse is represented from glottal opening to peak flow (the first zero crossing of the derivative glottal flow signal) by an inverted parabolic function, from peak flow to the minimum derivative flow by a cubic function, while the return phase is modeled using another parabolic function.

Causal-Anti-causal Linear Model In (Doval *et al.*, 2003), a new glottal source model was proposed. As the glottal flow signal exhibits some causal and anti-causal characteristics, the Causal-Anti-causal Linear Model (CALM) consists of the impulse responses of a mixed phase filter, using a single causal pole to represent the return phase of the pulse and a pair of complex conjugate anti-causal poles to model the open phase. The proposal of such a model combines the spectral and time-domain interpretations of the glottal flow signal with explicit emphasis upon the phase characteristics of the glottal source signal.

The time-domain representation of the pole configuration is calculated in two passes: the first calculates the impulse response of the anti-causal portion of the waveform backward in time from the impulse at 0. A second pass calculates the return phase. In order to generate the time derivative of the glottal flow, a real zero at 1 is simply added to the Z-plane. An example of the CALM glottal model is given in Figure 2.16.

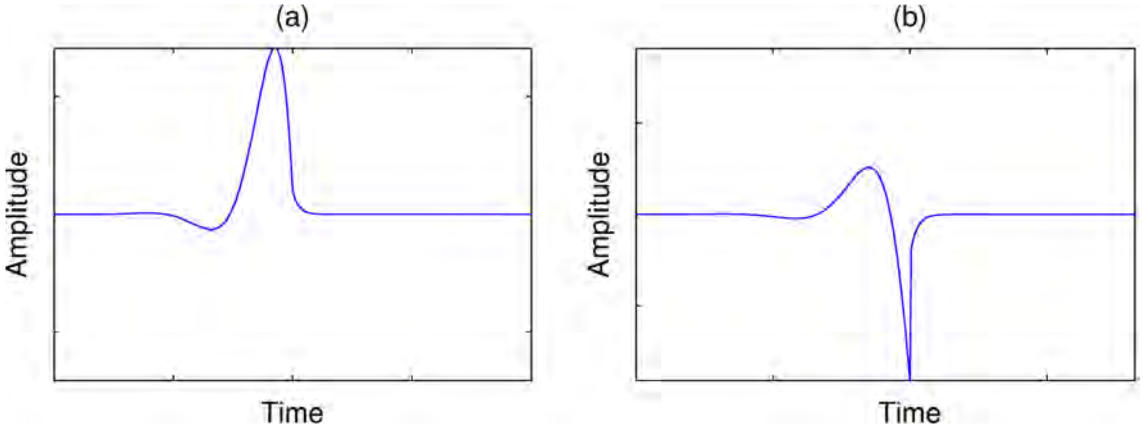


Figure 2.16: The above figures give the time-domain representations of the CALM model of (a) glottal flow and (b) derivative glottal flow.

Other Models Many other models for the glottal flow exist, including the Rosenberg++ model (Veldhuis, 1998), Fant model (Fant, 1979), Hedelin model (Hedelin, 1984), Childers polynomial model (Childers, 1995), *etc.* However, the models mentioned in this review are the most prevalent in the literature.

2.3.3.2 Glottal Noise

Turbulent noise theory dictates that noise occurs when a fluid passes through a narrow constriction. In terms of speech, this constriction may appear and thus noise may be generated when glottal aperture is small or at some other point in the vocal tract. Indeed, aspiration noise makes an important contribution to the perception of speech (Klatt and Klatt, 1990) and can enhance the naturalness of synthetic speech (Childers, 1995). In this study, experiments with synthetic speech employ a noise model to simulate this phenomenon.

Predominantly, time- and frequency-modulated Gaussian noise is used by researchers to model glottal noise. Such a model was implemented by (Hermes, 1991), who experimented with the perception of breathy vowels using a simplified source model. The source signal comprised a low-pass filtered pulse train combined with high-pass filtered Gaussian noise. The filter cut-off frequency of each filter was in the range 1.2 to 2 kHz, thus giving the excitation signal a flat frequency response. In the time domain, the noise was temporally modulated so that it appeared in bursts around the pulses. This step is important so that the noise signal perceptually fuses with the periodic portion of the waveform (Hermes, 1991).

Similar noise models have been employed by other researchers for speech synthesis utilising glottal signals. (Lu, 2002) models the aspiration signal as high-pass filtered Gaussian noise, temporally modulated by a scaled and modified Hann window, centred above a point a given lag away from the instant of glottal closure. The Hann window is modified by adding a constant, ensuring a noise floor during the signal. (del Pozo, 2008) notes the difficulty in obtaining these parameters from the speech signal and simply applies to scaled Gaussian noise the estimated glottal derivative pulse model waveform. Reportedly, this gave similar results to the model in (Lu, 2002). Other researchers have often used the model waveform in constructing the temporal envelope of the glottal noise signal (Agiomyrgiannakis and Rosec, 2008, 2009; Gobl, 2006).

2.4 Conclusions

Though speech results from complex anatomical motor-coordination, simplifying assumptions can be made to develop tractable models for speech signal processing. The prevalent source-filter model of speech has been widely adopted by speech engineers and is fundamental to many speech processing applications, including a focus of this study: voice-source estimation. This model interprets speech production as a linear time-invariant system, which can be separated into a phonatory source signal from within the larynx and filtering operations resulting from the geometric configuration of the articulators.

In order to model the acoustic behaviour of the vocal tract, one can adopt the all-pole model or other spectral approaches, *e.g.* cepstral-type envelopes. In this study, the all-pole filter is adopted. The main reason for this is that the all-pole filter logically follows from the acoustic tube physical model of the vocal tract. Additionally, the model has often previously been applied for this purpose, *e.g.* (Atal and Hanauer, 1971; Wong *et al.*, 1979; Markel and Gray, 1982; Alku, 1992; Lu, 2002; Vincent *et al.*, 2005).

This work will employ multiple models of the voice-source signal. For GCI estimation, the two pole glottal flow model is adopted in order to produce the linear predictive deconvolutive residual signal where the instants of glottal closure appear as discontinuities (see Chapters 4 and 7). For voice-source estimation, a more detailed model is desirable and therefore this study adopts the transformed LF model. As mentioned above, this model produces a subset of possible LF model shapes using a single parameter which predicts the full LF parameter set in an attempt to incorporate the natural covariation which exists between them (Fant, 1995). Thus, the shapes produced by the transformed LF model are more likely to be physiologically relevant. Additionally, in the context of voice-source estimation, the parameters of the full LF model are not independent and may introduce ambiguous unrealistic parameterisations (*e.g.* very high T_e and T_p values) (Fröhlich *et al.*, 2001; Vincent, 2007). Finally, regarding the ability of the transformed LF model to characterise the glottal signal, (Fant, 1995) qualified the waveshape parameter R_d as “the most effective single measure for describing voice qualities”.

Chapter 3

Voice-Source Estimation and Parameterisation

While the previous chapter discussed prevalent models of the components of the speech signal, this chapter discusses how those models are utilised for voice-source estimation and parameterisation. Based upon the source-filter theory of speech, state-of-the-art methods of voice-source estimation attempt to remove the spectral effects of the vocal tract in order to reveal the acoustic voice-source signal.

This chapter presents a review of state-of-the-art voiced speech estimation and parameterisation techniques, categorised according to the domain where they operate. It will be shown that time-domain techniques are well-suited for voice-source estimation and parameterisation, but that these approaches are sensitive to the location of the analysis frame and not robust to phase distortion that may be imparted on the signal *e.g.* during recording. Phase-based techniques also suffer from a similar lack of robustness for this phenomenon. Conversely, frequency-domain approaches are more robust to these issues. These observations form the basis for the PowRd method of voice-source estimation and parameterisation, proposed in Chapter 6.

3.1 Theory and Strategies

Voice-source estimation is the estimation of the glottal flow waveform (or its derivative) from the acoustic speech signal without the influence of the vocal tract (Miller, 1959). As the vocal tract and voice source are convolved together and both unknown, voice-source estimation is a blind deconvolution problem which can only be solved if certain assumptions are made about the nature of both the glottal source and vocal-tract filter.

The theoretical basis for glottal inverse filtering comes directly from a rearrangement of the linear source-filter theory of speech:

$$S(z) = G'(z)V(z) \tag{3.1}$$

$$\Rightarrow G'(z) = \frac{S(z)}{V(z)} \tag{3.2}$$

Thus, voice-source estimation is the inverse of the speech production process: the derivative glottal flow source is revealed by inverse filtering the speech signal by the vocal-tract filter. Figure 3.1 illustrates both speech production and glottal inverse filtering.

There are two basic strategies for glottal source estimation. The first is to make assumptions regarding some characteristic of the speech signal which can be attributed only to either the vocal tract or to the glottal flow, *e.g.* that the maximum phase components of speech result from the derivative glottal flow or that the glottal closed

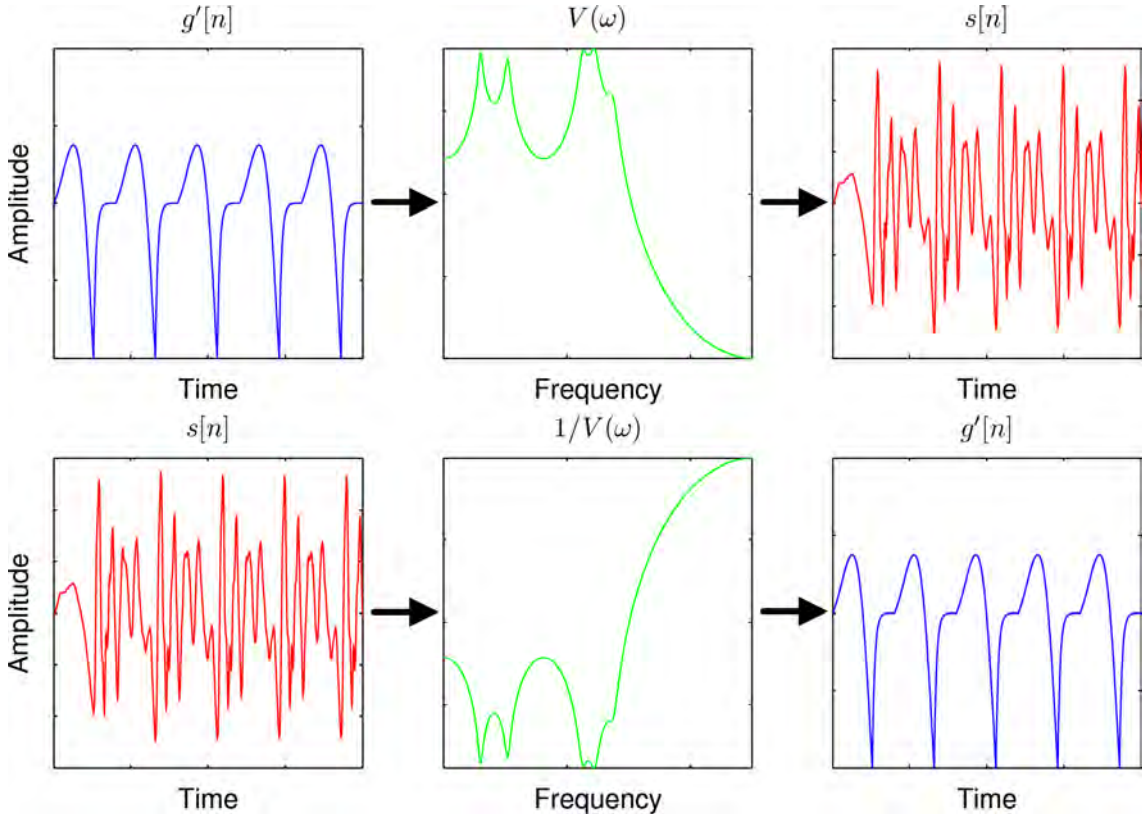


Figure 3.1: The above figures give a conceptual view of both the source-filter theory of speech production (top) and the process of glottal inverse filtering (bottom). For speech production, the time-domain glottal signal $g'[n]$ excites a vocal-tract filter $V(\omega)$ to produce $s[n]$ the speech signal. For glottal inverse filtering, the speech signal is inversely filtered by the vocal tract to give the derivative glottal flow signal.

phase represents only the decaying vocal tract resonances. Once this characteristic has been identified, the speech signal may be deconvolved by some means, *e.g.* by determining and inversely applying a parametric model.

The second general strategy of voice-source estimation involves the parametric modeling of the entire glottal contribution. By removing this model from the speech signal, the remaining vocal tract can be subsequently (or simultaneously) parameterised. The parameters modeling the voice-source signal can be fixed throughout analysis, adapted to measurements taken from the speech signal or varied until some optimum is found. Because these methods parameterise both the vocal tract and the voice source, these procedures are sometimes referred to as joint estimation techniques.

Following the estimation of the voice-source signal, it is often parameterised in order to quantify its features (Alku *et al.*, 2002). The voice-source estimate is parameterised by directly estimating the parameters from the characteristics of the signal, *e.g.* extracting signal landmarks, *etc.* Following direct estimation of the desired parameters, they can also be refined by minimising an error criterion using an optimisation algorithm. The means by which this is performed depends upon the domain where the parameterisation occurs.

3.2 Time-Domain-Based Approaches

This section will review voice-source estimation and parameterisation techniques based upon time-domain assumptions of the glottal flow signal. Firstly, the mathematical details behind these methods are explained in terms of linear systems. Following this, different state-of-the-art approaches to solving these equations are outlined and reviewed.

In the Z-domain, a generalised representation of the source-filter model of speech is the so-called AutoRegressive-Moving Average eXogenous (ARMAX) model:

$$\begin{aligned} S(z) &= G'(z)V(z) \qquad (3.3) \\ &= \frac{B(z)G'(z)}{A(z)} \qquad (3.4) \end{aligned}$$

where the vocal tract $V(z)$ comprises $B(z)$ and $A(z)$, the filter polynomials representing its zeros and poles respectively. The speech signal $S(z)$ is produced by the excitation of $V(z)$ by the voice-source signal $G'(z)$. Transforming into the time domain, $S(z)$ at time n becomes $s[n]$, and the above equation can be expressed:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + \sum_{j=0}^{q} b_j g'[n-j] + e[n] \qquad (3.5)$$

where a_k and b_j are the filter coefficients, p and q the filter orders, $g'[n]$ is the voice-source signal and $e[n]$ the residual modeling error signal, assumed to have a flat frequency spectrum.

The ARMAX speech model is a generalised speech representation which reduces to other speech models as special cases. If the all-pole assumption is imposed upon the vocal tract, the b_j coefficients can be reduced to a single gain parameter, b_0, yielding the AutoRegressive eXogenous (ARX) model of speech (Ding *et al.*, 1994):

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + b_0 g'[n] + e[n] \qquad (3.6)$$

As noted in Section 2.3.2, the nonuniform acoustic tube model of the vocal tract dictates that the unbranched vocal tract exhibits the characteristics of an all-pole filter, and so this model has seen wide application.

In the case that $g'[n] = 0$, the standard autoregressive speech model is obtained (Markel and Gray, 1982):

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n] \qquad (3.7)$$

A time-domain least-squares (TD-LS) solution for coefficients of the vocal-tract filter $V(z)$ of the general ARMAX model can be determined in terms of p, q, $s[n]$ and $u[n]$. Assembling the filtered signals to one side expresses the residual signal $e[n]$ in terms of the speech $s[n]$ and voice-source waveforms $u[n]$:

$$e[n] = s[n] - \mathbf{M_{n-1}x} \tag{3.8}$$

where

$$\mathbf{M_{n-1}} = [s[n-1] \; s[n-2] \; \cdots \; s[n-p] \; g'[n] \; g'[n-1] \; \cdots \; g'[n-q]] \tag{3.9}$$

$$\mathbf{x} = [a_1 \; a_2 \; \cdots \; a_p \; b_0 \; b_1 \; \cdots \; b_q]^T \tag{3.10}$$

Evaluation of Equation 3.8 over a certain interval $n = 0 \cdots N-1$ can be expanded into matrix form:

$$\mathbf{E} = \mathbf{S} - \mathbf{Dx} \tag{3.11}$$

where

$$\mathbf{E} = [e[0] \; e[1] \; \cdots \; e[N-1]]^T \tag{3.12}$$

$$\mathbf{S} = [s[0] \; s[1] \; \cdots \; s[N-1]]^T \tag{3.13}$$

$$\mathbf{D} = [\mathbf{M_{-1}}; \mathbf{M_{-2}}; \cdots; \mathbf{M_{N-2}}]^T \tag{3.14}$$

By minimising the energy of the error $||\mathbf{E}||^2$, the vocal-tract filter parameters $\mathbf{x}$ can be found. This solution is given by:

$$\mathbf{x} = \left(\mathbf{D}^T\mathbf{D}\right)^{-1} \mathbf{D}^T\mathbf{S} \tag{3.15}$$

This solution forms the basis for many time-domain voice-source estimation methods. However, of the elements of Equation 3.5, usually only the acoustic speech signal $s[n]$ is known. It is not possible to determine the solution $\mathbf{x}$ without estimates of:

- the filter orders p and q,
- the interval of analysis, n, and
- the time-domain shape of the glottal excitation signal within the analysis interval, $g'[n]$.

Under the all-pole vocal-tract filter assumption, the vocal-tract filter order p is related to its length and the sampling frequency, and is given by Equation 2.9. Exploiting this assumption, voice-source estimation methods determine $\mathbf{x}$ require knowledge of an appropriate analysis interval n and the voice-source shape $g'[n]$ within that interval.

3.2.1 Closed-Phase Inverse Filtering

Ultra high-speed cinematography of the larynx has enabled scientists to visually observe the cycle of the glottal pulses during voiced phonation (Childers, 2000). These films, utilising frame-rates of greater than 4,000 frames per second, have shown that for certain voices, the glottis is often fully sealed during the pulse cycle, during what is referred to as the glottal closed phase. If the vocal-tract filter is determined during this small interval, the estimated filter can be inversely applied to the speech signal in an operation known as Closed-Phase Inverse Filtering (CPIF)(Wong *et al.*, 1979).

A fully closed glottis implies that during the closed phase, the glottal contribution to the speech signal is zero, *i.e.* $g'[n] = 0$. Therefore, during this interval if the vocal tract is represented by an all-pole filter, the speech signal can be stated as Equation 3.7. In other words, the speech signal results solely from the decaying resonances of the vocal tract and residual error signal. If the glottal closed-phase interval can be located and an appropriate value for p is known, the vocal tract is parameterised according to Equation 3.15, where $\mathbf{D}$ and $\mathbf{x}$ are modified to account for the closed glottal phase and all-pole vocal tract assumptions. This method of all-pole filter coefficients parameterisation is often called the covariance method of linear prediction as the inversion of this equation involves the evaluation of a covariance matrix (Markel and Gray, 1982). Figure 3.2 shows an example of a speech signal, the demarcated closed phase and the estimated voice-source signal.

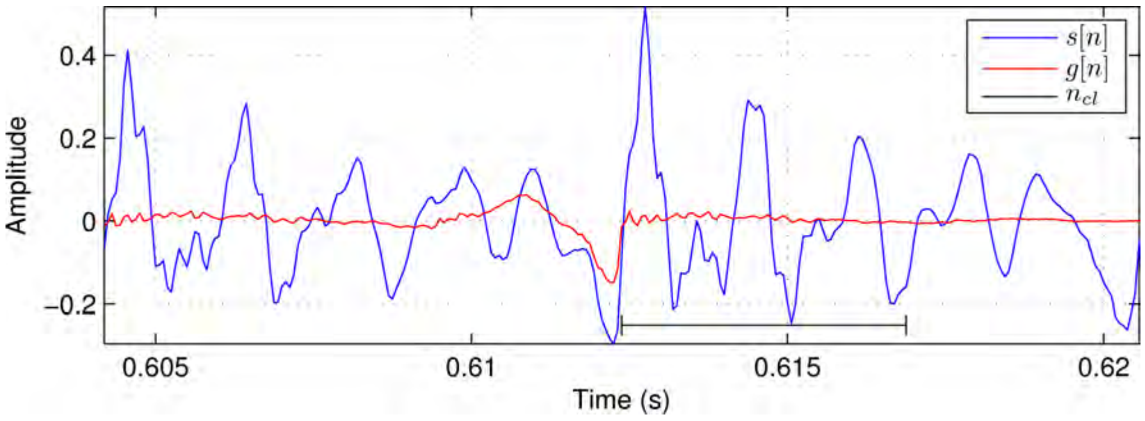


Figure 3.2: The figure above shows the derivative glottal flow waveform $g'[n]$ (red) estimated from the speech signal $s[n]$ (blue) by inverse filtering the speech segment using the all-pole filter coefficient determined by covariance linear prediction over the closed-phase interval n_{cl}, delimited in black.

Though speech engineers have been aware since at least the late 50's of the potential of the closed-phase condition for glottal inverse filtering (Miller, 1959), the theory behind closed-phase inverse filtering method was first formalised by (Wong *et al.*, 1979). It has since become a prevalent method of voice-source estimation (Larar *et al.*, 1985; Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986; Chan and Brookes, 1989; Brookes and Chan, 1994; Childers, 1995; Moore and Clements, 2004). In practice, however, some heuristic rules are imposed upon filter coefficients determined by the closed-phase analysis before it is taken as an estimate of the vocal tract parameters. The roots of the filter polynomial are obtained by factorisation, where certain properties are enforced:

Reflections of Maximum Phase Poles There is no guarantee that the vocal-tract filter estimated by covariance linear prediction is stable (Markel and Gray, 1982; Alku *et al.*, 2009) (contrary to autocorrelation linear prediction (Markel and Gray, 1982) or Stabilised Weighted Linear Prediction (SWLP) (Magi *et al.*, 2009)). However, the stable vocal tract system implies that all poles must lie within the unit circle. For this reason, any roots z of the filter polynomial with a magnitude great than one are replaced by their mirror image partners $\frac{1}{z^*}$, where z^* is the complex conjugate of z.

Removal of Real Poles (Alku *et al.*, 2009) remarks that "[poles on the positive real axis of the Z-plane are] unrealistic from the point of view of Fants source-tract theory of vowel production and its underlying theory of tube modeling". Because the theory cannot associate this pole with the vocal tract, any pole that appears on the positive real axis is removed (Wong *et al.*, 1979; Childers and Lee, 1991; Alku *et al.*, 2009). Failure to remove this pole can lead to distortions in the time-domain signal around the instant of glottal closure called "jags" (Wong *et al.*, 1979; Alku *et al.*, 2009). DC-constrained closed-phase linear prediction (Alku *et al.*, 2009), where the filter magnitude response is fixed a $0dB$ using constraints upon the solution of $\mathbf{x}$, is in part motivated to increase the likelihood that closed-phase analysis will yield pole locations at more realistic Z-domain coordinates.

Removal of Low Frequency Poles Other researchers also suggest the removal of poles whose centre frequency fall beneath a given threshold (Childers and Lee, 1991; Swerts and Veldhuis, 2001), which may result from an improperly placed analysis frame (thus corresponding to the glottal signal, not the vocal tract).

3.2.2 Time-Domain-Based Joint Estimation Techniques

In order to extend the analysis frame used for vocal tract estimation the closed-phase interval, the shape of the glottal excitation source during the glottal open phase, $g'[n]$ in the ARMAX and ARX speech models (Equations 3.5 and 3.6) above, must also be known. By assuming that $g'[n]$ is approximated by a voice-source model $g'_\theta[n]$ with parameters θ, the corresponding optimal vocal-tract filter coefficients $\mathbf{x}$ can be determined by solving the linear system according to Equation 3.15. Methods which attempt to determine the optimal vocal tract parameters $\mathbf{x}$ simultaneously with optimal voice-source parameters θ are called joint estimation techniques.

An example of a joint glottal source and vocal-tract filter estimation technique is illustrated in Figure 3.3. Like CPIF, time-domain joint estimation techniques operate in a pitch-synchronous manner, deconvolving and parameterising each voiced speech pulse. Also like CPIF, time-domain joint estimation techniques are sensitive to the placement of the analysis frame, which is usually supplied in terms of the glottal closing instant and the local pitch period T_0. However, determining the optimal solution for both θ and $\mathbf{x}$ is complicated by their nonlinear relationship, which can substantially increase computational requirements depending upon the degrees of freedom of $g'_\theta[n]$ (Fujisaki and Ljungqvist, 1986). Generally, this problem is solved by separating the optimisations into successive stages of voice-source parameterisation followed by vocal tract estimation. For this operation, researchers have adopted different strategies which are reviewed here. Note that for all methods no strategy is employed to determine p or q unless otherwise mentioned; it is assumed that this parameter is known *a priori* (*e.g.* by following Equation 2.9).

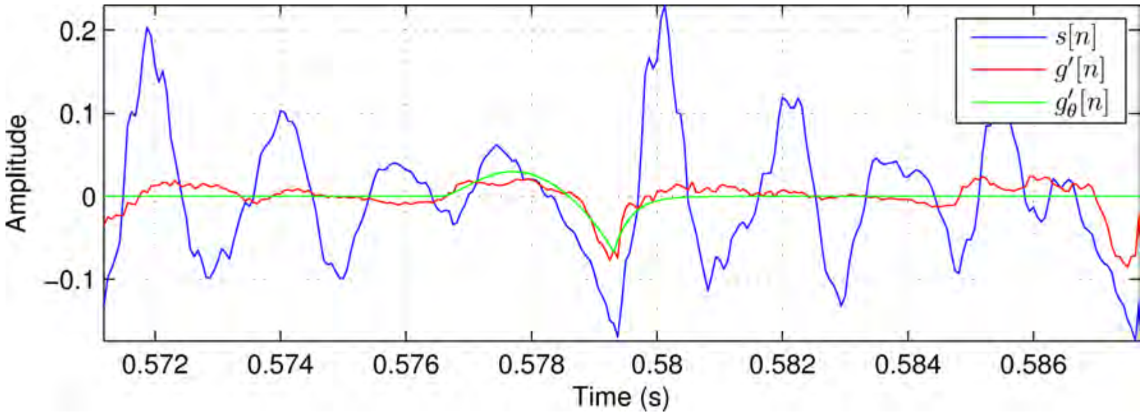


Figure 3.3: The figure above illustrates the convex optimisation time-domain joint estimation of the vocal-tract filter and the glottal excitation source $g'[n]$ from the speech signal $s[n]$, as described in (Lu, 2002). The method simultaneously fits a KLGLOTT88 model $g'_\theta[n]$ to the estimated voice source waveform $g'[n]$.

Hill Climbing (Fujisaki and Ljungqvist, 1986, 1987) In experiment comparing the abilities of different glottal models to satisfactorily model the speech signal, (Fujisaki and Ljungqvist, 1986) proposes a hill-climbing method to optimise the parameters of the ARX model. An iterative procedure generates the glottal model waveform according to θ, obtains a solution for $\mathbf{x}$ and modifies θ by changing a single element incrementally, searching for the minimum of the prediction error $||\mathbf{E}||$. In (Fujisaki and Ljungqvist, 1987), this method was extended to the ARMAX model, using the FL model for $g'_\theta[n]$.

Glottal LPC (Hedelin, 1986) An LPC-based vocoding system was described in (Hedelin, 1986) which determined the optimal parameters of their own glottal model along with the all-pole vocal tract. An iterative algorithm is described to solve the system, which is supervised to ensure that reasonable glottal parameters (*e.g.* no negative durations) and stable vocal-tract filters are obtained.

Simulated Annealing and Kalman Filtering (Ding *et al.*, 1994, 1997) Another ARMAX model parameterisation method is presented in (Ding *et al.*, 1994). The method approximates the voice-source signal using the KLGLOTT88 model, applies simulated annealing (Kirkpatrick *et al.*, 1983) to solve the nonlinear minimisation process for obtaining θ and the Kalman filtering algorithm (Kalman, 1960) to obtain the vocal tract system coefficients. The computational requirements of the algorithm was reduced in (Ding *et al.*, 1997), which also proposed a model order selection technique based on the assumption that the formants with centre frequencies below 3kHz should exhibit an average minimum bandwidth value not below a given threshold. $g'_\theta[n]$ is given by the KLGLOTT88 model.

Genetic Algorithm and Simulated Annealing (Funaki *et al.*, 1997) (Funaki *et al.*, 1997) proposes the "Glottal-ARMAX" speech model similar to the ARMAX speech model above, though the system has an additional noise signal $w[n]$ which is not assumed to be white rather is shaped by spectral envelope defined by a Moving-Average (MA) filter, defined by t number of coefficients c_h. The model may be expressed:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + \sum_{j=0}^{q} b_j g'[n-j] + \sum_{h=0}^{t} c_h w[n-h] + e[n] \tag{3.16}$$

The form of the solution of the system is therefore slightly different from Equation 3.15 and an additional layer of complexity it added with the addition of a further element of the speech model. The method adopts the KLGLOTT88 model to represent $g'_\theta[n]$ and employs a hybrid genetic algorithm (Holland, 1992) and simulated annealing approach to determine the optimal system parameters.

Convex Optimisation (Lu, 2002) A convex optimisation approach to glottal inverse filtering based upon the ARX was proposed in (Lu, 2002) and utilised and extended in (del Pozo, 2008; Pérez and Bonafonte, 2005, 2009). The glottal signal $g'_\theta[n]$ is approximated by the simplified Rosenberg model and the additional spectral tilt filter of the model $TL(z)$ which extends the Rosenberg model to the KLGLOTT88 model is incorporated into the estimated all-pole vocal-tract filter. Following analysis, the single real positive pole characterising $TL(z)$ is extracted from filter polynomial.

As review in Section 2.3.3.1, the glottal model has two parameters: the open quotient O_q and pitch period T_0. This simplicity yields a convex error function (Lu, 2002) ensuring that by varying O_q across its range, a global minimum can be found, see Figure 3.4.

In (del Pozo, 2008; Pérez and Bonafonte, 2009), $TL(z)$ is estimated from the speech signal by a first order linear predictive analysis and removed by inverse filtering before

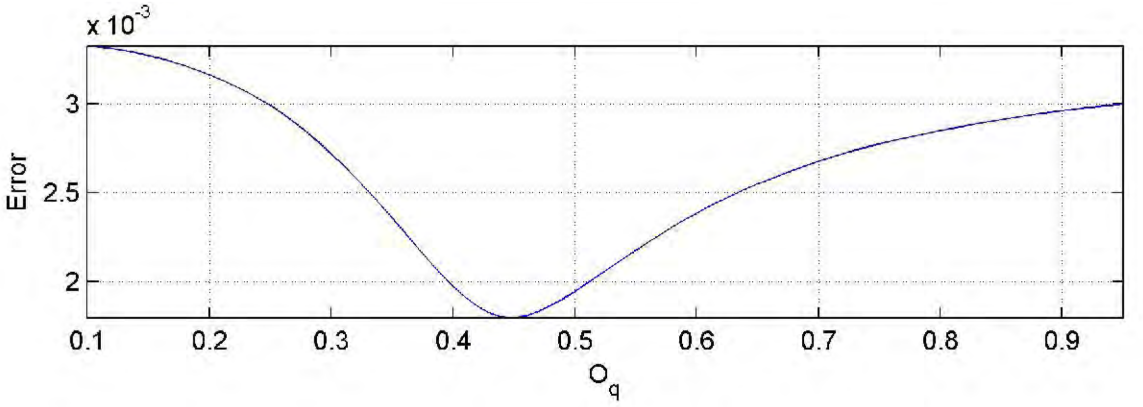


Figure 3.4: The above figure gives the error surface for the convex optimisation inverse filtering method (Lu, 2002) for a synthetic speech token with respect to the O_q parameter of the KLGLOTT88 model.

estimating the all-pole vocal-tract filter. This procedure, referred to as adaptive pre-emphasis (see Section 3.3.1), is used to increase robustness of analysis and improve the time-domain fitting of the return phase.

Low-band/Full-band ARX-LF (Vincent *et al.*, 2005) An iterative method to determine the ARX speech model parameters was proposed in (Vincent *et al.*, 2005), where $g'_\theta[n]$ is represented by the LF model, giving the ARX-LF speech model (ARX model of speech with LF model of the derivative glottal pulse). This method windows the speech signal $s[n]$ and glottal signal $g'[n]$ using a Hann function $2T_0+1$ in length, centred over the glottal closing instant. This ensures that, unlike other methods of ARX estimation, the analysis window does not change during estimation, which lends itself to increase computational speed (Vincent *et al.*, 2005).

Due to the complexity of the LF model, the method performs the vocal-tract filter estimation in two passes, a low frequency optimisation, followed by a full band optimisation. The first step estimates the open phase parameters of the LF model using the signal frame down-sampled to 2kHz. Decimating the frame in this manner serves to reduced the bandwidth of analysis so that the open phase parameters of the LF model waveform, which are mainly responsible for the glottal formant characteristics, can be estimated with decreased influence of the return phase characteristics. Initial estimates are taken from a discrete subset of LF model parameter configurations (Vincent *et al.*, 2005), and then refined using the simplex optimisation method (Nelder and Mead, 1965). Once the low bandwidth estimates of the open phase parameters are obtained, full bandwidth analysis is performed, optimising only for the return phase parameters.

The prediction error evaluated with a filter order p and glottal model parameters θ, $||E_\theta^p||$, will always decrease with increasing p. This makes the error function unsuitable for determining the optimal filter order. However, by normalising this error by the standard prediction error $||E_0^p||$ (calculated similarly to $||E_\theta^p||$ following from Equation 3.7), a new error function which does not exhibit this behaviour $\overline{E}_\theta^p$, termed the normalised prediction error, is obtained. This value is calculated as:

$$\overline{E}_\theta^p = \frac{||E_\theta^p||}{||E_0^p||} \tag{3.17}$$

As $||E_\theta^p||$ will always smaller than $||E_0^p||$, the normalised prediction error is always between 0 and 1 and independent of signal amplitude.

Iterative ARX (Fu and Murphy, 2006) Another two pass method for the ARX parameterisation was presented in (Fu and Murphy, 2006). First, the ARX problem is simplified by using the Rosenberg model to approximate the glottal flow, similar to the method of (Lu, 2002). The determined Rosenberg parameters are then used to obtain robust initial estimates of the more complex LF model. These parameters form the initial estimation for subsequent pass to identify the vocal tract and glottal model parameters using an interior trust-region to refine the LF model parameters, while the Kalman filtering algorithm adaptively identifies the vocal tract coefficients.

3.2.3 Time-Domain Voice-Source Parameterisation

Glottal inverse filtering yields an estimate of the voice-source signal: an example is given in Figure 3.5. This section discusses how time-domain parameterisation of a glottal model may be obtained from these estimates.

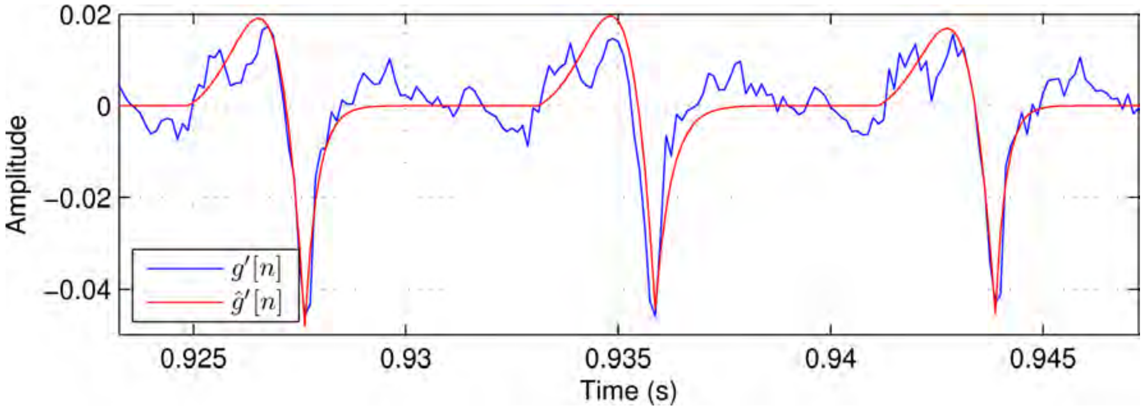


Figure 3.5: In the above figure are three derivative glottal flow pulses $g'[n]$, estimated from natural speech using CPIF and fitted by the LF model $\hat{g}'[n]$, using a method similar to (Strik, 1998).

Generally, time-domain glottal pulses are parameterised individually in two stages. The first stage obtains initial estimates of the model parameters. Once the parameters have been appropriately estimated, the parameters are then refined during a second optimisation. This section will detail these operations.

Initial parameter estimation Initial parameters are usually required in order to begin an optimisation procedure. For glottal signals, many initial parameters can be estimated directly from landmarks of the glottal pulse, though some parameters *e.g.* the LF models T_a parameter, may be more difficult (Strik, 1998; Lu, 2002). These landmarks correspond to various points along the time-domain signal, such as local extrema, zeros crossings, *etc.*

As estimated glottal flow waveforms are often corrupted by high frequency noise, the signal is often pre-processed by low-pass filter to reduce its effects (Strik, 1996). Convolution with a Blackman window 1ms in length is typical (Strik, 1998; Lu, 2002). It is important to note that any filtering operation will also affect the shape of the pulse - it is therefore advisable that the glottal pulse model to be fit on the signal also be filtered in the same way (Strik, 1996).

As many glottal source estimation techniques model the glottal pulse signal in some way, an alternative strategy to obtain the initial LF model parameters is to map them from the model used for glottal inverse filtering; for example, (Pérez and Bonafonte, 2005, 2009; Fu and Murphy, 2006) map the parameters of the KLGLOTT88 model used during glottal inverse filtering to LF model parameters.

Optimisation Once initial parameters have been estimated, they are then passed to an optimisation function which refines them by minimising an error criterion. The error criterion utilised is the energy of the residual signal (Strik, 1998; Lu, 2002; Fu and Murphy, 2006; Pérez and Bonafonte, 2005, 2009). Mathematically, the residual energy signal E_g between an estimated derivative glottal pulse $g'[n]$ spanning from glottal opening t_o to closure t_c and an derivative glottal model pulse model $g'_\theta[n]$ can be expressed:

$$E_g = \sum_{n=t_o}^{t_c} (g'[n] - g'_\theta[n])^2 \tag{3.18}$$

Attempts to use other perceptually weighted error and combined time/frequency-domain criteria have proved difficult (Strik, 1998).

The optimisation routines used for refining the initial estimations have varied. Regarding the fitting of the LF model to voice-source estimates, the work in (Strik, 1998) uses two algorithms: the simplex method (Nelder and Mead, 1965) to correct gross errors in the initial parameter estimations, followed by the Levenberg-Marquardt algorithm (Marquardt, 1963) to correct any final errors. However, (Tooher and McKenna, 2003) found that the fit was not improved and occasionally degraded by the second algorithm. (Lu, 2002; del Pozo, 2008) uses a constrained nonlinear optimisation which can be used to limit the parameter values to within a certain interval of the initial estimations.

Figure 3.5 shows the fitted LF model waveforms overlaid upon the estimated derivative glottal waveforms, following the method outlined in (Strik, 1998).

3.2.4 Discussion

As was shown in this section, the mathematical framework in which time-domain speech systems are described is straightforwardly posed, efficiently solvable and well-suited to the estimation and parameterisation of the voice-source signal. However, there are some issues which make time-domain methods difficult or inappropriate for this task.

Analysis Frame Location The main difficulty with time-domain voice-source estimation techniques is locating the analysis interval. For CPIF, determining the glottal closed phase from the speech signal requires the determination of two glottal events: glottal closure and glottal opening. As discussed in (Alku *et al.*, 2009) and illustrated in Figure 3.6, inaccuracies in the location of the closed phase can lead to large errors in the estimated glottal waveform. This is a result of the least squared energy criterion minimising the energy over the analysis interval when in actuality some energy from the glottal signal is present. While also requiring information of the glottal closing instant, joint estimation techniques do not require an estimate of the glottal opening instant, as this parameter will be estimated by the included glottal model, though they do impose the further assumptions about the glottal open phase and usually necessitate an estimate of the pitch period.

In order to identify the instants of glottal closure and opening, many researchers have utilised an ElectroGlottoGraph (EGG) signal synchronously recorded with the speech signal (Larar *et al.*, 1985; Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986; Chan and Brookes, 1989). Researchers have also developed algorithms to determine the glottal closed interval directly from the speech signal itself (see Chapter 4) - estimating the instant of glottal opening is a more difficult problem because it is usually a lower energy glottal event, though also for this reason, it has been claimed that knowledge of its exact location is not crucial (Brookes and Chan, 1994). (Plumpe *et al.*, 1997) proposes a closed-phase detection method based upon the stability of the estimation of the first formant of the vocal tract. An automatic

Kalman-filtering-based approach which excluded open phase data from the analysis was proposed in (McKenna, 2001). Other researchers have developed techniques which determine reasonable voice-source estimation results with imperfect glottal closed interval information. (Moore and Clements, 2004) proposed a fully automatic technique which would determine the best interval of analysis based upon the linear predictive analysis of the estimated glottal flow waveform given an approximate location of the glottal closing instant.

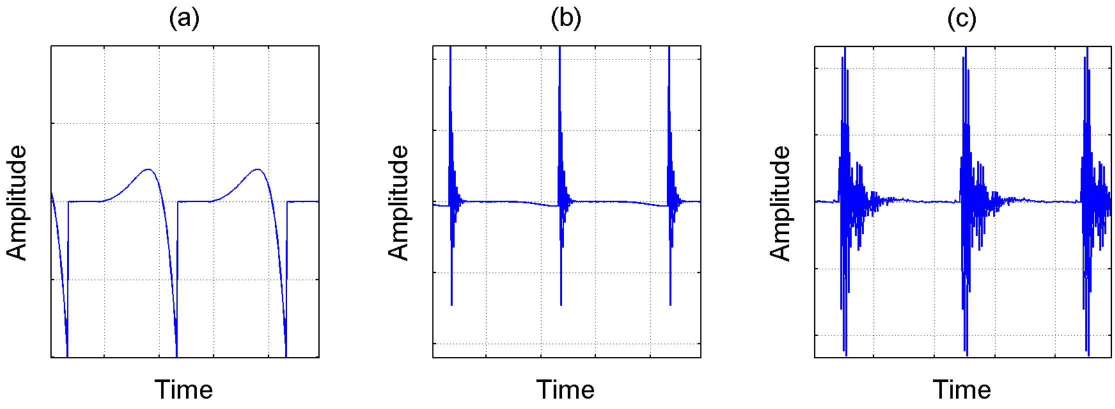


Figure 3.6: The figure above shows the CPIF-estimated glottal pulses using closed-phase interval offset by (a) 0 ms (b) +2.5 ms and (c) −2.5 ms.

Phase Distortion In addition to the position of the time-domain analysis frame, phase distortion from electro-acoustic equipment can also create difficulties for time-domain glottal estimation and parameterisation techniques (Holmes, 1975; Berouti *et al.*, 1977; Wong *et al.*, 1979; Hedelin, 1984; Akande, 2004; Walker and Murphy, 2007). A signal is phase distorted when the phase relationship between the components of the signal are altered in some way in a manner that is not necessarily perceived by a listener. Phase distortion is often attributed to audio equipment (Holmes, 1975; Berouti *et al.*, 1977; Strik, 1998) - (Doval and d'Alessandro, 2006) claims that most electro-acoustic equipment, *e.g.* studio microphones, non-anechoic chambers, tape recorders, *etc.* will introduce phase distortion. In addition to electro-acoustic equipment, phase distortion can also result from DSP operations subsequent to signal digitisation, *e.g.* high-pass filtering (Strik, 1996; Tooher and McKenna, 2003). This is a significant problem for time-domain-based analysis techniques.

Specifically in the context of voice-source analysis, the glottal waveform of phase-distorted signals may be have their shapes significantly changed, thus invalidating the assumptions that the voice-source signal exhibits a specific time-domain shape. Indeed, both (Berouti *et al.*, 1977) and (Akande, 2004) stated that phase distortion may eliminate the closed phase of a glottal cycle, particularly for low frequency voices. Time-domain-based error criteria for voice-source estimation (Equation 3.8) and parameterisation (Equation 3.18) are therefore inappropriate. Figure 3.7 shows the impact of a nonlinear linear recording system transfer function on synthetic glottal pulses.

In order to correct phase distortion, the transfer function of the recording system must be known and then inversely applied to the signal or implicitly canceled. It can be measured from the recording system itself using DSP approaches, *e.g.* via Maximum Length Sequences (Airas and Alku, 2007). Alternatively, the phase response can be explicitly canceled by inputting the time-reversed recorded signal into the recording system (Smith III, 2007). However, these methods require access to the recording system itself. If a reference signal of known signal shape (*e.g.* a square wave, an impulse train, *etc.*) has been recorded using the system, post-recording analysis can

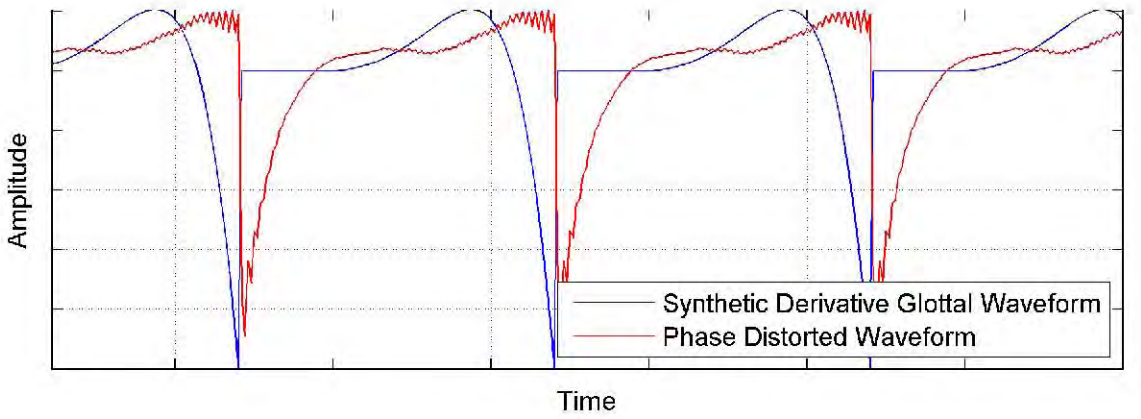


Figure 3.7: The above figure shows a synthetic glottal pulse overlaid with the corresponding phase-distorted pulse, obtained by filtering the original signal with the impulse response of a professional recording system. The glottal closed phase does not appear in the phase-distorted pulse.

be performed in order to construct a representation of the transfer function (Holmes, 1975; Berouti *et al.*, 1977; Brookes and Chan, 1994; Akande, 2004). These methods require the existence of a suitable reference.

(Hedelin, 1986) incorporates the estimation of a phase compensation filter along with the parameters of an ARX speech model. The filter $T(z)$ is assumed to be an all pass filter of the form:

$$T(z) = \frac{D(z^{-1})}{D(z)} \tag{3.19}$$

The coefficients of $D(z)$ are estimated by comparing the speech signal and the proposed ARX model using a linear prediction type approach - in this way it is similar to phase correction methods which use a reference signal. However, there are a number of issues with this approach. First, the order of $D(z)$ is unknown. Second, the filter stability is not assured. Third, the method cannot determine a linear phase offset (which relates to the problem of analysis frame position). Finally, though it is reported that $T(z)$ improves the perception of nasal sounds, its experimental utility for correcting phase distortion was not examined in (Hedelin, 1986). Instead, validation of the proposed approach utilises speech signals recorded using phase linear equipment. Therefore, rather than correcting phase distortion, the extra flexibility of this expanded speech model can compensate for deficiencies of the vocal tract and voice-source models and may then invalidate conclusions drawn from them.

3.3 Frequency-Domain-Based Approaches

Frequency-domain voice-source estimation methods work upon the second principle of glottal inverse filtering, where the derivative glottal flow signal is removed from the speech frame before estimation of the vocal tract. In this way, all of the following methods can be viewed as joint estimation approaches.

Removal of the glottal derivative signal can be performed by applying a transfer function representing the inverse of the derivative glottal flow model to the speech signal. Theoretically, this follows from source-filter theory:

$$S(z) = G'(z)V(z) \tag{3.20}$$

$$\Rightarrow V(z) = \frac{S(z)}{G'(z)} \tag{3.21}$$

Though based upon frequency-domain assumptions of the behaviour of the glottal signal, these algorithms can be performed in the time domain by utilising the difference equation representation of $G'(z)$.

A power spectrum interpretation representation of Equation 3.21 can be expressed by evaluating the Z-transform along the unit circle $z = e^{j\omega}$ and squaring:

$$|V(\omega)|^2 = \frac{|S(\omega)|^2}{|G'(\omega)|^2} \tag{3.22}$$

Therefore, source-filter theory dictates that power spectrum division can serves to remove the contribution of the derivative glottal flow from the speech signal.

The following glottal inverse filtering methods utilise these frequency-domain interpretation of source-filter theory as their foundations.

3.3.1 Pre-Emphasis Filtering Based Methods

Pre-emphasis-based voice-source estimation techniques use filter models of the glottal contribution, which are applied to the speech signal in order to cancel the glottal contribution $G'(z)$, as is outlined by Equation 3.21. The section describes these filtering based approaches.

First Order IIR Pre-Emphasis (Markel and Gray, 1982) In Section 2.3.3.1, a two pole model of the glottal flow signal $G(z)$ was reviewed. The poles of this model are of low frequency and near the unit circle (Markel and Gray, 1982; Doval and d'Alessandro, 2006). Following this model, $G(z)$ can be expressed:

$$G(z) = \frac{1}{(1 - e^{\alpha} z^{-1})^2} \tag{3.23}$$

where $\alpha \simeq 0$.

Recalling that the lip radiation characteristic $L(z)$ can be modeled as a first order differentiator, the linear source-filter of speech can be re-expressed:

$$S(z) = G(z)V(z)L(z) \tag{3.24}$$

$$= \frac{1}{(1 - e^{\alpha}z^{-1})^2}V(z)(1 - z^{-1}) \tag{3.25}$$

As the value of α is very small, one of the poles of the glottal flow is approximately canceled by lip radiation, and the equation can be expressed:

$$S(z) = \frac{1}{(1 - e^{\alpha}z^{-1})}V(z) \tag{3.26}$$

The remaining pole of $G(z)$ can therefore be canceled by a single pole high-pass filter similar to $L(z)$: this filter is referred to as a pre-emphasis filter $P(z)$. $P(z)$ is of the form $1-\beta z^{-1}$, where β is usually of the range $0.94 \leq \beta < 1$. The parameter β may be fixed at a particular value, or adaptive, where its value is estimated from the speech frame, *e.g.* using linear prediction techniques.

Pre-emphasis filters therefore flatten the contributions of the glottal derivative flow before estimating the vocal-tract filter (Markel and Gray, 1982). Essentially, this operation compensates for the influence of the spectral slope of the entire glottal derivative signal and thus improve the accuracy of the estimates of the vocal tract's upper formants. If it is assumed that the vocal tract impulse response is modeled as that of an all-pole filter, autoregressive filter estimator can be applied to the resulting signal frame in order to determine the optimal coefficients (Markel and Gray, 1982). Figure 3.8 illustrates the signals involved in this method of glottal inverse filtering.

IAIF (Alku and Laine, 1989) Iterative Adaptive Inverse Filtering (IAIF), introduced in (Alku and Laine, 1989), models the derivative glottal flow signal using a low-order all-pole filter, whose parameters are estimated in an iterative fashion

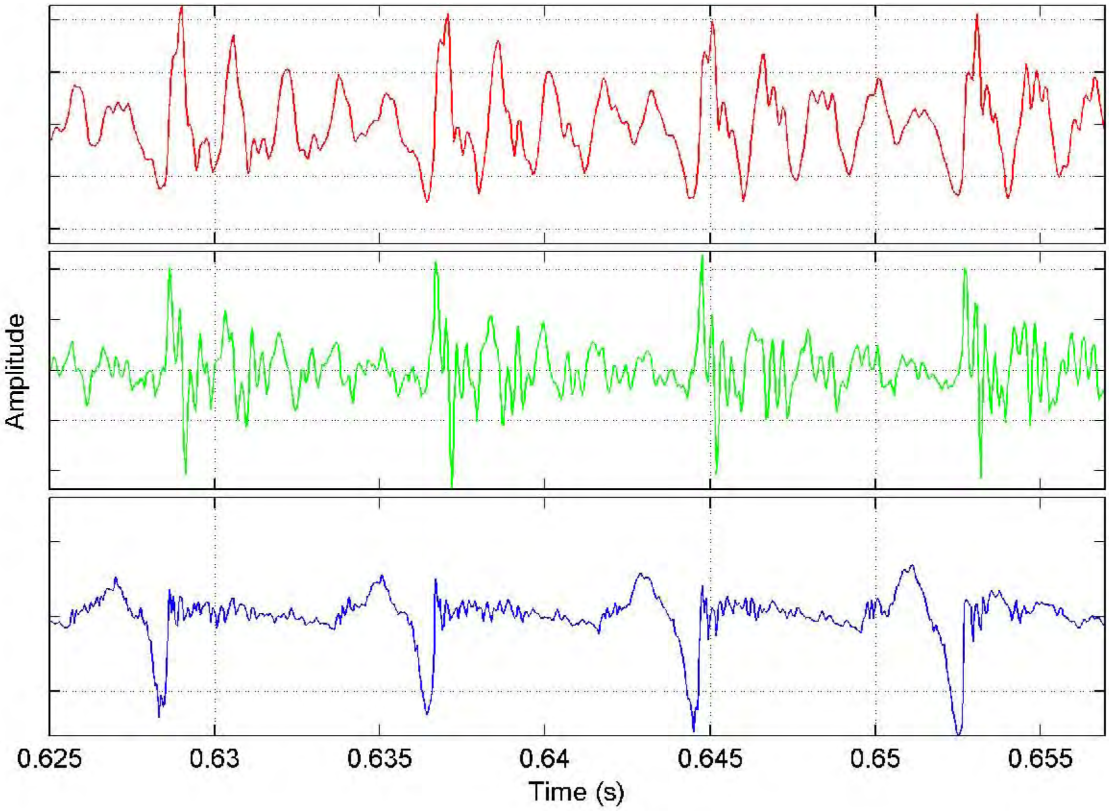


Figure 3.8: The above figures illustrate first-order FIR pre-emphasis filter voice-source estimation. (Top Panel) Speech signal. (Middle Panel) Pre-emphasised speech, using the filter $P(z) = 1 - 0.98z^{-1}$ (Bottom Panel) Voice-source waveform, obtained by inverse filtering the speech segment using the filter estimated from the pre-emphasised signal.

directly from the speech signal. The first stage follows the adaptive pre-emphasis glottal inverse filtering method described above, but the resulting derivative glottal flow estimate is then parameterised using another autoregressive analysis. This refined estimate of the glottal flow signal can then be used as a higher order pre-emphasis filter, which in a second pass further refines the vocal tract estimate. A figure illustrating the IAIF procedure is given in Figure 3.9.

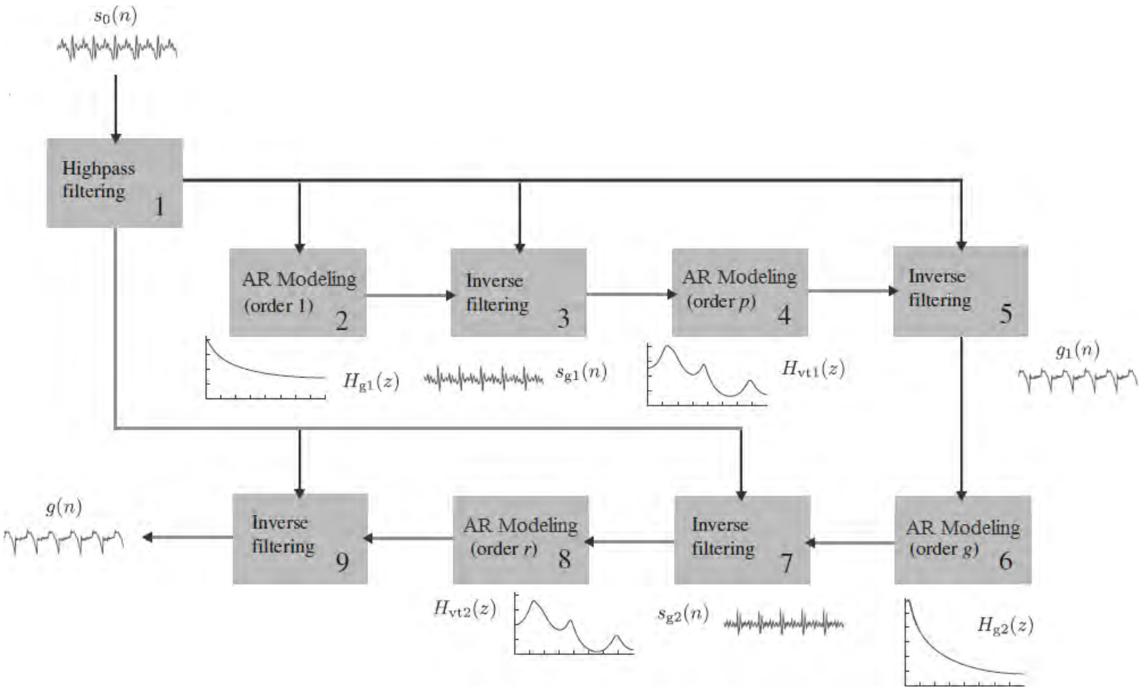


Figure 3.9: The above figure gives a schematic rendition of the IAIF algorithm, and is adapted from (Airas, 2008a). The blocks marked "AR Modeling" indicate an autoregressive analysis of given order. See text for details.

In (Alku and Laine, 1989), the autoregressive estimation was performed using autocorrelation method of linear prediction (Markel and Gray, 1982). However, this technique has a well known bias towards the peaks of the signal, due to the error cancelling property of the aliasing which occurs when the spectrum is sampled at discrete frequencies (El-Jaroudi and Makhoul, 1991; Makhoul, 1975). An improved IAIF algorithm using the Discrete All-Pole (DAP) modeling (El-Jaroudi and Makhoul, 1991) algorithm which avoids this error cancelation by employing a different error criterion, the Itakura-Saito distance measure (I-S) (Itakura and Saito, 1968), was proposed in (Airas, 2008a). Both the DAP and LPC algorithms and the Itakura-Saito error function are discussed in Appendix B.

AEVT (Akande and Murphy, 2005) The Adaptive Estimation of the Vocal Tract transfer function (AEVT) (Akande and Murphy, 2005) also uses pre-emphasis filters to remove the influence of the glottal signal. However, unlike the above pre-emphasis methods, the AEVT method uses a "dynamic, multi-pole, zero-phase lag high-pass" filter in order to remove the glottal contributions. The parameters of this filter are chosen according to low frequency gain criterion upon the resulting vocal tract estimate. If the glottal signal is appropriately canceled, the pre-emphasis gives an extended "pseudo-closed" phase, which can then be analysed using covariance linear prediction in order to estimate the vocal tract. The analysis window and filter order are also varied until certain criteria regarding the phase characteristics and formant bandwidths are satisfied.

Though this method is described as a pre-emphasis voice-source estimation technique because of the filters utilised to cancel the glottal signal, the subsequent pseudo-closed-phase analysis has much in common with CPIF.

3.3.2 Power-Spectrum-Based Joint Estimation Techniques

As opposed to transfer function representations of the derivative glottal flow, some researchers have opted to use power spectrum representations of time-domain models. As these methods assume an all-pole vocal-tract filter, these methods are similar to time-domain ARX approaches, albeit in a different domain.

Examples of this approach are given in (Fröhlich *et al.*, 2001; Arroabarren and Carlosena, 2003) which are broadly similar. They operate upon the power spectrum of a signal which is estimated using the Discrete Fourier Transform (DFT). Once in the frequency domain, the prominent peaks of the spectrum, representing samples of the spectral envelope, can then be extracted.

These spectral peaks represent the sampled speech signal envelope. The effect of a derivative glottal flow model of a given parameter configuration is removed by spectral division (or equivalently, log magnitude domain subtraction) yielding an estimate of the vocal tract. A p^{th} order all-pole model is then fit to the resulting spectrum using the DAP algorithm, which gives the Itakura-Saito error quantifying the goodness of fit of the envelope to the estimated vocal tract, and indirectly the glottal model parameter configuration.

If the derivative glottal flow model approximates the actual derivative glottal flow and the p^{th} order all-pole filter is appropriate for the vocal tract, the Itakura-Saito error will be minimised. Thus, similarly to time-domain methods of joint estimation glottal inverse filtering, the parameters of the glottal source model are searched over their ranges via an exhaustive search to yield a robust estimate of their parameters.

SIM (Fröhlich *et al.*, 2001) The Simultaneous Inverse filtering and Model matching (SIM) method (Fröhlich *et al.*, 2001) is a power-spectrum-based voice-source estimation/parameterization method which follows the above routine. It utilises the LF model for the derivative glottal flow and a modified version of the DAP algorithm where the initialisation and termination conditions of the iterative loop are slightly altered.

Additionally, following the exhaustive search of a discrete set of parameter configurations, the SIM method refines the LF model parameters using a two step approach due to the complex error surface of such an optimisation. First, the initial parameters controlling the pulse shape $\{t_p, t_e, t_a\}$ are refined using two optimisation algorithms: the simplex algorithm (Nelder and Mead, 1965) followed by Powell's method (Powell, 1964). Second, in order to estimate the LF model scale parameter E_e, a time-domain error criterion is used between the synthesised LF model pulses and estimated derivative glottal flow waveform. This error is then minimised using a Brent's method optimisation (Brent, 2002), while adjusting the lag to ensure the signals are matched in phase.

An example of the SIM algorithm is given in Figure 3.10.

GSBIF (Arroabarren and Carlosena, 2003) Glottal-Spectrum-Based Inverse Filtering (GSBIF) is another power-spectrum-based voice-source estimation/parameterization method, introduced in (Arroabarren and Carlosena, 2003). However, unlike the SIM method, this approach utilises the KLGLOTT88 model of glottal flow. Like the time-domain convex optimisation method (Lu, 2002), the spectral tilt filter of the KLGLOTT88 model $TL(z)$ is also estimated simultaneously with the vocal tract. Similarly, the model's open quotient parameter is determined via an exhaustive search.

3.3.3 Frequency-Domain Voice-Source Parameterisation

In contrast with time-domain methods, frequency-domain attempts to fit voice-source model signals are less prevalent in the literature. This is probably due to the fact that most derivative glottal flow models are defined in the time domain.

One method is described in (Swerts and Veldhuis, 2001), a study where the effect of the pitch contour of speech is compared against the characteristics of the source. This

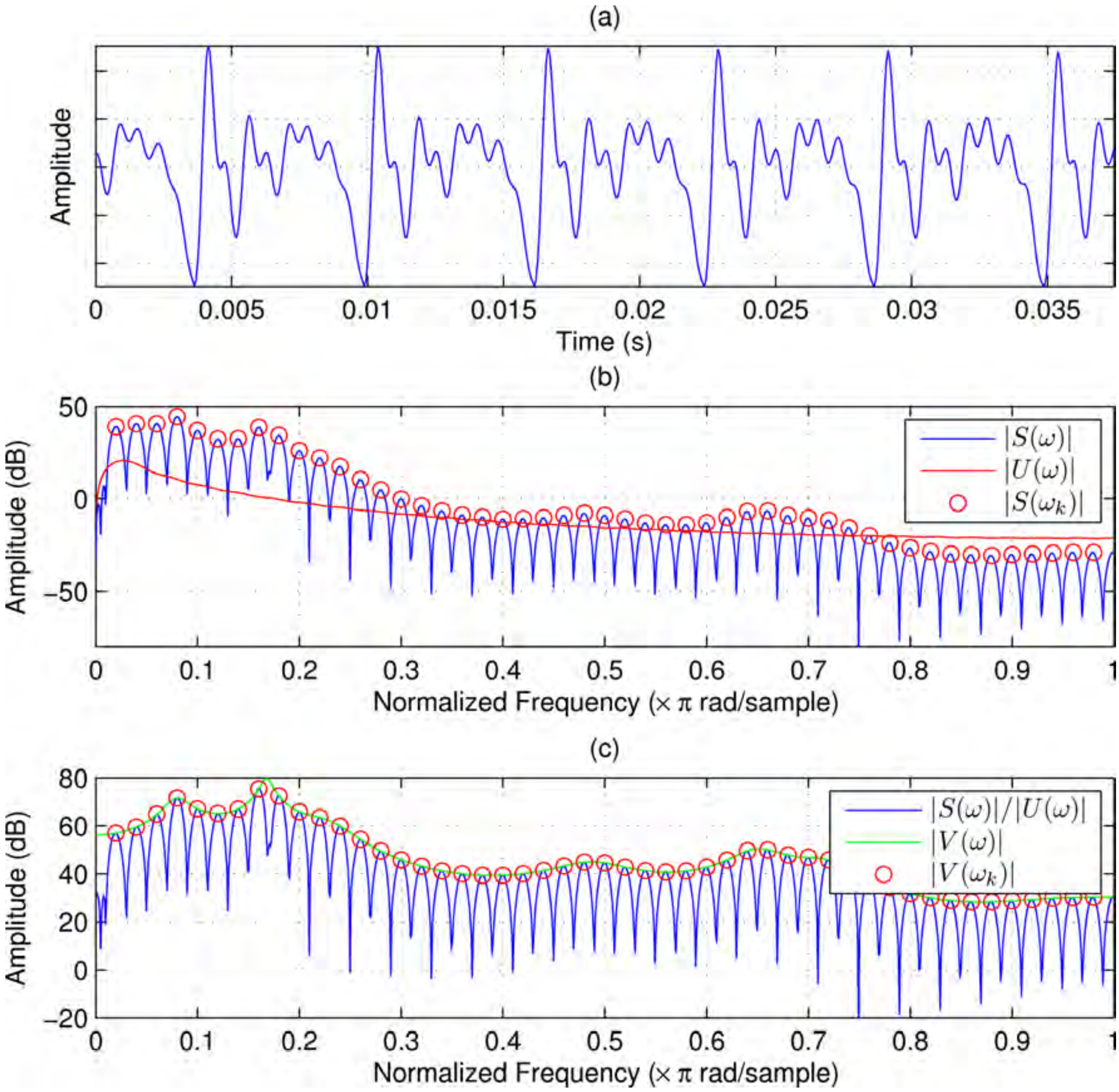


Figure 3.10: The above figures give a overview of the SIM method of voice-source estimation/parameterization (Fröhlich *et al.*, 2001). The figures shows: (a) The speech signal. (b) The speech and LF model spectra. The LF model spectrum approximates the derivative glottal flow signal. (c) The vocal tract spectrum and all-pole filter envelope.

method uses the Kullback-Leibler distance function (Kullback, 1987) for probability distributions to quantify the distance between the estimated glottal derivative flow spectrum $G'(\omega)$ and the model spectrum $G'_\theta(\omega)$, given by the LF model. The error is expressed:

$$E_{KL} = \sum_{k=1}^{L} |G'_\theta(\omega_k)|^2 \ln \frac{|G'_\theta(\omega_k)|^2}{|G'(\omega_k)|^2} \tag{3.27}$$

where w_k is the k^{th} harmonic of the spectrum and L represents the number of harmonics in the available bandwidth. Both spectra are power normalised before comparison, *i.e.* $\sum_{k=1}^{L} |G'(\omega_k)|^2 = \sum_{k=1}^{L} |G'_\theta(\omega_k)|^2 = 1$. This method therefore does not fully parameterise the glottal model signal: its scale factor is not estimated. However, not mentioned in (Swerts and Veldhuis, 2001) is the method of obtaining initial parameters or the algorithm used to optimise them.

The frequency-domain voice-source parameterisation method described in (Kane *et al.*, 2010) more closely follows the procedure of the time-domain methods above where initial values are found and then refined using an optimisation algorithm used to minimise some error. The voice-source estimate is fit using the LF model. The method determines an initial estimate of the LF model parameters by calculating $H1 - H2$, the difference between the first two harmonic of the voice-source estimate (Klatt and Klatt, 1990) supposedly indicative of the open quotient of the signal, and matching with values contained in a look-up table. Similarly to (Vincent *et al.*, 2005), a two stage low-band, full-band approach is then adopted to refine these initial parameters. The simplex optimisation method (Nelder and Mead, 1965) is used to minimise the residual difference error between the magnitude of the lowest 6 harmonics of the estimated and model voice source. The error is weighted so as to double the error the first two harmonics, which are known to be important for voice quality:

$$E_U = \sum_{k=1}^{6} \left(|G'(\omega_k)| - |G'_\theta(\omega_k)|\right)^2 + \sum_{k=1}^{2} \left(|G'(\omega_k)| - |G'_\theta(\omega_k)|\right)^2 \tag{3.28}$$

Once the low frequency portion of the frame has been optimised, the full bandwidth signal is refined to obtain a better match for the return phase value. Though this parameter mostly affects the upper frequencies, a certain effect is also imparted upon the lower frequency portion of the spectrum, due to the principle of area balance. In order to lessen the effect of the second high frequency optimisation upon the already-fitted lower frequencies, (Kane *et al.*, 2010) relaxes the area balance constraint. Figure 3.11 shows the estimated glottal source and fitted LF model spectra.

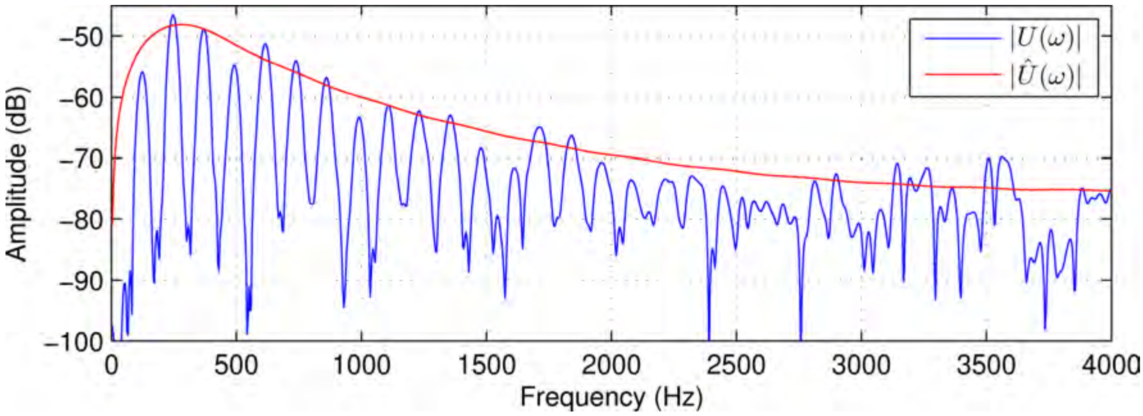


Figure 3.11: The figure above shows the amplitude spectrum of an estimated glottal waveform and a fitted LF model spectrum, following the procedure given in (Kane *et al.*, 2010).

3.3.4 Discussion

Frequency-domain approaches overcome the main drawback of time-domain voice-source estimation, namely sensitivity to the phase spectrum, whether in the position of the analysis frame or the presence of phase distortion. Time-domain voice-source estimation techniques demand a magnitude and phase match of the signal in order to parameterise the speech model, implicit in the time-domain residual energy criteria. Conversely, power spectrum and pre-emphasis filtering methods (AEVT excluded) model the magnitude spectrum of the signal only and are largely robust to the phase spectrum. This results from the determination of the all-pole envelope of the analysis signal from its power spectrum (utilised by DAP) or its autocorrelation coefficients (utilised by LPC and trivially calculated from the power spectrum) (Appendix B contains more details).

This make these approaches an attractive option for the robust analysis of the voice-source signal. However, other factors prove disadvantageous. While both the basic pre-emphasis and IAIF methods of glottal inverse filtering are computationally inexpensive and can yield reasonable results, the methods are not robust across all speech types and phonemes due to the determination of the glottal model parameters. Particularly with speech segments containing a first formant of low centre frequency (*e.g.* /i/), the first and glottal formants may overlap with the consequence that the first formant not be successfully removed from the voice-source estimate (Alku, 1992).

In order to achieve increased robustness in these scenarios, a brute force approach using a large set of different glottal models may be preferred. In this way, even if the spectral properties of the analysis signal are such that an accurate glottal model cannot be determined directly, an exhaustive search can potentially do so. Moreover, the IAIF and first order pre-emphasis methods use low order all-pole glottal model filters: it is more useful to simultaneously parameterise a more sophisticated glottal model, without requiring a separate optimisation stage.

This is the approach taken by the power-spectrum-based joint estimation techniques outlined in this section. However, these methods assumed that the filter order is known *a priori* (the SIM method follows the usual rule of thumb, while the method of (Arroabarren and Carlosena, 2003) utilises unusually high filter orders). Additionally, signal samples are obtained throughout the frequency band, without regards for any noise components which may corrupt the spectrum. Further, the scale factor of the incorporate model is not estimated by the SIM method in the frequency, rather determined in the time domain, where the approach adopts the usual time-domain least-squares error and implicitly the associated phase matching requirement.

3.4 Phase-Spectrum-Based Approaches

There is some evidence that the voiced speech signal may be interpreted as a mixed phase system, containing both minimum and maximum phase components (Bozkurt, 2005). Maximum phase characteristics are exhibited by a signal which exponentially increases with time, while a signal exhibits minimum phase characteristics when it decays with time (Quatieri, 2001). In the case of a pulse of voiced speech, the maximum phase components are attributed to the open phase of the derivative glottal flow signal, while the minimum phase components of speech signal are determined by the vocal tract and the return phase of the glottal signal. Thus, if these components of the speech signal can be separated, it also serves to deconvolve the speech signal into glottal open phase and vocal tract/return phase contributions. Figure 3.12 illustrates the principle behind these separation methods. These methods are also some times referred to as causal-anticausal decomposition approaches.

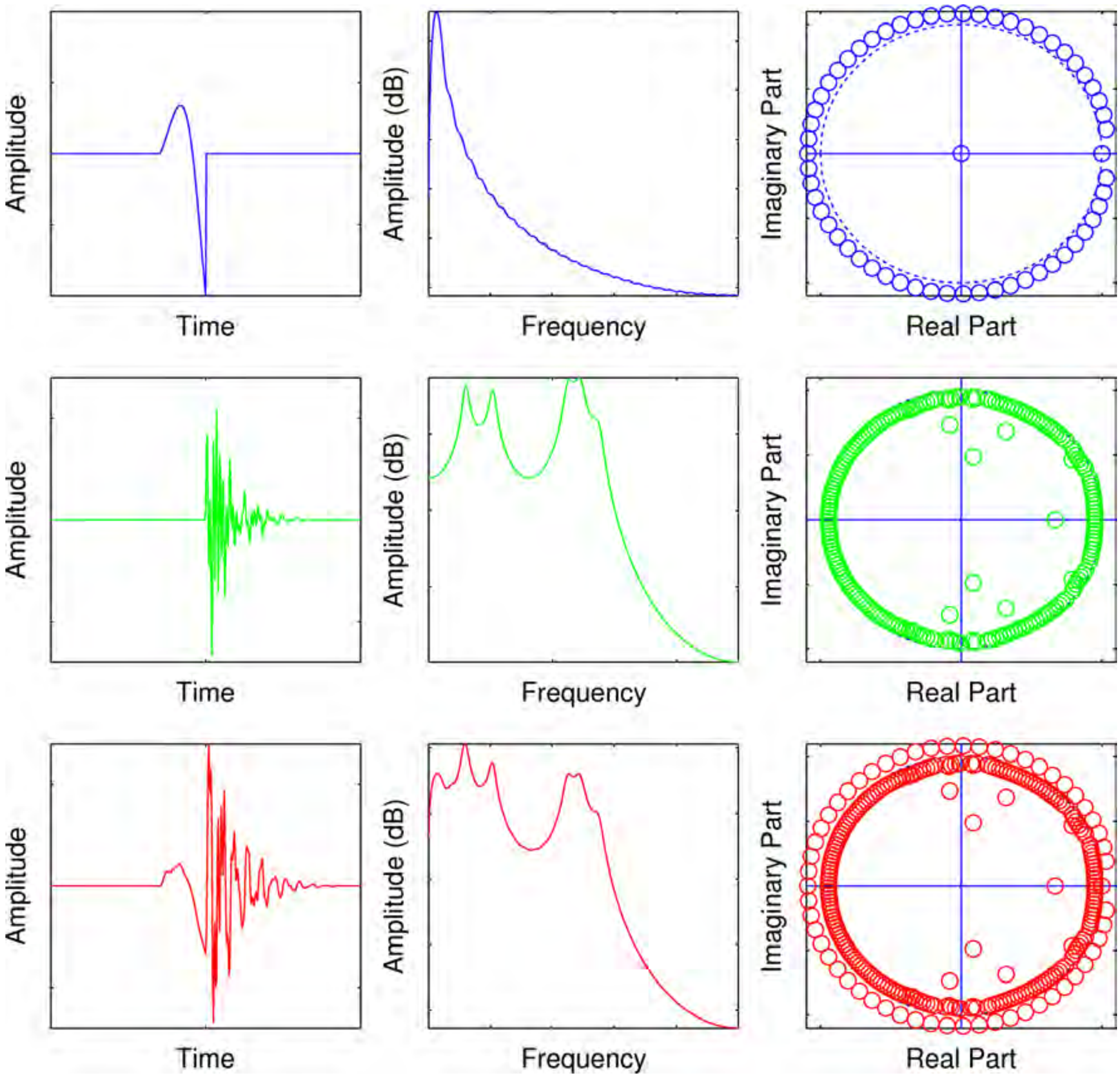


Figure 3.12: The above figures give the principles of minimum/maximum phase decomposition glottal inverse filtering. The above figures show, in three different domains (left, time domain; middle, log magnitude frequency-domain; right, Z-domain), representations of the glottal derivative flow (top), the vocal-tract filter (middle) and the speech signal (bottom).

3.4.1 Minimum/Maximum Phase Speech Decomposition

Two methods of source-filter decomposition based on the mixed-phase properties of speech have been proposed: the Zeros of the Z-Transform (ZZT) (Bozkurt, 2005) and Complex Cepstrum Decomposition (CCD) (Drugman *et al.*, 2009a). Both techniques achieve similar decompositions, but are based on slightly different mathematical approaches. Unlike other voice-source estimation methods, minimum/maximum phase separation techniques do not impose a parametric model upon the speech pulse: they only separate the estimated minimum and maximum phase components of the analysis frame.

ZZT (Bozkurt, 2005) The ZZT approach is based upon the fact that the unit circle separates the minimum and maximum phase components of a signal. Thus, by determining the zeros of a signal using the Z-transform, the signal decomposition can be performed. The Z-transform of the signal $x(n)$, spanning from $n = 0 \cdots N-1$, is calculated as follows:

$$X(z) = \sum_{n=0}^{N-1} x(n) z^{-n} \tag{3.29}$$

The zeros z_m of $X(z)$ can then be determined by factorisation:

$$\begin{aligned} X(z) &= \frac{x(0)\prod_{m=0}^{N-1}(z - z_m)}{z^{N-1}} & (3.30)\\ &= \frac{x(0)\prod_{k=0}^{N_i}(z - z_{in,k})\prod_{k=0}^{N_o}(z - z_{out,k})}{z^{N-1}} & (3.31) \end{aligned}$$

Once factorised, the zeros z_m are subdivided into two categories depending on their distance from the origin. The maximum phase zeros z_{out} are the subset of z_m which satisfy $|z_m| > 1$, while minimum phase zeros z_{in} are the subset of z_m which satisfy $|z_m| < 1$. Zeros which lie on the unit circle itself are said to be neither maximum nor minimum phase, but for the purposes of this kind of decomposition it is usual to associate it with the glottal source.

CCD (Drugman *et al.*, 2009a) CCD exploits the relationship between the complex cepstrum of a signal and the roots of the Z-transform. The complex cepstrum $\hat{x}_n$ of a signal x is defined as the inverse discrete-time Fourier transform (DTFT) of the complex logarithm of the $X(\omega)$, where $X(\omega)$ is the DTFT of the signal (Oppenheim and Schafer, 1975). Mathematically, it can be expressed:

$$\hat{x}_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln X(\omega) e^{j\omega n} d\omega \tag{3.32}$$

The complex logarithm function is calculated:

$$\ln z = \ln |z| + i\angle z \tag{3.33}$$

where $\angle z$ represents the unwrapped phase envelope of the set of complex numbers z.

It can be shown that the zeros of the Z-transform are related to the complex cepstrum by the following relationship (Oppenheim and Schafer, 1975):

$$\hat{x}_n = \begin{cases} -\sum_{k=1}^{N_i} \frac{z_{in,k}^n}{n}, & n > 0 \\ \sum_{k=1}^{N_o} \frac{z_{out,k}^n}{n}, & n < 0 \\ |x(0)| & n = 0 \end{cases} \tag{3.34}$$

Decomposition of the signal can be performed by inverting the cepstrum transform upon the appropriate portion of the complex cepstrum corresponding to the minimum and maximum phase components.

Both techniques are heavily dependent upon the careful placement and shape of the analysis window for proper separation (Bozkurt *et al.*, 2004; Drugman *et al.*, 2009a), as the signal frame must be such that the minimum/maximum phase model of the pulse is preserved. Blackman windows, $2T_0 + 1$ in length where T_0 is the local pitch period, centred over the instant of glottal closure are typically employed.

CCD is “sensibly” identical to the ZZT approach (Drugman *et al.*, 2009a), but because the approximation of the complex cepstrum can exploit the Fast Fourier Transform (FFT), it can be performed more efficiently than the ZZT approach, which depends on the factorisation of high order polynomials (Drugman and Dutoit, 2009; Pedersen *et al.*, 2010). However, differences between the ZZT and CCD are due to the time aliasing which may occur in the calculation of the complex cepstrum using the DFT, as an approximation of the DTFT. Additionally, the CCD requires the calculation of the complex logarithm which necessitates knowledge of the unwrapped phase envelope, which must be estimated from the wrapped version. These issues may be alleviated by sufficiently zero-padding the DFT (Drugman and Dutoit, 2009).

3.4.2 Phase-Spectrum-Based Joint Estimation Techniques

Similarly to the joint estimation techniques were proposed using power spectrum and time-domain signal information, several joint parameterisation approach based upon the minimum/maximum phase model of the speech signal have also been proposed (Degottex *et al.*, 2011). These methods are based upon the properties of the phase spectrum of a convolutive residual signal dependent only upon the parameters of the shape of the glottal signal and a linear offset. The minimum phase components of the speech signal can be canceled using a minimum phase envelope estimator while the maximum phase components are minimised using a glottal model. If the glottal model approximates the voice-source signal correctly, the phase spectrum of the convolutive residual is minimised. These methods therefore attempt to optimise glottal parameters based upon the least-squares phase (ϕ LS) of the convolutive residual and related error criteria.

Given a sinusoidal speech model $S(\omega_k)$ defined at harmonically related frequencies $\omega_k = k\omega_1$, the convolutive residual signal $R(\omega_k)$ is obtained by the spectral division by an approximated model:

$$R^{(\theta,\phi)}(\omega_k) = \frac{S(\omega_k)}{\hat{S}^{(\theta,\phi)}(\omega_k)} \tag{3.35}$$

$$= \frac{S(\omega_k)}{e^{jk\phi}G'^{\theta}(\omega_k)V_{-}^{\theta}(\omega_k)} \tag{3.36}$$

where $\hat{S}^{(\theta,\phi)}(\omega_k)$ is the modeled speech signal, comprising the derivative glottal flow model $G'^{\theta}(\omega_k)$ parameterised by shape parameter θ, minimum phase vocal tract model $V_{-}^{\theta}(\omega_k)$ and linear phase component $e^{jk\phi}$ to account for the difference in time position of the model. The minimum phase vocal tract model $V_{-}^{\theta}(\omega_k)$ is obtained by following equation:

$$V_{-}^{\theta}(\omega_k) = \mathcal{E}_{-}\left(\frac{S(\omega_k)}{G'^{\theta}(\omega_k)}\right) \tag{3.37}$$

where $\mathcal{E}_{-}$ is a function which calculates the minimum phase envelope of its argument. Throughout (Degottex *et al.*, 2011), this information is obtained via the real cepstrum (Oppenheim and Schafer, 1975).

Assuming that the magnitude spectrum of the estimated vocal tract is approximated sufficiently well, the convolutive residual is flat across all frequency bands, *i.e.*

$$|R^{(\theta,\phi)}(\omega_k)| = 1, \forall\omega_k \tag{3.38}$$

Therefore, any errors within the model of the speech spectrum $\hat{S}(\omega)$ can be attributed to the phase spectrum of $R^{(\theta,\phi)}(\omega_k)$.

The phase spectrum of $R^{(\theta,\phi)}(\omega_k)$ is minimised when θ approximates the actual derivative glottal flow and the ϕ matches the appropriate linear phase shift. If the phase of $R(\omega_k)$ is perfectly minimised, *i.e.* $\angle R(\omega_k) = 0$, as $|R(\omega_k)| = 1$ the convolutive residual signal is equivalent to the Dirac delta function $\delta[n]$, *i.e.*

$$\delta[n] = \mathcal{F}^{-1}\left(R(\omega_k)\right) \tag{3.39}$$

where

$$\delta[n] = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \tag{3.40}$$

and $\mathcal{F}^{-1}$ represents the inverse discrete Fourier transform.

(Degottex *et al.*, 2011) introduces several techniques which parameterise the speech signal based upon this approach, using the transformed LF model as the approximation of $G'^{\theta}(\omega_k)$.

MSP The first error criterion for the estimation of the glottal source parameters θ and offset ϕ is given by the Mean Squared Phase (MSP):

$$MSP(\theta,\phi) = \frac{1}{N}\sum_{k=1}^{N}\left(\angle R^{(\theta,\phi)}(\omega_k)\right)^2 \tag{3.41}$$

where N is the number of harmonics, *i.e.* $k = 1 \cdots N$. Figure 3.13(a) shows the gives the error surface of the MSP function of a synthetic signal frame.

MSPD If the parameters θ and ϕ exactly model the speech spectrum given by $S(\omega_k)$, the convolutive residual will represent a Dirac delta function. However, in the

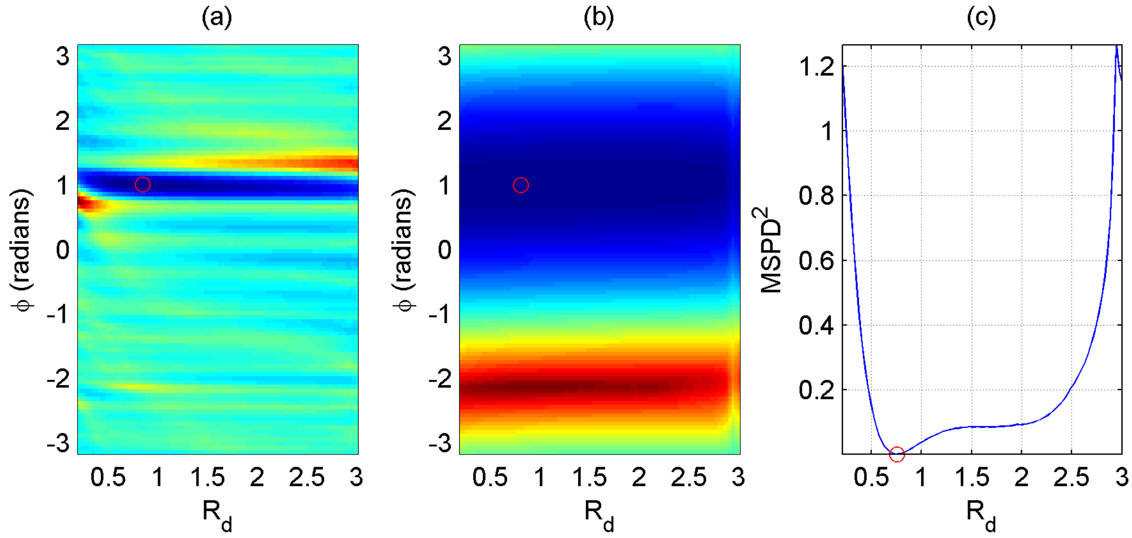


Figure 3.13: The above figures give the error surfaces of the (a) MSP, (b) MSPD and (c) MSPD^2 functions for a synthetic speech signal. For each analysis, $N = 12$ and the optimal parameters are $R_d = 0.8$ and $\phi = 1$.

case that the glottal model shape parameter θ is correct and ϕ is incorrectly estimated, $R^{(\theta,\phi)}(\omega)$ is given by an offset Dirac delta function, *i.e.* $\mathcal{F}^{-1}\left(R(\omega_k)\right) = \delta[n - n_0]$ where n_0 is the offset. Instead, the phase spectrum is given by a linear function, the slope of which is a constant. This is similar to the observations utilised for glottal closing instant estimation based upon the group delay function, reviewed in Section 4.2.

Thus, the sensitivity of the convolutive residual to the position of the analysis frame can be reduced by applying a difference operation (denoted by Δ) to the phase angle of the convolutive residual:

$$MSPD(\theta, \phi) = \frac{1}{N} \sum_{k=1}^{N} \left(\Delta\angle R^{(\theta,\phi)}(\omega_k)\right)^2 \tag{3.42}$$

The minimum of the $MSPD$ function is the same as MSP, however errors in the position of the pulse are now added to the error, instead of scale it. Figure 3.13(b) shows the gives the error surface of the MSPD function of a synthetic signal frame which shows less sensitivity to the analysis position.

MSPD^2 Though of reduced influence, the MSPD function is still sensitive to the position error, as the error corresponds to the (scaled) average group delay of $R^{(\theta,\phi)}(\omega_k)$. As noted above, if the group delay is the same throughout all frequency bands, *i.e.* $R^{(\theta,\phi)}(\omega_k)$ represents an offset Dirac function, that is sufficient condition to indicate a successful parameterisation of the glottal source. Therefore, by applying a subsequent difference operation to the MSPD function, the dependency upon ϕ can be removed. (Degottex *et al.*, 2011) also then performs a subsequent anti-difference operation (Δ^{-1}). This operation produces the MSPD^2 function:

$$MSPD^2(\theta) = \frac{1}{N}\sum_{k=1}^{N}\left(\Delta^{-1}\Delta^2\angle R^{\theta}(\omega_k)\right)^2 \tag{3.43}$$

Figure 3.13(c) gives the two dimensional error function of the MSPD^2 function for a synthetic speech segment.

3.4.3 Discussion

Methods which decompose the speech signal based upon its phase characteristics are promising considering the relatively few assumptions about the speech signal is required - unlike other methods which assume that the vocal tract is a minimum phase all-pole filter, these methods allow that it also contain zeros, an thus allow better modeling of nasals and other speech sounds known to invalidate the all-pole assumption. However, like time-domain glottal inverse filtering methods, phase-based methods of glottal source estimation also require precise information regarding the placement of the analysis frame and impose assumptions upon the time-domain shape of the signal.

The critical time reference point for these approaches is given by the GCI, as this instant demarcates the boundary between the maximum and minimum phase portions of the speech pulse. In order to satisfy the minimum/maximum phase speech model, the glottal contributions generally decrease in amplitude to the left of this point, while the minimum phase component decrease to the right of this point - windowing of the signal is also critical. Speech waveforms which do not adhere to this signal shape (*e.g.* phase-distorted waveform, incorrectly-placed analysis frames) may give spurious results. Methods based on the chirp Z-transform (where the Z transform is analysed at a boundary other than the unit circle (Rabiner *et al.*, 1969)) have also been proposed (Drugman and Dutoit, 2010) in order to increase the robustness of the method to the analysis frame time placement.

In addition to being sensitive to the position of analysis, the performance of minimum/maximum phase decomposition approaches also deteriorates significantly in the presence of noise (Drugman, 2011).

The maximum phase portion of the derivative glottal waveform is only attributed to the open phase of the glottal cycle - the return phase is included along with the vocal tract in the minimum phase portion of the waveform. For this reason, a strictly maximum/minimum phase decomposition is insufficient for obtaining a full representation of the glottal waveform. However, the method can be used in conjunction with other analysis algorithms, for example, it is used to narrow the necessary exhaustive parameter search of the ARX-LF time-domain joint estimation technique of glottal inverse filtering (Vincent *et al.*, 2005) in (d'Alessandro, 2009). Similarly, the joint estimation methods (MSP, MSPD, MSPD^2) utilise the transformed LF model to approximate the maximum phase characteristics of the signal, which then also imply a return phase.

Compared with the CCD and ZZT, the phase-based joint estimation methods are more robust to noise as these methods are based upon a sinusoidal model of the speech signal which can then be used to adopt a two band speech model. Additionally, these methods are also more robust to the position of the analysis frame. However, as the error criteria is based upon phase minimisation, accounting only for the minimum phase vocal tract and mixed phase glottal source model, inaccurate parameters will be obtained for phase-distorted signals.

3.5 Conclusions

This section has reviewed state-of-the-art techniques for the estimation and parameterisation of the voice-source signal. The theoretical and practical details of the operation of glottal estimation techniques were explained and reviewed, in addition to glottal model parameterisation methods.

Because it is a blind-deconvolution problem, methods of glottal source estimation are required to make certain assumptions regarding the glottal source and/or the vocal-tract filter. Time-domain methods of voice-source estimation and parameterisation, in addition to causal-anticausal based separation techniques, make assumptions about the time-domain shape of the signal. While these methods are useful, their sensitivity to the signal shape means that they are generally not robust to the position of the analysis frame or to phase distortion, which are common in otherwise high-quality recordings (Doval and d'Alessandro, 2006). Methods to correct phase distortion are generally unfeasible or prohibitively difficult without access to the original equipment or signals of known time-domain shape which have been passed through them. For these reasons, frequency-domain-based methods are an attractive option for robust voice-source estimation and parameterisation.

Of the frequency-domain glottal source estimation techniques, those utilising pre-emphasis filters to model the voice source can often fail in certain speech scenarios, due to the difficulties in differentiating between source and filter contributions (Alku, 1992). Those which adopt a codebook approach to the removal of the glottal contribution can avoid situations where these ambiguities may take place. Indeed, a joint power domain approaches, like (Fröhlich *et al.*, 2001) and (Arroabarren and Carlosena, 2003), is highly desirable as both the source and filter are both optimal in some sense. However, these methods, like many voice-source estimation methods, do not determine an optimum filter order. Additionally, they do not attempt to differentiate the noise portion of the waveform from the periodic part, which may introduce spurious spectral samples into the deconvolution operation. Finally, they also do not determine the scale of the glottal model in a phase-distortion-robust manner.

In light of these difficulties, Chapter 6 proposes a novel power-spectrum-based voice-source parameterisation technique called the PowRd method. Because it operates upon the power spectrum, the method can avoid unreliable phase information, to which time-domain and phase-based methods are sensitive. The method avoids high frequency information which has been corrupted by noise. Finally, the PowRd method utilises a novel error criterion, the Relative Itakura Saito error, which is suitable for determining the filter order and coefficients of the vocal tract and the scale parameter of the included voice-source model.

Chapter 4

Glottal Closing Instant Estimation

Glottal closing instant estimation is the determination of the time location of the instants of glottal closure, either from the acoustic speech signal or from other measurements of glottal activity, *e.g.* EGG signals. The relative timing of these instants, sometimes referred to as the glottal epochs, are important for the perception of pitch (de Cheveigné and Kawahara, 2002), linguistic cues (Klatt and Klatt, 1990) and voice pathologies (Silva *et al.*, 2009). Accordingly, the ability to locate these pulses in time is an important operation for many speech processing tasks, including voice-source estimation (Wong *et al.*, 1979; Bozkurt *et al.*, 2004), prosodic modifications (Charpentier and Moulines, 1989) and speech synthesis (Stylianou, 2001).

This chapter reviews methods of estimating the instant of glottal closure, firstly from EGG signals which is sometimes recorded synchronously with the speech signal, and secondly from the speech signal itself. Many techniques reviewed in this section and discussed in the following chapter are not explicitly robust to phase disturbances often experienced by speech signals. This deficiency lays the foundation for the FRESS algorithm proposed in Chapter 7, which is designed for generally recorded speech signals.

4.1 GCI Estimation from EGG Signals

Electroglottography, introduced in (Fabre, 1957), is a noninvasive technique used to measure glottal activity with a device called an electroglottograph[1]. This device measures the electrical impedance across the larynx using two electrodes, carefully placed on the outer surface of the neck. Figure 4.1 shows a voiced speech signal $s[n]$ and its corresponding EGG and Derivative ElectroGlottoGraph (DEGG) signals (denoted $l[n]$ and $l'[n]$ throughout this chapter). As the glottis opens and closes during phonation, the impedance measured by the EGG signal varies from large when the glottis is open to small when the glottis is closed. Sudden decreases in the measured impedance are thought to indicate vocal fold contact, while sudden increases indicate glottal opening - these instants appear as negative and positive discontinuities in the DEGG signal.

The simple sinusoidal variation of the EGG signal and impulse-like behaviour of the DEGG signal facilitates measuring glottal activity; the EGG is therefore a useful secondary signal when recorded synchronously with the acoustic speech waveform (Veeneman and BeMent, 1985; Krishnamurthy and Childers, 1986). The EGG signal has been used as a benchmark for voiced/unvoiced classification, pitch estimation (Bagshaw *et al.*, 1993; Camacho, 2007), and the estimation of glottal closing and opening instants (GCIs/GOIs)(Drugman and Dutoit, 2009; Thomas *et al.*, 2011). Note

[1] Also known as a laryngograph.

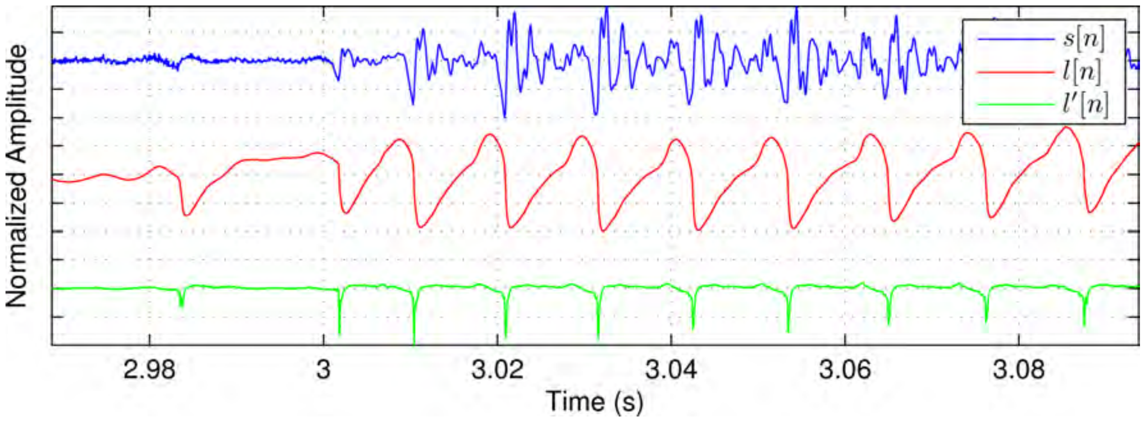


Figure 4.1: The above figure compares three signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded EGG signal $l[n]$ (middle) and DEGG signal $l'[n]$ (bottom).

that before the absolute time references can be extracted from the EGG signal and related to speech events, it is necessary to account for the acoustic wave propagation delay between the larynx and the microphone (Veeneman and BeMent, 1985).

Several different approaches are used to determine the GCI/GOIs from the EGG signal. Glottal closure can be assumed to be closed when the impedance indicated by the EGG signal exceeds a given threshold (Veeneman and BeMent, 1985). Establishing an absolute threshold to determine the glottal closed state can be avoided by observing the changing impedance of the EGG signal: sudden increases or decreases in impedance (corresponding to vocal fold contact) are easily identifiable on the DEGG signal and indicate glottal opening and closing respectively. These peaks can then be located by thresholding approaches (Acero, 1998), choosing extrema from parsed DEGG segments (Krishnamurthy and Childers, 1986), or by using a group delay function (Thomas and Naylor, 2009).

4.2 GCI Estimation from Speech Signals

For many speech recordings, an EGG signal is unavailable and the GCIs must be estimated from the speech signal alone. However, unlike the EGG signal, voiced speech is a more complex waveform and exhibits an elaborate time-domain structure which complicates GCI estimation. These elaborations are due to phenomena such as the resonances of the vocal tract, the presence of turbulent noise, *etc.* In order to facilitate GCI estimation, GCI estimation algorithms which operate on the speech signal alone utilise processes and transformations which generate simpler signals where the glottal activity information is less ambiguous.

One simplifying process adopted by many GCI estimation techniques is the removal of the effects of the vocal tract and glottal source (Childers and Lee, 1991; Smits and Yegnanarayana, 1995; Naylor *et al.*, 2007; Drugman and Dutoit, 2009). This operation yields the deconvolutive residual signal from which the GCI is a more easily identifiable landmarks. Removing the vocal tract alone (Degottex, 2010; Thomas *et al.*, 2011) reveals the voice-source signal from which GCIs are also more easily identifiable than the speech signal itself. Additional less complex signals which facilitate GCI estimation include the energy contour, the multiscale product, the fundamental sinusoidal signal, *etc.* Often, GCI estimation algorithms use these signals in combination to accurately determine the instant of glottal closure.

Following the determination of possible glottal closing instants, many estimation methods refine these candidates by employing dynamic programming algorithms to determine the most likely sequence of GCIs. These algorithms attempt to impose certain heuristic rules based upon characteristics of voiced speech signals which penalise unlikely GCI candidates or sequences of candidates. This operation is useful to discard candidates which in isolation appear as GCIs but fail to sufficiently qualify from a more general perspective.

The following is a review of the various signal transformations, speech signal characteristics and DSP methods utilised by various state-of-the-art methods to determine the GCI from the speech signal alone.

Deconvolutive Residual Signal Neglecting a noise term, a simple sinusoidal model interpretation of the source-filter theory of the voiced speech signal can be expressed in the frequency domain:

$$S(\omega_k) = e^{j(\omega_k + k\phi)} G'(\omega_k) V(\omega_k), \quad k = 1 \cdots N \tag{4.1}$$

where ω_k represents the N harmonically related angular frequencies and $e^{j(\omega_k + k\phi)}$ represents a harmonic comb which samples $G'(\omega)$ and $V(\omega)$ at frequencies ω_k and a linear phase offset.

If $S(\omega_k)$ is deconvolved into transfer functions $\hat{G}'(\omega)$ and $\hat{V}(\omega)$, a deconvolutive residual signal $R(\omega_k)$ can be formed via spectral division.

$$R(\omega_k) = \frac{S(\omega_k)}{G'(\omega_k)V(\omega_k)} \tag{4.2}$$

$$\Rightarrow R(\omega_k) = \frac{e^{j(\omega_k + k\phi)} G'(\omega_k) V(\omega_k)}{\hat{G}'(\omega_k)\hat{V}(\omega_k)} \tag{4.3}$$

Provided that the models $\hat{G}'(\omega)$ and $\hat{V}(\omega)$ approximate the actual $G'(\omega)$ and $V(\omega)$, these components will cancel producing a signal which, after transformation into the time domain, produces a periodic, bandlimited, impulse train $r[n]$:

$$R(\omega_k) = e^{j(\omega_k + k\phi)} \tag{4.4}$$

$$\Rightarrow r[n] = \sum_{k=1}^{N} \cos(\omega_k n + k\phi) \tag{4.5}$$

The discontinuities of this function are then indicative of the moment of excitation of the voice source, usually taken to be the instant of glottal closure (Childers and Lee, 1991; Smits and Yegnanarayana, 1995; Naylor *et al.*, 2007). Figure 4.2 shows the relationship between the deconvolutive residual signal and the speech waveform for a synthetic speech segment.

Many methods of glottal closing instant detection rely upon this deconvolutive residual signals in order to determine the GCI. The most prevalent example of the deconvolutive residual in the literature is the LPC residual signal, where $\hat{G}(\omega)$ and $\hat{V}(\omega)$ are simultaneously estimated using a linear predictive analysis. In (Degottex *et al.*, 2011), the deconvolutive residual signal is approximated by the transformed LF model for $\hat{G}(\omega)$ while $\hat{V}(\omega)$ is modeled by a minimum phase cepstral envelope. An example of voiced speech signal and its corresponding DEGG and LPC deconvolutive residual signal is given in Figure 4.3.

Though similar to GCI determination from the DEGG signal, extracting these discontinuities from the deconvolutive residual signal is complicated due to signal noise and imperfect separations. Methods have been devised which rely upon peak picking and thresholding (Childers and Lee, 1991), peak picking within a specifically parsed regions (Drugman and Dutoit, 2009), and group delay functions (Smits and Yegnanarayana, 1995; Naylor *et al.*, 2007).

Note that generally for these methods, the polarity of the signal must be known. If

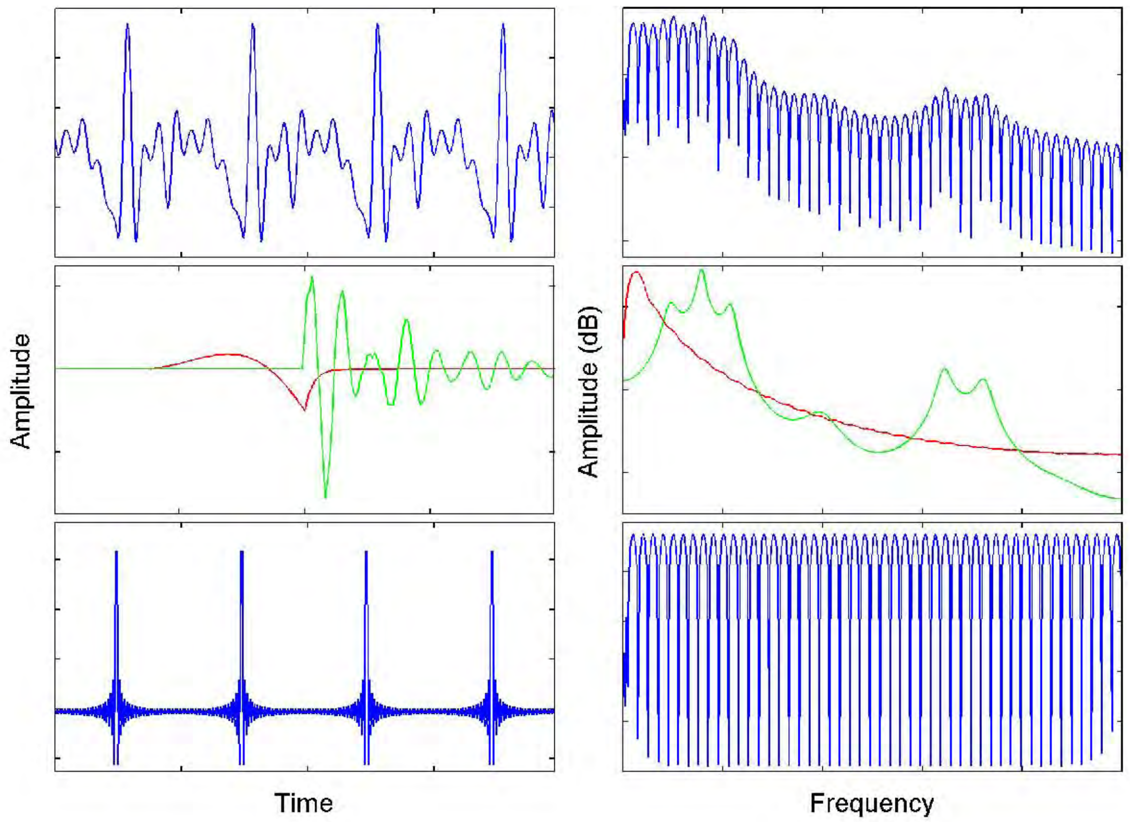


Figure 4.2: The above figures illustrate the construction of the deconvolutive residual from a synthetic speech signal, in both the time domain (left) and magnitude frequency domain (right). Top, the original speech signal. Middle, the deconvolved vocal tract (green) and voice source (red). Bottom, the deconvolutive residual signal.

the polarity is inverted, the local maxima of the deconvolutive residual signal appear instead as local minima. Several algorithms methods have been proposed for this purpose (Ding and Campbell, 1998; Saratxaga *et al.*, 2009; Drugman and Dutoit, 2011).

Energy Contour As the impulse response of the vocal-tract filter decays with time, high energy events are imparted in the speech signal following the points of excitation.

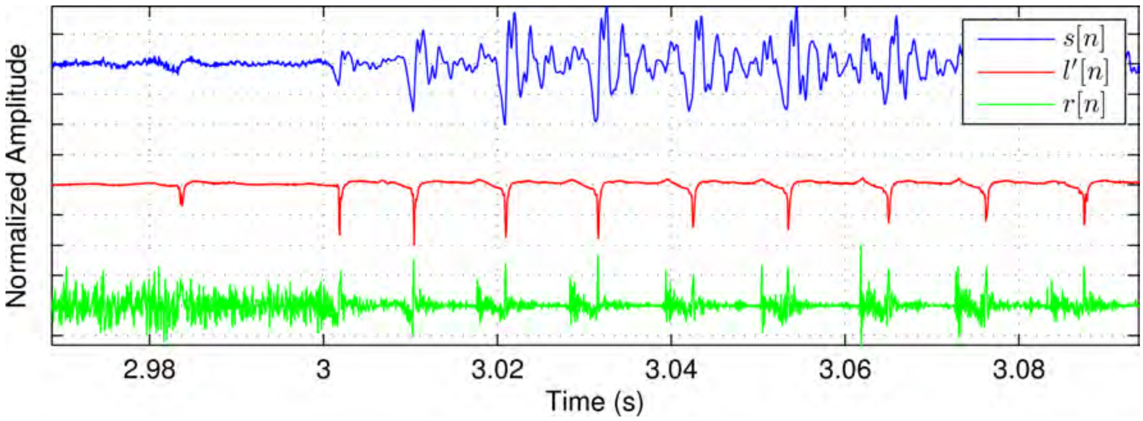


Figure 4.3: The above figure compares three signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded DEGG signal $l'[n]$ (middle) and LPC residual signal $r[n]$ (bottom).

Accordingly, the peaks of the energy function have been proposed as indicators of glottal closure[2] (Ma *et al.*, 1994). Pre-emphasising the speech signal may heighten these energy bursts, as it serves to partially remove the glottal contribution. The energy $e[n]$ of the speech signal $s[n]$ can be calculated:

$$e[n] = \sum_{m=-N}^{N} (w[m]s[n+m])^2 \tag{4.6}$$

where $w[n]$ is an appropriate window and N defines the window length. A variant of this signal is the Frobenius norm $F[n]$, which is can be calculated according to:

$$F[n] = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{p+1} s[n+i-j+p+1]^2} \tag{4.7}$$

[2] As the instant of excitation usually takes place some time before the peak of these energy functions, it is common to delay them by a small interval ($< 1ms$) so that their peaks may align more closely with the instants of excitation (Ma *et al.*, 1994).

where p and m define the size of the matrix of which the Frobenius norm is taken[3], recommended in (Ma *et al.*, 1994) to be $0.001f_s$ and $0.003f_s$, respectively, where f_s is the sampling frequency. An example of a voiced speech signal, its corresponding DEGG signal and Frobenius norm energy contour is given in Figure 4.4.

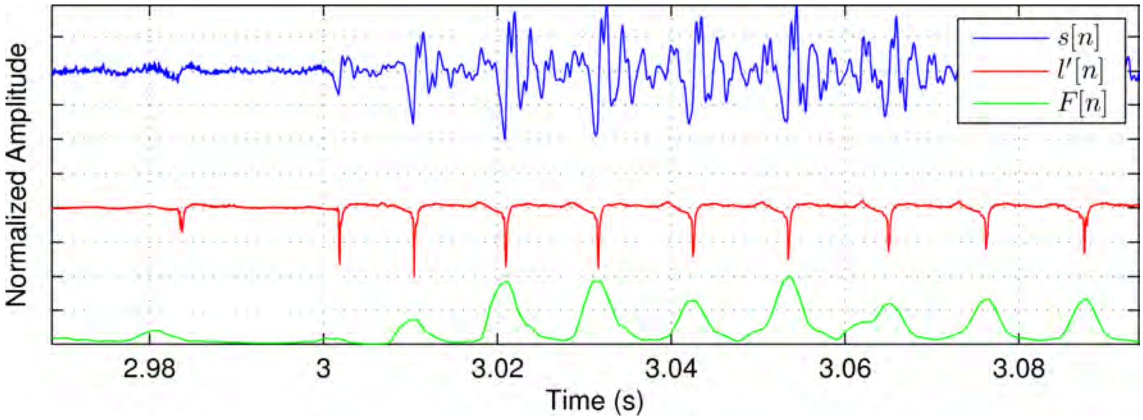


Figure 4.4: The above figure compares three signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded DEGG signal $l'[n]$ (middle) and Frobenius norm energy contour $F[n]$ (bottom).

Pitch Contour The pitch of a voiced speech pulse is often given by the distance between adjacent instants of glottal closure (Bagshaw *et al.*, 1993). Thus, an accurate estimation of the pitch contour of a speech signal gives the relative positions of the GCIs, which can then be utilised to estimate their absolute time position. The find_pmarks algorithm (Goncharoff and Gries, 1998) is based upon this concept. The algorithm first obtains an estimate of the pitch contour from the maxima of the energy

[3] A more efficient calculation of $F[n]$ is given by the square root of the convolution of $s[n]^2$ with an isosceles trapezoidal window $m + p + 1$ samples in length.

contour function. The relative positions defined by the pitch contour is then located upon the speech signal using a dynamic programming algorithm which maximises the amplitude of their samples locations.

Group Delay Function Given a frequency-domain signal $X_n(\omega)$ where

$$X_n(\omega) = \sum_{m=0}^{M-1} x_w[m]e^{-j\omega m} \tag{4.8}$$

and

$$x_w[m] = w[m]x[n+m] \tag{4.9}$$

where $w[m]$ is an appropriate window function, the group delay τ_n is the negative rate of change of phase with respect to frequency. It is calculated according to the following equation:

$$\tau_n(\omega) = -\frac{\Delta\angle X_n(\omega)}{\Delta\omega} \tag{4.10}$$

Note that the phase of signal $X_n(\omega)$ must be unwrapped.

If $x_w[m]$ contains a single impulse at $m = n_0$, it can be shown that $\tau_n(\omega) = n_0\ \forall\omega$. Because the group delay is independent of the magnitude of the signal, locating impulses using this signal avoids the necessity of establishing thresholds. By performing a sample-by-sample analysis upon a clean impulse train with an appropriately sized window, a time varying group delay function $\hat{\tau}[n]$ can be determined by setting

$$\hat{\tau}[n] = \tau_n(\omega) \tag{4.11}$$

$\hat{\tau}[n]$ then indicates the locations of the impulses as negative-going zero crossings[4].

[4] Some researchers utilise the "phase slope" function for this purpose which is distinguished from the group delay only by sign; impulses are indicated by *positive*-going zero crossings of this function.

(Smits and Yegnanarayana, 1995) proposed the use of a group delay function in order to determine the GCIs of a voiced speech signal from its LPC residual signal. However, when applied to noisy imperfect impulsive signals such as the LPC residual, $\tau_n(\omega)$ will generally not equal a constant for all ω. Thus, an averaging procedure is necessary to determine the value of the group delay function $\hat{\tau}[n]$. (Smits and Yegnanarayana, 1995) proposed fitting a linear function to $\tau_n(\omega)$, followed by a smoothing zero-phase Finite Impulse Response (FIR) filter upon $\hat{\tau}[n]$ itself before extraction of the GCIs.

Figure 4.5 illustrates the behaviour of the group delay function $\hat{\tau}[n]$ upon the LPC residual signal, compared with the voiced speech and DEGG signal.

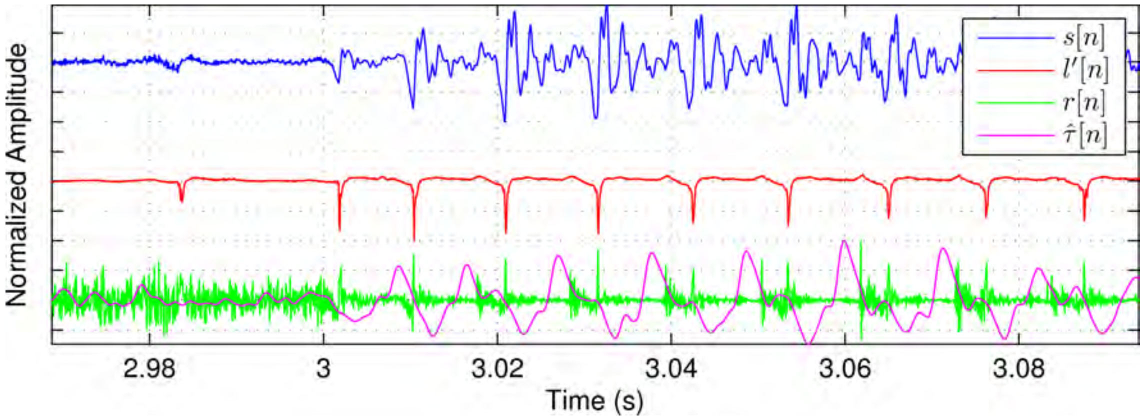


Figure 4.5: The above figure compares three signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded DEGG signal $l'[n]$ (middle) and LPC residual signal $r[n]$ and group delay function $\hat{\tau}[n]$ (bottom).

DYPSA While the approach of (Smits and Yegnanarayana, 1995) accurately identifies many epochs, it also exhibits many spurious zero-crossings for noisy signals and is computationally costly (Brookes *et al.*, 2006). Instead, (Brookes *et al.*, 2006) proposed the use of the energy-weighted group delay signal $\hat{\tau}_{EW}[n]$ to determine glottal closing instant candidates. This signal can be calculated as:

$$\hat{\tau}_{EW}[n] = \frac{\sum_{\omega} |X(\omega)|^2 \tau_n(\omega)}{\sum_{\omega} |X(\omega)|^2} \tag{4.12}$$

An alternative, more efficient formulation is given by:

$$\hat{\tau}_{EW}[n] = \frac{\sum_{m=0}^{M-1} m x_w[m]^2}{\sum_{m=0}^{M-1} x_w[m]^2} \tag{4.13}$$

where it is seen that the average energy-weighted group delay can be interpreted as the centroid of the analysis signal frame. Figure 4.6 gives this signal taken from the LPC residual, and the corresponding DEGG and voiced speech signals.

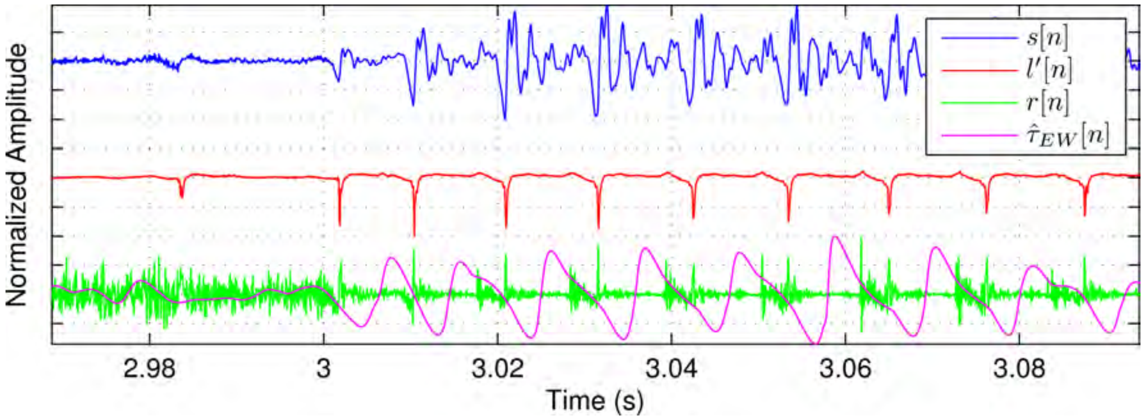


Figure 4.6: The above figure compares three signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded DEGG signal $l'[n]$ (middle) and LPC residual signal $r[n]$ and energy weighted group delay function $\hat{\tau}_{EW}[n]$ (bottom).

The DYnamic programming projected Phase Slope Algorithm (DYPSA) (Naylor *et al.*, 2007) relies upon this function to determine glottal closing instant candidates from the LPC residual signal. In order to capture those GCIs which may have been missed by $\hat{\tau}_{EW}[n]$, a phase projection technique is used where the midpoints between local maxima and minima of $\hat{\tau}_{EW}[n]$ which did not result in a zero-crossing are included. This results in a large set of GCI candidates, including many false positives.

In order to remove false positives from the final sequence of GCIs, an N-best dynamic programming algorithm (Chow and Schwartz, 1989) is applied to the candidates. This algorithm chooses the sequence of glottal closing instants which minimises a cost function based upon voiced speech heuristics. This cost function consists of five elements, the first three of which are fixed according to the characteristics of the glottal closing instant candidate c_r, while the last two correspond to the possible transitions of the glottal closing instant sequence. Additionally, these values are optimally weighted by training them upon a small speech data set (Naylor *et al.*, 2007). The costs are:

Projected Candidate Cost (C_J) Candidates which do not result from the zero-crossing of the group delay function (*i.e.* those obtained via phase slope projection) are attributed a cost of $C_J = 0.5$, while those that do are given no penalty, $C_J = 0$. The weight given to this cost is 0.4.

Ideal Phase Slope Deviation Cost (C_S) The group delay function of an ideal impulse train will cross the time axis with a slope of -1. The DYPSA algorithm penalties candidates which deviate from this ideal. The penalty is calculated according to the equation:

$$C_S(r) = 0.5 + m_\tau(c_r) \tag{4.14}$$

where $m_\tau(c_r)$ is the estimated slope of the $\hat{\tau}_{EW}[n]$ function at sample c_r calculated as:

$$m_\tau(c_r) = \frac{1}{\upsilon}(\hat{\tau}_{EW}[c_r + \frac{\upsilon}{2}] - \hat{\tau}_{EW}[c_r - \frac{\upsilon}{2}]) \tag{4.15}$$

where υ is an even length in samples. Candidates projected from the group delay function are given the ideal unit slope. The weight given to this cost is 0.1.

Normalised Energy Cost (C_f) As mentioned above, GCIs are thought to be the most significant excitation event during the pulse cycle, and thus correspond to high energy in the speech signal. Thus, it follows to penalise glottal closing instant candidates which do not correspond to high energy in the speech signal. Such a penalty is given by the equation:

$$C_f(r) = 0.5 - \frac{F[c_r]}{\breve{F}[c_r]} \tag{4.16}$$

where $F[n]$ is given by Equation 4.7 and the signal $\breve{F}[n]$ is a local maximum energy function, given by the following expression:

$$\breve{F}[n] = \max_k F[n-k], \quad 0 < k \leq L \tag{4.17}$$

L is chosen to be large enough to capture at least one excitation; DYPSA is implemented such that L is set to $\frac{60}{f_s}$, where f_s is the sampling frequency.

This cost will be small for those glottal closing instant candidates which exhibit $F[n]$ values which are similar to the local maximum $\breve{F}[n]$. The cost is large for those candidates which exhibit smaller $F[n]$ values compared with the local maxima and are likely attributable to other events such as glottal noise or glottal opening. The weight given to this cost is 0.3.

Waveform Similarity Cost (C_ρ) Speech signals are hypothesised to be relatively slowly changing, and therefore it follows that during speech, adjacent speech pulses should exhibit a high degree of similarity. This cost is given by the equation:

$$C_\rho(r, r-1) = -0.5\frac{\gamma_{r,r-1}}{\sqrt{\gamma_{r,r}\gamma_{r-1,r-1}}} \tag{4.18}$$

where $\gamma_{r,r-1}$ is a covariance function which compares the two windows extracted from speech signal x centred about the GCI candidates c_r and c_{r-1}:

$$\gamma_{r,r-1} = \sum_{n=-K}^{K} x[c_r + n]x[c_{r-1} + n] \tag{4.19}$$

The size of the windows over which to calculate these values $2K + 1$ is chosen to be 10ms. As $\frac{\gamma_{r,r-1}}{\sqrt{\gamma_{r,r}\gamma_{r-1,r-1}}}$ is limited between 1 and -1, the cost $C_\rho(r, r-1)$ will be smallest for those glottal closing instant candidates which are identical, with increasingly larger penalties for speech segment which are less similar. The weight given to this cost is 0.8.

Pitch Deviation Cost (C_P) As previously mentioned, the pitch of a voiced utterance is often defined as the distance between successive GCIs. While a certain amount of pitch deviation occurs during voicing, it is reasonable to assume that for normal voices this is a slowly changing parameter. Thus, it is appropriate to choose glottal closing instants which conform to a smoothly changing pitch contour. This is accomplished by the addition of the following penalty cost:

$$C_P(r, r-1, r-2) = 0.5 - e^{-(\psi(\Delta_P - 1))^2} \tag{4.20}$$

where

$$\Delta_P = \frac{\min(c_r - c_{r-1}, c_{r-1} - c_{r-2})}{\max(c_r - c_{r-1}, c_{r-1} - c_{r-2})} \tag{4.21}$$

The cost of pitch deviation then increases nonlinearly depending upon the value of ψ, which is 3.3, giving no penalty until pitch deviations surpass 25%. The weight given to this cost is 0.5.

ZFR A different approach to GCI estimation was taken in (Murty and Yegnanarayana, 2008). Excitation discontinuities like those which occur at glottal closure exhibit energy over all frequency bands, and thus (Murty and Yegnanarayana, 2008) proposed the output of a 0Hz resonator for the determination of the glottal epochs, the so-called Zero Frequency Resonator (ZFR) method. Deviations from this zero frequency gives an indication of the locations of the excitation events (Murty and Yegnanarayana, 2008). The speech signal $s[n]$ is passed through a filter with transfer function $H_0(z)$ chosen to resonate the zero frequency. The transfer function $H_0(z)$ is given by:

$$H_0(z) = \frac{1 - z^{-1}}{(1 - \alpha z^{-1})^4} \tag{4.22}$$

where α is very close to 1, *e.g.* $\alpha = 0.999$. The resonance frequency is much lower than the vocal tract resonances, thus the resulting signal is least effected by the them.

Such low pass filtering of the speech signal creates a signal $y[n]$ which may increases or decreases exponentially. In order to see the trend of the signal, the mean is removed using a sliding window:

$$\hat{y}[n] = y[n] - \frac{1}{2M+1} \sum_{k=-M}^{M} y[n+k] \tag{4.23}$$

where the window length $2M + 1$ is chosen be approximately 10ms. This operation may required repetitions in order to fully remove the DC component of the signal.

The total magnitude spectrum effect of the these filtering operations is given in

Figure 4.7. This filtering operation essentially determines a sinusoidal signal $\hat{y}[n]$ which oscillates with the fundamental frequency of voicing. Therefore, landmarks extracted from it can indicate the glottal pulse (though not necessarily the GCI). (Murty and Yegnanarayana, 2008) suggests that the positive zero crossings of $\hat{y}[n]$ can be taken as the glottal closing instants.

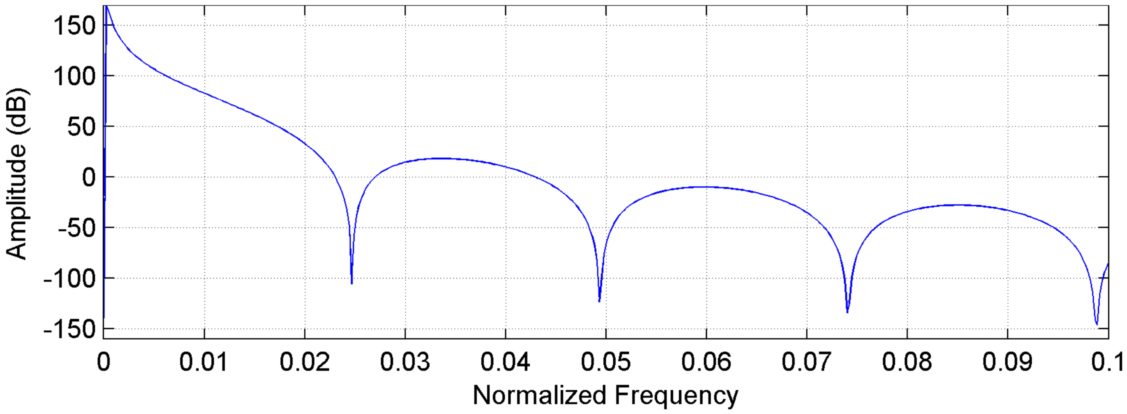


Figure 4.7: The above figure gives the low frequency detail of the magnitude spectrum of the filtering operations of the ZFR method for GCI estimation.

SEDREAMS The Speech Event Detection using the Residual Excitation And a Mean-based Signal (SEDREAMS) algorithm (Drugman and Dutoit, 2009) for GCI detection also uses a fundamental sinusoid signal (referred to as a “mean-based signal” in (Drugman and Dutoit, 2009)) in order to determine the regions within the LPC residual signal where the glottal closing instant is likely to occur. Once this region determined, the largest sample in the LPC residual during this interval is chosen as the glottal closing instant.

The mean-based signal used to determine the “fuzzy” regions of glottal closure is calculated from the speech signal $s[n]$ according to the following equation:

$$y[n] = \frac{1}{2N+1} \sum_{m=-N}^{N} w[m]s[n+m] \tag{4.24}$$

where $w[m]$ is typically a Blackman window function $2N+1$ samples in length. The parameter N controlling the length of the window must be chosen careful as for each local minima of $y[n]$, a glottal closing instant is chosen from the LPC residual signal. Thus, too short a window will lead to many false positives, while too long a window will increase the chance of missing a glottal cycle. N is chosen so as to minimise these opposing factors; (Drugman and Dutoit, 2009) chooses $N = \frac{7}{8}\overline{T}_0$ as a compromise, where $\overline{T}_0$ is the mean pitch period of the analysis utterance. Figure 4.8 shows the mean-based signal, in addition to the corresponding DEGG, voiced speech and LPC residual signals.

The relationship between the minima of the mean-based signal and the locations of glottal closing instant is found to be relatively static (Drugman and Dutoit, 2009). Given y^{c}_{min} and y^{c+1}_{min} the c^{th} and $c+1^{th}$ local minima of the mean-based signal $y[n]$, the following expression is used to locate c^{th} glottal closing instant n_c:

$$n_c = \arg\max_{n} r[n] \quad y^{c}_{min} \leq n \leq 0.65y^{c}_{min} + 0.35y^{c+1}_{min} \tag{4.25}$$

In this expression, $r[n]$ represents the LPC residual signal. These boundaries were determined by a survey of relative distance of mean-based signals from real speech data and GCIs as determined by synchronously recorded EGG signals (Drugman and Dutoit, 2009).

In a new version of the algorithm (Drugman, Retrieved October 8th, 2011), the search region is not located using this fixed relative distance from the minima. Instead, the median ratio η of the time interval of residual signal's prominent peaks[5] to the nearest minima of $y[n]$ relative to the distance between the minima and the following maxima is used to define a fixed interval. Thus, if y^c_{min} and y^c_{max} are the c^{th} local minimum and following maximum of the signal $y[n]$, the c^{th} glottal closing instant estimate n_c is located according to:

$$n_c = \arg\max_n r[n] \quad y^c_{min} - 0.25\eta\Delta y^c \leq n \leq y^c_{min} + 0.35\eta\Delta y^c \tag{4.26}$$

where $\Delta y^c = y^c_{max} - y^c_{min}$.

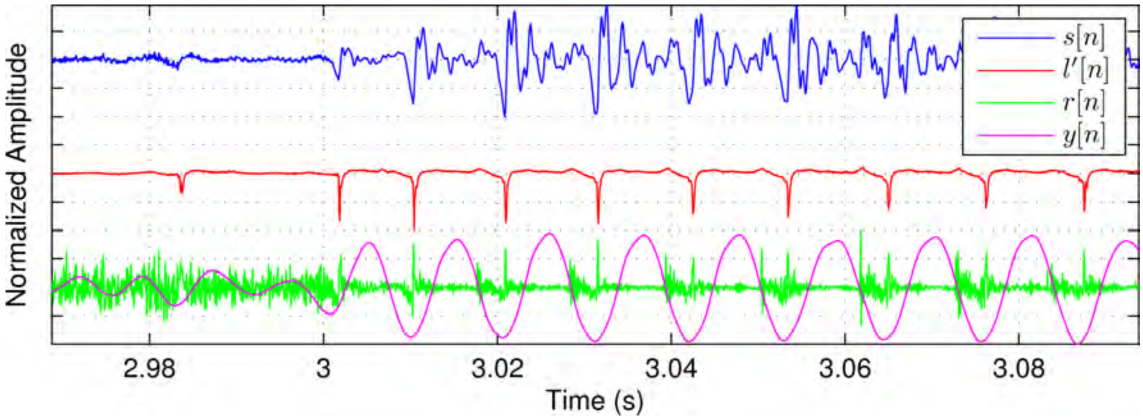


Figure 4.8: The above figure compares three signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded DEGG signal $l'[n]$ (middle) with LPC residual $r[n]$ and mean-based signal $y[n]$ (bottom).

YAGA The Yet Another GCI/GOI Algorithm (YAGA) (Thomas *et al.*, 2011) was proposed as a technique to estimate both GCIs and GOIs from voiced speech, based on a DYPSA-type framework. Like the DYPSA algorithm, the algorithm consists

[5] Prominent peaks are defined to be those samples of the LPC residual signal larger than 40% of the global maximum.

of two stages: first candidate detection, where glottal closing instants are estimated based upon the energy weighted average group delay function and missed candidates are added to the set using a phase slope projection technique, followed by candidate selection, where an N-best dynamic programming algorithm chooses the most likely sequence of glottal closing instants. However, the YAGA method contains significant differences to the DYPSA approach. Firstly, while DYPSA utilises the LPC residual in order to locate the glottal epochs, the YAGA algorithm utilises $p^{-}[n]$, the half-wave rectified, j_1^{th} root of the Stationary Wavelet Transform (SWT) multiscale product of the derivative glottal flow signal. SWT multiscale products have been previously applied to EGG signals (Bouzid and Ellouze, 2008) for similar purposes.

The $p^{-}[n]$ signal is formed as follows:

- Firstly, the derivative glottal flow signal $g'[n]$ is estimated using the pre-emphasis or IAIF technique (see Section 3.3).

- The discontinuities of this signal at glottal opening and closing are reinforced using the multiscale product of the SWT. The YAGA method utilises the biorthogonal spline wavelet with one vanishing moment, and corresponding detail and approximation filters $q[n]$ and $t[n]$.

  The SWT of $g'[n]$ is given by:

$$d_j[n] = \sum_k q_j[k] a_{j-1}[n-k] \tag{4.27}$$

$$a_j[n] = \sum_k t_j[k] a_{j-1}[n-k] \tag{4.28}$$

  The detail and approximation filter coefficients are up-sampled by a factor of 2^{j-1} at the j^{th} level of the SWT and $a_0[n] = g'[n]$.

The multiscale product $p[n]$ is calculated as the point multiplication of the output of the detail signal determined by the SWT:

$$p[n] = \prod_{j=1}^{J} d_j[n] \tag{4.29}$$

where J is the level of the transform, chosen in (Thomas *et al.*, 2011) to be 3.

- While $p[n]$ exhibits desirable impulsive behaviour, performance of the YAGA algorithm is improved by taking the J^{th} root of the half-wave rectified $p[n]$ signal. The resulting signal $p^{-}[n]$ is given by

$$p^{-}[n] = \begin{cases} \sqrt[J]{p[n]} & p[n] \leq 0 \\ 0 & p[n] > 0 \end{cases} \tag{4.30}$$

Figure 4.9 shows the signal $p^{-}[n]$, in addition to the corresponding DEGG, voiced speech and estimated voice-source signals.

Candidates are detected in this signal by determining its energy weighted group delay function, followed by phase slope projection technique in order to supply additional possibly missed candidates. As this approach determines both GCI and GOI candidates, two sequential dynamic programming stages are used to distinguish them: first, to determine the GCIs from all glottal event candidates, and subsequently to determine the GOIs.

Like the DYPSA algorithm, the N-best dynamic programming algorithm is applied in order to determine the likely sequence of GCIs which minimises a set of costs. In addition to the costs utilised by the DYPSA algorithm outlined above, a closed-phase energy cost is added in order to further differentiate GCIs from GOIs.

Closed-Phase Energy Cost, (C_C) Glottal closure cause the energy between the glottal closure instant and its the following opening instant to be low. Thus, an appropriate cost to distinguish glottal closing instants from adjacent glottal opening instants is that the interval following the GCI must contain low energy:

$$C_C(r) = -0.5 + \frac{\sum_{n=c_r}^{c_{r+1}-1} g'[n]^2}{\max_k \sum_{n=c_{r+k}}^{c_{r+k+1}-1} g'[n]^2}, \quad k = 0, 1, \cdots, \breve{R} \tag{4.31}$$

where c_r is the glottal event extracted from the signal $p^-[n]$ and $\breve{R}$ is the number of intervals under analysis.

This cost has the effect of penalising candidates which delimit the beginning of interval of high energy in the estimated glottal source signal, *i.e.* glottal opening instants.

Finally, once the glottal closing instants have been selected from the global set, they are adjusted to coincide with the peaks of the $p^-[n]$ function. Note that this refinement is performed only when the negative peak of the $p^-[n]$ closest to the glottal closing instant estimate falls below a certain predefined threshold. This stage is necessary due to the behaviour of the energy weighted group delay function to non-ideal impulse signals (Thomas *et al.*, 2011).

4.3 Discussion

As phase distortion has a dramatic impact upon the results of voice-source estimation, it is interesting to analyse the effect of this phenomenon upon the common voice-source analysis technique of GCI estimation. Below is a discussion of the different key

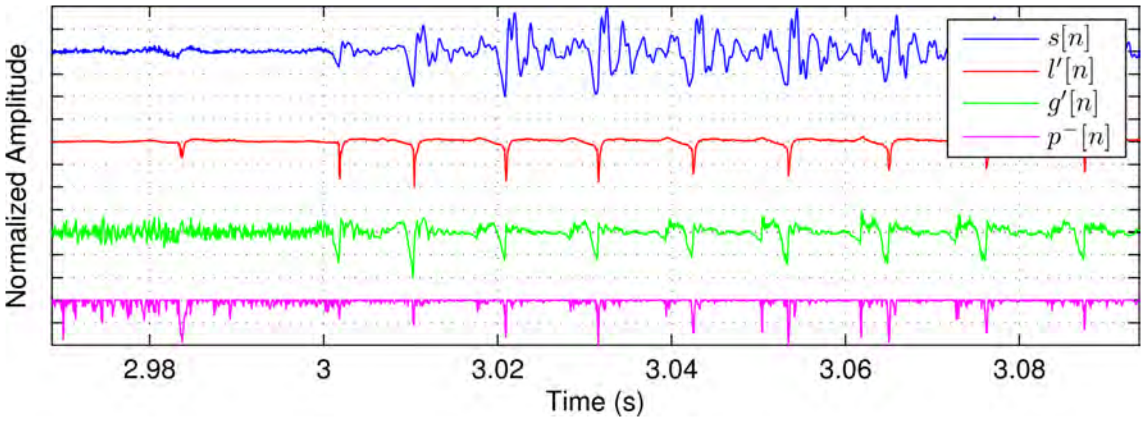


Figure 4.9: The above figure compares four signals: (Top) a speech signal $s[n]$ during an unvoiced-to-voiced transition, the synchronously recorded DEGG signal $l'[n]$ (middle-top), IAIF-estimated voice-source signal $g'[n]$ (middle-bottom) and rectified 3^{rd} root multiscale product of the voice-source estimate $p^-[n]$ (bottom).

signals utilised by glottal closing instant estimation techniques and their robustness to low frequency phase distortion, and general suitability for GCI estimation.

Deconvolutive Residual Signal During voiced speech, deconvolutive residual signals, of which the LPC residual is an example, approximate an impulse train function, with impulses at the instants of glottal closure, which are then searched for by many different glottal closing instant algorithms (Smits and Yegnanarayana, 1995; Naylor *et al.*, 2007; Drugman and Dutoit, 2009). How closely aligned the peaks of the residual signal are with the actual instants of glottal closure is determined by the suitability of the adopted speech model, in addition to noise present in the signal. Therefore, the usefulness of the residual peaks for glottal closing instant detection is speaker- and situation-dependent. However, for approximate speech models like the all-pole model, it is often the case that this instant is in the region of the glottal closing instant, though because of the noisy nature of the residual signal, it does not appear as a peak in the signal. Simple low pass filtering can reduce the noisiness.

The effect of phase distortion upon LPC residual signals can be understood if one considers an ideal impulse train. Because these signals require that its sinusoidal components align in phase at the beginning of each period, any degree of phase distortion will alter the signal. However, if the phase disturbance confined to a specific spectral region e.g low frequencies, a certain degree of phase alignment will occur and thus the impulsive character of the residual will be retained.

Energy Contour The excitation of the vocal-tract filter can result in large amplitude pulse in the voiced speech signal. Assuming these pulses are the result of glottal closure, peaks in the speech signal energy contour are indicative of glottal closure (see Section 4.2). However, the relationship of these peaks to the instant of glottal closure is dependent upon the speaker. Additionally, the peaks are not reliable by themselves as an indicator of glottal closure and often glottal closure may not correspond to a contour peak, *e.g.* during regions of rapidly changing amplitude.

In terms of phase distortion, the energy contour signal is relatively robust, provided that the window size used to generate it (see Equation 4.6) is of appropriate size.

Fundamental Frequency Sinusoid This signal is obtained by low pass filtering the voice speech signal above the fundamental frequency, and is utilised by the

SEDREAMS and ZFR algorithms for glottal closing instant estimation. However, if phase distortion is present at this frequency (via recording conditions or the variety of filter used to obtain the signal), the relationship of this signal to glottal closing instants will change. Thus, the performance of the SEDREAMS[6] and ZFR algorithms will be affected by this type of distortion, the extent of which is dependent upon the degree of phase distortion.

Glottal Source Signal As discussed in Section 3.2.4, the glottal source signal can significantly change in the presence of phase distortion. This has the implication that algorithms which place an expectation upon the time-domain shape of the estimated voice-source signal, like the YAGA method, may experience a degradation in performance.

Following on from this discussion, it can be postulated that certain GCI estimation technique will be relatively unaffected by phase distortion phenomena. The group delay method of (Smits and Yegnanarayana, 1995) and the DYPSA algorithm, which rely upon the LPC residual signal and group delay derived functions, will be relatively unaffected. Additionally, the find_pmarks algorithm which derived source information from the energy contour signal, will also be relatively immune to phase disturbances.

Conversely, some GCI estimation methods may be severely affected. The original SEDREAMS algorithm is unlikely to operate correctly in the case of low frequency phase distortion because the relationship of the fundamental sinusoid signal minima

[6] The most recent version of the SEDREAMS algorithm, available from (Drugman, Retrieved October 8th, 2011), will exhibit improved performance in the case of phase-distorted speech because it does not assume a fixed relationship between the minima of the fundamental sinusoid and the peaks of the LPC residual signal, rather a flexible one which is determined at analysis time.

to the discontinuities of the LPC residual signal may have changed. Additionally, during the dynamic programming stage, the YAGA algorithm weights candidates according to the energy found between them in the estimated glottal source waveform. If the waveform does not approximate the expected shape, this weighting will penalise potentially appropriate GCI candidates. As these algorithms have produced the best results in comparative experiments (Cabral *et al.*, 2011; Drugman, 2011; Thomas *et al.*, 2011), it is useful to develop extensions and modifications which make them more robust for this kind of error.

4.4 Conclusions

This chapter has reviewed state-of-the-art GCI estimation techniques. The methods, along with the various operations, transforms and DSP techniques involved, were given in detail. Additionally, some of these techniques were observed to be sensitive to the common phase distortion that can be imparted by electro-acoustic equipment.

Interestingly, the YAGA and SEDREAMS methods which are judged to be amongst the best performing of GCI estimation methods in comparative experiments (Cabral *et al.*, 2011; Drugman, 2011; Thomas *et al.*, 2011) are also those not robust to phase distortions. While the most recent version of the SEDREAMS algorithm attempts to locate the regions of glottal closure dynamically, the original version assumed that it was a fixed distance from the minima of the mean-based signal. The YAGA algorithm penalties GCI candidates based upon the amount of energy between it and the following glottal event. These and similar assumptions can be invalidated if the speech signal are recorded using non-linear phase equipment.

From these observations, it is concluded that a method of GCI estimation which is explicitly robust to phase disturbances and can perform similarly to these algorithms would be useful to the speech research community. In Chapter 7, a new GCI estimation technique which exhibits these characteristics is proposed.

Chapter 5

Review Conclusions

The previous chapters have reviewed state-of-the-art voice-source estimation and parameterisation techniques in addition to glottal closing instants estimation methods. In those chapters, certain deficiencies of those methods were highlighted. This chapter recapitulates and summaries those deficiencies, before beginning the investigation section of this study.

5.1 Voice-Source Estimation and Parameterisation

As voice-source estimation is a blind deconvolution problem, assumptions must be made about the glottal source and the vocal-tract filter in order to separate them. Table 5.1 summaries the reviewed methods of voice-source estimation and their assumptions and error criteria. The following is a discussion of these methods and a justification of the chosen approach.

Table 5.1: The table below gives a summary of voice-source estimation techniques, documenting the assumptions of the methods regarding the vocal tract and glottal source, in addition to the error criterion used to determine the optimal parameters. $*$ indicates no error criterion used by the ZZT and CCD methods, as no parametric model is estimated.

Method	Vocal Tract Assumption	Voice-Source Assumption	Error Criterion
CPIF (Wong *et al.*, 1979; Plumpe *et al.*, 1997; McKenna, 2001)	AR	Closed Phase Existence	TD-LS
(Fujisaki and Ljungqvist, 1986, 1987)	AR/ARMA	FL	TD-LS
(Hedelin, 1986)	AR	Hedelin	TD-LS
(Ding *et al.*, 1994, 1997)	ARMA	KLGLOTT88	TD-LS
(Funaki *et al.*, 1997)	ARMA	KLGLOTT88	TD-LS
(Lu, 2002; del Pozo, 2008; Pérez and Bonafonte, 2005)	AR	KLGLOTT88	TD-LS
(Fu and Murphy, 2006)	ARMA	KLGLOTT88/LF	TD-LS
ARX-LF (Vincent *et al.*, 2005)	AR	LF	TD-LS
Pre-Emphasis (Markel and Gray, 1982)	AR	First Order AR	TD-LS
IAIF (Alku, 1992; Airas, 2008a)	AR	Low Order AR	TD-LS / I-S
SIM (Fröhlich *et al.*, 2001)	AR	Power Domain LF	I-S
GSBIF (Arroabarren and Carlosena, 2003)	AR	Power Domain KLGLOTT88	I-S
ZZT (Bozkurt, 2005)	Minimum Phase	Mixed Phase Signal	$*$
CCD (Drugman *et al.*, 2011b)	Minimum Phase	Mixed Phase Signal	$*$
MSP (Degottex *et al.*, 2011)	Minimum Phase	Transformed LF	ϕ LS
MSPD (Degottex *et al.*, 2011)	Minimum Phase	Transformed LF	$\Delta\phi$ LS
MSPD^2 (Degottex *et al.*, 2011)	Minimum Phase	Transformed LF	$\Delta^{-1}\Delta\Delta\phi$ LS

Voice-Source Assumptions As the voice-source signal is generally parameterised following estimation, joint parameterisation approaches are preferable as they can avoid the necessity of a second separate parameterisation stage. Additionally, the source and filter parameterisation are both optimal in the same sense, whether giving the minimal Itakura-Saito distance, least-squares time-domain energy, *etc.* This work will therefore adopt a joint approach for voice-source estimation and parameterisation.

Ideally, the model used to parameterise the glottal signal should be interpretable in ways which are useful. For this reason, specific voice-source models such as the LF or KLGLOTT88 models are preferable to the all-pole filter model of the IAIF method because they give an direct indication of salient glottal features and characteristics, *e.g.* the pulse's open quotient. In this study, the transformed LF model is adopted. This model produces physiologically relevant LF model pulses based on statistical analyses undertaken to identify the covariations and characteristic trends of a large database of LF model parameters fitted onto real voice-source signals (Fant *et al.*, 1994; Fant, 1995). Additionally, it avoids certain unrealistic parameterisations which may occur when using LF model (Fröhlich *et al.*, 2001). Finally, its single parameter R_d was qualified in (Fant, 1995) as "the most effective single measure for describing voice qualities".

Vocal Tract Assumptions The assumption that the vocal tract can be represented by an all-pole filter is prevalent, and as can be seen in Figure 5.1, is utilised by many voice-source estimation methods. For phonemes produced by an unbranched vocal tract, this assumption is well justified, following from the geometry of a concatenated tube, as discussed in Section 2.3.2. The assumption is also convenient for time-domain analysis due to the tractable linear systems that derive from all-pole system analysis.

However, phonemes produced using a more complex geometrical configuration will require spectral zeros in addition to poles (Markel and Gray, 1982). Additionally, the planar wave assumption, necessary for the application of the all-pole vocal tract model, becomes invalidated at frequencies above 4kHz. ARMA vocal tract models are therefore more appropriate for these sounds, as they can represent the zeros in the signal, though it is not clear how to estimate q, the order of the FIR filter polynomial of the ARMA model. Alternatively, the assumption that the vocal tract can be represented by a minimum phase system is also capable of modeling these sounds, provided that both the poles and zeros of the vocal tract lie within the boundaries of the unit circle.

However, when operating only upon power spectrum information to decompose the speech signal (for reasons of an unreliable phase spectrum), it is always possible to recover a minimum phase spectral envelope *e.g.* using cepstral techniques (Cappé *et al.*, 1995; Röbel and Rodet, 2005; Degottex *et al.*, 2011). With ARMA modeling it is difficult to predict the number of zeros that may be necessary, while the order of the all-pole vocal tract model is related to its length (Markel and Gray, 1982). ARMA modeling is also a difficult nonlinear optimisation problem (Makhoul, 1975). Finally, to re-iterate the conclusions of Chapter 2, the acoustic tube model of the vocal tract implies an all-pole spectrum. For these reasons, it is adopted by this study.

Error Functions Voice-source parameterisation methods determine optimal parameters by minimising a certain error criteria. However, not all of these functions are robust to nonlinear phase recording conditions, as is discussed here.

A common function to minimise for glottal inverse filtering and parameterisation is the time-domain least-squares energy. The use of this error function implies a time-domain matching of the signal and thus an expectation of the time-domain shape is imposed upon the signal, as is discussed in Section 3.2. For example, closed-phase inverse filtering assumes null signal flow during the closed phase of the glottal cycle. Because a phase distortion may alter the time-domain shape of the signal ((Akande, 2004) specifically notes that it may eliminate the closed phase), it is not reasonable to estimate the glottal source signal in this fashion. The phase sensitivity of time-domain-based glottal source estimation methods also explains the lack of robustness to the position of the analysis frame (Wong *et al.*, 1979; Alku *et al.*, 2009).

Similarly, voice-source estimation methods based upon the phase criteria are adversely affected. The necessity of ideal phase conditions for these types of analyses is briefly mentioned in (Degottex, 2010), while the sensitivity of the maximum/minimum phase model to the analysis window is referred to in (Drugman *et al.*, 2009a). An alternative interpretation of the time-domain signal shape expectation is an expectation upon both the magnitude and phase spectrum of the signal. If this phase coherence has been disturbed, the phase minimisation is no longer reliable indicator of an appropriate parameterisation.

5.1.1 Robust Voice-Source Estimation

Of the methods of voice-source estimation, the only general approach which jointly estimates and parameterises the voice-source signal without relying on a phase-related error criterion is the power-spectrum-based joint estimation techniques, of which SIM (Fröhlich *et al.*, 2001) and GSBIF (Arroabarren and Carlosena, 2003) are examples. These methods utilise a power spectrum signal representation and the Itakura-Saito distance function for voice-source estimation; the transformation and subsequent separation of signal magnitude and phase information makes the approach robust to phase-related errors. However, a number of issues arise with the methods which make it difficult or inappropriate for the analysis of continuous speech:

Analysis Frame Size As the SIM method was designed for clinical analysis of sustained vowels, throughout (Fröhlich *et al.*, 2001) a frame size of 200ms is suggested. For continuous speech, however, the usual time interval assumed for stationarity is approximately 25ms. Thus, in order to make the method more appropriate for continuous speech, the size of the analysis frames should be of a similar scale of duration. In order to determine the spectral information accurately using the DFT, frame sizes should be approximately 3 to 4 local pitch periods in length (Harris, 1978; Serra, 1989) or alternatively, if an accurate estimate of the fundamental frequency is available, using a least-squares harmonic analysis (Laroche *et al.*, 1993). Both methods of sinusoidal model parameterisation are discussed in Appendix C.

Vocal-Tract Filter Order The GSBIF method uses inappropriately large filter orders, *e.g.* $p = 17$ with a sampling frequency of 12kHz. The SIM method is validated on experiments with synthetic speech, where the vocal-tract filter order was known *a priori*. Though it is not mentioned explicitly, during real speech experiments, it is assumed that the filter order was chosen following the usual rule of thumb discussed in Section 2.3.2. However, during running speech, the length of the vocal tract is unlikely to be constant, due to the extending of the lips or rising of the larynx, thus an alternative solution which permits a dynamic filter order is desirable. According to Equation 2.9, at a sampling rate of 10kHz, a filter order suitable for a vocal tract length 14cm and 20cm is approximately 8 and 12, respectively.

Glottal Model Scale Parameter For certain applications, *e.g.* speech coding, speech synthesis, the speech signal must be fully parameterised, including its scale. However, GSBIF does not parameterise the scale factor of the voice-source model. In addition, the SIM method estimates the scale factor of the glottal model following the determination of the its shape parameters using a *time-domain* error criteria. This implies both a phase and amplitude match of the signal. The SIM method accommodates linear phase offsets by employing an exhaustive search, but is otherwise not robust to more complex phase distortions which may be present.

Speech Signal Noise Noise will be present in every speech signal, particularly at the upper frequencies. This is a well-known problem for voice-source estimation, where the noise components of the speech signal make it more difficult to estimate various parameters, particularly the spectral slope of the glottal signal (Strik, 1998). The situation is illustrated in Figure 5.1. When extracting spectral samples for voice-source estimation, neither the SIM of GSBIF methods make a distinction between harmonic and noise peaks which may lead to inaccurate parameterisations, particularly the underestimation of the spectral tilt.

In (Strik, 1998; Lu, 2002), the effect of this noise upon time-domain glottal model fitting is reduced by a low pass filter - in the spectral domain, this is equivalent to choosing only low frequency spectral samples for source estimation/parameterization. An equivalent perspective is to adopt a harmonic-plus-noise type model, where the speech signal is divided into two frequency bands, the lower consisting of harmonic sinusoids and the upper consisting of noise. Such schemes have been adopted by other voice parameterisation methods, *e.g.* phase minimisation techniques (see Section 3.4.2). However, this solution presents another problem, intertwined with the issue discussed above: the filter order appropriate for the lower band is unknown.

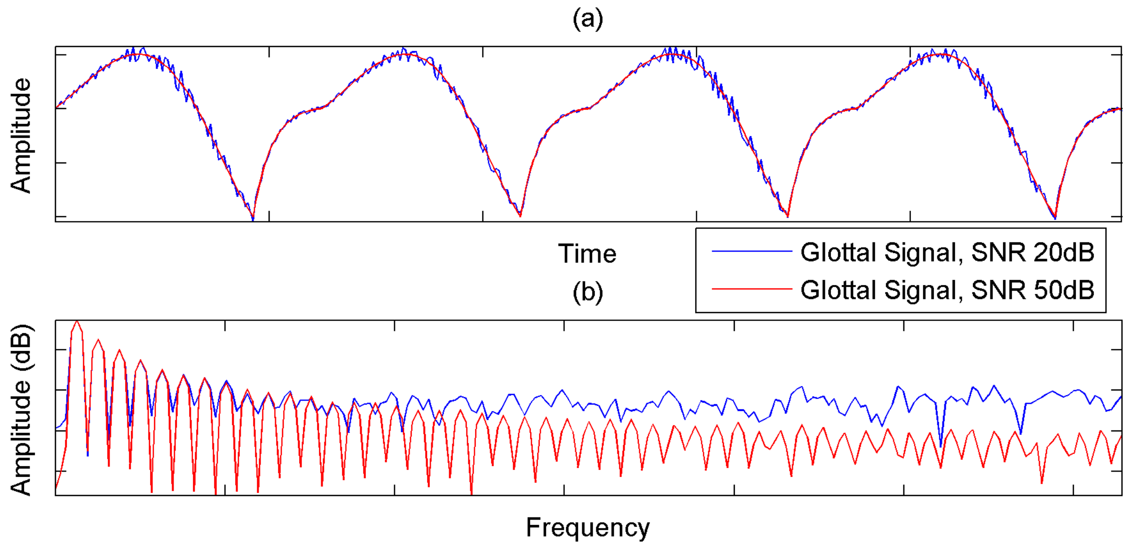


Figure 5.1: The figure above shows two glottal signals with differing noise content in (a) the time-domain glottal signals and (b) the magnitude frequency domain. The intrusion of high frequency noise can prevent accurate parameterisation of the spectral slope.

Following on from this discussion, a new voice-source parameterisation scheme, representing the first contribution of this thesis listed in Section 1.2.1, is proposed in Chapter 6: the PowRd method. The method adopts a power spectrum approach in order to avoid issues of phase distortion and the position of the analysis frame. The vocal tract, represented by an all-pole filter, and glottal flow signal, which is modeled by the transformed LF model, are simultaneously estimated. In these ways, the technique is similar to the SIM method. However, unlike the SIM method, the PowRd method minimises a novel error function, the Relative Itakura-Saito error. This new function is similar to the usual Itakura-Saito error, yet it is scaled by a factor based on a relationship between the analysis frame and filter order which makes it appropriate for determining an optimal filter order value, the details are given in Section 6.2.2. The use of the Relative Itakura-Saito error gives the analysis the flexibility to avoid high frequency noise present in the signal, not require *a priori* knowledge of the filter order, and can be utilised to determine also the scale parameter of the transformed LF model.

5.2 GCI Estimation

Phase distortion disrupts the time-domain signal shape of the speech signal which has the consequence that various derived signals utilised for GCI estimation behave in a manner which is unreliable. Section 4.3 gives a detailed discussion of this phenomenon, which is summarised here.

The characteristics of certain signals are only slightly impacted by nonlinear phase disturbances. In order to exhibit impulsive time-domain behaviour, the components of the deconvolutive residual signal must be in phase. Although phase distortion can by alter this relationship, numerous components may still align and the signal will retain its impulsive character. The energy contour of the signal is similarly robust to phase distortion, provided that the window used to generate the signal is an appropriate length.

Conversely, other signals derived from speech are adversely affected by phase distortion. Specifically, the phase of the fundamental sinusoid signal, obtained by low pass filtering the speech signal and utilised by the ZFR and SEDREAMS methods, can be significantly changed, depending upon the filter used to generate it and the degree of phase distortion. Additionally, as mentioned in the previous section above, the estimated voice-source signal utilised by the YAGA algorithm can also be significantly altered by phase distortion.

According to comparative experiments (Cabral *et al.*, 2011; Drugman, 2011; Thomas *et al.*, 2011), the SEDREAMS and YAGA algorithms are both among the best performing GCI estimation methods. As these methods are affected by phase distortion, a GCI estimation technique which is robust to this phenomenon is potentially very useful.

5.2.1 Phase-Distortion-Robust GCI Estimation

This work proposes a new method for glottal closing instant determination in Chapter 7, the second contribution of this work listed in Section 1.2.1. The new method utilises the LPC residual searching strategy of the SEDREAMS algorithm to determine likely regions of glottal closure, combined with the dynamic programming algorithm of the

DYPSA and YAGA algorithms to determine the most likely sequence of glottal closing instants. However, in order to ensure that the algorithm is explicitly robust to phase distortion, the search regions are aligned with the peaks of the energy contour. In this way, the likely regions of glottal closure have a dynamic relationship with the signal under analysis in a manner that is robust to phase distortion.

5.3 Voiced Speech Analysis/Synthesis

Voice-source estimation and parameterisation methods have been applied to speech synthesis and related areas (Lu, 2002; Vincent, 2007). As a third contribution of this study, this application is demonstrated using a power-spectrum-based voice-source estimation and parameterisation approach similar to the PowRd method in Chapter 8. The analysis/synthesis system has at its core a modified PowRd method which accommodates a larger variety of voice-source shapes and a parameter smoothing procedure which utilises dynamic programming and filtering operations to obtain continuous parameters from voiced speech utterances.

In a preference test of 50 listeners, the synthetic speech produced by this system is compared with a state-of-the-art time-domain procedure parameterisation system, similar to the systems presented in (Lu, 2002), (del Pozo, 2008) and (Pérez and Bonafonte, 2011). The approach proposed by this study is generally preferred in both linear phase and nonlinear phase recording conditions, justifying the power-spectrum-based approach taken by this study and giving encouraging results for future work.

Chapter 6

Power-Spectrum-Based Voice-Source Parameterisation

This chapter proposes a new method to parameterise the glottal source signal which is robust to noise and phase distortion, introduced as Contribution 1 of this thesis. The proposed method, referred to as the Power spectrum determination of the R_d parameter (PowRd) method, estimates the optimal waveshape parameter of the transformed LF model R_d from a speech frame in a manner which is robust to the time-domain location and phase spectrum of the analysis frame. Additionally, it is also robust to noise which can dominate the high frequency regions of the speech signal.

The PowRd method draws from the power-spectrum-based approach adopted by (Fröhlich *et al.*, 2001). The SIM method, reviewed in Section 3.3.2, is robust to both phase disturbances which may have been imparted to the speech signal and also the location of the analysis frame. However, a number of factors make it unsuitable for the analysis of many speech signals. Firstly, the method is not robust to the presence of high frequency noise in the speech signal, as the approach extracts spectral samples from across the entire bandwidth of analysis, including noisy high frequency regions. Secondly, the method assumes that vocal-tract filter order is known and relies upon the usual rule of thumb (given in Section 2.3.2), which may be inappropriate. Finally, the SIM method does not estimate the LF model scale parameter E_e in the power domain, switching instead to the time-domain and a least-squares error criterion to determine it, which as discussed in Section 5.1 is inappropriate for phase-distorted recordings.

Conversely, the PowRd method retains the advantages of the SIM method and offers solutions to the above issues. As signal noise in the speech signal is often confined to the high frequency portions of the signal, the method adopts a two-band Harmonic plus Noise Model (HNM) type speech model. The upper noise portion of the spectrum beyond a maximum voiced frequency is ignored and the lower band alone undergoes analysis. The appropriate filter order for this new lower bandwidth analysis is simultaneously determined with the optimal R_d parameter using a novel error criterion based on the Itakura-Saito error proposed in this chapter named the Relative Itakura-Saito error function. The Relative Itakura-Saito error has the same characteristics as the usual Itakura-Saito error, yet it can determine the optimal filter order. Finally, this chapter also describes a method for the determination of E_e parameter in the power spectrum alone, without resorting to the time domain.

In order to validate the PowRd method, experiments are performed under linear phase and simulated nonlinear phase conditions. Comparative experiments with synthetic speech with 4 other state-of-the-art glottal source parameterisation methods will show that the PowRd method is the most robust voice-source parameterisation algorithm under the tested conditions. Experiments are also performed with real speech, where the open quotients predicted from the R_d parameters are compared with those obtained from simultaneously recorded EGG signals. The PowRd method is shown to perform similarly to other state-of-the-art approaches, yet not requiring strict time placement of the signal frame, phase linear recording conditions nor a pre-determined filter order.

This chapter begins by restating the theory behind the power-spectrum-based voice-source parameterisation, and a detailed description of the PowRd approach. The new error function, the Relative Itakura-Saito error, used to determine the parameters is also described. Experiments on synthetic speech validate the technique, and additional experimental results on real speech signals are then given. Finally, the results are discussed, summarising the main strengths and weaknesses of the various approaches, and the chapter is concluded.

6.1 Theory

The theoretical underpinnings of the PowRd algorithm are the same as other power-spectrum-based joint estimation methods. This approach is briefly recapitulated here.

The relationship between the vocal tract power spectrum $|V(\omega)|^2$ and the power spectrum of the speech signal $|S(\omega)|^2$ and the underlying glottal contributions $|G'(\omega)|^2$ is restated:

$$|V(\omega)|^2 = \frac{|S(\omega)|^2}{|G'(\omega)|^2} \tag{6.1}$$

It is clear from this equation that for any given voice-source contribution, a corresponding vocal tract can be calculated. Thus, power-spectrum-based methods of voice-source estimation and parameterisation require strict assumptions regarding the magnitude behaviour of both the vocal tract and glottal source: the minimum phase envelope assumption is not sufficiently restrictive to determine a solution. For this reason, the all-pole assumption is imposed upon the estimated vocal tract spectrum.

If $|S(\omega)|^2$ can be approximated by a transformed LF model pulse exciting an all-pole vocal-tract filter, there exists a parameter value R_d and a set of all-pole filter coefficients a_k which characterises $|G'(\omega)|^2$ and $|V(\omega)|^2$ respectively. Thus, given an R_d value, the corresponding derivative glottal flow model power spectrum $\left|G'^{R_d}(\omega)\right|^2$ can be calculated and the vocal tract power spectrum $\left|V^{R_d}(\omega)\right|^2$ necessary to produce $|S(\omega)|^2$ from $\left|G'^{R_d}(\omega)\right|^2$ can be determined. This is expressed in the following equation:

$$\left|V^{R_d}(\omega)\right|^2 = \frac{|S(\omega)|^2}{|G'^{R_d}(\omega)|^2} \tag{6.2}$$

If the given R_d parameter characterises the actual derivative glottal flow signal, $\left|V^{R_d}(\omega)\right|^2$ can be seen as the scaled power spectrum of an all-pole filter. This filter can be determined by applying an all-pole envelope estimator to $\left|V^{R_d}(\omega)\right|^2$, the fitting error of which will be small. However, if the R_d parameter is not suitable or the order of the filter is inappropriate, then the error determined by the envelope-fitting operation will be large. Thus, this error quantifies the suitability of both the given glottal shape parameter R_d and the filter order to approximate the speech signal. Minimising this error will serve to optimally parameterise the speech signal in terms of the assumed speech model.

6.2 The PowRd Algorithm

The practical implementation of the PowRd algorithm can be described in four general steps:

- A power spectrum representation $|S(\omega)|^2$ of a voiced speech signal frame is obtained. The pitch period of the signal within the analysis frame is assumed to be known *a priori* and denoted T_0.

- The R_d parameter is discretely sampled across its range and, together with the estimated T_0 value, is used to generate the power spectra of the corresponding voice-source signals, $|G'^{R_d}(\omega)|^2$. These representations are then inversely applied to the speech power spectrum $|S(\omega)|^2$ according to Equation 6.2 in order to yield an estimate of the corresponding vocal tract power spectrum $|V^{R_d}(\omega)|^2$.

- All-pole spectral envelopes are then fit to each $|V^{R_d}(\omega)|^2$, and the fitting error is calculated using the Relative Itakura-Saito distance function.

- The lowest error initial estimates from this approach are then refined using an optimisation procedure, yielding the lowest error R_d parameter for the signal frame.

This section will discuss each of these steps in detail.

6.2.1 Power Spectrum Representations

The periodicity of the time-domain voiced speech signal $s[n]$ has the consequence that its spectrum $S(\omega)$ can only be reliably estimated at discrete quasi-harmonic frequencies. Typical methods to obtain this information include peak selection from the magnitude spectrum or using a least-squares analysis. These methods are briefly summarised here, and more details are given in Appendix C.

Signal periodicity implies strong correlation with the basis sinusoids of the discrete Fourier transform, which then sample the spectrum at quasi-harmonic frequencies. These sinusoidal components of signals appear as prominent peaks in the DFT magnitude spectrum, which can then be selected using a simple search algorithm (McAulay and Quatieri, 1986; Serra, 1989). To ensure that these peaks can be properly resolved following the transformation into the frequency domain, a minimum of 3 to 4 periods are needed in the analysis frame depending on the windowing function used (Harris, 1978; Serra, 1989). Additional spectral resolution may be obtained by interpolation procedures, *e.g.* zero-padding the time-domain signal frame or fitting a parabola to peaks of the log power spectrum and their neighbouring points.

An alternative method for determining spectral samples is possible if the fundamental frequency of the signal is known. By assuming the signal is composed of harmonically-related sinusoids, the complex amplitudes of the harmonics may be determined more accurately and with smaller frame sizes than DFT approaches by using a least-squares approach (Laroche *et al.*, 1993; Stylianou, 1996). This method is more accurate than peak picking from the DFT magnitude spectrum as the spectral smearing introduced by time-domain windowing is not accounted for by this approach. With the least-squares harmonic approach, as the frequencies of the harmonics are known *a priori*, their mutual influence upon each other can be considered in their estimation. Smaller frame sizes can be utilised as the number of unknown variables depends upon the bandwidth of the signal and fundamental frequency of the analysed signal frame.

Two Band Speech Model Researchers have previously utilised a two band model of the speech signal for speech synthesis and modification purposes (Stylianou, 2001; Pantazis *et al.*, 2008). It is assumed that the voiced speech signal is composed of two bands, the lower band containing quasi-harmonic sinusoids, the upper band containing noise. The frequency separating these two bands is referred to as the maximum voice frequency F_m, which can be determined algorithmically (Stylianou, 1996; Erro *et al.*, 2011) or fixed at specific frequency (Drugman *et al.*, 2009b). By adopting this signal model and ignoring spectral information beyond F_m, the PowRd algorithm avoids the noisy signal components which may adversely affect the glottal signal. This spectral perspective of avoiding noisy high frequency information is equivalent to the low pass filtering operation taken by other authors also for the purposes of voice-source parameterisation, *e.g.* (Strik, 1998) and (Lu, 2002).

The new signal only contains frequency components ω_k in the bandwidth $0 \leq \omega_k \leq \omega_m$, where $\omega_m = 2\pi \frac{F_m}{f_s}$. A frequency transformation function is then used to map these boundaries of spectral envelope to the range between 0 and π. A straight forward transformation function can be used for this purpose:

$$\omega'_k = \frac{\pi}{\omega_m}\omega_k \tag{6.3}$$

The operation is similar to the selective linear prediction technique discussed in (Makhoul, 1975), where a portion of the spectrum is isolated by a frequency transformation and fit by the usual full band linear predictive techniques. It allows better approximation of the frequencies closer to F_m. Following this transformation, the assumptions of the PowRd method are more stated as the voice-source signal is approximated by the transformed LF model and the *bandlimited* vocal tract is modeled by an all-pole filter.

$G'^{R_d}(\omega)$ **Determination** In order to determine $V^{R_d}(\omega)$ from the estimated speech power spectrum, the transformed LF model spectrum $G'^{R_d}(\omega)$ is required. As the pulse can be generated without periodic interference and windowing artifacts, its harmonics can be accurately calculated using a phasor correlation approach. However, as the formula of the LF model maybe be re-expressed solely in terms of exponentials, the phase correlations can be reexpressed as scaled and summed geometric summations. Compared with the phasor correlation approach, this method reduces the computational effort by more then 85%. Additionally, a faster though inexact approximation of the spectral parameters can be achieved by interpolating the DFT. This constitutes a minor contribution of this thesis, the details of which are given in Appendix D.

6.2.2 Vocal-Tract Filter Estimation

Once the power spectrum of the vocal-tract filter has been estimated, it is assumed that it represents a sampled all-pole filter envelope. A prevalent approach for the determination of all-pole filter coefficients from discrete spectral samples is linear prediction (Makhoul, 1975). However, it is well known that the all-pole envelopes estimated by linear prediction contains a bias towards the harmonics of the spectrum, which is particularly an issue for female speech where harmonic spacing is generally wider. For this reason, Discrete All-Pole (El-Jaroudi and Makhoul, 1991) is preferred.

The DAP algorithm obtains a more accurate filter measurement by refining the filter estimated by spectral linear prediction by using an iterative algorithm and a different error criterion. The new error criterion is the discretised Itakura-Saito error function (Itakura and Saito, 1968), which has been qualified as a "subjectively meaningful measure of speech distortion" (Gray *et al.*, 1980). Given two power spectra $P(\omega_n)$ and $\hat{P}(\omega_n)$ defined at a set of discrete frequencies ω_n for $n = 1 \cdots N$, the discretised Itakura-Saito error is calculated according to following equation:

$$E_{IS} = \frac{1}{N}\sum_{n=1}^{N} W(\omega_n)\left(\frac{P(\omega_n)}{\hat{P}(\omega_n)} - \ln\frac{P(\omega_n)}{\hat{P}(\omega_n)} - 1\right) \tag{6.4}$$

where $W(\omega_n)$ is a frequency dependent weighting function. The technical details of both DAP and spectral linear prediction are discussed in Appendix B.

By applying the DAP algorithm to an estimated vocal tract power spectrum $\left|V^{R_d}(\omega)\right|^2$, the Itakura-Saito error $E_{IS}^{R_d,p}$ which quantifies the goodness of fit of a p^{th} order all-pole filter a_k to $\left|V^{R_d}(\omega)\right|^2$ can be calculated according to the following equation:

$$E_{IS}^{R_d,p} = \frac{1}{N}\sum_{n=1}^{N} W(\omega_n)\left(\frac{\left|V^{R_d}(\omega_n)\right|^2}{\hat{P}(\omega_n)} - \ln\frac{\left|V^{R_d}(\omega)\right|^2}{\hat{P}(\omega_n)} - 1\right) \tag{6.5}$$

where $\hat{P}(\omega_n)$ is the power spectrum of the DAP-optimised all-pole envelope sampled at frequencies ω_n. As it is assumed that an all-pole envelope will fit the vocal-tract filter, $E_{IS}^{R_d,p}$ gives an indication of the goodness of fit of the transformed LF model parameter R_d used to generate the power spectrum.

Determination of the Optimal Vocal-Tract Filter Order While the Itakura-Saito distance function is useful for quantifying the distance between two spectra, it is not appropriate for choosing the order of the vocal-tract analysis since increasing the order of analysis will generally decrease the distance between the original and modeled spectra. For this reason, a novel error criterion is proposed for determining the filter order, named the Relative Itakura-Saito distance function.

The Relative Itakura-Saito distance function is calculated by the following equation:

$$E_{rIS}^{R_d,p} = \frac{E_{IS}^{R_d,p}}{E_{IS}^{0,p}} \tag{6.6}$$

where $E_{IS}^{0,p}$ is the minimum Itakura-Saito distance between the *speech* power spectrum and the sampled power spectrum of the DAP-estimated best fitting p^{th} order all-pole envelope.

As it is simply a scaled version of the Itakura-Saito distance function, for fixed filter orders the error surface described by Relative Itakura-Saito error function possesses the same characteristics of that error. However, since both $E_{IS}^{R_d,p}$ and $E_{IS}^{0,p}$ generally decrease together, $E_{rIS}^{R_d,p}$ has the key property that it no longer decreases as the filter order increases, and therefore it can be used to choose p. This behaviour is due to the fact that the glottal contributions to the speech signal can be approximated by a low order all-pole filter (Markel and Gray, 1982; Alku and Laine, 1989; Doval and d'Alessandro, 2006); once the order of analysis is sufficiently high, the general shape of the speech spectrum is captured by the all-pole approximation and any further order increase serves only to model the finer spectral details. However, if the glottal derivative power spectrum is approximated by R_d, $E_{IS}^{R_d,p}$ will be small before the filter order is large enough to capture both the vocal tract and the glottal model.

Figure 6.1 shows the behaviour of $E_{IS}^{0,p}$, $E_{IS}^{R_d,p}$ and $E_{rIS}^{R_d,p}$ for a synthetic speech frame, with fixed R_d parameter and increasing filter order.

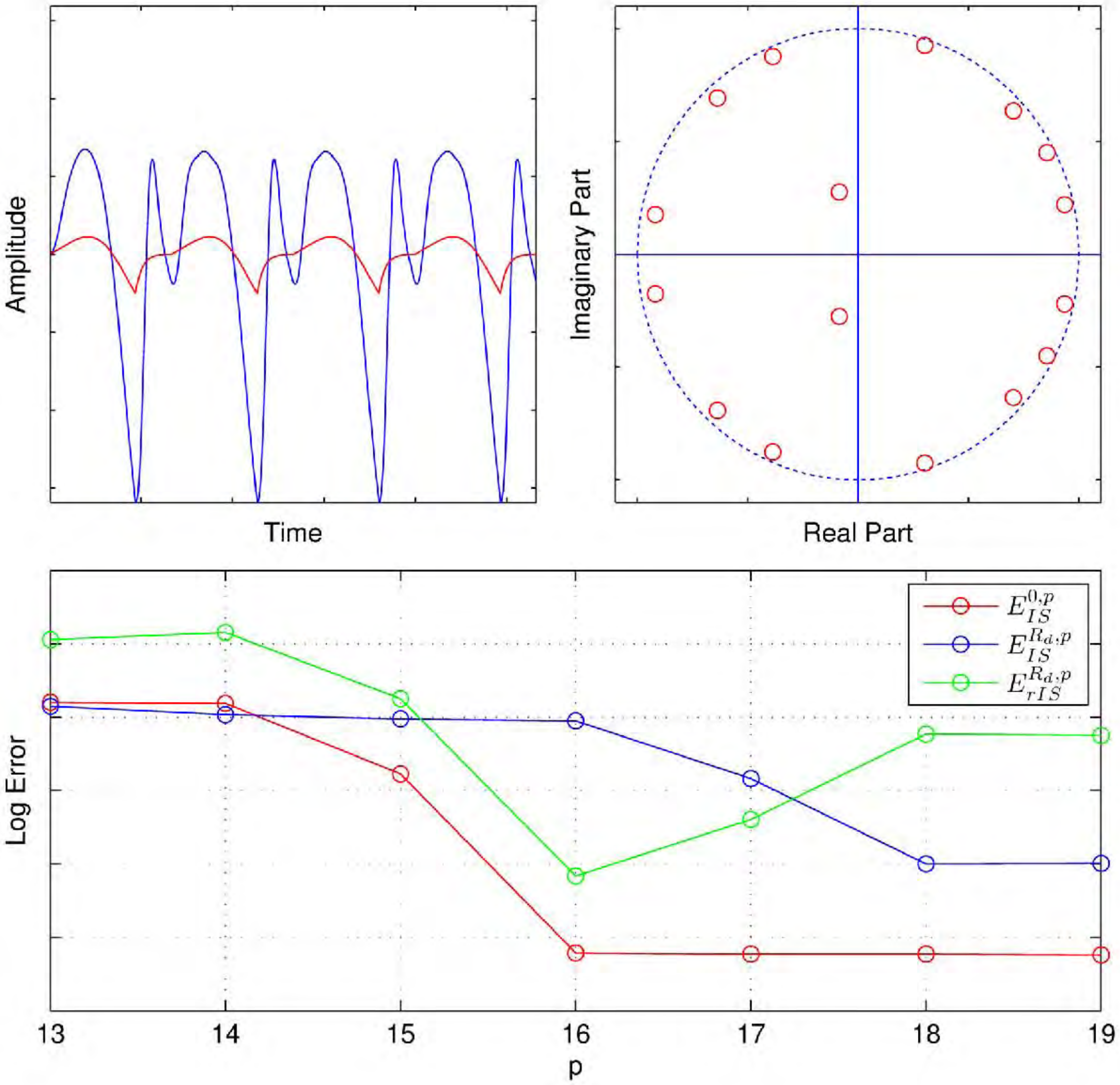


Figure 6.1: The figures above help to illustrate the operation of the Relative Itakura-Saito error for determining the optimal filter order. In the top-left figure is a voiced speech segment overlaid with the appropriate glottal signal. The 16^{th} order all-pole vocal-tract filter is represented by the top-right figure. In the bottom figure, the behaviour of the $E_{IS}^{R_d,p}$ and $E_{IS}^{0,p}$ can both be seen to decrease with increasing filter order. However, the Relative Itakura-Saito error (the ratio of these two errors) indicates the correct filter order (16).

E_e **Determination** The gains within with the speech signal are interdependent, meaning that an increase of the gain of the vocal tract can be offset by a decrease of the gain upon the voice-source signal. This makes it difficult to estimate the gain upon the source signal accurately. However, if the vocal-tract filter is normalised in some respect (for example, such that the first filter coefficient $a_0 = 1$), the amplitude variations of the signal can be attributed solely to the gain upon the voice source.

Because the DAP algorithm necessarily uses non-normalised filter coefficients (*i.e.* $a_0 \neq 1$) to match the scale and shape of the discrete spectrum under analysis, normalising the filter (by dividing all coefficients by a_0) gives the factor which must be applied to the input spectrum in order to match the sampled all-pole filter envelope. This gain factor b_0 is simply the reciprocal of the first filter coefficient:

$$b_0 = \frac{1}{a_0} \tag{6.7}$$

However, this is only correct in relation to the amplitude of the speech power spectrum, the determined vocal-tract filter, and the gain of the LF model spectrum used to calculate the voice-source model spectra. Two conditions are necessary for accurate determination of the E_e parameter:

Speech spectrum normalisation. The amplitude speech spectrum must be normalised by sum of the analysis window, as the amplitude of the spectral components is dependent upon the size of the window.

LF model spectrum normalisation. In addition to also being normalised by the length of the analysis window used to calculate its spectrum, the LF model spectra $G'^{R_d}(\omega)$ used to estimate the vocal tract all-pole filter $V^{R_d}(\omega)$ must be determined such that its scale parameter $E_e = 1$.

If these conditions hold, then $b_0 = E_e$. This novel perspective avoids the necessity for any time-domain parameter estimation, as necessitated by the SIM technique.

6.2.3 R_d Determination

The previous sections discuss how the all-pole filter coefficients, filter order and gain parameter of the incorporated glottal model may be determined from the vocal-tract filter power spectrum which has been estimated using a given transformed LF model shape parameter R_d. This section will discuss how this parameter is determined.

Brute Force Initialisation In order to obtain initial estimates of the parameters of a voice-source model, researchers have often utilised a codebook approach (Fröhlich *et al.*, 2001; Lu, 2002; Vincent *et al.*, 2005). The codebook contains a pre-determined set of glottal model parameter configurations which cover a subset of the available glottal shapes of that model. To obtain initial estimates of the R_d parameter, this work uses a similar brute force approach.

A codebook of R_d parameters is generated by sampling the span of the parameter across its range, *i.e.* $0.209 \leq R_d \leq 3$. Two opposing practical considerations must be taken into account when designing such a codebook:

- Adequate coverage over the subspace is required such that the initial parameter supplied to the optimisation procedure is close to the actual parameter.
- The codebook should not contain more entries than necessary to ensure efficient calculation.

These considerations oppose one another in that the first tends to increase the number of codebook entries, while the second demands a smaller subset. This work has found that 100 entries is adequate.

The PowRd algorithm determines the filter order in a similar, brute force fashion. As previously discussed, the order of the vocal-tract filter is related to the bandwidth of the speech signal. As the PowRd method essentially bandlimits the speech signal to the maximum voiced frequency F_m, the filter order is related to this frequency. The following equation, related to the usual rule of thumb where each kilohertz of bandwidth yields another formant, can be derived:

$$p_{F_m} = \lfloor \frac{2F_m}{1000} + 0.5 \rfloor \tag{6.8}$$

$$= \lfloor \frac{F_m}{500} + 0.5 \rfloor \tag{6.9}$$

where $\lfloor x \rfloor$ maps the real number x to the largest previous integer. However, as previously mentioned, the filter order is also related to the geometry of the vocal tract. Thus, the PowRd method tests over a range of filter orders in the region of p_{F_m}. This work utilises the range $p_{F_m} - 2 \leq p \leq p_{F_m} + 2$.

Vocal Tract Heuristics During the brute force initialisation, the estimated vocal-tract filters are factorised and the parameter configurations (shape parameter R_d and filter order p) corresponding to those tract containing positive real poles are removed from further consideration. This removal of positive real poles from estimated vocal-tract filters is a common procedure (Wong *et al.*, 1979; Childers and Lee, 1991; Alku *et al.*, 2009) as the vocal tract is assumed to be a resonating system and real poles contribute to the spectral tilt and not the filters resonant characteristics.

Additionally, there is no guarantee of filter stability. If the final estimated filter has

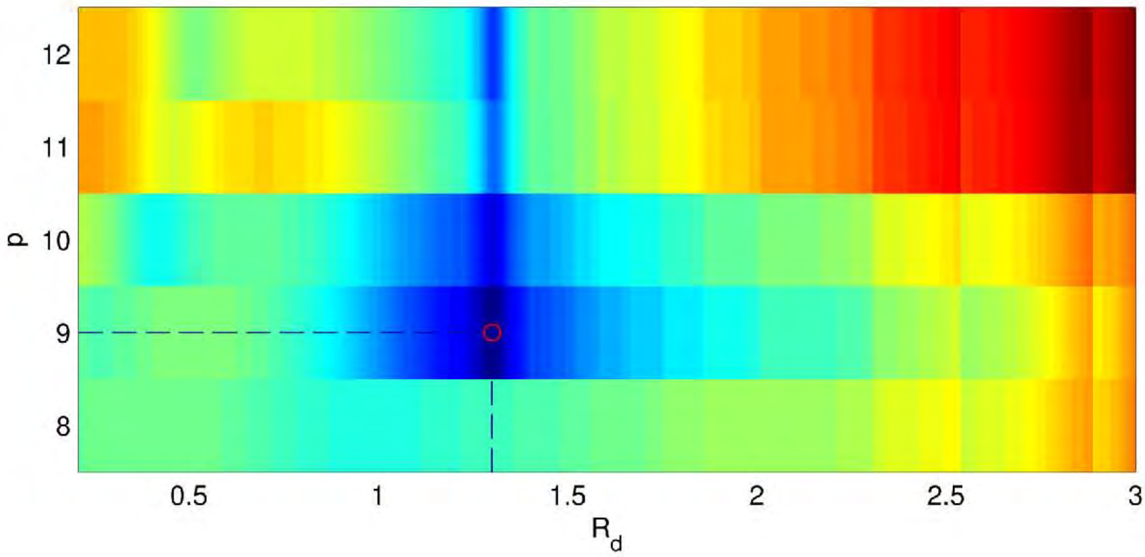


Figure 6.2: The above figure shows the extraction of the initial parameters from the Relative Itakura-Saito error matrix. Colder colours represent lower error values, and the optimal $R_d = 1.3$.

pole outside the unit circle, they are simply replaced by their mirror image partners, as explained in Section 3.2.1.

Refining the Estimates Following the brute force processing of the R_d codebook and filter order range, the initial estimates of the waveshape parameter and filter order, $\hat{R}_d$ and $\hat{p}$ respectively, can be extracted from the error surface, see Figure 6.2. The initialisation procedure gives a robust estimation of filter order, though the voice-source parameter may need to be refined in order to compensate for the discrete nature of the codebook. Thus, fixing the filter order $\hat{p}$, the Itakura-Saito error function $E_{IS}^{R_d,\hat{p}}$ is minimised using $\hat{R}_d$ as the starting point of an optimisation routine.

This parameter is refined using the downhill simplex method algorithm (Nelder and Mead, 1965), which has been applied successfully in other voice-source parameterisation methods (Strik, 1998; Tooher and McKenna, 2003; Vincent *et al.*, 2005; Kane *et al.*, 2010). The method functions without derivatives and is reputed to be relatively robust against bad initial estimates (Press, 2007). The implementation of this algorithm was performed by the Matlab (MATLAB, 2010) function `fminsearch`. Experiment has found that it suffices to allow a maximum of 100 iterations of the algorithm, with a function tolerance of 0.001.

6.2.4 PowRd Algorithm Summary

The PowRd method can be summarised as follows:

Algorithm 1: The PowRd algorithm.

```
Input: Speech frame $s[n]$, pitch period $T_0$, maximum voiced frequency $F_m$
Output: Shape parameter $R_d$, scale parameter $E_e$, filter coefficients $a_k$
Transform $s[n]$ into power spectrum representation $|S(\omega_k)|^2$, up to $F_m$;
Transform $\omega_k$ to $\omega'_k$ using Eq. 6.3;
Calculate $p_{F_m}$ according to Eq. 6.8;
foreach $p$ from $p_{F_m} - 2$ to $p_{F_m} + 2$ do
    Use DAP algorithm to calculate $E_{IS}^{0,p}$ using $|S(\omega'_k)|^2$;
foreach $R_d$ in codebook do
    Calculate $\left|G'^{R_d}(\omega_k)\right|^2$;
    Calculate $\left|V^{R_d}(\omega'_k)\right|^2 = \frac{|S(\omega'_k)|^2}{|G'^{R_d}(\omega_k)|^2}$;
    foreach $p$ from $p_{F_m} - 2$ to $p_{F_m} + 2$ do
        Use DAP algorithm to calculate $\hat{a}_k$ and $E_{IS}^{R_d,p}$ using $\left|V^{R_d}(\omega'_k)\right|^2$;
        Factorise $\hat{a}_k$ into roots $z$;
        if Any $\angle z = 0$ then
            Set $E_{rIS}^{R_d,p} = \infty$;
        else
            Calculate $E_{rIS}^{R_d,p}$ using Eq. 6.6;
Locate minimum of $E_{rIS}^{R_d,p}$ to obtain $\hat{p}$ and $\hat{R}_d$ ;
Use simplex search algorithm to minimise $E_{IS}^{R_d,\hat{p}}$ with initial parameter $\hat{R}_d$;
Given final $R_d$ value, calculate $a_k$ and $E_e$ with DAP algorithm.
```

6.3 Validation/Testing

In order to test its performance, the PowRd algorithm is compared with four other state-of-the-art voice-source estimation/parameterization methods. These methods are CPIF[1] (Wong *et al.*, 1979), IAIF[2] (Alku, 1992; Airas, 2008b), CCD[3] (Drugman *et al.*, 2009a) and an adapted SIM method[4] (Fröhlich *et al.*, 2001). As the first three methods are voice-source estimation algorithms, a second parameterisation stage is necessary to obtain an estimate for the voice-source parameter. A time-domain method similar to the one described in (Strik, 1998) was used. Furthermore, as the original description of the SIM method utilised the LF model rather than the transformed LF model, it was adapted slightly from the one described in (Fröhlich *et al.*, 2001).

6.3.1 Synthetic Speech

Because synthetic speech are generated under fully controlled circumstances, the sensitivities of each voice-source parameterisation algorithm to various phenomena can be observed. In this work, different scenarios were undertaken to test each algorithm. Those scenarios are:

- fundamental frequency, f_0
- glottal noise, SNR_g
- first formant centre frequency, F_1

[1] Own implementation.

[2] Own implementation.

[3] Implementation available from (Drugman, Retrieved October 8th, 2011).

[4] Own implementation.

- filter order, p
- interaction effects
- phase distortion

Further details of each phenomenon is given in the results section below.

The performance of the different algorithms upon synthetic speech is quantified by the difference between the estimated R_d^{est} parameter and its actual value R_d^{act}. The error is calculated simply as:

$$\Delta_{R_d} = R_d^{act} - R_d^{est} \tag{6.10}$$

By experimenting over a wide variety of scenarios, the mean $\mu_{\Delta_{R_d}}$ and standard deviation $\sigma_{\Delta_{R_d}}$ of this error is a good indicator of the performance of each method.

The synthetic speech tokens were generated using 500 randomly generated R_d parameters across its range of variation from 0.209 to 3, and different combinations of three variables: fundamental frequency, glottal noise level and vocal-tract filter. For each token, the amplitude parameter E_e was fixed at 1, though the estimation error upon this parameter is not measured in this chapter[5]. Additionally, for the experiments where the impact of a specific phenomenon is being observed, the other variables are set to fixed values - f_0 was fixed at $140Hz$, the glottal noise to signal ratio was fixed at $30dB$, and the vocal-tract filter was filter (xiii) (see Appendix A). Unless otherwise specified, each experiment was conducted using the non-interactive speech model under linear phase conditions and, with the exception of the PowRd method, given the correct filter order.

[5] Experiments validating the PowRd approach to E_e estimation are presented in Chapter 8

6.3.2 Real Speech

Objective quantification of the quality of the voice-source parameterisation of real speech is a difficult problem as in this case there is no ground truth available. However, the waveshape parameter R_d and local pitch period T_0 can be used to predict the open quotient $O_q^{R_d}$ of the waveform under analysis (using the equations found in Section 2.3.3.1), which can be corroborated with the open quotient parameter estimated from a synchronously recorded EGG signal, O_q^{EGG}. The open quotient error Δ_{O_q} is calculated as the difference between the open quotient estimates:

$$\Delta_{O_q} = O_q^{EGG} - O_q^{R_d} \tag{6.11}$$

Conversely to the previous experiment which utilised the mean and standard deviation of the calculated error, for the real speech experiments the performance of each approach is indicated by the median $\mu_{\frac{1}{2}}$ and the interquartile range iqr of Δ_{O_q}. The median and interquartile range are used in this case because they are more robust to outliers in the data which are more likely to appear in the real speech experiments due to, *e.g.* inaccurate glottal closing instant estimations from the SIGMA analysis of the EGG signal, noisy analysis frames, aperiodicities, *etc.*

The algorithms were tested upon 15 utterances spoken by three voices taken from the CMU-ARCTIC database (Kominek and Black, 2003): two male voices (*bdl* and *jmk*) and one female voice (*slt*). These voices were recorded in a sound-proof booth at 32kHz sampling rate with simultaneous EGG measurements, and their lexical contents represents phonetically balanced American English. Before testing, both signals were down-sampled to 16kHz following processing using a zero phase low pass anti-aliasing filter. Additionally, the speech and EGG signals were time-aligned to compensate for delay introduced by the propagation of the acoustic signal from the glottis to the microphone (Veeneman and BeMent, 1985). This delay was assumed constant over the entire database for each speaker and was estimated to be approximately 0.94ms for speakers *bdl* and *slt* while a delay of approximately 0.69ms was suitable for speaker *jmk*. This delay is similar to the one used by another study (Cabral *et al.*, 2011).

Following pre-processing, the SIGMA algorithm[6] (Thomas and Naylor, 2009) is applied to the EGG signal to determine the instants of glottal closure which are used to centre the analysis frames for each technique. The reference open quotient values O_q^{EGG} are calculated from the EGG signal using the thresholding method described in (Howard, 1995). Using the glottal closing instants which can be robustly detected from the DEGG, this method establishes a threshold which is used to determine the time during which the glottis is open, see Figure 6.3. The threshold τ_{O_q} was set to $\frac{3}{7}(max - min)$, as has been established experimentally by (Davies *et al.*, 1986). This robust open quotient estimation method does not rely on the presence of strong peaks in the DEGG waveform to estimate the glottal opening instant (Howard, 1995).

[6] Available from (Brookes, Retrieved January 22nd, 2009).

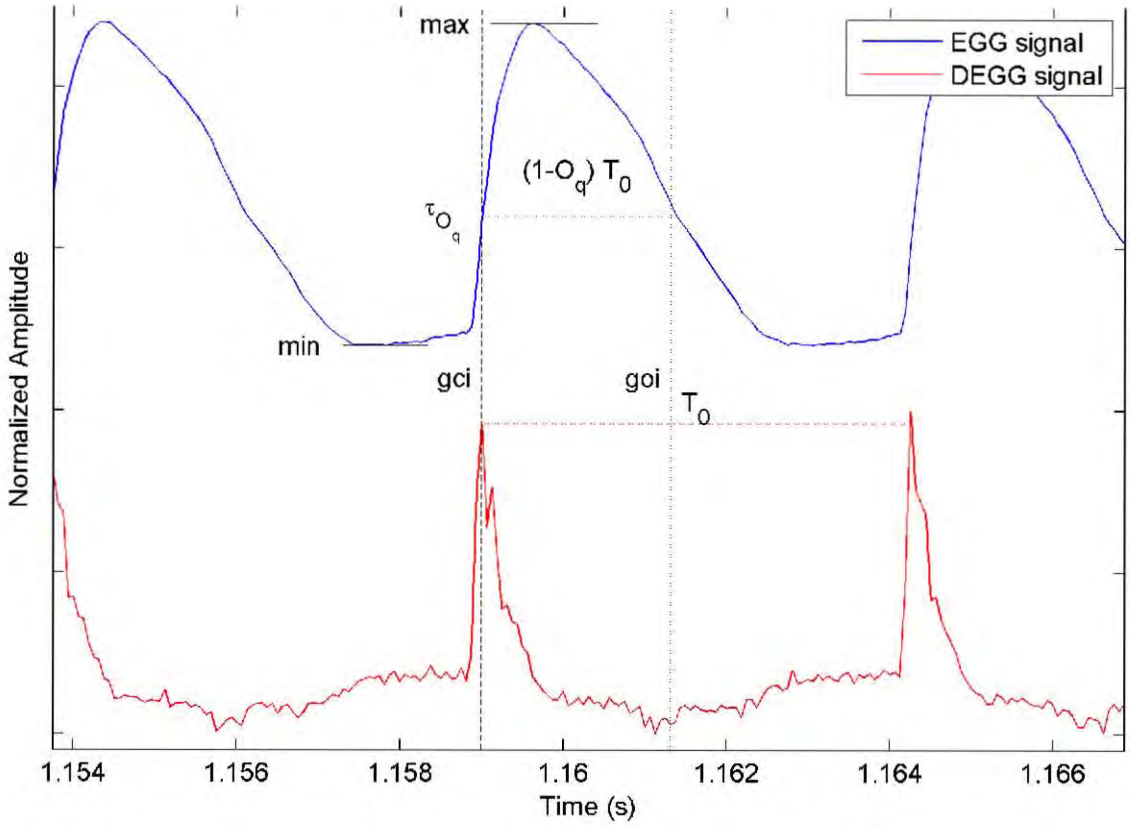


Figure 6.3: The above figure illustrates the thresholding method of glottal cycle open quotient estimation from the EGG and DEGG signals, as described in (Howard, 1995).

6.4 Results and Discussion

This section discusses and interprets the results of the various experiment undertaken. Figures 6.4 to 6.8 illustrate the performance of the source estimation algorithms upon synthetic speech, while Figure 6.10 to 6.13 and Tables 6.1 to 6.2 gives the results for analysis of real speech signals.

6.4.1 Synthetic Speech

Glottal Noise A certain level of noise will be found in real speech. In order to observe the impact of this phenomenon, modulated Gaussian noise was added to the source signal at six different signal to noise ratios, from 60dB SNR to 10dB SNR in steps of -10dB SNR, while maintaining the pitch and vocal-tract filter constant. The noise was amplitude modulated using a raised glottal flow signal determined by the LF model parameters, and passed through a differentiation filter representing lip radiation, in a manner similar to (Agiomyrgiannakis and Rosec, 2008). The results of this experiment can be seen in Figure 6.4.

It is clear that all voice-source parameterisation methods suffer from by the addition of noise, and generally follow the trend that the greater the amount of the noise, the further degradation to the parameterisation. In particular, CCD and CPIF suffer from this phenomenon, while conversely the IAIF method does not exhibit pronounced performance degradation with increasing noise levels. This corroborates the findings in (Drugman *et al.*, 2011a), which also observed the sensitivity of the CCD and CPIF methods to noise and robustness of the IAIF method.

With increasing noise, both power-spectrum-based approaches PowRd and SIM tend to underestimate the correct R_d parameter. This is due to the intrusion of high frequency noise, which boosts the amplitude of high frequency spectral samples, and thus underestimates the spectral slope. However, due to the HNM-type model adopted by the PowRd algorithm, it does not experience such a loss of performance until the noise levels reach a certain level. It is expected that if the maximum voiced frequency parameter F_m, which controls the analysis signal bandwidth, was assigned in a dynamic fashion and not fixed (at 3kHz, also the cutoff frequency of the low-pass utilised during time-domain parameterisation) as it was in these experiments, improved performance would be observed.

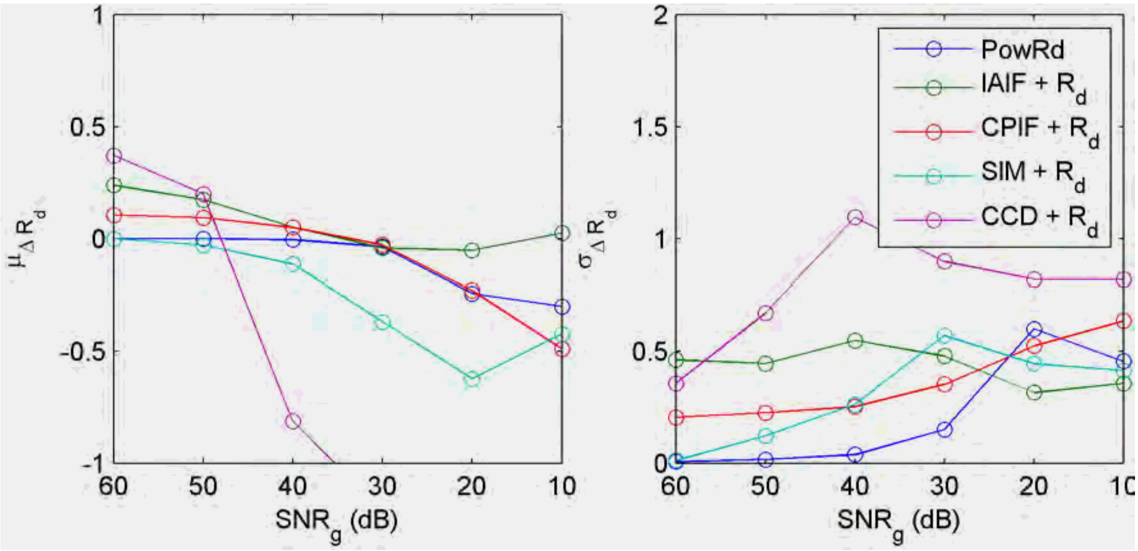


Figure 6.4: The above figure gives the impact of glottal noise upon voice-source parameterisation methods.

Fundamental Frequency In the frequency domain, higher fundamental frequencies imply wider spacing between the spectral samples and thus less information within a given bandwidth. Similarly in the time domain, higher pitch implies shorter pulse periods. While keeping the vocal-tract filter and noise characteristics of the signal the same, the fundamental frequency of synthetic speech is varied in order to observe the performance of each algorithm relating to this variable. The range of fundamental frequency are from 80Hz to 240Hz in steps of 20Hz. The results of which can be seen in Figure 6.5.

Though is was expected that a decrease in accuracy would be witnessed with increasing frequency, due to the shorter analysis frames, the data implies that almost all of the methods under analysis are relatively robust to increases in fundamental frequency. However, the PowRd method does see a jump in the standard deviation of the parameterisation error Δ_{R_d} once the fundamental frequency increases beyond 180Hz, possible due to the decreasing number of spectral samples. Another exception can be made for the CCD approach which actually *improved* performance with increasing frequency, contradicting the findings of (Drugman *et al.*, 2011a), where increasing f_0 had little impact.

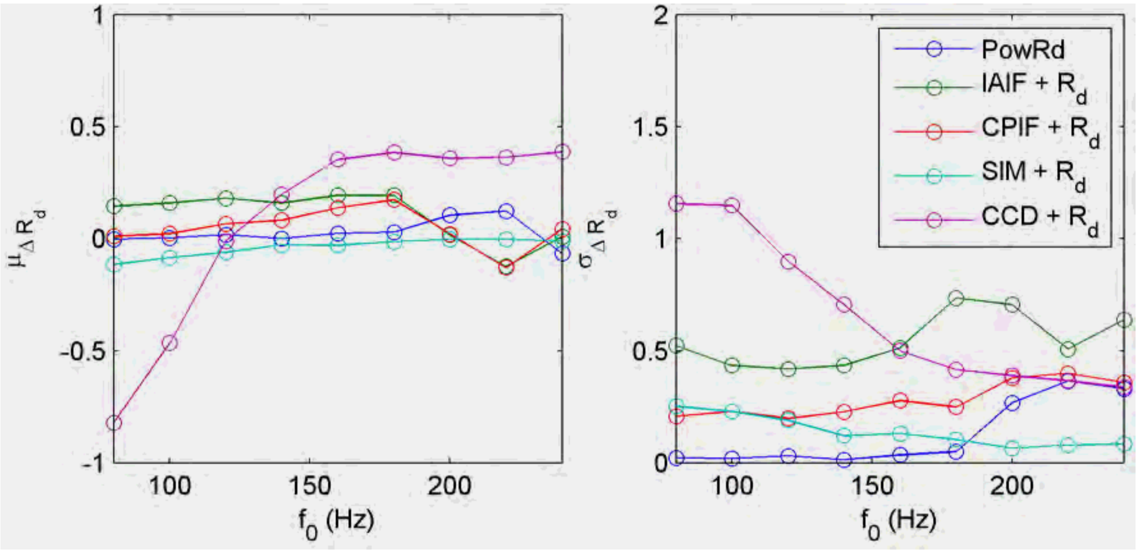


Figure 6.5: The above figure gives the impact of fundamental frequency upon voice-source parameterisation methods.

First Formant It is well known that interference between the first formant F_1 and the glottal formant can prevent successful deconvolutions of the source and filter.

Twenty-two different vocal tract functions covering the vocalic trapezoid with varying first formant frequencies are used during experimentation. The parameters and diagrams of these functions are given in Appendix A, and the results of the experiments is contained in Figure 6.6.

The glottal formant of the voice-source signal depends on the waveshape parameter of the glottal flow, and spans the approximate range from $0.9f_0$ to $3.5f_0$. In these experiments where the fundamental frequency is held constant at 140Hz, the glottal formant and first formant will overlap at frequencies below 500Hz. As can be seen in Figure 6.6, all methods show some influence of this effect, though it is most pronounced for the IAIF method. This iterative procedure adopted by this approach attempts to remove low frequency information of the speech signal which it is assumed, represents the glottal contributions. However, the assembled low order model can occasionally "lock" onto a low frequency formant instead. Unlike the IAIF approach, the power-spectrum-based approaches are relatively robust to these low first formant errors as they avoid confusion with the vocal tract by brute force, and do not use the characteristics of the speech signal itself to determine an appropriate glottal model.

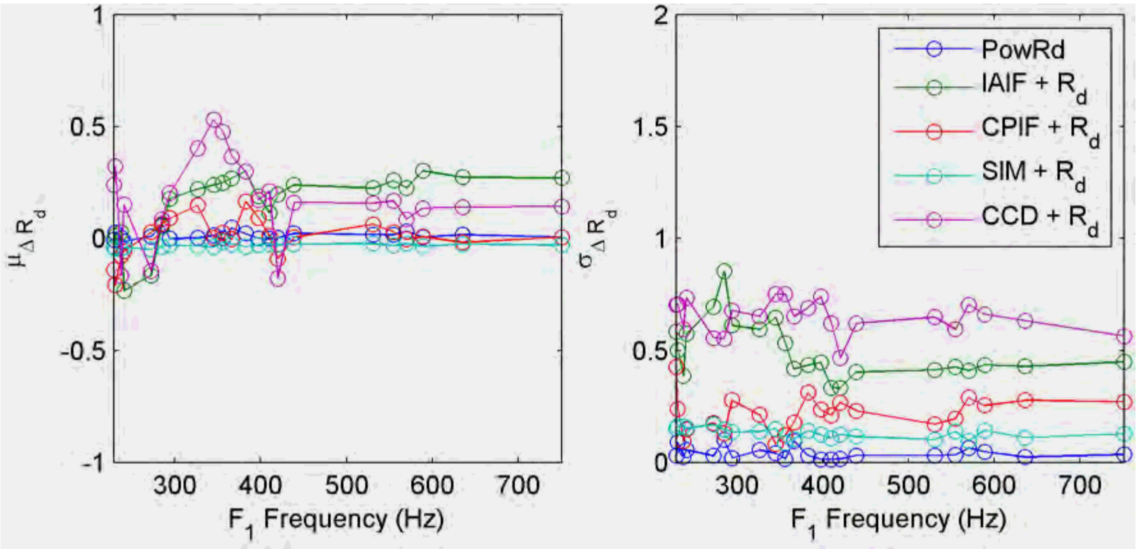


Figure 6.6: The above figure gives the impact of the first formant centre frequency upon voice-source parameterisation methods.

Filter Order The order of the vocal-tract filter is an unknown parameter and estimated using the usual rule of thumb. An experiment measuring the impact of varying this parameter is undertaken using a range of filter orders, from $p\pm2$ where p is the ideal filter order of 16. Obviously because the CCD does not impose a parametric model upon the vocal tract, that method is excluded from this experiment. As the Relative Itakura-Saito error is capable of recovering the parameters nearly perfectly for the PowRd algorithm, it is included for reference purposes.

As can be seen in Figure 6.7, choosing the incorrect filter order does have some impact upon the results. This is particularly noticeable when the order is underestimated, which can give substantial degradation to the effectiveness of each algorithm, particularly regarding the spread of the results.

The closed-phase inverse filtering method actually achieves optimal filter order at

$p+1$. This is most likely because the extra flexibility of the additional pole allows the method to model a nonzero mean in the analysis frame caused by, for example, the return phase of the glottal cycle. The closed-phase inverse filtering routine used in this work removes real poles from the estimated filter for this reason.

Finally, it is also noted that the frequency-domain approaches to vocal-tract filter estimation (SIM and IAIF) achieve similar performance once the filter order rises to the correct filter order.

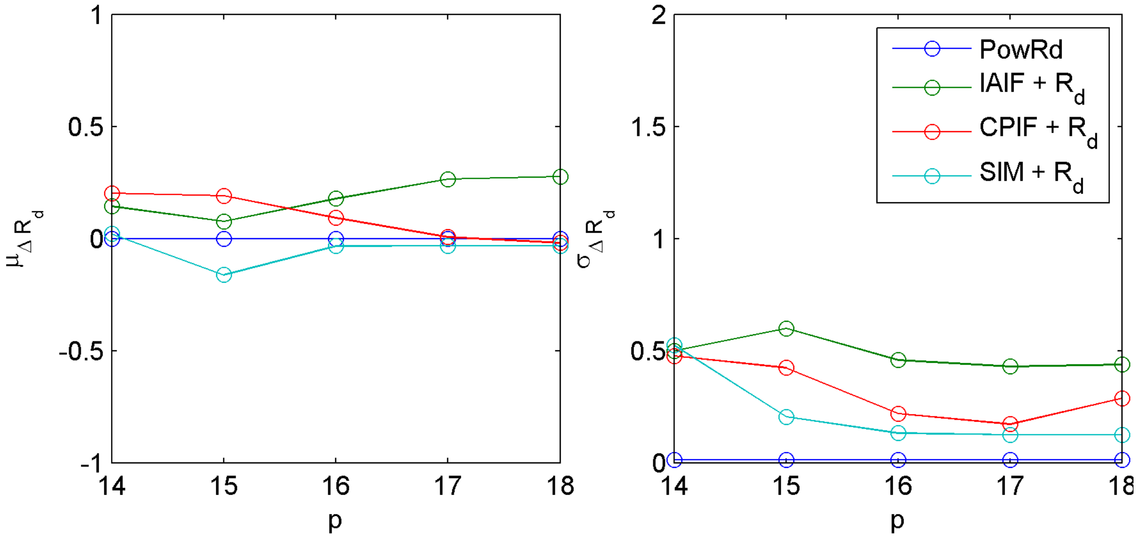


Figure 6.7: The above figure gives the impact of the filter order upon voice-source parameterisation methods.

Interaction Effects As mentioned in Section 2.2.1, interaction between the behaviour of the glottal source and the resonances of the vocal tract are known to be present in real speech. Thus, a basic modeling of the effect was introduced in a more realistic speech model by implementing a different vocal-tract filter during the open phase of the glottal signal. The open-phase filter was determined by interpolation from the closed-phase filter to another such that the centre frequency and bandwidth of the first formant are increased by 10%, consistent with the experimental findings (Ananthapadmanabha and Fant, 1982; Krishnamurthy, 1992). The interpolation was performed linearly using line spectral frequencies (Itakura, 1975). The results of simulating this natural speech effect is seen in the middle row of Figure 6.8. When compared with the reference performance in the top row of Figure 6.8, it can be seen that it degrades slightly the performance of all methods.

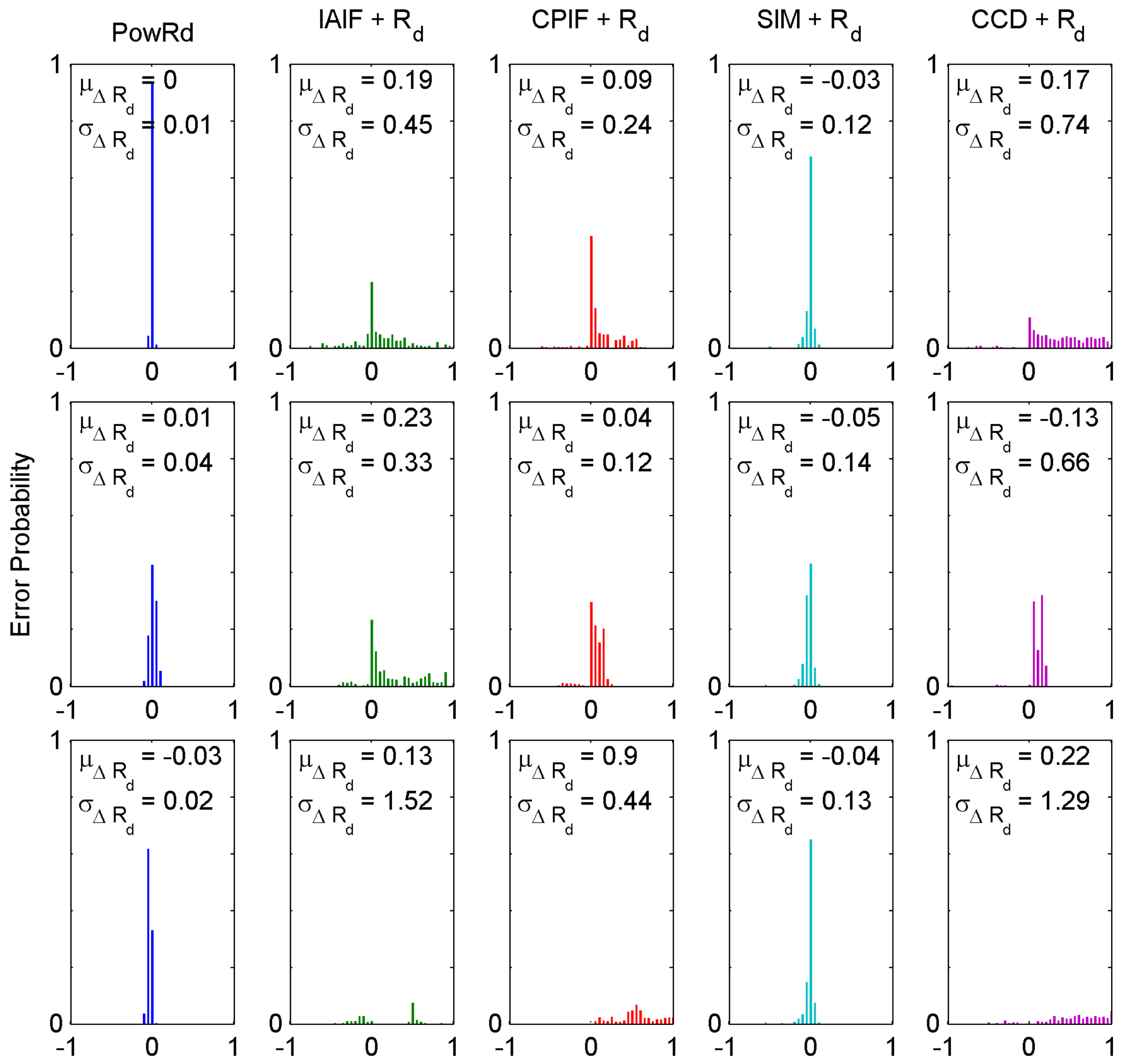


Figure 6.8: The above figures show the Δ_{R_d} error distributions for three scenarios, using 500 randomly generated glottal pulses. (Top) Reference set, (middle) source-filter interaction effects, and (bottom) phase distortion.

Phase Distortion Finally, the effect upon the algorithms of phase distortion is also observed in the bottom row of Figure 6.8. Methods relying on error functions which are sensitive to phase disturbances are expected to be compromised by this common phenomenon. When compared with the reference performance under phase linear conditions in the top of Figure 6.8, it can be seen that the power-spectrum-based methods are virtually unaffected by the introduced distortion, while unsurprisingly, the effectiveness of the IAIF, CPIF and CCD methods are severely compromised.

In this experiment, phase distortion is imposed upon synthetic speech segments by spectral multiplication of the estimated spectral nonlinear phase characteristics of a professional recording studio. These characteristics were measured using reference waveforms of known shape, similar to the manner proposed in (Holmes, 1975) and (Berouti *et al.*, 1977). This transfer function, a low frequency detail of which is given in Figure 6.9, affects mainly the low frequencies of the signal and severely alters the signal shape but has little to no impact upon the signal perceptually. As characteristics of the phase distortion are system-dependent, it is expected that the performance of phase sensitive algorithms will differ if different recording equipment is utilised. However, the performance exhibited by the power-spectrum-based approaches can be generalised to clean speech signal recordings, as they explicitly ignore phase information as much as possible.

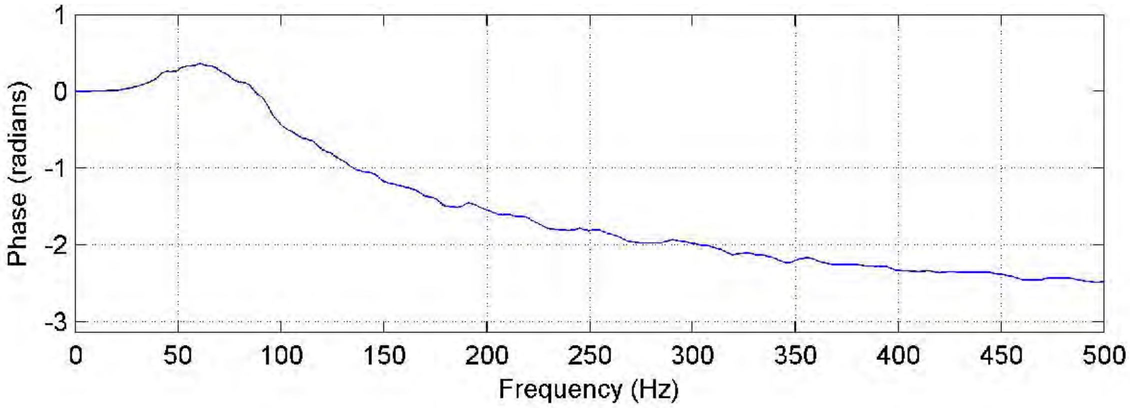


Figure 6.9: The above figure shows the low frequency detail of a transfer function estimated from a professional studio recording equipment used for distorting the phase of the speech segments in the experiments.

6.4.2 Real Speech

The results of the real speech experiments are presented as the distribution of the open quotient error Δ_{O_q} in Figures 6.10 to 6.13, which are also described by their medians ($\mu_{\frac{1}{2}}$) and interquartile ranges (iqr) in Tables 6.1 and 6.2.

Under phase linear speech recordings, where the phase conditions of the signal were not perturbed in any way, a similar performance in accuracy is noted between all methods for speakers *bdl* and *slt*. For these two speakers, only the closed-phase inverse filtering method distinguishes itself by producing the most consistent (*i.e.* exhibiting the lowest iqr) results in both cases. Indeed, closed-phase inverse filtering also produces the most consistent result for the all voices including speaker *jmk*, however for this voice, it is the power-spectrum-based approaches which produce markedly the most accurate results.

Table 6.1: The table below gives the O_q estimation performance, for three speakers with utterances recorded under linear phase conditions.

Speaker		PowRd	SIM+R_d	IAIF+R_d	CPIF+R_d	CCD+R_d
bdl	$\mu_{\frac{1}{2}}$	-0.052	-0.060	-0.066	-0.070	-0.075
	iqr	0.135	0.143	0.111	0.100	0.283
jmk	$\mu_{\frac{1}{2}}$	-0.053	-0.065	-0.119	-0.141	-0.137
	iqr	0.132	0.146	0.142	0.107	0.449
slt	$\mu_{\frac{1}{2}}$	-0.197	-0.198	-0.201	-0.214	-0.198
	iqr	0.135	0.113	0.150	0.084	0.155

Table 6.2: The table below gives the O_q estimation performance, for three speakers with utterances recorded under simulated nonlinear phase conditions.

Speaker		PowRd	SIM+R_d	IAIF+R_d	CPIF+R_d	CCD+R_d
bdl	$\mu_{\frac{1}{2}}$	-0.037	-0.047	-0.136	-0.144	0.165
	iqr	0.134	0.142	0.243	0.264	0.109
jmk	$\mu_{\frac{1}{2}}$	-0.031	-0.044	0.106	-0.152	0.295
	iqr	0.144	0.162	0.406	0.227	0.075
slt	$\mu_{\frac{1}{2}}$	-0.191	-0.192	-0.197	-0.242	0.165
	iqr	0.140	0.112	0.213	0.157	0.112

As shown in Figures 6.10 and 6.11, the CCD method yields a bimodal distribution for both male speakers. Distributions of this type have been observed previously with approaches based upon the minimum/maximum phase model of the voiced speech signal *e.g.* (Drugman and Dutoit, 2010), where it is presumed that the minor mode of the distribution is indicative of incorrectly decomposed analysis frames. Incorrectly

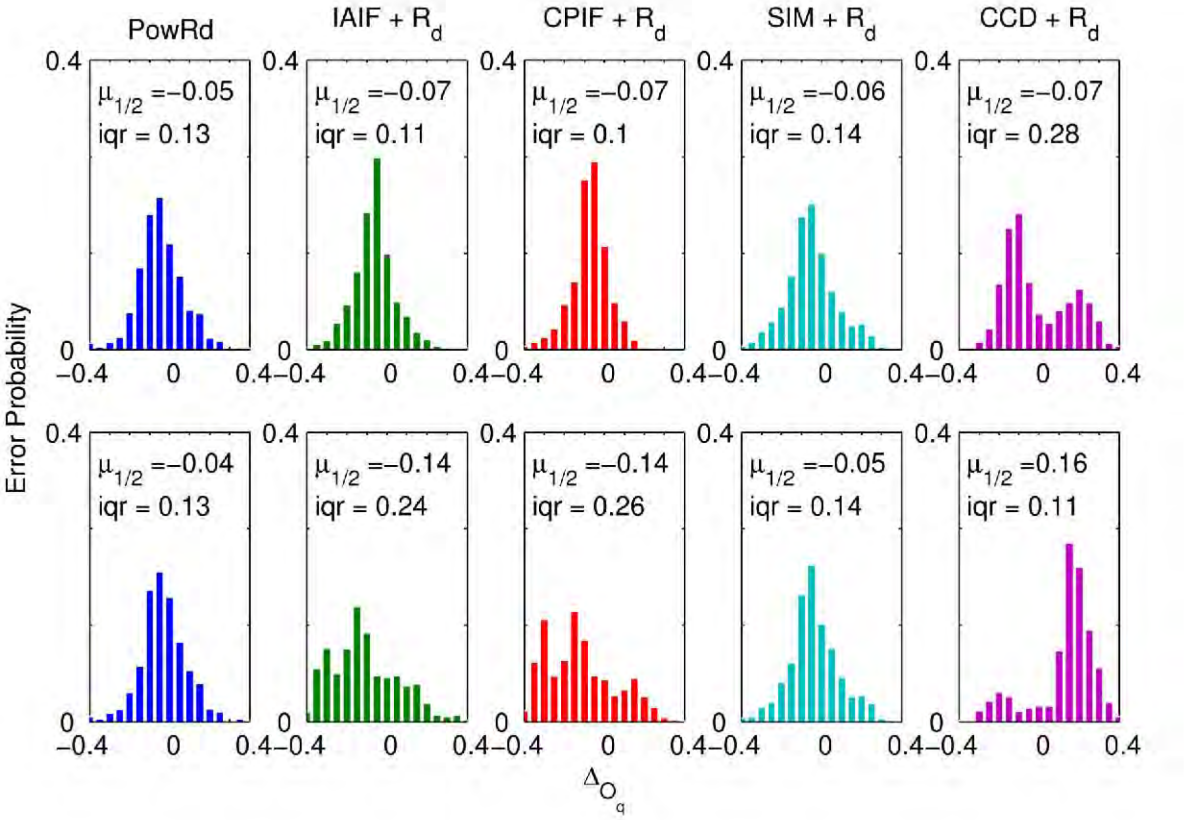


Figure 6.10: The above figures give the O_q estimation performance upon speaker *bdl*: (Top) phase linear conditions, (bottom) nonlinear phase conditions.

decomposed frames generally give noise-type waveforms, see Figure 6.12; the subsequent glottal model fitting operation will tend to low R_d values, thus yielding high Δ_{O_q} values when compared with the reference. The results presented here seem to corroborate this conclusion as it is noted that if these samples were excluded from the analysis, error distribution similar to those exhibited by the other methods would be obtained.

Though generally similar to the closed-phase inverse filtering method, the iterative adaptive inverse filtering is generally less precise than that method giving larger *iqr*

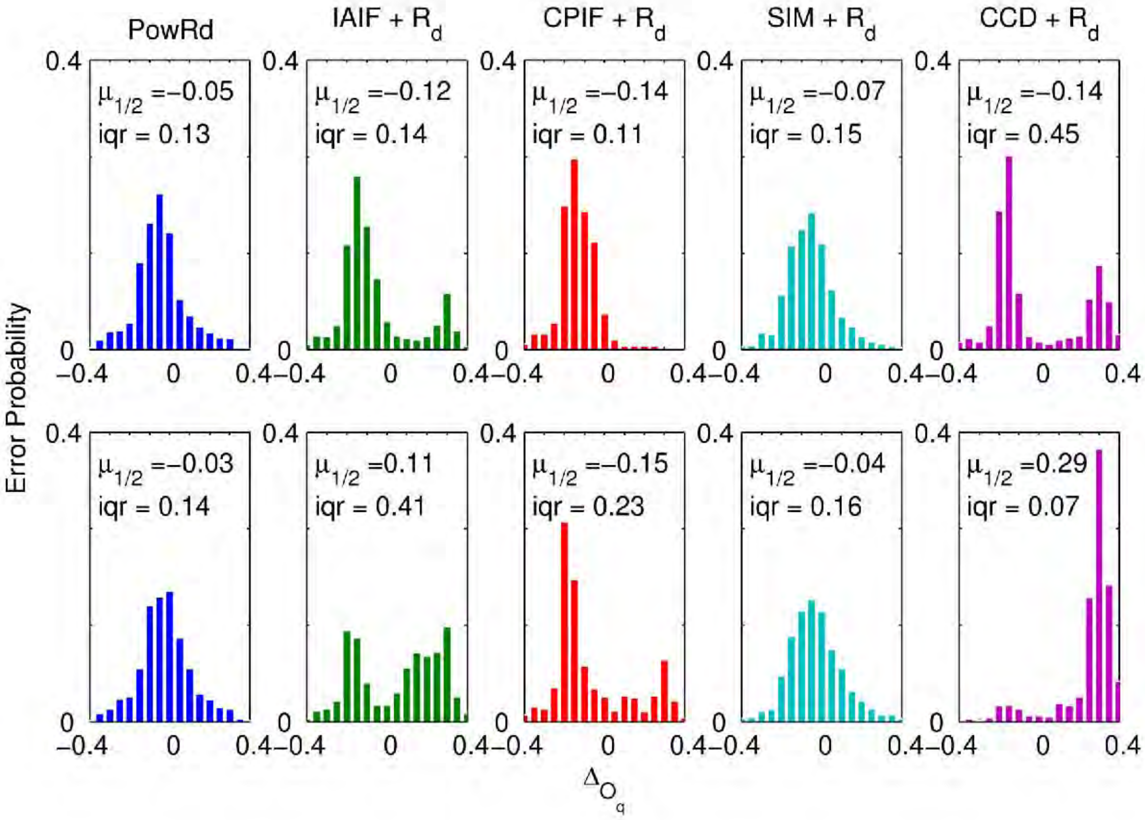


Figure 6.11: The above figures give the O_q estimation performance upon speaker *jmk*: (Top) phase linear conditions, (bottom) nonlinear phase conditions.

values and, like the CCD approach, yielding a bimodal distribution for the speakers *jmk* and slightly *slt*.

It is clear that the power-spectrum-based methods give the most accurate results in all cases, particularly in the case of speaker *jmk*, where the difference between the PowRd and SIM methods is notably lower that the other approaches. Between themselves, the PowRd method gives generally a slight improvement in accuracy and precision over the SIM method, though it is less precise with the female voice analysed in this experiment ($iqr_{PowRd} = 0.14$ versus $iqr_{SIM} = 0.11$).

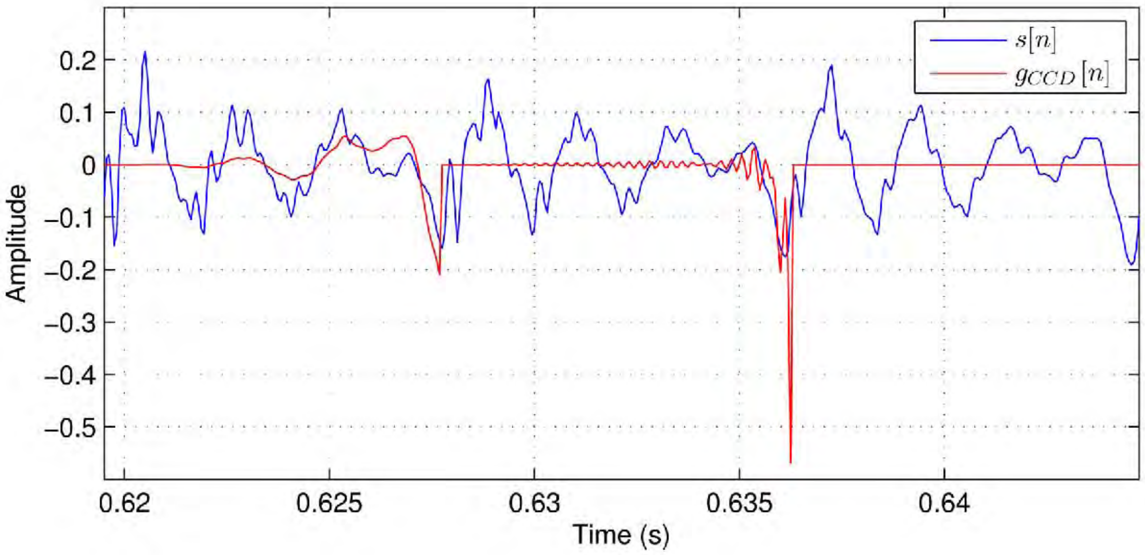


Figure 6.12: The above figure shows two adjacent voiced speech pulses decomposed by CCD, correctly (left) and incorrectly (right).

When comparing the error distributions in both phase scenarios, it is clear that, as expected, the power-spectrum-based approaches are robust to the phase distortions as the overall shape and position of the distributions is very similar. Similarly unsurprising, the disturbed phase conditions can adversely affect the performance of the other approaches relying on time-domain parameterisation; this follows from the marked dissimilarity of the error distributions to the phase-linear case.

Curiously, the precision of the CCD method increases when analysing phase-distorted speech. This is probably indicative of consistently incorrectly decomposed speech frames.

The IAIF and CPIF approaches yield the error distributions are more widely spread and less accurate than the phase-linear case. However, there is a less dramatic shape change in the case of the female voice *slt*. This is perhaps due to the low frequency impact of the simulated phase distortion which would impact lower pitched voices more greatly than higher pitched ones.

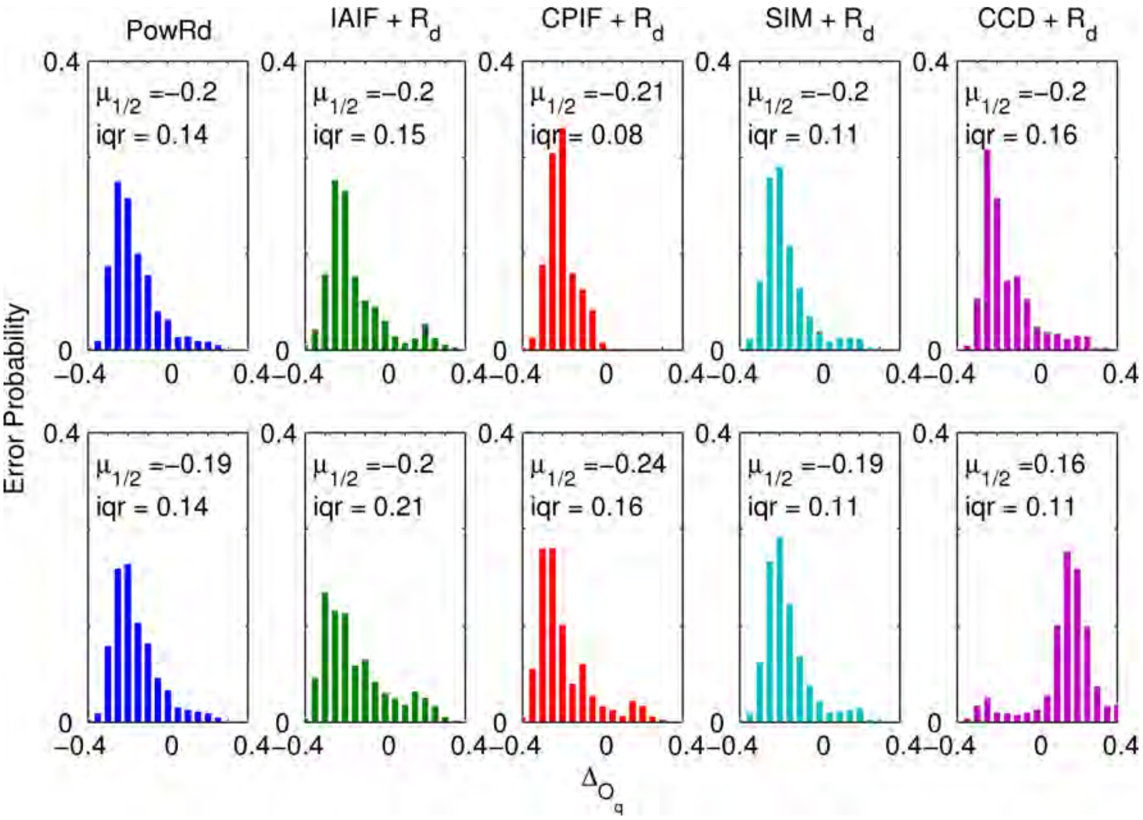


Figure 6.13: The above figures give the O_q estimation performance upon speaker *slt*: (Top) phase linear conditions, (bottom) nonlinear phase conditions.

6.5 Conclusions

This chapter has described a new method of voice-source parameterisation, referred to in the introductory chapter as Contribution 1, called the PowRd method. The PowRd method is closely related to the SIM method, with certain extensions which:

- improves the performance of the algorithm in noise,
- determines the filter order automatically using a novel error criterion, the Relative Itakura-Saito error, and
- estimates the optimal E_e parameter simultaneously with the vocal-tract filter.

Experimental testing confirms the improvements of the PowRd method over the SIM method and also tested against 3 other state-of-the-art voice parameterisation methods: IAIF, CPIF and CCD. The efficacy of the methods were tested upon synthetic speech considering different phenomena likely to be encountered when analysing natural speech. Those phenomena were:

- differing levels of glottal noise,
- a range of fundamental frequencies,
- different vocal tract configurations,
- incorrectly supplied filter orders,
- simulated source-filter interaction and
- the presence of phase distortion effects.

The addition of noise degraded all methods, though the two-band model adopted by the PowRd approach demonstrated increased robustness than the SIM technique. The CCD technique was found to be particularly sensitive to noise. All techniques showed a similar robustness to rising fundamental frequencies except for CCD which actually improved, while only the lowest formant affected the voice parameterisation methods, in particular the IAIF technique. Interaction effects degrade all algorithms slightly, while only the power-spectrum-based methods demonstrated robustness to simulated phase distortion. Of the two power spectrum methods, the PowRd technique consistently shows superior performance in synthetic testing in its ability to recover the R_d parameter, accurately and precisely.

Due to the more challenging conditions of natural speech, for example, the stationarity of the speech frame, pitch irregularities *etc.* real speech experiments are more difficult to interpret. However, using the EGG signal as a benchmark from which open quotients could be estimated, the real speech experiments broadly confirm the findings of the previous synthetic speech tests, particularly the robustness of the power-spectrum-based voice-source approaches to phase distortion. In the phase-linear case, closed-phase inverse filtering was found to give the most consistent results, provided that accurate glottal closing instant information can be provided. While the results obtained were similar in the case of speakers *bdl* and *slt*, the power spectrum approaches were markedly more accurate for speaker *jmk*. Finally, the complex-cepstrum-based decomposition is determined to be the least robust of the other tested voice-source estimation approaches for open quotient parameterisation, in almost every case yielding the largest interquartile range values.

Chapter 7

Phase-Distortion-Robust Glottal Closing Instant Estimation

This chapter introduces Contribution 2, a new method for the determination of the glottal closing instants of a voiced speech signal, called the Fundamental RESidual Search (FRESS) method. The FRESS method is inspired by other extant methods, however, it is specifically designed to handle speech signals that may have been recorded using nonlinear phase equipment. Like the SEDREAMS and ZFR methods, it passes the speech signal through a low pass filter of very low cut-off frequency, so as to reveal only the first harmonic of the speech signal. However, as identified in Chapter 5, the fundamental frequency sinusoidal is not robust to phase distortions which may be have imparted upon the signal. In order to produce a GCI estimation method robust to possible phase disturbances, landmarks extracted from this sinusoidal signal are aligned with the peaks of the speech signal's energy contour, a signal relatively more robust to phase distortion.

Realigned landmarks from the fundamental sinusoid signal indicate likely intervals of glottal closure, removing the necessity to make assumptions about the GCIs as does both the SEDREAMS and ZFR algorithms. Searching these regions for peaks of the LPC residual signal yields a set of glottal closing instant candidates. Like the DYPSA and YAGA methods, dynamic programming is used to select the likely sequence of candidates based upon certain speech heuristics.

In this chapter, the FRESS method is compared with eight state of the art glottal epoch estimation techniques. The experiment shows that comparable performance is obtained for the case where the signal has been recorded using phase linear equipment, while in the case of a simulated nonlinear recording environment, the FRESS algorithm determines the instants of glottal excitation more reliably and accurately than other approaches.

The chapter first describes the two stages of the FRESS algorithm, comprising glottal epoch selection and dynamic programming. Following this section, the comparative experiment is described, after which the results are discussed and conclusions given.

7.1 The FRESS Algorithm

Briefly stated, the FRESS method, like the SEDREAMS (Drugman and Dutoit, 2009) method and ZFR methods, uses a mean-based signal to determine glottal epoch candidates and then, like the DYPSA (Naylor *et al.*, 2007) and YAGA (Thomas *et al.*, 2011) methods, refines the candidates using the N-best dynamic programming algorithm. This step penalises candidates which deviate from heuristic properties of the glottal epochs of speech. This section will detail each stage of implementation.

7.1.1 GCI Candidate Selection

The first stage of the FRESS algorithm is the selection of likely glottal epoch candidates. In the DYPSA and the YAGA methods, these candidates are initially found by locating where a group delay function crosses the time axis in response to impulse-like behaviour found in the signal under analysis. Unfortunately, this leads to many spurious candidates, which must then be rejected by the subsequent dynamic programming algorithm. For this reason, the SEDREAMS algorithm has the advantage whereby many spurious peaks are avoided by searching for the prominent impulse events in isolated regions of the LPC residual waveform. The FRESS algorithm attempts to exploit the ability of the SEDREAMS algorithm to locate likely regions of the glottal epochs in such a way that the spurious candidates are minimised.

The ability of the SEDREAMS technique to partition the LPC residual signal into the regions likely to contain glottal epochs come from a mean-based signal. As its construction is equivalent to zero phase low pass filtering, the mean-based signal can be interpreted as the fundamental sinusoid of the voiced speech signal. This signal is assumed to have a relatively static relationship with the position of the glottal closing instants[1]. However, if the signal has been phase distorted, imparting a certain delay, this relationship can be altered. Therefore, a fixed search area in relation to the mean signal may not necessarily contain the glottal epoch.

[1] This is true for the original SEDREAMS algorithm. As mentioned in Section 4.2, a new implementation uses an median-based approach to determine the relationship of the minima of the mean-based signal to the likely regions of glottal closure.

The FRESS algorithm attempts to realign the minima extracted from the fundamental harmonic signal with an energy function which is more robust to the position of the glottal epochs and less affected by phase distortion.

Thus, the first stage of the FRESS algorithm is to generate the fundamental sinusoidal signal. Because the minima of this signal will be realigned after filtering, the signal can be formed by using any low pass filter which has the ability to isolate only the fundamental harmonic of the voiced speech segment, *i.e.* the zero-phase property of the filter is unimportant. Should the filtered signal introduce any harmonics above this single fundamental, the resulting signal will exhibit additional extrema which may increase the false positives identified by the approach. In this work, the fundamental harmonic signal is created using a sixth order low pass Butterworth filter. A sixth order filter provide ensure good rejection of the unnecessary high frequency signal information. The only remaining parameter is the cut-off frequency f_c of this filter. Given an estimate of the mean fundamental frequency of the signal $\overline{f_0}$, a cut-off frequency of $1.1\overline{f_0}$ gives a good compromise between the false positive, miss and identification rates (as defined below in Section 7.2), see Figure 7.1.

Landmarks are then extracted from this oscillating signal. If it is assumed that each glottal pulse is represented by a single sinusoidal cycle, then many landmarks are appropriate: like the SEDREAMS algorithm, the each local minima of the filtered speech signal are then located, and are denoted y_{min}. An impulse signal $i[n]$, the same length of the speech signal, is constructed such that:

$$i[n] = \begin{cases} 1 & n \in y_{min} \\ 0 & n \notin y_{min} \end{cases} \tag{7.1}$$

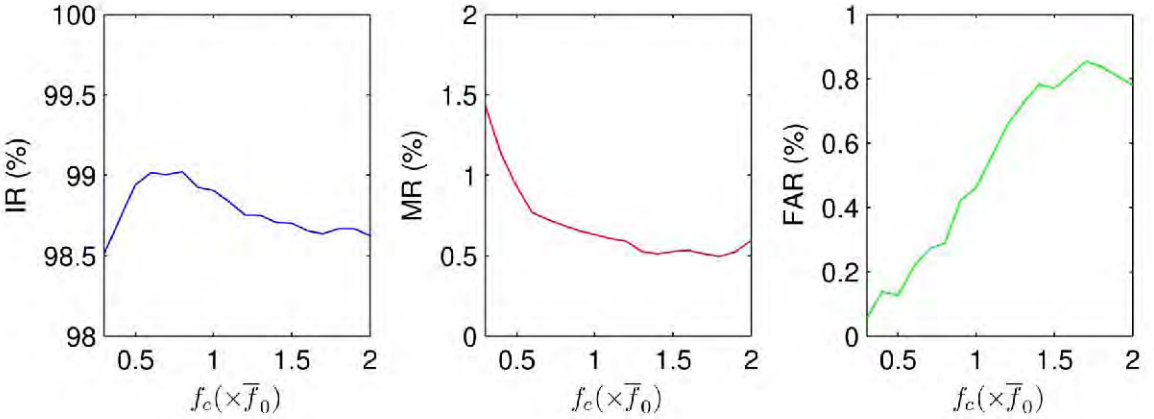


Figure 7.1: The above figure illustrates the effect of the filter cut-off frequency f_c utilised by FRESS algorithm to generate the fundamental sinusoid signal upon the various GCI estimation errors.

These peaks of this signal give an indication of the relative positions of the glottal epochs, including some spurious peaks due to improper filtering and others due to unvoiced regions of the analysed signal. In order to align this signal with the likely regions of glottal excitation, it is cross-correlated with a normalised energy signal. It is this realignment step which makes the FRESS method more robust to phase distortions as the peaks of the function $i[n]$ are explicitly re-positioned such that they are situated in regions of likely glottal closure.

The normalised energy signal $F_N[n]$ is determined such that

$$F_N[n] = \frac{F[n]}{\breve{F}[n]} \tag{7.2}$$

where $F[n]$ and $\breve{F}[n]$ are defined as in Section 4.2. This function is normalised between 0 and 1 and is maximal near the peaks of the energy contour, near the instants of vocal tract excitation.

The $F_N[n]$ and $i[n]$ signals are correlated according to the equation:

$$R[\tau] = \sum_{n=-\infty}^{\infty} F_N[n]i[n-\tau] \tag{7.3}$$

where the signals $F_N[n]$ and $i[n]$ are defined to be zero outside their lengths. The lag corresponding to the maximum of the correlation function R is the amount of delay necessary that the signal i maximally aligns with the signal $F_N[n]$.

Once the signal $i[n]$ has been realigned, each peak of the signal is used as an anchor point about which to search for the glottal epoch in the LPC residual signal, much like the SEDREAMS algorithm. However, due to the noisy character of the LPC residual signal, the largest value within this area does not necessarily correspond to the actual glottal epoch. Thus, the residual signal is smoothed using a low pass zero phase FIR filter, the cut-off frequency of which has been set to 4kHz. Once smoothed, the 5 most prominent candidates are selected from the signal. Five candidates are chosen because experimentally the number was found to offer a compromise between the accuracy of the results and computational load of the subsequent dynamic programming algorithm.

7.1.2 Dynamic Programming

Once the set of candidate glottal epochs is given, the sequence of most likely glottal epochs is determined using the N-best dynamic programming approach, previously utilised for the same purpose by the DYPSA and YAGA algorithms. The algorithm attempts to minimise a set of costs which are attributed to the candidates themselves and the transitions between candidates. For the FRESS algorithm, the cost attributed to the r^{th} candidate at sample number c_r is given by:

- Waveform Similarity Cost, C_ρ.
- Pitch Deviation Cost, C_P.
- Normalised Residual Amplitude Cost, C_r.

The first two costs are calculated as with the DYPSA algorithm, described in Section 4.2. The third cost, the Normalised Residual Amplitude cost, is attributed each glottal closing instant candidate by the following equation:

$$C_r = -0.5\hat{r}[c_r] \tag{7.4}$$

where $\hat{r}$ is the amplitude-normalised LPC residual signal. This cost has the effect of penalising glottal epoch candidates which represent low amplitude samples of the LPC residual signal, as from the theory of deconvolutive signal, they should be high amplitude events.

The costs are weighted before they are input into the dynamic programming algorithm. The weighting factors were optimised by brute force upon 15 sentences from the CMU-ARCTIC database not used for testing, and set to: $[0.4, 0.5, 1]$ for C_ρ, C_P, and C_r, respectively.

7.2 Validation/Testing

An experiment was undertaken comparing the FRESS algorithm against six other state of the art glottal epoch estimation algorithms: the Group Delay method[2] (Smits

[2] The implementation utilised is available at (Fernandez, Retrieved November 14th, 2010)

and Yegnanarayana, 1995), the find_pmarks algorithm[3] (Goncharoff and Gries, 1998), the DYPSA method[4] (Naylor *et al.*, 2007), the ZFR method[5] (Murty and Yegnanarayana, 2008), the SEDREAMS method[6] (Drugman and Dutoit, 2009), and the YAGA method[7] (Thomas *et al.*, 2011).

Two additional methods are also included, a new SEDREAMS method (Drugman, Retrieved October 8th, 2011) and a modified YAGA method, which are more robust to phase-distorted speech, giving a total of eight algorithms against which the FRESS algorithm is compared. The modified YAGA algorithm is as follows. As mentioned in Section 4.2, the algorithm weights each glottal closing instant candidate according to the closed-phase energy between successive cycles. As phase distortion can eliminate the closed phase of the cycle, this additional cost can mislead the subsequent dynamic programming algorithm by penalising actual glottal closing instants inappropriately. A modified YAGA algorithm is then proposed which does not impose any assumption upon the shape of the glottal source signal by setting the weighting factor upon this cost to zero. In the comparative experiment, this algorithm is denoted YAGA*. Similarly, the new SEDREAMS algorithm is denoted SEDREAMS*.

Like the experiments in the previous chapter, the algorithms were tested upon synthetically-generated speech segments and real speech utterances spoken by three voices (two male, *bdl* and *jmk*, and one female, *slt*) taken from the CMU-ARCTIC

[3] The algorithm was utilised was an implementation available from (Goncharoff, Retrieved July 6th, 2011), though slight modification was necessary to operate upon signal with sampling rates above 8kHz.

[4] Algorithm available at (Brookes, Retrieved January 22nd, 2009)

[5] Own implementation.

[6] Own implementation.

[7] Original author's implementation, not publicly available.

database (Kominek and Black, 2003). Also as with the experiment of Chapter 6, in order to simulate nonlinear recording conditions, the transfer function estimated from a nonlinear recording system was applied to the test signals.

For the synthetic speech experiment, 1000 utterances were generated, using randomly-chosen vocal-tract filters (from those listed in Appendix B) and glottal source parameters (R_d, f_0 and signal-to-noise ratio). All parameters were held constant for the duration of each synthetic utterance, which was set to be 100 pulses in length. Because the glottal closing instants of the synthetic speech are known, imprecisions and inaccuracies of the results are determined only by the approaches themselves and any subsequent phase-distorting filter operation.

For the real speech signals, these instants are determined from the synchronously recorded EGG signal. For the real speech signals, all pairs of EGG and speech signals (of which there were 3300) were down-sampled to 16kHz and time-aligned as in the previous chapter. In both phase conditions, the simultaneously recorded EGG signals served as the benchmark against which to compare the estimations of the glottal epochs. The reference glottal epochs n_c were identified in the EGG signal using the SIGMA algorithm (Thomas and Naylor, 2009).

The description of the performance of each algorithm follows the convention established in (Naylor *et al.*, 2007), which uses a three way classification scheme for each reference larynx cycle. The c^{th} larynx cycle spans a range of samples to either side of the c^{th} glottal epoch n_c and is given by:

$$\frac{1}{2}(n_c + n_{c-1}) < n \leq \frac{1}{2}(n_{c+1} + n_c) \qquad (7.5)$$

where n_{c-1} and n_{c+1} are left and right neighbouring glottal epochs respectively. In the case where one of these adjacent points is unavailable, *e.g.* at the beginning or end of a voiced segment, the length of the search interval is taken to be the same on each side of the epoch in question.

Each reference larynx cycle is classified according to one of three categories, depending on the detection of a glottal closing instant in its larynx cycle: identified, missed and false alarm. The total performance of each algorithm can then be categorised by three percentages:

Identification Rate (IR) The percentage of larynx cycles for which exactly one GCI is detected;

Miss Rate (MR) The percentage of larynx cycles for which no GCI is detected; and,

False Alarm Rate (FAR) The percentage of larynx cycles for which more than one GCI is detected.

Additionally to these classifications, an indication of the identification accuracy (ζ) of the identified glottal epochs is given by the distribution of the time intervals between estimated and actual glottal epochs, characterised by mean μ_ζ and standard deviation σ_ζ. Figure 7.2 illustrates this scheme.

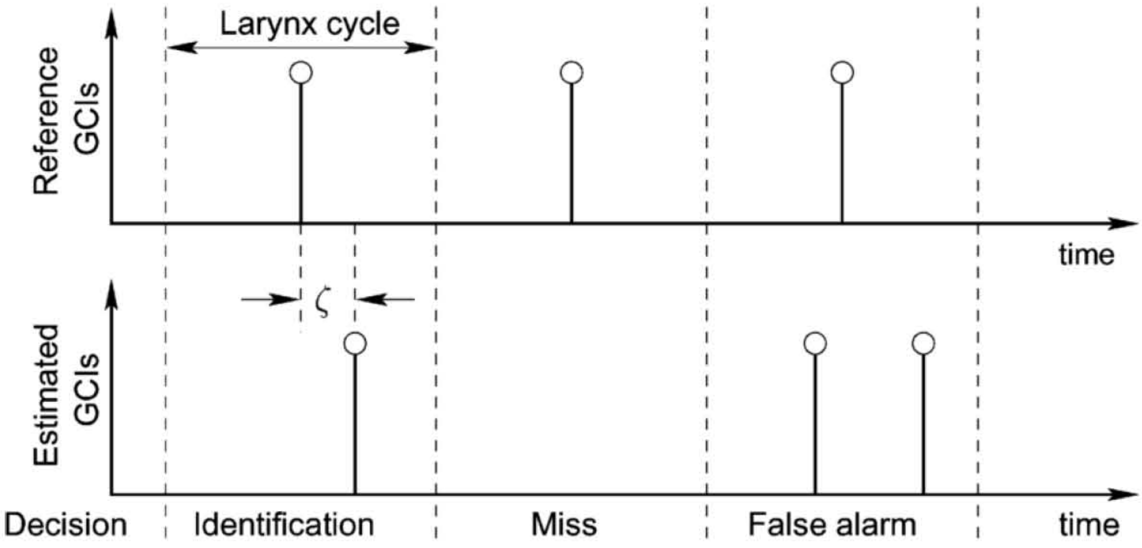


Figure 7.2: The figure above shows the classification of estimated glottal closing instants following testing, taken from (Naylor *et al.*, 2007).

7.3 Results and Discussion

7.3.1 Synthetic Speech

The results of the synthetic speech experiments under phase linear and nonlinear phase recording conditions are given in Tables 7.1 and 7.2, respectively.

As can be seen in Table 7.1, all of the methods except for the find_pmarks method, achieve extremely high identification rates of greater than 98.5% in the linear phase synthetic speech case. Indeed, the SEDREAMS* technique successively identifies all of the glottal closing instants. Such high rates are unsurprising as the model upon which these GCI estimation methods are based are followed exactly in these cases. For the find_pmarks approach, this experiments show that the use of the peaks of the energy contour of the speech signal alone are inappropriate for accurate glottal

Table 7.1: The table below gives the performance of GCI estimation methods upon synthetic speech under linear-phase recording conditions.

Method	IR(%)	MR(%)	FAR(%)	σ_ζ(ms)	μ_ζ(ms)
Group Delay	98.65	0.00	1.35	0.68	-0.33
find_pmarks	83.13	0.01	16.86	0.11	0.09
DYPSA	99.02	0.05	0.93	0.71	-0.02
ZFR	99.70	0.00	0.30	0.44	-0.35
SEDREAMS	99.78	0.12	0.11	0.73	0.46
SEDREAMS*	100.00	0.00	0.00	0.60	-0.07
YAGA	99.82	0.00	0.18	0.11	0.10
YAGA*	99.56	0.00	0.43	0.36	0.09
FRESS	99.81	0.16	0.03	0.57	0.10

closing instant estimation for a wide range of source-filter based synthetic speech.

The table of results for the nonlinear synthetic speech experiment, Table 7.2, similarly high identification rates are seen. However, the precision and accuracy of the identified glottal closing instants of many of the approaches suffer. The ZFR, original SEDREAMS and YAGA algorithms in particular see large increases in the standard deviations of the identification accuracy ζ. This is expected as these methods were reviewed in Chapter 4 as being particularly sensitive to the shape of the speech signal. Curiously, the find_pmarks approach actually improves in this scenario, probably due to the more precise alignment in the nonlinear phase case of the glottal instant and local energy maximum of the speech signal. Conversely, the FRESS algorithm sees no large deviation from its performance in the nonlinear phase case from the linear phase scenario.

Table 7.2: The table below gives the performance of GCI estimation methods upon synthetic speech under nonlinear-phase recording conditions.

Method	IR(%)	MR(%)	FAR(%)	σ_ζ(ms)	μ_ζ(ms)
Group Delay	98.81	0.00	1.19	0.64	−0.22
find_pmarks	92.60	0.01	7.39	0.21	0.09
DYPSA	98.28	0.04	1.68	0.50	0.02
ZFR	98.56	0.37	1.07	3.52	−0.81
SEDREAMS	99.94	0.03	0.03	1.38	−1.31
SEDREAMS*	99.99	0.01	0.00	0.92	−0.14
YAGA	98.78	0.34	0.88	1.72	0.05
YAGA*	99.30	0.00	0.69	0.50	0.16
FRESS	99.91	0.07	0.02	0.39	0.10

7.3.2 Real Speech

The results of the classifications of the glottal epochs for each method are given in Tables 7.3 and 7.4, for the linear phase and nonlinear phase conditions, respectively. Additionally, the distributions of the identification accuracy ζ are given for each algorithm in Figures 7.3 to 7.11.

In the case of speech recorded using nonlinear phase equipment, as anticipated the phase disturbance imparts significant performance degradation for the ZFR, original SEDREAMS and YAGA algorithms. This is clearly noticed viewing the differences in ζ distributions, Figures 7.6, 7.7 and 7.9, respectively. In particular with the speaker *jmk*, the SEDREAMS algorithm suffers a reduction of almost 20% in the identification rate. This results from the demarcated LPC residual search regions capturing impulsive events other than the peak corresponding to glottal closure. This incorrect

Table 7.3: The table below gives the performance of GCI estimation methods upon real speech under linear-phase recording conditions.

Speaker	Method	IR(%)	MR(%)	FAR(%)	σ_ς(ms)	μ_ς(ms)
bdl	Group Delay	90.97	5.41	3.62	0.66	0.28
	find_pmarks	92.72	5.14	2.14	1.14	−0.14
	DYPSA	90.71	6.21	3.08	0.63	0.02
	ZFR	89.81	3.50	6.70	0.42	−0.66
	SEDREAMS	93.81	3.73	2.46	0.45	0.12
	SEDREAMS*	93.69	3.71	2.60	0.47	0.06
	YAGA	93.47	3.21	3.31	0.41	0.08
	YAGA*	93.12	3.36	3.52	0.45	0.10
	FRESS	94.70	3.47	1.83	0.42	0.13
jmk	Group Delay	81.89	6.76	11.35	1.06	0.28
	find_pmarks	93.06	5.62	1.32	1.53	0.27
	DYPSA	92.54	6.01	1.46	0.69	0.15
	ZFR	94.56	4.90	0.53	0.70	−1.31
	SEDREAMS	94.79	4.97	0.24	0.54	0.12
	SEDREAMS*	94.80	4.96	0.24	0.62	0.04
	YAGA	94.25	4.85	0.91	0.53	0.14
	YAGA*	94.38	4.91	0.71	0.54	0.19
	FRESS	94.62	5.04	0.33	0.47	0.20
slt	Group Delay	95.96	2.19	1.84	0.52	0.10
	find_pmarks	98.02	0.50	1.48	0.52	0.06
	DYPSA	97.06	1.75	1.19	0.46	0.06
	ZFR	99.18	0.38	0.44	0.24	−0.57
	SEDREAMS	99.11	0.28	0.61	0.32	0.13
	SEDREAMS*	99.09	0.30	0.61	0.33	0.12
	YAGA	98.76	0.37	0.86	0.28	0.07
	YAGA*	98.63	0.38	0.99	0.29	0.08
	FRESS	99.00	0.41	0.59	0.25	0.07

decision has the consequence that adjacent glottal epochs are inconsistently misidentified, thus increasing the missed and false alarm rates. In the scenario where the search region is capable of uniquely identifying each laryngeal cycle, the identifica-

Table 7.4: The table below gives the performance of GCI estimation methods upon real speech under nonlinear-phase recording conditions.

Speaker	Method	IR(%)	MR(%)	FAR(%)	σ_ζ(ms)	μ_ζ(ms)
bdl	Group Delay	91.04	4.97	3.98	0.63	0.50
	find_pmarks	92.92	5.09	1.99	1.09	−0.09
	DYPSA	90.59	6.15	3.25	0.63	0.09
	ZFR	82.44	8.28	9.28	3.37	−0.90
	SEDREAMS	91.09	5.55	3.37	1.31	−2.05
	SEDREAMS*	94.18	3.53	2.28	0.48	0.04
	YAGA	84.79	3.68	11.53	1.15	−0.32
	YAGA*	92.84	3.19	3.97	0.45	0.03
	FRESS	94.94	3.31	1.75	0.40	0.06
jmk	Group Delay	84.64	6.42	8.94	1.01	0.46
	find_pmarks	93.03	5.68	1.30	1.54	0.18
	DYPSA	91.92	6.21	1.86	0.81	0.24
	ZFR	89.49	7.02	3.49	2.85	2.13
	SEDREAMS	75.91	14.03	10.06	2.73	−1.24
	SEDREAMS*	94.02	5.31	0.67	1.16	−0.28
	YAGA	87.45	7.26	5.29	2.74	0.61
	YAGA*	94.21	4.93	0.86	0.62	0.07
	FRESS	94.59	5.05	0.36	0.45	0.09
slt	Group Delay	96.41	1.88	1.71	0.50	0.35
	find_pmarks	98.08	0.50	1.42	0.49	0.01
	DYPSA	97.09	1.70	1.21	0.47	0.12
	ZFR	96.95	1.71	1.33	0.77	−2.08
	SEDREAMS	98.84	0.45	0.71	0.71	−0.48
	SEDREAMS*	99.12	0.28	0.60	0.31	0.07
	YAGA	97.16	0.77	2.07	0.98	−0.63
	YAGA*	98.57	0.34	1.09	0.29	−0.01
	FRESS	99.01	0.36	0.63	0.22	0.02

tion, missed and false alarm rates will be similar to the linear phase case, but the identification accuracy ζ will suffer because the wrong impulsive events are selected, as can be seen with female speaker *slt*.

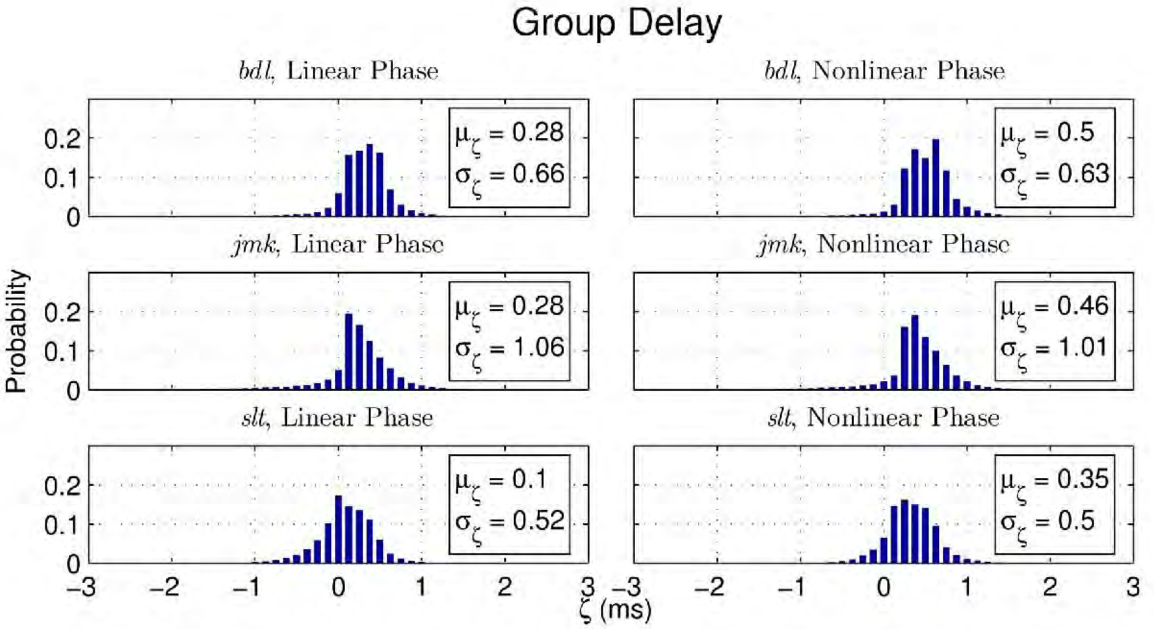


Figure 7.3: The figures above show for each speaker the ζ distributions for the Group Delay algorithm.

Similar multi-modal distributions are observed for the other phase-sensitive algorithms. Because the ZFR algorithm uses a fixed landmark of the fundamental harmonic sinusoid signal to identify the glottal epochs, its accuracy is directly influenced by phase disturbances that may have been applied to the speech signal. For that reason, as the speaker changes pitch, a consistent performance of the algorithm cannot be expected. The original YAGA algorithm also experiences difficulties with phase disturbances. As explained above, the glottal epoch candidates are weighted by the algorithm according to the presence of an adjacent closed phase. This cost can serve to penalise otherwise accurate epoch candidates under nonlinear phase conditions.

On the other hand, some of the algorithms are relatively unaffected by this dis-

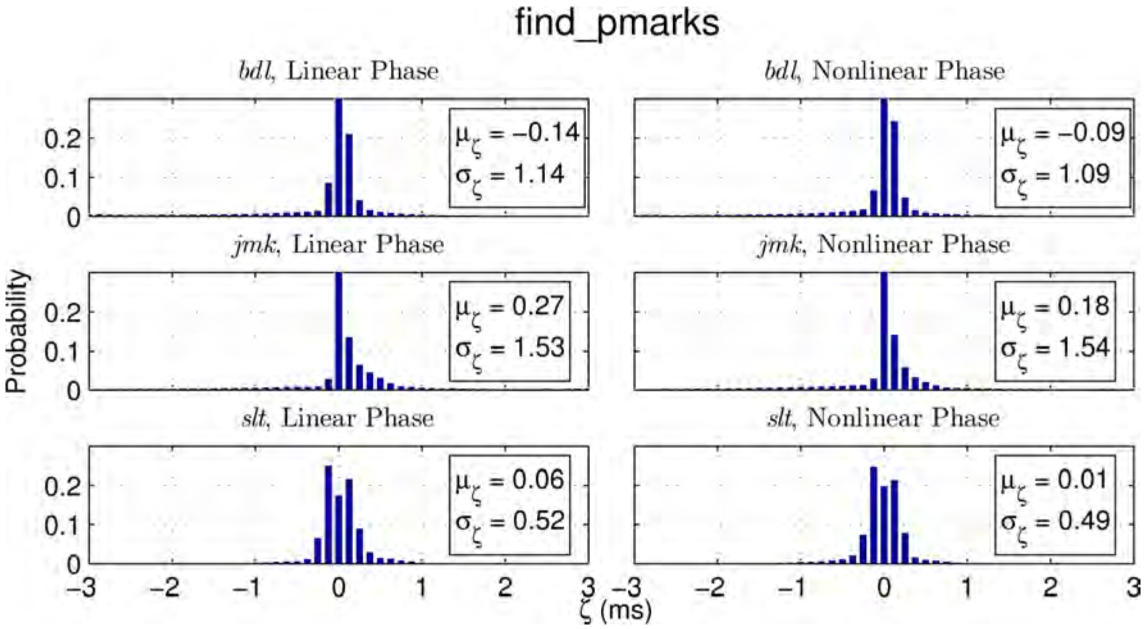


Figure 7.4: The figures above show for each speaker the ζ distributions for the find_pmarks algorithm.

tortion. As can been from the shape of the ζ distributions in Figures 7.3, 7.4, 7.5, 7.8, 7.10, and 7.11, the Group Delay, find_pmarks, DYPSA, SEDREAMS*, YAGA* and FRESS algorithms are reasonably static, though the bias of the distribution may shift. This offset is due to the effect of the phase distortion upon the signal from which the algorithms determines the glottal closing instant (Section 5.2 discusses this in more detail). The modified SEDREAMS* algorithm is somewhat capable of correcting the alignment introduced by phase distortion by dynamically determining the search regions, but not with equal success across all speakers (*e.g.* speaker *jmk*). Similarly, excluding the closed-phase energy weighting of the YAGA algorithm serves to remove assumptions regarding the signal shape.

The Group Delay algorithm consistently under-performs in comparison with the all

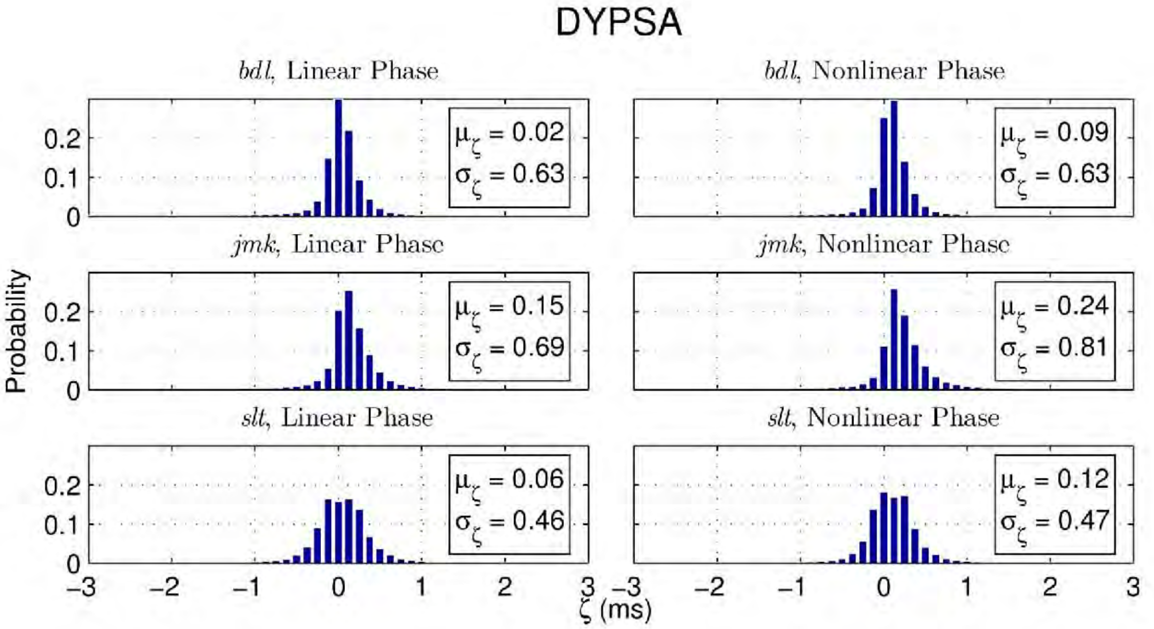


Figure 7.5: The figures above show for each speaker the ζ distributions for the DYPSA algorithm.

other approaches, often exhibiting the highest miss and false alarm rates in addition to the lowest hit rate. As was reviewed in Section 4.2, the group delay function operates best upon ideal impulse train type signals, and thus is not very robust to the noisy conditions common with LPC residual signals. Generally the energy-weighted group delay function utilised by the DYPSA algorithm, combined with dynamic programming and phase slope projection technique, improves the accuracy of the basic group delay approach, in addition to lowering miss and false alarm rates, particulary for the speaker *jmk*.

The find_pmarks algorithm identifies many glottal cycles successfully, but is imprecise regarding the location of the glottal closing instant, with the exception of the female voice. Its operation does not degrade much with phase distortion, which

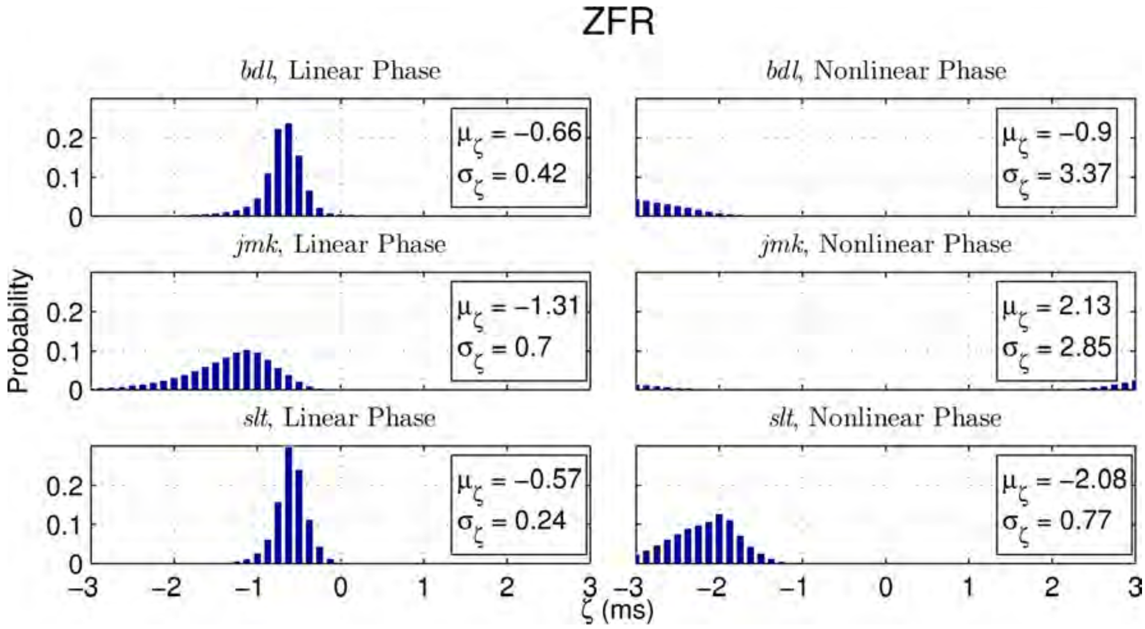


Figure 7.6: The figures above show for each speaker the ζ distributions for the ZFR algorithm.

is essentially the same under both conditions. This is unsurprising as the relative positions of the glottal epochs is first determined from the energy contour of the speech signal. However, its accuracy is improved under nonlinear phase conditions, probably because the large amplitude samples of the analysed speech signals which the algorithm attributed to glottal closure more closely aligns with the epochs in that scenario.

The overall results of the phase linear experiment corroborate those reported in (Drugman, 2011; Cabral *et al.*, 2011), in that the SEDREAMS and YAGA algorithms perform better in comparison with other existing approaches, with both methods exhibiting high identification rates. To this group the FRESS method can be added as all three techniques usually within 0.5% of each other in the phase linear case.

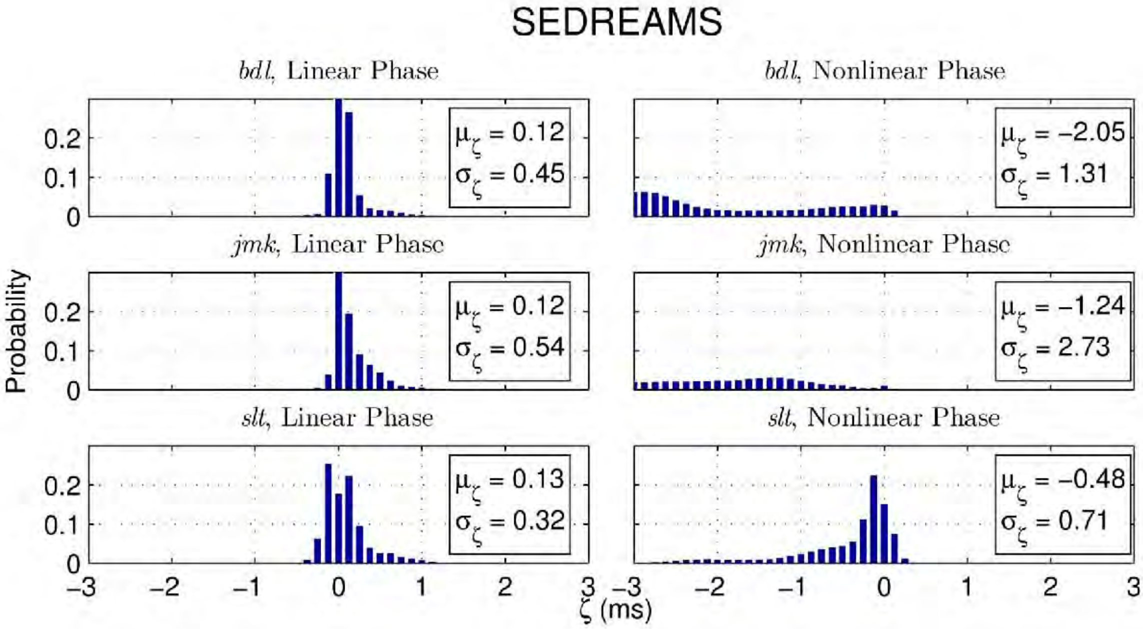


Figure 7.7: The figures above show for each speaker the ζ distributions for the SEDREAMS algorithm.

In fact, referring to Figures 7.8, 7.10 and 7.11, the distributions are broadly very similar. Of these approaches, no one method is clearly the most appropriate in terms of accuracy as this result is seemingly speaker-dependent. However, for all voices and phase conditions, the SEDREAMS* method exhibits a mean accuracy within 0.28ms, the YAGA* method a mean accuracy within 0.19ms, and the FRESS method a mean accuracy of less than or equal to 0.2ms.

The FRESS algorithm offers the highest precision compared with other approaches under both phase conditions, achieving the narrowest ζ distributions when compared with all other glottal epoch estimation algorithms (excepting the phase-linear *bdl* results by the YAGA algorithm). This is probably a result of choosing a number of glottal epochs from the LPC residual signal; thus, the most consistent candidates

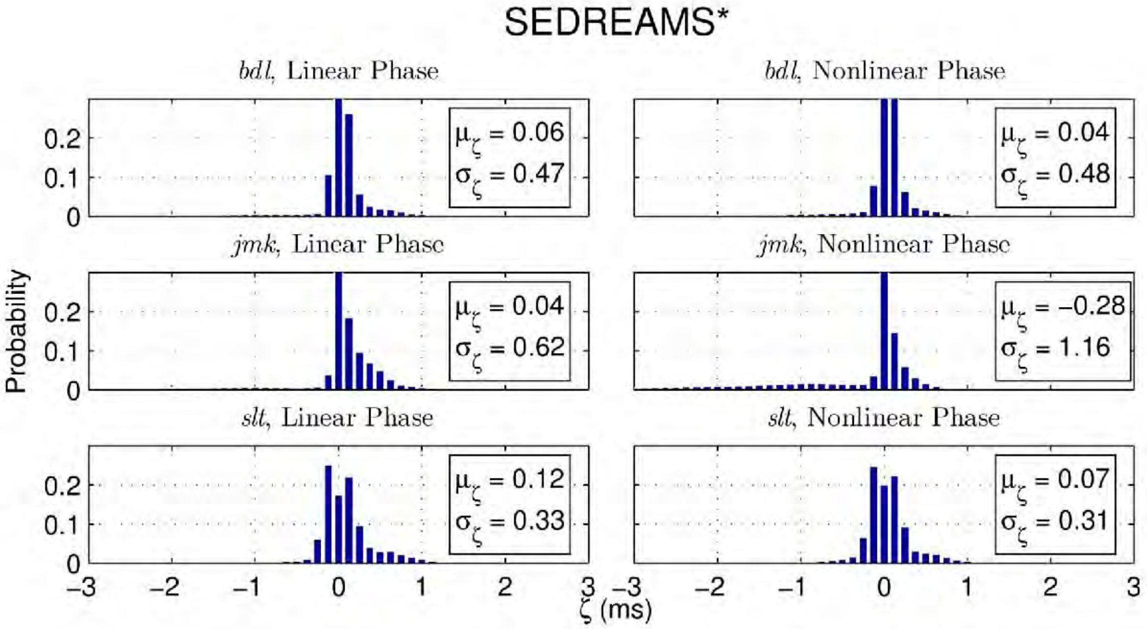


Figure 7.8: The figures above show for each speaker the ζ distributions for the SEDREAMS* algorithm.

can be selected according to the heuristics imposed by the dynamic programming algorithm.

Finally, at least some errors of the experiment can be attributed to the SIGMA algorithm utilised to establish the reference glottal epochs from the EGG signal. The performance of the algorithm was found to be dependent upon the speaker. Additionally, as with the previous chapter, nonlinear phase recording conditions were simulated using measurements taken from a professional recording studio set-up. In the case of a different systems, different results can be expected, though the FRESS algorithm has been designed to be robust to these conditions.

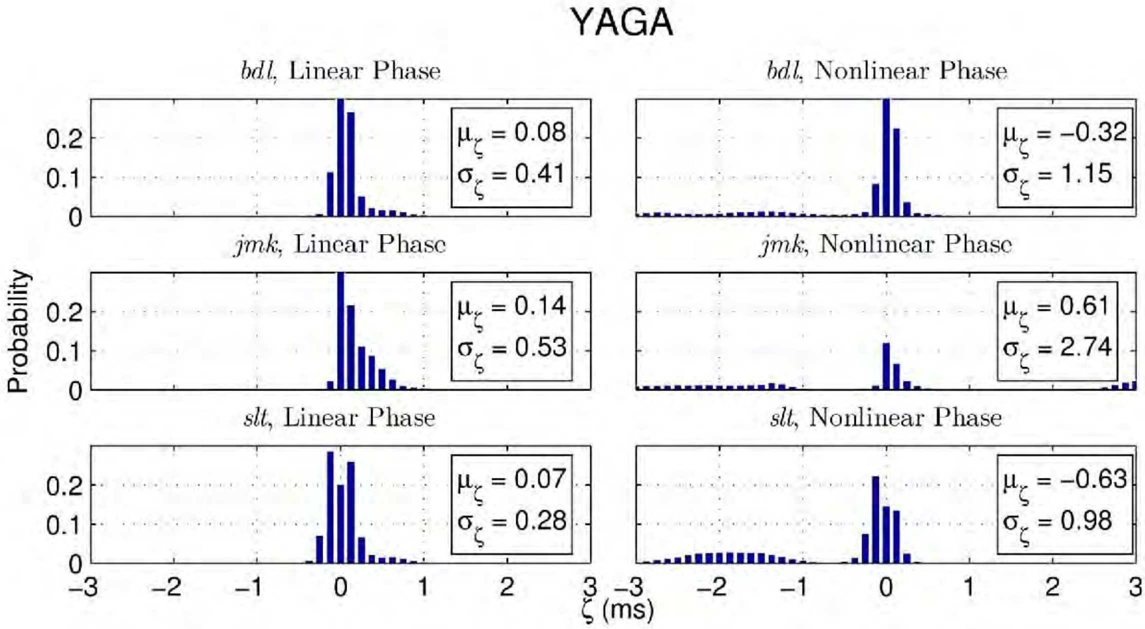


Figure 7.9: The figures above show for each speaker the ζ distributions for the YAGA algorithm.

7.4 Conclusions

This chapter introduces the second major contribution of this work outlined in introductory chapter, a glottal epoch estimation approach specifically designed to be robust to phase distortions, the FRESS method. The approach draws from existing algorithms in that a fundamental frequency sinusoid obtained by low pass filtering the speech signal is used to locate the regions of glottal closure. Landmarks from this signal are used as an indication of the relative locations of the glottal epochs, and a correlation operation is used to align the points with the peaks of the normalised energy contour signal. A dynamic programming algorithm is then used to select the most likely sequence of glottal epochs, based on continuities of pitch, waveform similarity and the amplitude of the LPC residual signal.

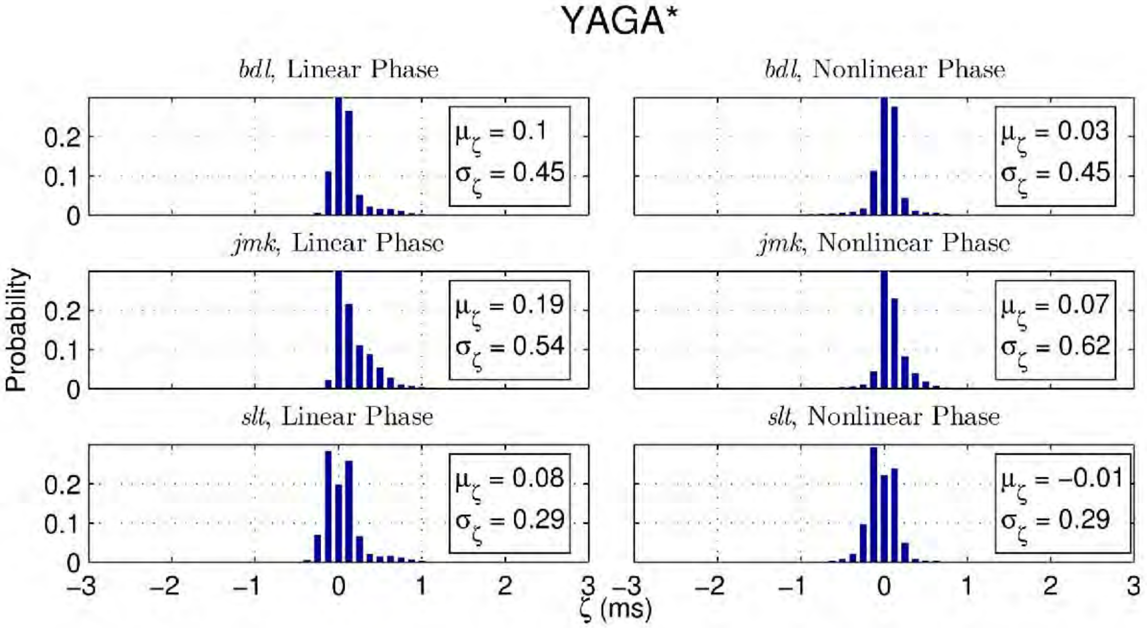


Figure 7.10: The figures above show for each speaker the ζ distributions for the modified YAGA algorithm.

A comparative experiment was then undertaken where the FRESS method was compared with eight other state of the art algorithms. Simulated phase distortion was shown to adversely affect some of the algorithms, while others were relatively robust. While it has comparable identification rate and accuracy to other methods, the FRESS algorithm is the most precise, probably due to the fact that a number of candidates are extracted from the LPC residual signal in the region of glottal closure.

Following the results of the comparative experiment, making a qualified judgement regarding the presented algorithms is dependent upon the desirable attributes the algorithm must have. For example, it is feasible that for certain speech applications that the identification rate be very high and the accuracy of the method is relevant. Under phase linear conditions, the YAGA and SEDREAMS* algorithms may offer increased

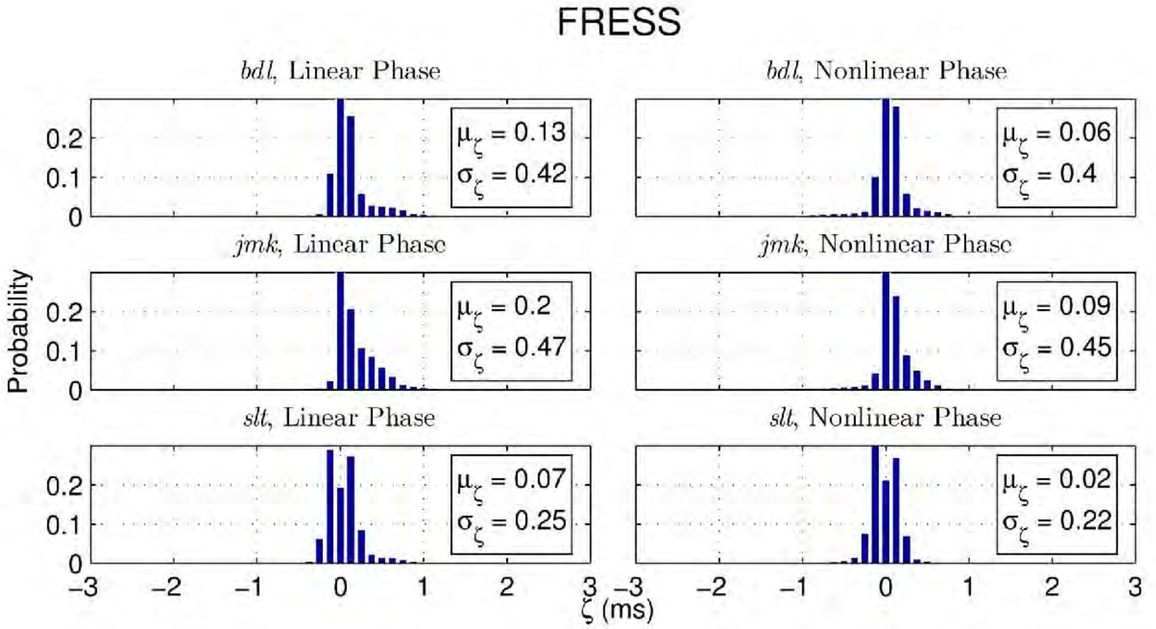


Figure 7.11: The figures above show for each speaker the ζ distributions for the FRESS algorithm.

accuracy yet with slightly inferior identification rates than the FRESS method. Due to the lack of a dynamic programming stage, the SEDREAMS algorithm is the most appropriate for real-time implementations. Finally, it is concluded that the FRESS algorithm represents the best choice of glottal epoch estimation algorithms in all phase conditions if precise estimation is desired.

Chapter 8

Voiced Speech Analysis/Synthesis

In this chapter, synthetic speech generated by a power-spectrum-based parameterisation technique closely related to the PowRd method is compared against an existing time-domain-based, voiced-speech parameterisation approach in a listening test. The experiment and system constitute Contribution 3 of this study. The experiment asks listeners to compare the synthetic speech signal of both approaches and opine their general preference. Linear phase equipment was utilised for the recordings of the analysis speech signals, which were also tested following the application of phase distortion. It was found that the power spectrum approach was generally preferred, particularly in the phase-distorted case for which the method was specifically designed. Essentially, this chapter performs two functions: it presents an extrinsic evaluation of the power spectrum approach to voice-source parameterisation, while simultaneously demonstrating an application of the tools described in this study.

In order to distinguish it from the PowRd technique, this alternative method is referred to as the PowARXLF method, Power-Spectrum-based ARX-LF parameterisation of voiced speech. This alternative technique is closely related to the PowRd technique in that it also parameterises the voiced speech signal by operating upon the power spectrum and using the Relative Itakura-Saito error function, yet differs in that the voice-source signal is assumed to be approximated by the LF model, and not the simpler transformed LF model. The subtle differences in its formulation of the algorithm are described below.

Before describing the experiment, an overlap-add technique similar to existing synthesis methods is detailed, which both the time and power spectrum methods utilise to generate the speech segments. Additionally, as the parameters were estimated using disjoint frames, parameter smoothing stages are described which are useful to impose certain speech heuristics upon the synthetic signal.

8.1 Analysis/Synthesis

This section elaborates the differences between PowARXLF and PowRd techniques, in addition to describing the parameter smoothing stages, and synthesis procedures necessary for the experiment. The time-domain ARX-LF parameterisation method used for experiment, based upon the method proposed in (Lu, 2002) and extended by (del Pozo, 2008) and (Pérez and Bonafonte, 2009), is also described.

8.1.1 The PowARXLF Method

The experiment described in this chapter utilises the PowARXLF method, rather than the PowRd method described in Chapter 6. This alternative voice-source parameterisation method differs only in one respect: the assumptions placed upon the voice-source signal. While the PowRd method assumes that the voice-source signal can be parameterised using the transformed LF model of a single parameter, the PowARXLF method assumes that the voice source is parameterised by the original LF model, which as reviewed in Section 2.3.3.1. This lends the PowARXLF method greater flexibility regarding the characteristics of the source signal, yet Algorithm 1 is changed only in that the codebook of initial parameters is larger.

Codebook Assembly The PowARXLF codebook is constructed by sampling the LF model shape parameters $\{O_q,\ \alpha_m,\ Q_a\}$ over their entire ranges. The shape parameters are the previously defined open quotient $O_q = \frac{T_e}{T_0}$, an asymmetry coefficient $\alpha_m = \frac{T_p}{T_e}$ which indicates the skewness of the glottal pulse, in addition to the return phase coefficient $Q_a = \frac{T_a}{(1-O_q)T_0}$. The extrema of each parameters are:

$$O_q = \{0.3,\ 0.95\} \tag{8.1}$$

$$\alpha_m = \{0.65,\ 0.95\} \tag{8.2}$$

$$Q_a = \{0,\ 0.95\} \tag{8.3}$$

Parameter samples are taken at each point separated by a step-size of 0.01. This leads to a codebook with almost 200,000 entries. In order to reduce its size, any parameter set within it which generates an LF model pulse which is deemed too similar to any other pulse within the codebook is removed.

The normalised correlation coefficient was proposed for determining similarity of codebook entries in (Vincent *et al.*, 2005) and is employed again here. The normalised correlation coefficient ρ calculates the similarity between two signals x and y of length

N using the following equation:

$$\rho = \frac{\sum_{n=0}^{N-1} x[n]y[n]}{\sum_{n=0}^{N-1} x[n]^2 \sum_{n=0}^{N-1} y[n]^2} \tag{8.4}$$

When signals x and y are identical in shape and position, ρ takes the value 1 and is less than 1 otherwise.

For the large codebook, the similarity coefficients are calculated between all pulses. If the normalised correlation coefficient between any two pulses is larger than a threshold value $\hat{\rho}$, it is discarded. The value of $\hat{\rho}$ is chosen to give a compromise between the two opposing considerations mentioned previously in Section 6.2.3: the computation efficiency and adequate subset coverage. In this work, like (Vincent *et al.*, 2005), $\hat{\rho} = 0.99$, leading to a final codebook size of 630 entries.

Full Band Analysis Both the PowRd and PowARXLF methods bandlimit the analysis frame in order to avoid high frequency noise components of the speech signal. Thus, in order to obtain a full band representation of the vocal tract, following the bandlimited analysis, the PowARXLF method estimates the full-band all-pole envelope of the vocal tract spectrum using voice-source parameters obtained during the initial bandlimited analysis. The estimated filter order is set to $p = \lfloor \frac{f_s}{1000} + 0.5 \rfloor + 2$. Note that fixing the filter order in this case does not affect the voice-source parameterisation procedure as its contributions are removed from the signal before analysis.

8.1.2 Parameter Smoothing

The parameters determined by the PowARXLF method may not necessarily represent the best parameters if a requirement is also that the parameters of the speech segment change smoothly. For this reason, two smoothing operations are performed upon the results of the analysis: firstly, a dynamic programming algorithm chooses the parameters which represent the smoothest changes based upon certain criteria, followed by an simple averaging operation.

Dynamic Programming Following the PowARXLF initial brute force initialisation of each signal frame, the usual procedure for each frame is to choose the LF model parameter configuration which gives the lowest Itakura-Saito error and refine using an optimisation algorithm. However, in this work, as smoothly changing parameters are also desirable, an additional transition cost is added to this initial error such that quickly changing parameters are penalised.

This is implemented using a dynamic programming method. Firstly, for each voiced speech segment, the brute force initialisation is performed. Each analysis frame then has associated with it, for each LF model parameter configuration within the codebook:

- an all-pole filter representing the vocal tract, and
- an Itakura-Saito error measure, quantifying the goodness of fit.

A cost matrix is then created, populated by estimated Itakura-Saito errors. A dynamic programming method is then implemented upon this matrix, which imposes an additional discontinuity penalty between adjacent frames dependent upon the distance between parameter configurations. These penalties are:

- The distance between adjacent LF model pulse signals λ_ρ, calculated as:

$$\lambda_\rho = 1 - \rho \tag{8.5}$$

As explained above, if the pulses are exactly identical, $\rho = 1$ and thus the cost is λ_ρ will be 0.

- The distance between the estimated vocal-tract filters λ_k calculated as:

$$\lambda_k = \sqrt{\sum_{n=1}^{N} (k_n^i - k_{n+1}^i)} \tag{8.6}$$

where k_n^i and k_n^{i+1} are n^{th} reflection coefficients of the adjacent estimated vocal-tract filters (Wakita, 1973). Reflection coefficients relate to the dimensions of the simplified acoustic tube model of the vocal tract, which should change shape slowly between analysis points. They have been previously used for purpose of parameter smoothing in (Lu, 2002).

The dynamic programming algorithm then gives the most likely sequence of initial voice-source and vocal-tract-filter parameters. The vocal-tract-filter parameters are then refined using the simplex algorithm (Nelder and Mead, 1965).

Smoothing Filters Once the refined voice-source parameters are obtained, they are further smoothed before re-synthesis by using a filtering operation, much like the ones performed in (Lu, 2002; del Pozo, 2008). A three-point moving average filter is used for this purpose, and are applied to the line spectral frequency representations of the vocal tract and the R parameters of the LF model pulses (while retaining the end-points). The line spectral frequencies (Itakura, 1975) of the vocal tract all-pole filter are used because of their desirable interpolation properties compared with other filter representations (Paliwal and Kleijn, 1995).

8.1.3 Time-Domain ARX-LF Parameterisation

The PowARXLF method is compared with a two stage time-domain ARX-LF parameterisation method, based upon the method proposed in (Lu, 2002). This method has found application in synthesis of singing voice with quality control (Lu, 2002), analysis/synthesis of vowel segments (Pérez and Bonafonte, 2009) and speaker conversion (del Pozo, 2008; Pérez and Bonafonte, 2011).

The first stage of the algorithm utilised the convex optimisation approach described in Section 3.2.2 to jointly estimate the all-pole vocal-tract filter and parameters of the KLGLOTT88 model by minimising the squared error of the residual signal in the time domain using of KLGLOTT88 parameters. Adaptive pre-emphasis was used to increase robustness (del Pozo, 2008; Pérez and Bonafonte, 2009). A dynamic programming algorithm, using the reflection coefficients of the vocal-tract filter and the parameters of the KLGLOTT88 model as described in (Lu, 2002), is applied to find the lowest overall error. Like the method above, the vocal-tract filter coefficients are then smoothed.

Once the vocal tract is estimated, it is used to inverse filter the analysis speech signal and estimate the voice source. Each source pulse is then re-parameterised using the LF model in an approach similar to (Pérez and Bonafonte, 2009): the initial LF model parameters used to begin the refinement are mapped from the previously optimised KLGLOTT88 parameters (which is more robust than direct estimation) and a constrained optimisation algorithm is used to fine the optimal fit. Finally, similar to the method described above, the R parameters of the final sequence of LF parameters are also smoothed.

8.1.4 Overlap-Add Synthesis

Like analysis, synthesis of speech from ARX-LF parameters is performed in a pitch synchronous scheme, similar to other existing methods. Each speech pulse is generated and overlap-added to produce the speech segment. The process is outlined here.

The instants of synthesis correspond to the analysis instants, though the pitch and duration may be easily changed using a simple mapping scheme (Stylianou, 1996). For each synthesis point, the LF model pulse is generated and placed in frames $2T_0+1$ in length such that the glottal closure instant coincides with the center of the frame. The pulse is convolved with the associated vocal-tract filter and windowed using a Hann function. The frame is then overlap-added to the output signal, centred above the synthesis instant. A diagram showing the scheme is given in Figure 8.1.

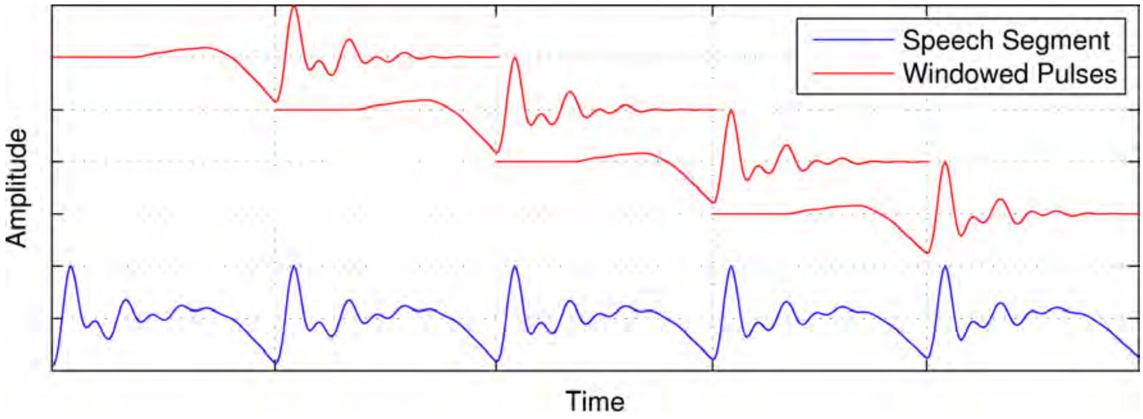


Figure 8.1: The above figure illustrates overlap-add synthesis from ARX-LF parameters: LF model pulses are convolved with the estimated vocal-tract filter, windowed using a Hann function, and then added using overlapping windows into the synthetic voiced speech segment.

8.2 Preference Test

Three experiments were performed to test the PowARXLF method against the time-domain-based approach described above. The purpose of these tests is to investigate whether a power-spectrum-based approach of voiced-speech analysis is preferred over a time-domain-based method when the goal is to re-synthesise voiced speech. In undertaking these experiments, this work aims to discover the utility of a power-spectrum-based approach of voiced-speech analysis beyond the quantitative parameterisation advantages as shown in Chapter 6, but additionally on a perceptual level. In order to achieve these goals, a perceptual experiment was designed where listeners could compare speech synthesised using parameters obtained by both time-domain and frequency-domain approaches and rate them upon a Likert-type scale.

The first experiment compared the performance of both parameterisation algorithms using signals recorded using phase linear equipment. The second experiment focused on the phase robustness of each technique and accordingly the test signals were convolved with the impulse response of a non-linear phase recording device, taken from the description given in (Berouti *et al.*, 1977). These two experiments used the same test data, the first five sentences from the CMU-ARCTIC database for two speakers, one male (*bdl*) and one female (*slt*). A third experiment was performed on speech obtained using an inexpensive headset microphone and a laptop computer which does not exhibit linear phase characteristics. Five sentences from a male and female speaker were recorded.

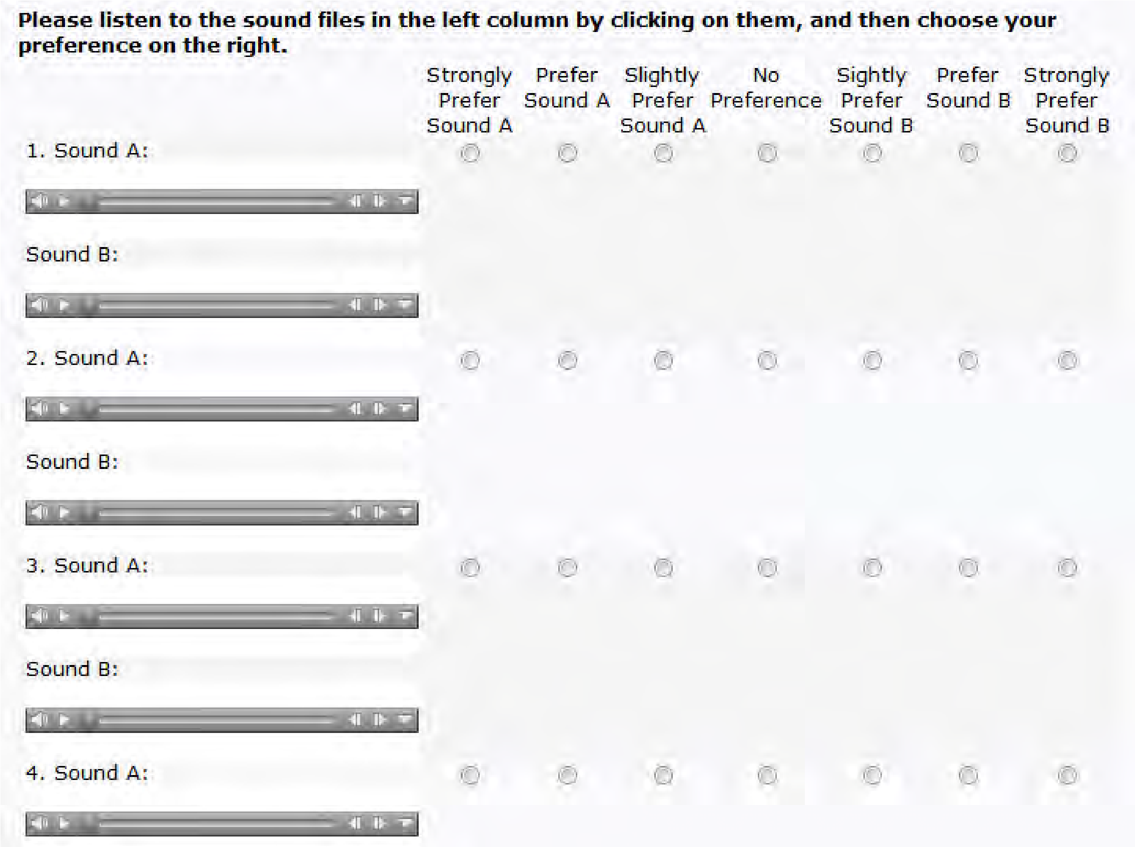
Both voice-source parameterisation methods were used to analyse the voiced speech segments of the signals. The obtained parameters were then used to synthesise speech segments. As they are not parameterised by either method, unvoiced speech segments are simply added into the output signal.

The glottal closing instant information required for the time-domain algorithm was obtained from the DYPSA algorithm[1] (Naylor *et al.*, 2007). As this information is particularly crucial for the time-domain algorithm, the glottal closing instants are refined following a first pass of the algorithm by choosing the glottal derivative flow signal minimum close the initial estimation. Pitch was estimated by the SWIPE' algorithm (Camacho, 2007).

The signals from the time-domain and frequency-domain approaches were then compared with each other using a listening test. The participants, of which there were 50, were asked to listen to both versions of the sentence and to give a score on a 7-point Likert-type scale, according to their general preference. The preference scores ranged from -3 to $+3$ corresponding to a strong preference for either the time-domain or frequency-domain method, while a 0 score denoted no preference for either technique. A screen shot of the web interface used for testing is given in Figure 8.2.

The parameterisation methods are fully automatic and no further processing was performed on the signals other than described. Due to some errors in the deconvolution procedure, disagreeable discontinuity-type artifacts were generated from the time-domain parameters of 4 sentences of the male speaker *bdl*, two from both the linear phase and nonlinear phase experiments. These utterances were removed from the data set. The mean preferences of the remaining signals and their 95% confidence intervals are presented within Figure 8.3. Table 8.1 gives the means μ and standard

[1] The FRESS algorithm was not used as it was not mature at the time when this experiment was undertaken.



Please listen to the sound files in the left column by clicking on them, and then choose your preference on the right.

	Strongly Prefer Sound A	Prefer Sound A	Slightly Prefer Sound A	No Preference	Sightly Prefer Sound B	Prefer Sound B	Strongly Prefer Sound B
1. Sound A:	○	○	○	○	○	○	○
Sound B:							
2. Sound A:	○	○	○	○	○	○	○
Sound B:							
3. Sound A:	○	○	○	○	○	○	○
Sound B:							
4. Sound A:	○	○	○	○	○	○	○

Figure 8.2: The above figure gives a screen shot of the web interface used for the perceptual test.

deviations σ of the preference scores for each experiment, and their corresponding t-scores and p-values. The full listening test results are given in Appendix E.

8.3 Discussion

The data from the listening tests clearly show a tendency of the participants to significantly ($p < 0.05$) prefer the speech synthesised with parameters of the PowARXLF approach over the time-domain ARX-LF parameterisation method for almost all

Table 8.1: This table contains the means and standard deviations of the preference scores of the perceptual experiment and their corresponding t-statistics and p-values.

Gender	Experiment	μ	σ	$t(49)$	p
Male	Linear Phase	0.06	1.84	0.23	0.41
	Nonlinear Phase	0.77	1.52	3.57	4×10^{-4}
	Laptop	0.33	1.42	1.66	0.05
Female	Linear Phase	1.24	1.59	5.52	1×10^{-6}
	Nonlinear Phase	1.36	1.39	6.89	$< 1 \times 10^{-6}$
	Laptop	1.23	1.23	7.04	$< 1 \times 10^{-6}$

recording conditions scenarios and both sexes.

Under phase linear conditions, one would expect that there would be generally no preference for either ARX-LF parameterisation method, as neither method has an obvious theoretical advantage. Indeed, this is what is observed for the male speaker under linear phase conditions, where the preference for the PowARXLF approach is slight. However, unexpectedly, the data shows that the synthetic speech of female speakers generated using the frequency-domain approach was particularly preferred over the time-domain method under phase linear conditions. This may be due to the difficulty in obtaining accurate glottal closing instant for these voices, which, as discussed in Chapters 3 and 5, is critical for time-domain voice-source parameterisation. As previously mentioned, the PowARXLF method is robust to the position of the analysis frame.

Under nonlinear phase conditions, *i.e.* the scenarios where the signals were convolved with the impulse response from a nonlinear phase recording system and recorded with inexpensive equipment, the PowARXLF approach was unsurprisingly superior to

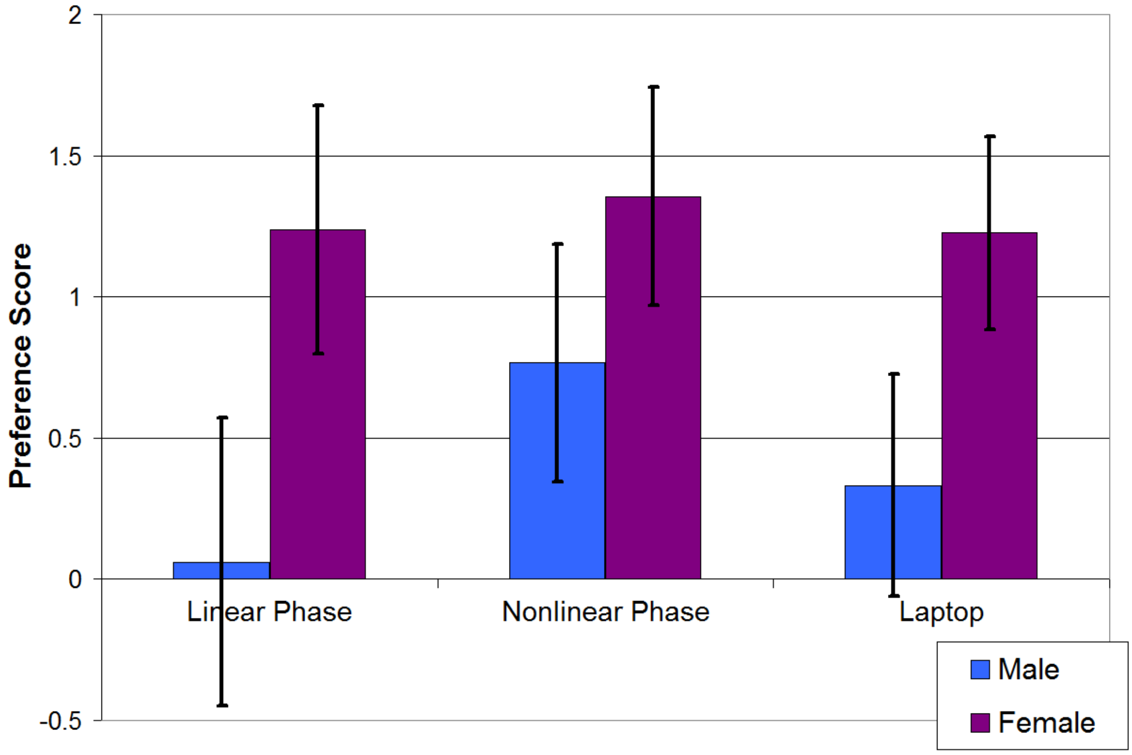


Figure 8.3: The above chart shows the average of the preference test results for the 3 experiments, separated into both male and female speakers. A positive score indicates a preference for the frequency-domain approach to ARX-LF parameter estimation.

the time-domain-based method. PowARXLF parameterisation is robust to non-ideal phase conditions by simply ignoring phase information. Conversely, time-domain approaches are not very robust to the time placement of the analysis frame. While efforts were made to mitigate this error in this experiment, this was almost certainly a source of some audible artifacts.

It is interesting to note the discrepancy in the relative preference increase in the first two tested scenarios between the male and female speakers. The preference for the power spectrum method for the male speaker increases substantially, while the increased preference for the female speakers is less. This can be understood by considering the phase response utilised by this experiment to corrupt the phase spectrum of the analysis utterances.

The response described in (Berouti *et al.*, 1977) is most distorted at very low frequencies ($< 45Hz$) and subsequently approaches linearity as frequency increases. In this case, the system would introduce more time-domain changes to low frequency signal components. Lower pitch male voices are therefore more likely to be affected by this kind of distortion, meaning that in the nonlinear phase case, the synthetic male speech would be more distorted. Therefore, it is then unsurprising that the PowARXLF method preference increase for male voices is higher than the female case.

8.4 Conclusions

This chapter presents Contribution 3, an extrinsic evaluation of a power spectrum approach to voice-source parameterisation and demonstration of the potential of this approach for speech synthesis and related applications. In order to accomplish this, a power-spectrum-based voice-source parameterisation was extended using two smoothing operations in order to obtain continuous parameters for an ARX-LF speech model. A comparative experiment was then undertaken to compare the synthetic speech generated by the parameters estimated from real speech signals by this approach against a reference system, a state-of-the-art time-domain-based ARX-LF parameterisation technique. The analysis signals were both recorded using linear phase equipment, convolved with a impulse response of a non-linear phase system, and also with signals recorded using generic non-linear phase audio recording equipment.

The experiment found that the power-spectrum-based approach to ARX-LF parameterisation is preferred over other techniques in all recording scenarios for all voices, though the preference was slight for the case with male speakers under phase linear conditions. This is an important and encouraging result, which justifies power-spectrum-based approaches to voice-source parameterisation.

Chapter 9

Conclusions and Future Work

9.1 Conclusions

This study has focussed upon speech analysis, particularly voice-source estimation and parameterisation. A literature review of these techniques finds that many of them are not robust to the time position of the analysis frame and phase distortion of the signal. Phase distortion is a prevalent phenomena affecting otherwise well-recorded signals. Methods of voice-source parameterisation which are robust to phase disturbances require assumptions of the filter order and do not attempt to avoid high frequency noise, which can degrade the accuracy of these methods.

For these reasons, the first contribution of the study, a robust glottal source parameterisation technique, is proposed in Chapter 6. The novel PowRd technique operates on the power spectrum of the speech signal and avoids high frequency noise by adopting a two band, HNM-type speech model. The lower band is fit with an all-pole filter envelope and transformed LF model glottal pulse using a new error criterion, the Relative Itakura-Saito error. The error is minimised using the DAP algorithm. Testing the algorithm with synthetic data showed comparable performance with other state of the art algorithms and superior robustness in the case of phase distorted speech, for which the algorithm was specifically designed. Interpreting the results of voice-source parameterisation algorithms on real speech is more difficult due to the distant relationship between the estimated acoustic waveform of the voice source and the interpretation of EGG data. However, the results indicate that the PowRd method is at least as good as the existing voice-source parameterisation methods, with the considerable advantage of robustness to the shape and position of the analysis frame.

The problem of glottal epoch estimation was also addressed in this work. A review of the literature that nonlinear phase recording conditions also affect those techniques, though to a lesser degree than voice-source estimation methods. It is also observed that the application of dynamic programming techniques to the output of a modified version of the SEDREAMS algorithm would give improved results for generally recorded speech.

These observations led to the proposal of a new method of glottal epoch estimation in Chapter 7, the second contribution of this work. Like the SEDREAMS method, the FRESS algorithm searches for peaks of the LPC residual signal within regions defined by the fundamental harmonic of the signal. Like the DYPSA and YAGA methods, the FRESS technique then uses dynamic programming to determine the most likely sequence of glottal epochs according to speech heuristics. In order to improve the robustness of the approach to phase disturbances, the fundamental frequency signal which defines the LPC residual search regions is aligned with the peaks of the normalised energy contour. The new technique is compared with other methods of glottal closure estimation by testing over a database of three speakers under real and simulated phase conditions. It was found that, under phase linear recording conditions, the FRESS method determines the glottal epochs with comparable accuracy and improved precision. Simulated nonlinear phase conditions are demonstrated to adversely affect other methods of glottal closure estimation, and not the FRESS method.

Finally, the third contribution was in the form of a speech analysis/synthesis system, which demonstrates the potential of a power-spectrum-based voice-source parameterisation approach for speech synthesis applications. A perceptual experiment was undertaken to compare the synthetic speech generated using the parameters obtained by a method similar to the PowRd approach with speech utterances synthesised using parameters obtained by a time-domain speech-parameterisation technique. The experiment analysed signals that were both recorded using phase linear and phase nonlinear equipment and also those ideally recorded speech convolved with a impulse response of a non-linear phase system. In order to synthesise the signals, an overlap-add synthesis scheme similar to one utilised by other existing methods was employed. The results of the listening test found that the power-spectrum approach is preferred over the time-domain technique in both recording scenarios for all voices, justifying the power-spectrum-based approach to voice-source parameterisation and giving encouraging results for future research.

9.2 Future Work

Some areas for future work are outlined in this section.

Voice-source estimation based on phase spectrum and Itakura-Saito minimisation The joint phase-spectrum-based voice-source parameterisation methods (Degottex *et al.*, 2011) discussed in Section 3.4.2 determine the optimum parameters by minimising the phase spectrum of the analysis frame. Instead of the imposing the assumption of an all-pole vocal tract, the methods assume that the vocal tract can be described by minimum phase spectral envelope, which is determined by the real cepstrum. This is a more general assumption because it permits spectral zeros, however, it has the consequence that the magnitude of the resulting deconvolutive residual signals $R(\omega) = 1, \forall\omega$. Therefore, magnitude-spectrum information of this signal is useless for parameterisation purposes.

However, for vowel sounds representing an unbranched vocal tract, the all-pole filter vocal tract assumption is suitable and can be utilised for voice-source parameterisation, as was demonstrated in this thesis. In this case, for signals recorded under ideal conditions, both the magnitude and phase spectra indicate the suitability of the speech model. This may give rise to new methods of voice-source parameterisation which operate on the principle of phase minimisation in addition to, or in combination to, the minimum Itakura-Saito error. Agreement between the two approaches may strengthen confidence in a particular glottal source estimate. A lack of agreement could be used to indicate *e.g.* phase disturbances or the lack of generality in the adopted models.

Vocal tract ARMA model Throughout this work, the all-pole filter was adopted as a model of the acoustic behaviour of the vocal tract. However, as discussed in Chapter 2, the assumption of the all-pole vocal tract filter is based upon the unbranched acoustic tube model. Certain sounds couple the nasal cavity with the main vocal tract and produce sounds which no longer respect the acoustic tube behaviour and may introduce zeros into the spectral envelope. Thus, the PowRd algorithm in its present form is ill-designed to approximate nasal sounds.

However,despite the nonlinearities encountered in determining spectral zeros (Makhoul, 1975), spectral methods to determine ARMA parameters have been developed (Badeau and David, 2008). Replacing the DAP algorithm in the PowRd method would enable it to more closely approximate nasalised phonemes. This is particularly interesting to experiment with the behaviour of the error criteria in this case, as it would be necessary to determine the number of poles *and* zeros.

Incorporation into voice coding, modification and synthesis systems Chapter 8 shows the potential of the described voice-source parameterisation algorithm for speech synthesis purposes. Indeed, the need for accurate voice-source parameterisation is becoming more prevalent with rising interest in emotional speech synthesisers (Cabral, 2010; Lanchantin *et al.*, 2010), in addition to pseudo-physical voice modification schemes (Lu, 2002; Vincent *et al.*, 2007; Agiomyrgiannakis and Rosec, 2009; Degottex, 2010). Because of the robustness of the PowRd method to phase distortion, a wider range of speech signals are suitable for analysis. Furthermore, the fully parametric nature of the voiced speech representation makes the method a suitable candidate for low bit rate speech coding applications (Spanias, 1994).

Bibliography

Acero, A., 1998. Source-filter models for time-scale pitch-scale modification of speech. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 2. Seattle, Washington, pp. 881–884.

Agiomyrgiannakis, Y., Rosec, O., 2008. Towards Flexible Speech Coding for Speech Synthesis: an LF + Modulated Noise Vocoder. In: Proceedings of the 2008 International Conference on Speech Communication and Technology (INTERSPEECH). Brisbane, Australia, pp. 1849–1852.

Agiomyrgiannakis, Y., Rosec, O., 2009. ARX-LF-based source-filter methods for voice modification and transformation. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Taipei, Taiwan, pp. 3589–3592.

Airas, M., 2008a. Methods and studies of laryngeal voice quality analysis in speech production. PhD thesis, Helsinki University of Technology.

Airas, M., 2008b. TKK Aparat: An environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology 33 (1), 49–64.

Airas, M., Alku, P., 2007. Comparison of Multiple Voice Source Parameters in Different Phonation Types. In: Proceedings of the 2007 International Conference on Speech Communication and Technology (INTERSPEECH). Antwerp, Belgium, pp. 1410–1413.

Akande, O. O., 2004. Speech Analysis Techniques for Glottal Source and Noise Estimation in Voice Signals. PhD thesis, University of Limerick.

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

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11 (2), 109–117.

Alku, P., Laine, U. K., 1989. A New Glottal LPC Method for Voice Coding and Inverse Filtering. In: Proceedings of the 1989 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1831–1834.

Alku, P., Magi, C., Yrttiaho, S., Bäckström, T., Story, B. H., 2009. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. The Journal of the Acoustical Society of America 125 (5), 3289–3305.

Alku, P., Vintturi, J., Vilkman, E., 2002. Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. Speech Communication 38 (3-4), 321–334.

Ananthapadmanabha, T., Fant, G., 1982. Calculation of true glottal flow and its components. KTH Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR) 23 (1), 1–30.

Arroabarren, I., Carlosena, A., 2003. Glottal spectrum based inverse filtering. In: Proceedings of the 2003 International Conference on Speech Communication and Technology (INTERSPEECH). Geneva, Switzerland, pp. 57–60.

Atal, B. S., Hanauer, S., 1971. Speech analysis and synthesis by linear prediction of the speech wave. The Journal of the Acoustical Society of America 50 (2), 637–655.

Bäckström, T., 2004. Linear predictive modelling of speech-constraints and line spectrum pair decomposition. PhD thesis, Helsinki University of Technology.

Badeau, R., David, B., 2008. Weighted maximum likelihood autoregressive and moving average spectrum modeling. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, Nevada, pp. 3761–3764.

Bagshaw, P., Hiller, S., Jack, M., 1993. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In: Proceedings of the 1993 European Conference on Speech Communication and Technology (EUROSPEECH). Berlin, Germany, pp. 1003–1006.

Berouti, M., Childers, D. G., Paige, A., 1977. Correction of tape recorder distortion. In: Proceedings of the 1977 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hartford, Connecticut, pp. 397–400.

Birkholz, P., Jackèl, D., Kroger, K., 2006. Construction and control of a three-dimensional vocal tract model. In: Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 1. Toulouse, France.

Bouzid, A., Ellouze, N., 2008. Electroglottographic measures based on GCI and GOI detection using multiscale product. Journal of Electrical and Computer Engineering 3 (1), 21–32.

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

Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T., 2004. Appropriate windowing for group delay analysis and roots of Z-transform of speech signals. In: Proceedings of the 2004 European Signal Processing Conference (EUSIPCO). Vienna, Austria, pp. 733–736.

Brent, R. P., 2002. Algorithms for Minimization Without Derivatives. Courier Dover Publications.

Brookes, D., Retrieved January 22nd, 2009. Voicebox: Speech Processing Toolbox for Matlab. `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

Brookes, D., Chan, D., 1994. Speaker Characteristics from a Glottal Airflow Model using Robust Inverse Filtering. Proceedings of the Institute of Acoustics 16, 501–508.

Brookes, D., Naylor, P. A., Gudnason, J., 2006. A Quantitative Assessment of Group

Delay Methods for Identifying Glottal Closures in Voiced Speech. IEEE Transactions on Audio, Speech, and Language Processing 14 (2), 456–466.

Cabral, J. P., 2010. HMM-based Speech Synthesis Using an Acoustic Glottal Source Model. PhD thesis, University of Edinburgh.

Cabral, J. P., Kane, J., Gobl, C., Carson-Berndsen, J., 2011. Evaluation of Glottal Epoch Detection Algorithms on Different Voice Types. In: Proceedings of the 2011 International Conference on Speech Communication and Technology (INTERSPEECH). Florence, Italy, pp. 1989–1992.

Camacho, A., 2007. SWIPE: A Sawtooth Waveform Inspired Pitch Estimator. PhD thesis, University of Florida.

Cappé, O., Laroche, J., Moulines, E., 1995. Regularized Estimation of Cepstrum Envelope from Discrete Frequency Points. In: Proceedings of the 1995 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz, New York, pp. 213–216.

Chan, D., Brookes, D., 1989. Variability of excitation parameters derived from robust closed phase glottal inverse filtering. In: Proceedings of the 1989 European Conference on Speech Communication and Technology (EUROSPEECH). Paris, France, pp. 2199–2202.

Charpentier, F., Moulines, E., 1989. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In: Proceedings of the 1989 European Conference on Speech Communication and Technology (EUROSPEECH). Paris, France, pp. 2013–2019.

Childers, D. G., 1995. Glottal source modeling for voice conversion. Speech Communication 16, 127–138.

Childers, D. G., 2000. Speech processing and synthesis toolboxes. John Wiley & Sons, Inc.

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

Chow, Y., Schwartz, R., 1989. The n-best algorithm: An efficient procedure for finding top n sentence hypotheses. In: Proceedings of the 1989 Workshop on Speech and Natural Language. pp. 199–202.

Cooley, J. W., Tukey, J. W., 1965. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation 19 (90), 297–301.

d'Alessandro, N., 2009. Realtime and accurate musical control of expression in singing synthesis. PhD thesis, University of Mons.

Davies, P., Lindsey, G., Fuller, H., Fourcin, A., 1986. Variation of glottal open and closed phases for speakers of English. Proceedings of the Institute of Acoustics 8 (7), 539–46.

de Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America 111 (April), 1–14.

Degottex, G., 2010. Glottal source and vocal-tract separation. PhD thesis, UPMC.

Degottex, G., Röbel, A., Rodet, X., 2011. Phase minimization for glottal model estimation. IEEE Transactions on Audio, Speech, and Language Processing 19 (5), 1080–1090.

del Pozo, A., 2008. Voice Source and Duration Modelling for Voice Conversion and Speech Repair. PhD thesis, Cambridge University.

Ding, W., Campbell, N., 1998. Determining polarity of speech signals based on gradient of spurious glottal waveforms. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, Washington, pp. 857–860.

Ding, W., Campbell, N., Higuchi, N., Kasuya, H., 1997. Fast and robust joint estimation of vocal tract and voice source parameters. In: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Munich, Germany, pp. 1291–1294.

Ding, W., Kasuya, H., Adachi, S., 1994. Simultaneous estimation of vocal tract and voice source parameters with application to speech synthesis. In: Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP). pp. 159–162.

Doval, B., d'Alessandro, C., 2006. The spectrum of glottal flow models. Acta Acustica united with Acustica 92 (6), 1026–1046.

Doval, B., d'Alessandro, C., Henrich, N., 2003. The voice source as a causal/anticausal linear filter. In: Proceedings of the 2003 ISCA Tutorial and Research Workshop on Voice Quality (VOQUAL'03). Geneva, Switzerland.

Drioli, C., Avanzini, F., 2000. Model-based synthesis and transformation of voiced sounds. In: Proceedings of the 2000 International Conference on Digital Audio Effects (DAFx). Verona, Italy, pp. 7–10.

Drugman, T., 2011. Advances in Glottal Analysis and its Applications. PhD thesis, Univeristy of Mons.

Drugman, T., Retrieved October 8th, 2011. GLOAT Toolbox. `http://tcts.fpms.ac.be/~drugman/Toolbox/`.

Drugman, T., Bozkurt, B., Dutoit, T., 2009a. Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation. In: Proceedings of the 2009 International Conference on Speech Communication and Technology (INTERSPEECH). Brighton, England, pp. 116–119.

Drugman, T., Bozkurt, B., Dutoit, T., 2011a. A comparative study of glottal source estimation techniques. Computer Speech & Language 26 (1), 20–34.

Drugman, T., Bozkurt, B., Dutoit, T., Jul. 2011b. Causal/anticausal decomposition of speech using complex cepstrum for glottal source estimation. Speech Communication 53 (6), 855–866.

Drugman, T., Dutoit, T., 2009. Glottal Closure and Opening Instant Detection from Speech Signals. In: Proceedings of the 2009 International Conference on Speech Communication and Technology (INTERSPEECH). Brighton, England, pp. 2891–2894.

Drugman, T., Dutoit, T., 2010. Chirp Complex Cepstrum-based Decomposition for

Asynchronous Glottal Analysis. In: Proceedings of the 2010 International Conference on Speech Communication and Technology (INTERSPEECH). Chiba, Japan, pp. 657–660.

Drugman, T., Dutoit, T., 2011. Oscillating Statistical Moments for Speech Polarity Detection. In: Proceedings of the 2011 ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NOLISP). Las Palmas, Gran Canaria, pp. 48–54.

Drugman, T., Wilfart, G., Dutoit, T., 2009b. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In: Proceedings of the 2009 International Conference on Speech Communication and Technology (INTERSPEECH). Citeseer, Brighton, England, pp. 1779–1782.

El-Jaroudi, A., Makhoul, J., 1991. Discrete All-Pole Modeling. IEEE Transactions on Signal Processing 39 (2), 411–423.

Erro, D., Sainz, I., Navas, E., Hernáez, I., 2011. Improved HNM-based Vocoder for Statistical Synthesizers. In: Proceedings of the 2011 International Conference on Speech Communication and Technology (INTERSPEECH). Florence, Italy, pp. 1809–1812.

Fabre, P., 1957. Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation: glottographie de haute fréquence. Bulletin de l'Académie Nationale de Médecine 141, 66.

Fant, G., 1970. Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations, 2nd Edition. Mouton De Gruyter.

Fant, G., 1979. Glottal source and excitation analysis. KTH Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR) 20 (1), 85–107.

Fant, G., 1995. The LF-model revisited. Transformations and frequency domain analysis. KTH Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR) 36 (2-3), 119–156.

Fant, G., Kruckenberg, A., Liljencrants, J., Bavegard, M., 1994. Voice source parameters in continuous speech, transformation of LF parameters. Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP), 1451–1454.

Fant, G., Liljencrants, J., Lin, Q., 1985. A four-parameter model of glottal flow. KTH Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR) 26 (4), 1–13.

Fernandez, R., Retrieved November 14th, 2010. Speech Prosody Analysis Tools. `http://affect.media.mit.edu/software.php`.

Flanagan, J. L., 1972. Speech analysis: Synthesis and Perception, 2nd Edition. Springer-Verlag.

Fröhlich, M., Michaelis, D., Strube, H. W., 2001. SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. The Journal of the Acoustical Society of America 110 (1), 479–488.

Fu, Q., Murphy, P. J., 2006. Robust Glottal Source Estimation Based on Joint Source-filter Model Optimization. IEEE Transactions on Audio, Speech, and Language Processing 14 (2), 492–501.

Fujisaki, H., Ljungqvist, M., 1986. Proposal and evaluation of models for the glottal source waveform. In: Proceedings of the 1986 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Tokyo, Japan, pp. 1605–1608.

Fujisaki, H., Ljungqvist, M., 1987. Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. In: Proceedings of the 1987 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, Texas, pp. 637–640.

Funaki, K., Miyanaga, Y., Tochinai, K., 1997. A time varying ARMAX speech modeling with phase compensation using glottal source model. In: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 2. IEEE, Munich, Germany, pp. 1299–1302.

Gobl, C., 2003. The voice source in speech communication-production and perception experiments involving inverse filtering and synthesis. PhD thesis, KTH Royal Institute of Technology.

Gobl, C., 2006. Modelling Aspiration Noise During Phonation Using the LF Voice Source Model. In: Proceedings of the 2006 International Conference on Speech Communication and Technology (INTERSPEECH). Pittsburgh, USA, pp. 965–968.

Goncharoff, V., Retrieved July 6th, 2011. Matlab File Exchange: find_pmarks Algorithm. `http://www.mathworks.com/matlabcentral/fileexchange/18497-speech-processing-tool`.

Goncharoff, V., Gries, P., 1998. An algorithm for accurately marking pitch pulses in speech signals. In: Proceedings of the 1998 IASTED Signal and Image Processing Conference. Las Vegas, Nevada.

Gray, R., Buzo, A., Gray, A. H., Matsuyama, Y., Aug. 1980. Distortion measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (4), 367–376.

Harris, F., 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE 66 (1), 51–83.

Hedelin, P., 1984. A glottal LPC-vocoder. In: Proceedings of the 1984 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). San Diego, California, pp. 21–24.

Hedelin, P., 1986. High quality glottal LPC-vocoding. In: Proceedings of the 1986 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Tokyo, Japan, pp. 465–468.

Henrich, N., Sundin, G., Ambroise, D., d'Alessandro, C., Castellengo, M., Doval, B., 2003. Just noticeable differences of open quotient and asymmetry coefficient in singing voice. The Journal of Voice 17 (4), 481–494.

Hermes, D. J., 1991. Synthesis of breathy vowels: Some research methods. Speech Communication 10, 497–502.

Holland, J. H., May 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. Bradford Books.

Holmes, J. N., Sep. 1975. Low-frequency phase distortion of speech recordings. The Journal of the Acoustical Society of America 58 (3), 747–9.

Honda, K., 2007. Physical Processes of Speech Production. In: Benesty, J., Sondhi, M., Huang, Y. (Eds.), Springer Handbook of Speech Processing. Springer-Verlag, Heidelberg, pp. 7–26.

Howard, D. M., Jun. 1995. Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. The Journal of Voice 9 (2), 163–72.

Hu, H.-T., Wu, H.-T., 2000. A Glottal-Excited Linear Prediction (GELP) Model for Low-Bit-Rate Speech Coding. In: Proceedings of the 2000 Republic of China National Science Council. pp. 134–142.

Ishizaka, K., Flanagan, J. L., 1972. Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell System Technical Journal 51 (6), 1233—-1268.

Itakura, F., Apr. 1975. Line spectrum representation of linear predictor coefficients of speech signals. The Journal of the Acoustical Society of America 57 (S1), S35.

Itakura, F., Saito, S., 1968. Analysis synthesis telephony based on the maximum likelihood method. In: Proceeding of the International Congress on Acoustics. Tokyo, Japan.

Kalman, R., 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82 (Series D), 35–45.

Kane, J., Kane, M., Gobl, C., 2010. A spectral LF model based approach to voice source parameterisation. In: Proceedings of the 2010 International Conference on Speech Communication and Technology (INTERSPEECH). Chiba, Japan, pp. 2606–2609.

Kirkpatrick, S., Gelatt, C., Vecchi, M., 1983. Optimization by simulated annealing. Science 220 (4598), 671.

Klatt, D. H., 1987. Review of text-to-speech conversion for English. The Journal of the Acoustical Society of America 82, 737.

Klatt, D. H., Klatt, L. C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. The Journal of the Acoustical Society of America 87 (2), 820–857.

Kominek, J., Black, A., 2003. CMU ARCTIC databases for speech synthesis. Tech. rep., CMU Language Technologies Institute.

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

Krishnamurthy, A. K., Childers, D. G., 1986. Two-channel speech analysis. IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (4), 730–743.

Kullback, S., 1987. The Kullback-Leibler Distance. The American Statistician 41 (4), 340–341.

Laine, U. K., 1982. Modelling of lip radiation impedance in Z-domain. In: Proceedings of the 1982 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Paris, France, pp. 1992–1995.

Lanchantin, P., Degottex, G., Rodet, X., Mar. 2010. A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, Texas, pp. 4630–4633.

Larar, J. N., Alsaka, Y. A., Childers, D. G., 1985. Variability in closed phase analysis of speech. In: Proceedings of the 1985 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Tampa, Florida, pp. 1089–1092.

Laroche, J., Stylianou, Y., Moulines, E., 1993. High-quality speech modification based on a harmonic+noise model. In: Proceedings of the 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. pp. 550–553.

Lehto, L., Airas, M., Björkner, E., Sundberg, J., Alku, P., 2007. Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types. The Journal of Voice 21 (2), 138–50.

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

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

Magi, C., Pohjalainen, J., Bäckström, T., Alku, P., May 2009. Stabilised weighted linear prediction. Speech Communication 51 (5), 401–411.

Makhoul, J., 1975. Spectral linear prediction: properties and applications. IEEE Transactions on Acoustics, Speech, and Signal Processing 23 (3), 283–296.

Markel, J. D., Gray, A. H., May 1982. Linear Prediction of Speech. Springer-Verlag, New York, New York.

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

MATLAB, 2010. Version 7.10.0 (R2010a). The Mathworks Inc., Natick, Massachusetts.

McAulay, R. J., Quatieri, T. F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (4), 744–754.

McKenna, J. G., 2001. Automatic glottal closed-phase location and analysis by Kalman filtering. In: Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis. Pitlochry, Scotland.

Miller, R. L., 1959. The Nature of the Vocal Cord Wave. The Journal of the Acoustical Society of America 31 (6), 667–677.

Moore, E., Clements, M. A., 2004. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Montreal, Canada, pp. 101–104.

Moulines, E., Laroche, J., 1995. Non-parametric techniques for pitch-scale and time-scale modification of speech. Speech Communication 16 (2), 175–205.

Mullen, J., 2006. Physical modelling of the vocal tract with the 2D digital waveguide mesh. PhD thesis, University of York.

Murty, K. S. R., Yegnanarayana, B., Nov. 2008. Epoch Extraction From Speech Signals. IEEE Transactions on Audio, Speech, and Language Processing 16 (8), 1602–1613.

Musicus, B. R., 1988. Levinson and fast Choleski algorithms for Toeplitz and almost Toeplitz matrices. Tech. Rep. 538, Massachusetts Institute of Technology, Research Laboratory of Electronics.

National Cancer Institute, Retrieved September 12th, 2009. Larynx & Trachea Anatomy. `http://training.seer.cancer.gov/anatomy/respiratory/passages/larynx.html`.

Naylor, P. A., Kounoudes, A., Gudnason, J., Brookes, D., 2007. Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm. IEEE Transactions on Audio, Speech, and Language Processing 15 (1), 34–43.

Nelder, J. A., Mead, R., 1965. A simplex method for function minimization. The Computer Journal 7, 308–313.

Nord, L., Ananthapadmanabha, T., Fant, G., 1984. Signal analysis and perceptual tests of vowel responses with an interactive source filter model. KTH Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR) 2 (3), 1984.

Ó Cinnéide, A., Dorran, D., Gainza, M., Coyle, E., 2010. On the appearance of a positive real pole in the results of glottal closed phase linear prediction. In: Proceedings of the 2010 European Signal Processing Conference (EUSIPCO). Aalborg, Denmark.

Oppenheim, A. V., Schafer, R. W., 1975. Digital Signal Processing. Prentice Hall.

Paliwal, K., Kleijn, W. B., 1995. Quantization of LPC parameters. Speech Coding and Synthesis, 433–466.

Pantazis, Y., Rosec, O., Stylianou, Y., 2008. On the Estimation of the Speech Harmonic Model. In: Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Speech Analysis and Processing for Knowledge Discovery. Aalborg, Denmark, pp. 5–8.

Pedersen, C., Andersen, O., Dalsgaard, P., 2010. Separation of mixed phase signals by zeros of the Z-transform - A reformulation of complex cepstrum based separation by causality. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, Texas, pp. 5050–5053.

Pérez, J., Bonafonte, A., 2005. Automatic voice-source parameterization of natural speech. In: Proceedings of the 2005 International Conference on Speech Communication and Technology (INTERSPEECH). Lisbon, Portugal, pp. 1065–1068.

Pérez, J., Bonafonte, A., 2009. Towards robust glottal source modeling. In: Proceedings of the 2009 International Conference on Speech Communication and Technology (INTERSPEECH). Brighton, England, pp. 68–71.

Pérez, J., Bonafonte, A., 2011. Adding Glottal Source Information to Intra-lingual Voice Conversion. Proceedings of the 2011 International Conference on Speech Communication and Technology (INTERSPEECH), 2773–2776.

Plumpe, M. D., Quatieri, T. F., Reynolds, D. A., 1997. Modeling of the glottal flow derivative waveform with application to speaker identification. MSc thesis, Massachusetts Institute of Technology.

Powell, M., 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. The Computer Journal 7 (2), 155.

Press, W., 2007. Numerical recipes: the art of scientific computing. Cambridge University Press.

Quatieri, T. F., 2001. Discrete-Time Speech Processing: Principles and Practice. Prentice Hall.

Rabiner, L. R., Schafer, R. W., 1978. Digital Processing of Speech Signals. Prentice Hall, Englewood Cliffs, New Jersey.

Rabiner, L. R., Schafer, R. W., Rader, C., 1969. The chirp Z-transform algorithm. IEEE Transactions on Audio and Electroacoustics 17 (2), 86—-92.

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P., 2011. HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Transactions on Audio, Speech, and Language Processing 19 (1), 153–165.

Röbel, A., Rodet, X., 2005. Real time signal transposition with envelope preservation in the phase vocoder. In: Proceedings of the 2005 International Computer Music Conference (ICMC). Citeseer.

Rodet, X., 1997. Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models. In: Proceedings of the 1997 Time-Frequency and Time-Scale Workshop. Vol. 4. pp. 1–10.

Rosenberg, A. E., 1970. Effect of Glottal Pulse Shape on the Quality of Natural Vowels. The Journal of the Acoustical Society of America 94 (2), 583–590.

Saratxaga, I., Erro, D., Hernáez, I., Sainz, I., Navas, E., 2009. Use of harmonic phase information for polarity detection in speech signals. In: Proceedings of the 2009 International Conference on Speech Communication and Technology (INTERSPEECH). Brighton, England, pp. 1075–1078.

Sciamarella, D., d'Alessandro, C., 2004. On the acoustic sensitivity of a symmetrical two-mass model of the vocal folds to the variation of control parameters. Acta Acustica united with Acustica 90 (4), 746–761.

Serra, X., 1989. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. PhD thesis, Stanford University.

Silva, D. G., Oliveira, L. C., Andrea, M., 2009. Jitter Estimation Algorithms for Detection of Pathological Voices. EURASIP Journal on Advances in Signal Processing 2009, 1–10.

Smith III, J. O., 2007. Introduction to Digital Filters: with Audio Applications. W3K Publishing.

Smits, R., Yegnanarayana, B., 1995. Determination of Instants of Significant Excitation in Speech Using Group Delay Function. IEEE Transactions on Speech and Audio Processing 3 (5), 325–333.

Spanias, A., 1994. Speech coding: A tutorial review. Proceedings of the IEEE 82 (10), 44.

Story, B. H., 2003. Physical modeling of voice and voice quality. In: Proceedings of the 2003 ISCA Tutorial and Research Workshop on Voice Quality (VOQUAL'03). Geneva, Switzerland.

Strik, H., 1996. Comments on "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering" [J. Acoust. Soc. Am. 98, 763-767 (1995)]. The Journal of the Acoustical Society of America 767 (1995), 1246–1249.

Strik, H., 1998. Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. The Journal of the Acoustical Society of America 103 (5), 2659–2669.

Stylianou, Y., 1996. Harmonic plus Noise Methods for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, ENST-Telecom.

Stylianou, Y., Mar. 2001. On the implementation of the harmonic plus noise model for concatenative speech synthesis. IEEE Transactions on Speech and Audio Processing 9 (3), 232–239.

Sündermann, D., 2008. Text-Independent Voice Conversion. PhD thesis, Bundeswehr University of Munich.

Swerts, M., Veldhuis, R. N. J., 2001. The effect of speech melody on voice quality. Speech Communication 33, 297–303.

Thomas, M. R. P., Gudnason, J., Naylor, P. A., 2011. Estimation of Glottal Closing and Opening Instants in Voiced Speech Using the YAGA Algorithm. IEEE Transactions on Audio, Speech, and Language Processing 6 (99), 1–1.

Thomas, M. R. P., Naylor, P. A., Nov. 2009. The SIGMA Algorithm: A Glottal Activity Detector for Electroglottographic Signals. IEEE Transactions on Audio, Speech, and Language Processing 17 (8), 1557–1566.

Tooher, M., McKenna, J. G., 2003. Variation of glottal LF parameters across F0, vowels, and phonetic environment. In: Proceedings of the 2003 ISCA Tutorial and Research Workshop on Voice Quality (VOQUAL'03). Geneva, Switzerland, pp. 41–46.

Van den Berg, J. W., 1958. Myoelastic-Aerodynamic Theory of Voice Production. The Journal of Speech, Language, and Hearing Research 1 (3), 227–244.

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

Veldhuis, R. N. J., 1998. A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. The Journal of the Acoustical Society of America 103 (1), 566–571.

Vincent, D., 2007. Analyse et contrôle du signal glottique en synthèse de la parole. PhD thesis, Rennes University.

Vincent, D., Rosec, O., Chonavel, T., 2005. Estimation of LF glottal source parameters based on an ARX model. In: Proceedings of the 2005 International Conference on Speech Communication and Technology (INTERSPEECH). Lisbon, Portugal, pp. 333–336.

Vincent, D., Rosec, O., Chonavel, T., 2007. A new method for speech synthesis and tranformation based on an ARX-LF source-filter decomposition and HNM modeling. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu, Hawaii, pp. 525–528.

Wakita, H., 1973. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Transactions on Audio and Electroacoustics 21 (5), 417–427.

Walker, J., Murphy, P. J., 2007. A Review of Glottal Waveform Analysis. In: Stylianou, Y., Faundez-Zanuy, M., Esposito, A. (Eds.), Progress in Nonlinear Speech Processing. Vol. 4391 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–21.

Wells, J., Ramsaran, S., Ladefoged, P., Retrieved November 19th, 2011. UCLA Phonetics Lab Data. http://hctv.humnet.ucla.edu/departments/linguistics/VowelsandConsonants/course/chapter1/wells/wells.html.

Wong, D. Y., Markel, J. D., Gray, A. H., 1979. Least squares glottal inverse filtering from the acoustic speech. IEEE Transactions on Acoustics, Speech, and Signal Processing 27 (4), 350–355.

Appendix A

Vocal-Tract Filters

This appendix contains diagrams and parameters of the all-pole spectral envelopes used for representing the vocal tract in Chapters 6 and 7. Uttered by trained phoneticians covering the IPA vocalic trapezoid (Figure A.1), all spectral envelopes were estimated from real speech signals using the IAIF method and are given in Figures A.2 and A.3; the recordings are available at (Wells *et al.*, Retrieved November 19th, 2011). The bandwidth of each analysis signal was 8kHz.

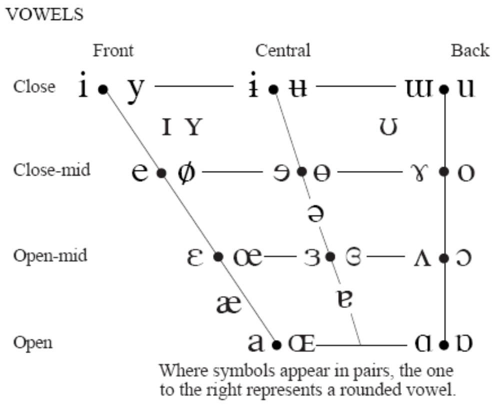


Figure A.1: Above is a diagram of the IPA vocalic trapezoid which provided the vocal tract filters utilised within this work.

Table A.1: The table below gives the all-pole vocal-tract filter parameters used in Chapters 6 and 7.

Filter		Formant							
		1	2	3	4	5	6	7	8
(i)	F_k (Hz)	228.84	295.95	2213.42	3607.46	4516.20	5618.05	6354.66	7279.11
	B_k (Hz)	70.73	1204.73	277.44	494.04	135.42	1331.25	1349.15	952.80
(ii)	F_k (Hz)	230.94	1503.45	1942.68	3190.99	4273.56	5106.34	6643.11	7122.32
	B_k (Hz)	26.30	179.27	280.33	300.81	393.52	412.18	789.47	1043.51
(iii)	F_k (Hz)	237.76	1776.89	2126.43	3611.60	4326.63	5113.57	6459.14	7594.27
	B_k (Hz)	14.78	222.92	116.49	140.50	315.44	461.28	820.84	733.26
(iv)	F_k (Hz)	240.89	1151.11	2090.51	3357.41	4038.45	5528.84	6300.14	7326.56
	B_k (Hz)	102.73	197.17	279.89	640.06	153.96	160.58	945.05	552.72
(v)	F_k (Hz)	272.63	1550.68	1977.15	3326.32	3990.32	4744.12	6059.92	7497.41
	B_k (Hz)	45.73	128.92	346.28	2321.59	1383.86	569.94	589.93	602.49
(vi)	F_k (Hz)	285.04	1296.68	2003.54	3103.16	4184.74	5159.35	6705.98	8000.00
	B_k (Hz)	82.39	37.18	144.70	107.50	383.56	222.84	598.96	529.69
(vii)	F_k (Hz)	294.33	724.07	2509.82	3349.70	4619.19	5157.64	6639.28	7034.30
	B_k (Hz)	188.67	161.55	146.11	505.47	409.80	196.67	1801.25	26.64
(viii)	F_k (Hz)	326.37	994.56	2228.28	3331.74	4116.59	5298.55	5795.11	7056.05
	B_k (Hz)	105.11	110.58	80.82	474.97	231.16	407.26	900.88	192.32
(ix)	F_k (Hz)	345.01	1520.98	1998.31	3249.68	4326.12	5063.71	5948.31	7683.52
	B_k (Hz)	34.90	83.69	194.21	69.37	1681.61	265.58	1145.34	1037.56
(x)	F_k (Hz)	356.12	1834.92	2519.09	3185.47	4876.93	5141.56	5679.61	7433.97
	B_k (Hz)	31.68	370.36	224.30	244.60	505.53	1310.71	547.99	270.66
(xi)	F_k (Hz)	366.28	2002.29	2657.72	2792.77	3928.55	5042.55	5779.97	7231.58
	B_k (Hz)	97.03	179.44	227.62	3935.92	80.89	394.77	759.52	310.88
(xii)	F_k (Hz)	382.45	603.91	2417.36	3796.29	4267.19	5283.86	6517.42	7180.48
	B_k (Hz)	136.78	263.20	113.60	1211.92	214.38	384.09	1013.11	338.76

Table A.2: The table below gives the all-pole vocal-tract filter parameters used in Chapters 6 and 7 (cont'd).

Filter		Formant							
		1	2	3	4	5	6	7	8
(xiii)	F_k (Hz)	397.45	805.36	2511.31	3382.55	4486.01	5381.45	6619.74	8000.00
	B_k (Hz)	237.60	220.98	163.61	359.86	359.80	228.53	271.80	1078.48
(xiv)	F_k (Hz)	409.91	1142.92	2490.62	3412.60	4363.78	5473.25	5532.66	7243.60
	B_k (Hz)	143.43	87.86	108.45	150.42	283.56	292.35	853.97	723.66
(xv)	F_k (Hz)	420.13	1314.11	2334.64	3503.30	4438.27	5215.35	5502.84	7613.07
	B_k (Hz)	32.84	82.90	149.95	409.04	298.24	3734.32	175.50	378.06
(xvi)	F_k (Hz)	438.78	779.95	2497.28	3636.53	4464.95	5523.44	5849.79	7296.37
	B_k (Hz)	105.26	256.87	115.47	557.63	320.61	408.74	1885.98	919.26
(xvii)	F_k (Hz)	530.42	1042.32	2637.03	3507.72	4363.58	5338.43	5709.63	7509.99
	B_k (Hz)	181.58	92.39	149.38	260.53	860.81	1438.62	369.78	678.71
(xviii)	F_k (Hz)	554.39	1154.28	2624.00	3507.30	4788.89	5630.93	6215.78	8000.00
	B_k (Hz)	274.26	234.21	308.13	145.42	709.58	342.10	1355.89	91.80
(xix)	F_k (Hz)	569.62	1746.89	2293.08	2545.74	4254.93	5126.68	5694.44	7237.06
	B_k (Hz)	361.11	250.15	2025.81	510.67	443.16	364.04	244.56	225.87
(xx)	F_k (Hz)	588.44	1232.80	1869.99	3198.58	4760.92	5057.42	5894.51	7518.31
	B_k (Hz)	203.52	173.15	275.49	138.64	6325.87	221.18	312.55	344.54
(xxi)	F_k (Hz)	635.31	1342.04	1841.56	3169.04	3892.81	5187.06	5767.76	7581.03
	B_k (Hz)	178.82	92.74	432.85	3480.60	554.07	351.39	466.35	1233.19
(xxii)	F_k (Hz)	750.73	1262.18	1740.85	3189.57	4847.91	5146.95	5795.29	7058.78
	B_k (Hz)	202.05	112.23	160.76	518.40	3073.75	200.29	234.29	1407.98

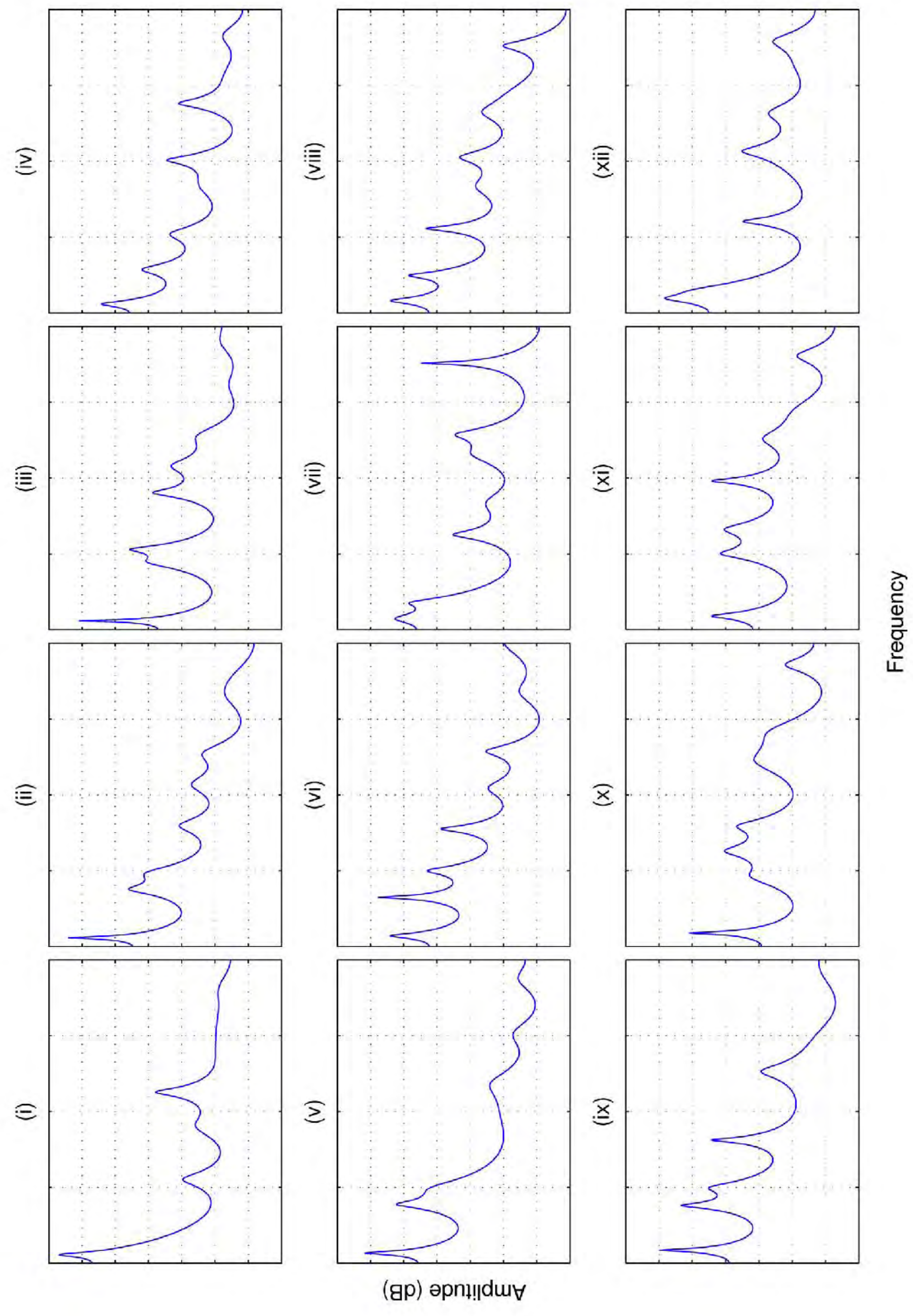


Figure A.2: This figure contains the vocal-tract filters utilised in the synthetic speech experiments outlined in this work.

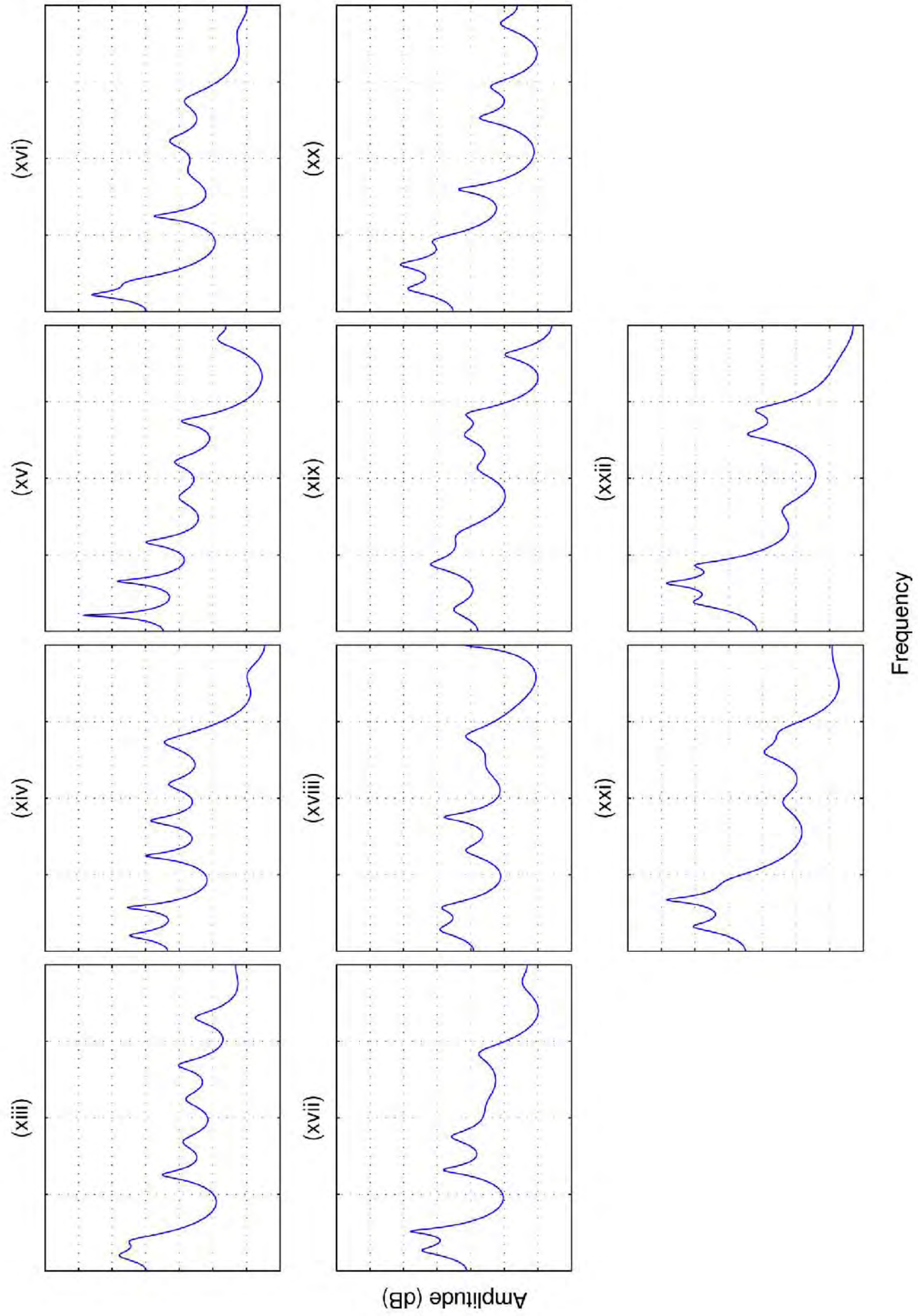


Figure A.3: This figure contains the vocal-tract filters utilised in the synthetic speech experiments outlined in this work (cont'd).

Appendix B

All-Pole Filter Envelope Estimation of Discrete Power Spectra

As discussed in Chapter 2, the all-pole model has been used to approximate both the spectral envelope of the vocal tract and the spectral characteristics of the glottal signal. Thus, for speech and many other signals, all-pole filter parameterisation is an important and useful task.

For signals that are periodic *e.g.* voiced speech, the spectral envelope information may only be available at discrete points. This appendix discusses two methods to determine the optimum all-pole filter fitting a discrete spectrum: linear prediction and Discrete All-Pole modeling (DAP).

B.1 Spectral Linear Prediction

Given a set of N power spectrum samples at frequencies ω_n, spectral linear prediction determines the best fitting p^{th} order all-pole envelope which minimises E_{LP}, the mean ratio between the given discrete spectrum $P(\omega_n)$ and the spectrum of the all-pole filter sampled at the same frequency points $\hat{P}(\omega_n)$. E_{LP} is given by:

$$E_{LP} = \frac{1}{N} \sum_{n=1}^{N} \frac{P(\omega_n)}{\hat{P}(\omega_n)} \tag{B.1}$$

The power spectrum $\hat{P}(\omega_n)$ of a p^{th} order all-pole filter a_k may be calculated at angular frequency ω_n according to:

$$\hat{P}(\omega) = \frac{1}{\left|\sum_{k=0}^{p} a_k e^{-j\omega k}\right|^2} \tag{B.2}$$

Note that for all-pole filters determined by linear prediction, $a_0 = 1$.

The all-pole filter yielding the minimum E_{LP} is determined by solving the normal equations (Makhoul, 1975), which are solved according to:

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r} \tag{B.3}$$

where

$$\mathbf{a} = [a_1 \; a_2 \; \cdots \; a_p]^T \tag{B.4}$$

$$\mathbf{R} = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \tag{B.5}$$

$$\mathbf{r} = -\,[R_1 \; R_2 \; \cdots \; R_p]^T \tag{B.6}$$

The function R_k is the autocorrelation function corresponding to the signal spectrum $P(\omega_n)$. It is calculated:

$$R_k = \frac{1}{N} \sum_{n=1}^{N} P(\omega_n) \cos(k\omega_n) \tag{B.7}$$

Due to its Toeplitz symmetry, the inversion of the autocorrelation matrix $\mathbf{R}$ can be efficiently performed (Musicus, 1988). The spectral linear prediction algorithm is summarised in Algorithm 2.

Algorithm 2: Spectral linear prediction algorithm.

```
Input: Analysis signal, $s[n]$, and filter order, $p$
Output: Optimal linear predictive filter, $a_k$
Determine the power spectral samples $P(\omega_n)$ of signal $s[n]$ using e.g. a
sinusoidal model of the signal (see Appendix C);
foreach $k$ from 0 to $p$ do
    Compute $R_k$ using Eq. B.7;
Assemble $\mathbf{R}$ and $\mathbf{r}$;
Determine $a_k$ by solving Eq. B.3, using e.g. the Levinson algorithm (Musicus,
1988);
```

B.2 Discrete All-Pole Modeling

Unfortunately, the criterion given by Equation B.1 contains an error cancelation property due to the aliasing in the autocorrelation domain which occurs due to spectral sampling (El-Jaroudi and Makhoul, 1991). Performance is particularly degraded when spectral sampling is low, *i.e.* in the case of high pitched voices. In these cases, the filters estimated by the linear prediction approach tend towards the positions of the harmonics, see Figure B.1.

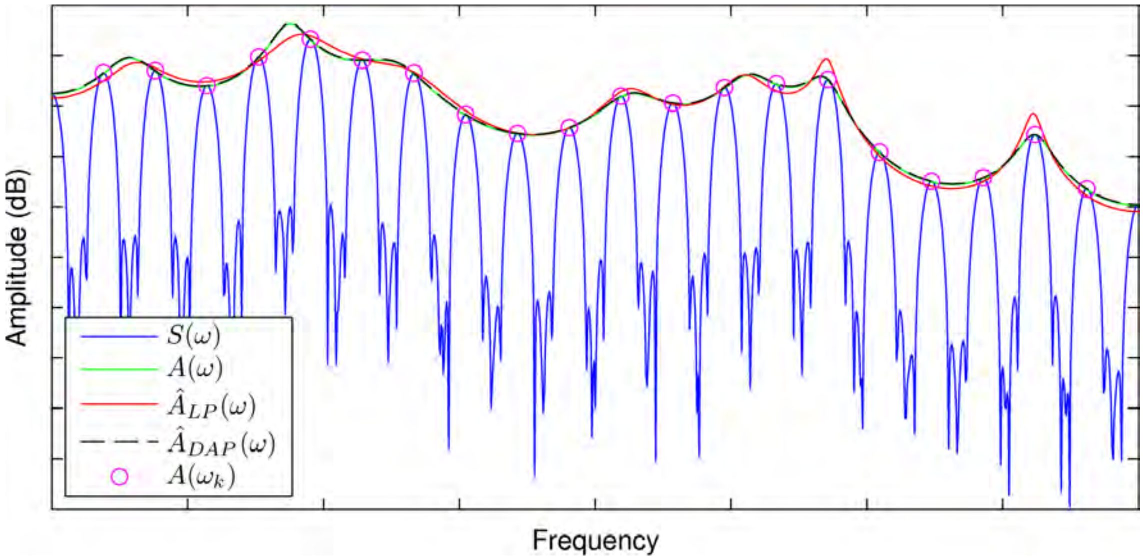


Figure B.1: The above figure gives the log magnitude spectrum of a Hann-windowed periodic all-pole signal, and three spectral envelopes: the actual (green), LP-estimated (red) and DAP-estimated (dashed black). The DAP algorithm is able to recover an all-pole envelope more similar to the original than linear prediction.

The DAP approach (El-Jaroudi and Makhoul, 1991) obtains a more accurate all-pole filter estimate by refining the linear prediction filter using an iterative algorithm and a different error criterion. The new error criterion is the discretised Itakura-Saito error function which was introduced in (Itakura and Saito, 1968) and followed from the estimation of short-time speech spectra using all-pole modeling. The error has been qualified as a "subjectively meaningful measure of speech distortion" (Gray *et al.*, 1980).

Given two power spectra $P(\omega_n)$ and $\hat{P}(\omega_n)$ defined at a set of N discrete frequencies ω_n, the Itakura-Saito error E_{IS} is calculated according to:

$$E_{IS} = \frac{1}{N} \sum_{n=1}^{N} \frac{P(\omega_n)}{\hat{P}(\omega_n)} - \ln \frac{P(\omega_n)}{\hat{P}(\omega_n)} - 1 \tag{B.8}$$

The value of E_{IS} is always non-negative and is only zero in the case where $P(\omega_n)$ equals $\hat{P}(\omega_n)$ for all frequencies ω_n. This function has the additional property that it is amenable to frequency dependent weighting (El-Jaroudi and Makhoul, 1991).

In order to determine an all-pole filter which minimises this error function, (El-Jaroudi and Makhoul, 1991) exploits the property that the autocorrelation function $\hat{R}_k$ of an all-pole filter $\hat{a}_k$ relates to its time-reversed impulse response $\hat{h}[-i]$ by the following equation:

$$\sum_{k=0}^{p} \hat{a}_k \hat{R}_{i-k} = \hat{h}[-i] \;\; \forall\, i \tag{B.9}$$

This equation holds when $\hat{R}_k$ is the true autocorrelation function of $\hat{a}_k$, calculated using Equations B.7 and B.2. However, given a desired autocorrelation function, Equation B.9 can be used to refine an all-pole filter estimate.

Re-arranging Equation B.9 in order to solve for $\hat{a}_k$, substituting $\hat{R}_k$ by R_k and restating in matrix form yields:

$$\hat{\mathbf{a}} = \mathbf{R}^{-1}\hat{\mathbf{h}} \tag{B.10}$$

where $\mathbf{R}$ is defined as in Section B.1, and

$$\hat{\mathbf{a}} = [a_0 \; a_1 \;\; \cdots \;\; a_p]^T \tag{B.11}$$

$$\hat{\mathbf{h}} = [h[0] \; h[-1] \;\; \cdots \;\; h[-p]]^T \tag{B.12}$$

Thus, given any time-reverse impulse response $\hat{\mathbf{h}}$ and an autocorrelation function $\mathbf{R}$, the corresponding all-pole filter can be determined. In (El-Jaroudi and Makhoul, 1991), an iterative algorithm is proposed to determine a solution which terminates when E_{IS} falls below a threshold τ_{IS}, which is given below in Algorithm 3.

Algorithm 3: DAP algorithm.

```
Input: Analysis signal, $s$, filter order, $p$, and threshold $\tau_{IS}$
Output: Optimal linear predictive filter, $a_k$
Determine the power spectral samples $P(\omega_n)$ of signal $s$;
Calculate LP all-pole filter estimate $a_k$ using Algorithm 2;
Initialise $E_{IS}$ to $\infty$;
while $E_{IS} > \tau_{IS}$ do
    Determine $\hat{\mathbf{h}}$ using Eq. B.9;
    Calculate new $\hat{\mathbf{a}}$ by solving Eq. B.10;
    Sample the power spectrum of $\hat{\mathbf{a}}$ to determine $\hat{P}(\omega_n)$ using Eq. B.2;
    Calculate $E_{IS}$ according to Eq. B.8;
Normalise $\hat{\mathbf{a}}$ by dividing all coefficients by $a_0$ to give $a_k$;
```

Appendix C

Sinusoidal Model Parameterisation

Based upon Fourier theory, the modeling of signals as a sum of sinusoidal components is a powerful representation which can facilitate a great many applications. This appendix describes two methods the amplitudes, frequencies and phases of a signal's component sinusoids may be estimated. The first method is based upon peak picking the magnitude spectrum of the analysis signal, while the other reformulates the problem into a linear system, the solution of which minimises the energy between the analysis signal and its sinusoidal model.

C.1 Discrete Fourier Transform

The DFT transforms the time-domain signal $s[n]$ of N equally-spaced time samples into an frequency-domain representation of N equally-spaced frequency bins. The k^{th} frequency bin $S[k]$ is calculated according to the equation:

$$S[k] = \sum_{n=0}^{N-1} w[n]s[n]e^{-i\frac{2\pi k}{N}n} \tag{C.1}$$

where $w[n]$ is an appropriate window function and k is an integer in the range $0 \leq k \leq N-1$. Each frequency bin $S[k]$ is in general a complex number containing both magnitude and phase information of the underlying sinusoidal basis function.

The time-domain periodicity of sinusoidal components in the analysis frame will manifest itself as a spectral peak in the magnitude frequency domain at the frequency of the sinusoid. Thus, the location of the bin where the peak is found gives an estimate of the sinusoid's frequency. Its amplitude and phase are given by the magnitude and angle of peak bin's complex amplitude (McAulay and Quatieri, 1986).

However, as the frequency resolution of the transform is related to the length of the analysis frame, the certain frequency components, particularly low frequency components, may not have been satisfactorily resolved. In order to increase the frequency resolution, the spectrum can be interpolated by zero-padding the time-domain signal before frequency-domain transformation. Furthermore, increased resolution can be provided by parabolically interpolating the log magnitude spectrum and linearly interpolating the unwrapped phase spectrum (Serra, 1989). Figure C.1 shows an typical example.

Finally it is noted that while this simple approach is straightforward, not all peaks of the magnitude spectrum can be classified as sinusoids and some will be attributable to noise or other signal components such as windowing artifacts. In order to counter this ambiguity, a more sophisticated peak pick schemes can be employed, *e.g.* the sinusoidal likeness measure (Rodet, 1997).

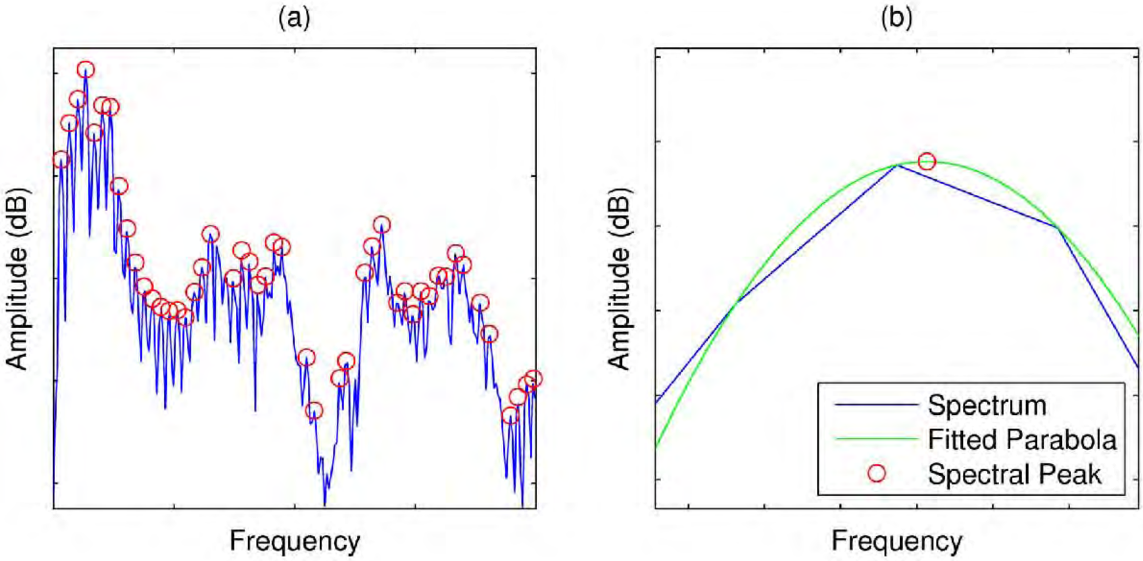


Figure C.1: The above figures show (a) a log-magnitude spectrum of a voiced speech frame with the spectral peaks highlighted in red and (b) the effects of quadratic interpolation for refining an estimate of a single spectral peak.

C.2 Least-Squares Analysis

In the previous section, the parameters of a signal's sinusoidal components were estimated by searching for the spectral peaks in the magnitude spectrum. The magnitude domain search is required because the frequencies of the sinusoidal components are unknown. However, if these frequencies are given *a priori*, a least-squares analysis can be performed in order to locate them more accurately (Laroche *et al.*, 1993; Stylianou, 1996).

It is assumed that the signal under analysis $s[n]$ can be approximated by a sinusoidal model, $\hat{s}[n]$. The time-domain manifestation of $\hat{s}[n]$ is given by:

$$\hat{s}[n] = \sum_{k=-L}^{L} A_k e^{i\omega_k n} \tag{C.2}$$

where w_k are the angular frequencies of the sinusoidal components, L the number of components and A_k is the complex amplitude of the k^{th} sinusoid, containing both amplitude and phase information. The angular frequencies ω_k are assumed to be known, *e.g.* by imposing a harmonic model and determining a maximum voiced frequency (Stylianou, 1996) or some other method.

An expression for the modeling error $e[n]$ is then constructed:

$$e[n] = (s[n] - \hat{s}[n]) \tag{C.3}$$

Summing the energy of this error signal over a given analysis window yields the equation:

$$\begin{aligned}\sum_{n=-N}^{N} e[n]^2 &= \sum_{n=-N}^{N} w[n]^2 \left(s[n] - \hat{s}[n]\right)^2 \\ &= \sum_{n=-N}^{N} w[n]^2 \left(s[n] - \sum_{k=-L}^{L} A_k e^{i\omega_k n}\right)^2 \end{aligned} \tag{C.4}$$

where $w[n]$ is an appropriate window function and $2N+1$ is the window length.

Equation C.4 defines a system of linear equations, the least-squares solution of which is given by:

$$\mathbf{A} = \left(\mathbf{E_0}^T\mathbf{W}^T\mathbf{W}\mathbf{E_0}\right)^{-1}\mathbf{E_0}^T\mathbf{W}^T\mathbf{W}\mathbf{s} \tag{C.5}$$

$$= \mathbf{R}^{-1}\mathbf{b} \tag{C.6}$$

where

- $\mathbf{A}$ is a $(2L+1)\times 1$ vector containing the sought sinusoidal complex amplitudes,
- $\mathbf{E_0}$ is a $(2N+1)\times(2L+1)$ matrix where each element on the n^{th} row and k^{th} column $E_0^{n,k}$ is given by

$$E_0^{n,k} = e^{i\omega_k n} \tag{C.7}$$

- $\mathbf{W}$ is a $(2N+1) \times (2N+1)$ matrix containing the window function $w[n]$ across the main diagonal, and
- $\mathbf{s}$ is the $(2N+1) \times 1$ vector containing the analysis frame $s[n]$.

In the case where the set of frequencies ω_k are harmonically related, the matrix $\mathbf{R}$ exhibits Toeplitz symmetry and can be efficiently inverted (Musicus, 1988).

Appendix D

Generation of the LF Model Pulse

The LF model is utilised throughout this work as a model of the derivative glottal flow signal. This appendix discusses the generation of the model in the time and frequency domains.

Firstly, the LF model formulation is re-stated, and how the time-domain waveform is constructed from the timing parameters is explained. It is illustrated how improper sampling of the LF model pulse in the time domain can introduce significant distortions into the waveform. Attempts to solve this problem by quantising the timing parameters to integer values produce undesirable properties upon the error functions necessary for optimisation routines. An alternative method based upon the shifting of the time samples which produces smoother error functions is described.

Secondly, the computation efficiency of three methods to estimate the frequency-domain information of the LF model is discussed. The fast determination of the frequency-domain parameters are important for the PowRd algorithm, described in Chapter 6, which undertakes a brute force initialisation and optimisation procedure which demands the determination many LF model spectra. For this, an analytical expression for the spectral information of the LF model is derived. An informal experiment shows that the new method can determine the exact solution for any frequency with increased computational efficiency over a phasor correlation approach. An alternative method using the Fast Fourier Transform (Cooley and Tukey, 1965) and spectral interpolation the spectrum is also proposed, which further reduces the computational time, yet provides an approximate solution.

D.1 Time-Domain Generation of the LF Model

The LF model is formulated in the time domain as a piecewise mathematical function:

$$u_{LF}(n) = \begin{cases} E_0 e^{\alpha n} \sin \omega_g n & 0 \leq n < T_e \\ \frac{-E_e}{\epsilon T_a}\left(e^{-\epsilon(n-T_e)} - e^{-\epsilon(T_c-T_e)}\right) & T_e \leq n \leq T_c \\ 0 & T_c \leq n < T_0 \end{cases} \tag{D.1}$$

When T_a is small relative to $T_c - T_e$, the return phase of the signal is closely approximated by an decreasing exponential curve. As this function returns asymptotically to zero, it is convenient to combine the return and closed phases together such that T_c coincides with T_0. Under this formulation, both the return and closed phase can be very closely approximated as the truncated impulse response of a single positive real pole IIR filter. This arrangement is a suitable approximation for many voice types, and relates the filter pole position μ_{ret} to the T_a parameter by the equation $T_a = \frac{-1}{\ln \mu_{ret}}$ (Ó Cinnéide *et al.*, 2010), which can be related to the $TL(z)$ filter of the KLGLOTT88 model.

As mentioned in Section 2.3.3.1, the LF model pulse is generated in the time domain using its direct synthesis parameters $\alpha, \omega_g, \epsilon, E_0$ which are calculated from its time parameters T_p, T_e, T_p, T_c and scale parameter E_e. The direct parameters ω_g and E_0 are determined by the following identities:

$$\omega_g = \frac{\pi}{T_p} \tag{D.2}$$

$$E_0 = \frac{-E_e}{e^{\alpha T_e} \sin \omega_g T_e} \tag{D.3}$$

while ϵ and α are given by the nonlinear equations:

$$\epsilon T_a = 1 - e^{-\epsilon(T_c - T_e)} \tag{D.4}$$

$$\sum_{n=0}^{T_c} u_{LF}[n] = 0 \tag{D.5}$$

This second requirement is sometimes referred to the area balance of the LF model, as the area of the return phase equals the area of the open phase, ensuring that there no baseline drift over the course of pulse cycle. An algorithm to determine these parameters is given below.

Algorithm 4: Algorithm to determine the LF model direct synthesis parameters from its timing parameters.

```
**Input**: LF model timing parameters $T_p, T_e, T_p, T_c$ and scale parameter $E_e$
**Output**: LF model direct synthesis parameters: $\alpha, \omega_g, \epsilon, E_0$
Calculate $\omega_g$ using Eq. D.2;
Eq. D.4 is solved using e.g. Newton's method to give $\epsilon$;
Using $E_e = 1$, solve Eq. D.5 using a root finding algorithm;
Calculate $E_0$ following Eq. D.3;
```

This algorithm dictates that the LF pulse be sampled at various time points. However, depending on the value of the pulse's timing parameters, the breakpoints defining the boundaries of the different phases of the glottal cycle may fall between sample points therefore cause waveform discontinuities, as shown in Figure D.1. These discontinuities may introduce audible artifacts.

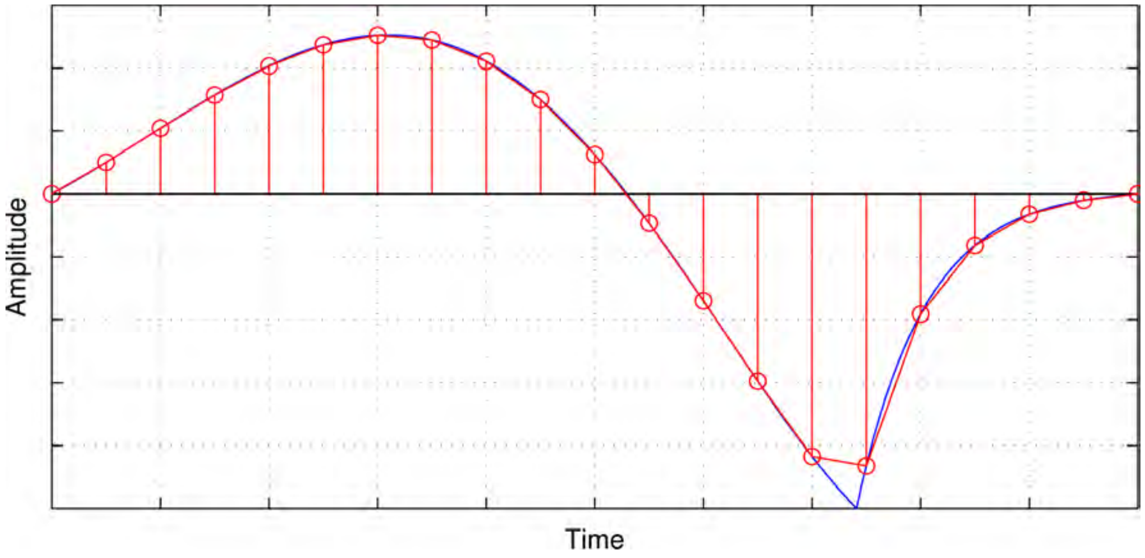


Figure D.1: The above figure shows how sampling the LF model waveform can introduce discontinuities in the waveform around the breakpoints of the signal. In the above scenario, a large discontinuity is introduced around the instant of maximum negative amplitude.

This kind of distortion is inherent when using digital system where time segments can only be expressed in integer sample lengths. In order to avoid introducing discontinuities, the timing parameters of the model which define non-integer length segments can be snapped to the nearest integer value. While this may be acceptable for LF model generation as the quantisation effect is unlikely to be perceptual (Henrich *et al.*, 2003), it causes problems for LF model fitting algorithms as it produces a "staircase" error function, which may have the consequence that the optimisation algorithm becomes stuck in a local minimum (Strik, 1998), as shown in Figure D.2.

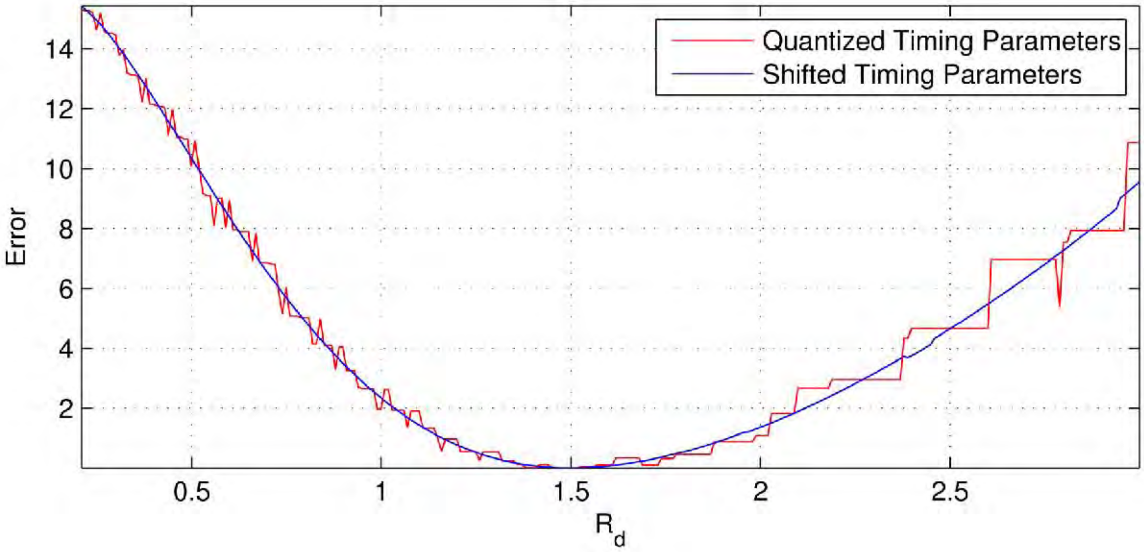


Figure D.2: The above figure shows the error surface of two LF model fitting techniques. The first uses an LF model generating routine which quantises the timing parameters which results in an undesirable staircase-like error function. This may cause the optimisation program to become stuck in a local minimum. The other utilises a routine based upon the timing parameter shifting described in the text, which produces a smoother error function and is therefore more likely to converge to the global minimum.

An alternative solution is one which accepts the inherent fact of this distortion and attempts to minimise it. The discontinuities introduced by incorrectly sampling the LF model waveform occur when there is a large amplitude difference between successive signal points. Thus, the greatest distortion is introduced when the signal changes from between its open and return phases, in the interval of the cycle where it reaches its maximum negative amplitude. By ensuring that this important instant is captured with a sample point, the inherent discontinuities of the waveform are effectively shifted to the fringes of the pulse, at instants T_o and T_c, where the signal amplitude is small and unlikely to introduce as large amplitude discontinuities. The "staircase" phenomenon associated with parameterisation quantisation is eliminated (see Figure D.2 in blue).

D.2 Frequency-Domain Generation of the LF Model

The PowRd method described in Chapter 6 necessitates the calculation of the frequency-domain information of the LF model at specific frequencies. This information can be obtained from the time-domain LF model pulse by correlation of the pulse with a complex phasor signal at the desired frequency. However, because of the formulation of the LF model, the problem can be re-expressed as a set geometric summations, which can be solved more efficiently.

Frequency information can also be determined from the FFT of the pulse. However, the FFT determines this information at fixed frequencies equally spaced across the signal bandwidth. In order to determine frequency information of specific frequencies, interpolating procedures can be utilised. This method offers faster performance at determining spectral information than the geometrical sum method, though at the expense of decreased accuracy.

D.2.1 Phasor Correlation

The complex amplitude of a sinusoidal component with angular frequency ω of an N length signal x can be obtained by correlating the signal with a phasor at that frequency:

$$X(\omega) = \sum_{n=0}^{N-1} x(n)e^{-i\omega n} \tag{D.6}$$

Thus, for a single LF model pulse T_0 samples in length, this sinusoidal component is given by:

$$U_{LF}(\omega) = \sum_{n=0}^{T_0-1} u_{LF}(n)e^{-i\omega n} \tag{D.7}$$

The phases values of $U_{LF}(\omega)$ are dependent on the time positions of the complex exponential basis functions, and are therefore a function of n. For the purposes of generality, Equation D.7 can be multiplied by another complex exponential representing a general phase shift of n_0 samples. This alters the summation above to the following:

$$U_{LF}(\omega) = e^{-i\omega n_0} \sum_{n=0}^{T0-1} u_{LF}(n)e^{-i\omega n} \tag{D.8}$$

This shift may be necessary if it is desirable that the zero point align with an alternative time reference than T_o, *e.g.* T_e (Degottex *et al.*, 2011).

D.2.2 Geometric Summation

Equation D.8 takes into account the entire cycle of the glottal pulse, but the signal can also be separated into its different phases:

$$U_{LF}(\omega) = e^{-i\omega n_0} \sum_{n=0}^{T_0-1} u_{LF}(n) e^{-i\omega n} \tag{D.9}$$

$$= e^{-i\omega n_0} \sum_{n=0}^{T_e} u_{LF}^{open}(n) e^{-i\omega n} + e^{-i\omega n_0} \sum_{n=T_e+1}^{T_0-1} u_{LF}^{ret}(n) e^{-i\omega n} \tag{D.10}$$

where $u_{LF}^{open}(n)$ and $u_{LF}^{ret}(n)$ refer to the open and return phase, respectively. The sinusoidal component $U_{LF}(\omega)$ can then be calculated as the sum of the separate contribution of each phase of the model, which can be reformulated as the scaled sum of geometric summations.

The Open Phase By utilising the relationship

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \tag{D.11}$$

the open phase portion of the equation can also be rewritten as a sum of complex exponentials:

$$E_0 e^{\alpha n} \sin \omega_g n = E_0 e^{\alpha n} \frac{e^{i\omega_g n} - e^{-i\omega_g n}}{2i} \tag{D.12}$$

$$= \frac{E_0}{2i} (e^{\alpha n + i\omega_g n} - e^{\alpha n - i\omega_g n}) \tag{D.13}$$

$$= \frac{E_0}{2i} (e^{n(\alpha + i\omega_g)} - e^{n(\alpha - i\omega_g)}) \tag{D.14}$$

The expression for $U_{LF}^{open}(\omega)$ can therefore be expressed and simplified.

$$\begin{aligned} U_{LF}^{open}(\omega) &= e^{-i\omega n_0}\sum_{n=0}^{T_e} u_{LF}^{open}(n)e^{-i\omega n} && \text{(D.15)} \\ &= \frac{E_0}{2i}e^{-i\omega n_0}\sum_{n=0}^{T_e}\left(e^{n(\alpha+i\omega_g)} - e^{n(\alpha-i\omega_g)}\right)e^{-i\omega n} && \text{(D.16)} \\ &= \frac{E_0}{2i}e^{-i\omega n_0}\sum_{n=0}^{T_e}(e^{n(\alpha+i\omega_g)-i\omega n} - e^{n(\alpha-i\omega_g)-i\omega n}) && \text{(D.17)} \\ &= \frac{E_0}{2i}e^{-i\omega n_0}\left(\sum_{n=0}^{T_e} e^{n(\alpha+i(\omega_g-\omega))} - \sum_{n=0}^{T_e} e^{n(\alpha-i(\omega_g+\omega))}\right) && \text{(D.18)} \end{aligned}$$

The complex amplitude of the sinusoidal component of $U_{LF}^{open}(\omega)$ is now re-expressed as the scaled sum of two geometric series. The sum of a geometric series can be obtained analytically by the following equation:

$$\sum_{n=0}^{N} ar^n = a\frac{1-r^{N+1}}{1-r} \qquad \text{(D.19)}$$

Equation D.19 can be applied to the Equation D.18 to give:

$$U_{LF}^{open}(\omega) = \frac{E_0}{2i}e^{-i\omega n_0}\left(\frac{1-e^{(T_e+1)(\alpha+i(\omega_g-\omega))}}{1-e^{(\alpha+i(\omega_g-\omega))}} - \frac{1-e^{(T_e+1)(\alpha-i(\omega_g+\omega))}}{1-e^{(\alpha-i(\omega_g+\omega))}}\right) \qquad \text{(D.20)}$$

The Return/Closed Phase $U_{LF}^{ret}(\omega)$ can be calculated in a similar fashion as the open phase component.

$$U_{LF}^{ret}(\omega) = e^{-i\omega n_0} \sum_{n=T_e+1}^{T_0-1} u_{LF}^{ret}(n+T_e) e^{-i\omega n} \tag{D.21}$$

$$= e^{-i\omega n_0} \sum_{n=T_e+1}^{T_0-1} \frac{-E_e}{\epsilon T_a} \left(e^{-\epsilon(n-T_e)} - e^{-\epsilon(T_c-T_e)} \right) e^{-i\omega n} \tag{D.22}$$

$$= e^{-i\omega n_0} \frac{-E_e}{\epsilon T_a} \sum_{n=T_e+1}^{T_0-1} e^{-\epsilon(n-T_e)-i\omega n} - e^{-\epsilon(T_c-T_e)} e^{-i\omega n} \tag{D.23}$$

$$= e^{-i\omega n_0} \frac{-E_e}{\epsilon T_a} \sum_{n=T_e+1}^{T_0-1} e^{n(-\epsilon-i\omega)+\epsilon T_e} - e^{-\epsilon(T_c-T_e)} e^{-i\omega n} \tag{D.24}$$

$$= e^{-i\omega n_0} \frac{-E_e}{\epsilon T_a} \left(e^{\epsilon T_e} \sum_{n=T_e+1}^{T_0-1} e^{n(-\epsilon-i\omega)} - e^{-\epsilon(T_c-T_e)} \sum_{n=T_e+1}^{T_0-1} e^{-i\omega n} \right) \tag{D.25}$$

In order to determine the analytic result, the limits of each summation are adjusted and D.19 is applied.

$$U_{LF}^{ret}(\omega) = e^{-i\omega n_0} \frac{-E_e}{\epsilon T_a} \left(e^{\epsilon T_e} \sum_{n=0}^{T_0-T_e-2} e^{(n+(T_e+1))(-\epsilon-i\omega)} - e^{-\epsilon(T_c-T_e)} \sum_{n=0}^{T_0-T_e-2} e^{-i\omega(n+(T_e+1))} \right) \tag{D.26}$$

$$= e^{-i\omega n_0} \frac{-E_e}{\epsilon T_a} \left(e^{\epsilon T_e+(T_e+1)(-\epsilon-i\omega)} \sum_{n=0}^{T_0-T_e-2} e^{n(-\epsilon-i\omega)} - e^{-\epsilon(T_c-T_e)-i\omega(T_e+1)} \sum_{n=0}^{T_0-T_e-2} e^{-i\omega n} \right) \tag{D.27}$$

$$= e^{-i\omega n_0} \frac{-E_e}{\epsilon T_a} \left(e^{-\epsilon-i\omega(T_e+1)} \frac{1-e^{(T_0-T_e-1)(-\epsilon-i\omega)}}{1-e^{(-\epsilon-i\omega)}} - e^{-\epsilon(T_c-T_e)-i\omega(T_e+1)} \frac{1-e^{-i\omega(T_0-T_e-1)}}{1-e^{-i\omega}} \right) \tag{D.28}$$

D.2.3 FFT/Interpolation

A third approach to determining frequency-domain information of the LF model pulse exploits the computation efficiency of the FFT algorithm. The FFT algorithm computes the DFT of a signal more efficiently that the usual correlation method by exploiting certain redundancies.

For an N-point signal $x[n]$, the DFT is defined:

$$X[\omega_k] = \sum_{n=0}^{N-1} x[n]e^{-i\omega_k n} \tag{D.29}$$

where

$$\omega_k = \frac{2\pi k}{N}, \quad 0 \leq k \leq N-1 \tag{D.30}$$

However, as the complex amplitudes $X[\omega_k]$ are available only for the harmonically related frequencies ω_k, an estimate of the value of the complex amplitude of the sinusoid with general angular frequency ω can only be approximated by interpolation. The errors introduced by this approximation can be reduced by increasing the frequency resolution of the spectrum via zero-padding, but this operation also increases the computational load.

D.2.4 Computational Comparison

An experiment was performed to compare the computational efficiency of the three different of methods of determining the frequency information of the LF model pulse. One thousand frequency points were generated randomly and determined from the parameters of 1000 LF pulses. Table D.1 below summaries the results of the experiment.

The results of the experiment clearly show that the method based upon phasor correlation gives the slowest results and that both other methods offer large improvements in computational efficiency - the geometric-sum-based method offering the exact solution with an computational decrease of 85%, and the FFT-interpolation-based

Table D.1: The table below gives the computational load of different LF model frequency-domain estimation techniques, relative to the correlation-based method.

Method	Relative Computation Time
Correlation	1.00
Geometric Sum	0.14
FFT/Interpolation	0.04

method giving a 96% improvement, with an approximate solution.

Appendix E

Results of Synthetic Speech Perceptual Experiment

This appendix contains the full results from the listening test described in Chapter 8 in Tables E.1 to E.3. A positive score indicates a preference for the synthetic speech segment produced using the techniques proposed in this thesis. Additionally, for reasons discussed in Chapter 8, the results of Comparisons 3, 4, 11 and 14 were ultimately removed from the final analysis.

Table E.1: The table below gives the full listener preference scores of the perceptual comparison of the analysis/re-synthesis of speech recorded using linear phase equipment.

Speaker, Sentence	Percentage	Score	Speaker, Sentence	Percentage	Score
bdl, Arctic 1	14.00%	3	*jmk, Arctic 1*	8.51%	-3
	12.00%	2		8.51%	-2
	16.00%	1		10.64%	-1
	4.00%	0		6.38%	0
	6.00%	-1		14.89%	1
	34.00%	-2		29.79%	2
	14.00%	-3		21.28%	3
bdl, Arctic 2	6.00%	-3	*jmk, Arctic 2*	2.08%	-3
	16.00%	-2		4.17%	-2
	12.00%	-1		8.33%	-1
	6.00%	0		6.25%	0
	24.00%	1		18.75%	1
	24.00%	2		35.42%	2
	12.00%	3		25.00%	3
bdl, Arctic 3	28.00%	3	*jmk, Arctic 3*	34.69%	3
	22.00%	2		28.57%	2
	22.00%	1		24.49%	1
	8.00%	0		4.08%	0
	6.00%	-1		2.04%	-1
	12.00%	-2		0.00%	-2
	2.00%	-3		6.12%	-3
bdl, Arctic 4	2.04%	-3	*jmk, Arctic 4*	34.69%	3
	6.12%	-2		26.53%	2
	2.04%	-1		16.33%	1
	4.08%	0		14.29%	0
	18.37%	1		2.04%	-1
	28.57%	2		2.04%	-2
	38.78%	3		4.08%	-3
bdl, Arctic 5	2.08%	3	*jmk, Arctic 5*	12.00%	3
	12.50%	2		16.00%	2
	31.25%	1		30.00%	1
	29.17%	0		24.00%	0
	6.25%	-1		12.00%	-1
	6.25%	-2		4.00%	-2
	12.50%	-3		2.00%	-3

Table E.2: The table below gives the full listener preference scores of the perceptual comparison of the analysis/re-synthesis of speech when simulating nonlinear phase equipment.

Speaker, Sentence	Percentage	Score	Speaker, Sentence	Percentage	Score
	6.25%	3		20.00%	3
	22.92%	2		26.00%	2
	16.67%	1		32.00%	1
bdl, Arctic 1	33.33%	0	*jmk, Arctic 1*	10.00%	0
	16.67%	-1		6.00%	-1
	4.17%	-2		4.00%	-2
	0.00%	-3		2.00%	-3
	2.00%	-3		2.13%	-3
	4.00%	-2		8.51%	-2
	10.00%	-1		14.89%	-1
bdl, Arctic 2	10.00%	0	*jmk, Arctic 2*	6.38%	0
	22.00%	1		23.40%	1
	28.00%	2		25.53%	2
	24.00%	3		19.15%	3
	14.00%	3		32.00%	3
	28.00%	2		34.00%	2
	30.00%	1		20.00%	1
bdl, Arctic 3	10.00%	0	*jmk, Arctic 3*	4.00%	0
	10.00%	-1		4.00%	-1
	4.00%	-2		4.00%	-2
	4.00%	-3		2.00%	-3
	28.00%	3		14.00%	3
	34.00%	2		42.00%	2
	22.00%	1		26.00%	1
bdl, Arctic 4	10.00%	0	*jmk, Arctic 4*	12.00%	0
	2.00%	-1		0.00%	-1
	2.00%	-2		4.00%	-2
	2.00%	-3		2.00%	-3
	5.88%	3		0.00%	-3
	9.80%	2		2.00%	-2
	21.57%	1		6.00%	-1
bdl, Arctic 5	33.33%	0	*jmk, Arctic 5*	4.00%	0
	9.80%	-1		30.00%	1
	15.69%	-2		38.00%	2
	3.92%	-3		20.00%	3

Table E.3: The table below gives the full listener preference scores of the perceptual comparison of the analysis/re-synthesis of speech recorded using non-linear phase equipment.

Speaker, Sentence	Percentage	Score	Speaker, Sentence	Percentage	Score
Male, Sentence 1	2.04%	-3	*Female, Sentence 1*	14.00%	3
	4.08%	-2		28.00%	2
	4.08%	-1		32.00%	1
	12.24%	0		22.00%	0
	34.69%	1		2.00%	-1
	28.57%	2		2.00%	-2
	14.29%	3		0.00%	-3
Male, Sentence 2	2.00%	3	*Female, Sentence 2*	22.00%	3
	8.00%	2		26.00%	2
	16.00%	1		24.00%	1
	38.00%	0		20.00%	0
	16.00%	-1		6.00%	-1
	16.00%	-2		0.00%	-2
	4.00%	-3		2.00%	-3
Male, Sentence 3	4.00%	3	*Female, Sentence 3*	0.00%	-3
	10.00%	2		2.08%	-2
	18.00%	1		2.08%	-1
	38.00%	0		10.42%	0
	22.00%	-1		43.75%	1
	6.00%	-2		29.17%	2
	2.00%	-3		12.50%	3
Male, Sentence 4	2.00%	-3	*Female, Sentence 4*	14.00%	3
	10.00%	-2		30.00%	2
	8.00%	-1		26.00%	1
	28.00%	0		14.00%	0
	16.00%	1		10.00%	-1
	28.00%	2		4.00%	-2
	8.00%	3		2.00%	-3
Male, Sentence 5	8.00%	3	*Female, Sentence 5*	12.00%	3
	12.00%	2		38.00%	2
	22.00%	1		22.00%	1
	14.00%	0		18.00%	0
	24.00%	-1		8.00%	-1
	14.00%	-2		2.00%	-2
	6.00%	-3		0.00%	-3