ESTIMATION OF THE GLOTTAL PULSE
FROM SPEECH OR SINGING VOICE

Sandra de Oliveira Dias

Monograph
Masters Degree in Biomedical Engineering

July, 2011
ESTIMATION OF THE GLOTTAL PULSE
FROM SPEECH OR SINGING VOICE

Sandra de Oliveira Dias

Mathematics Education Degree by University of Minho (2003)

Supervisor: Prof. Dr. Aníbal João de Sousa Ferreira

Assistant Professor of the Department of Electrical and Computer Engineering
School of Engineering of the University of Porto
Abstract

The human speech production system is, briefly, the result of the convolution between the excitation signal, the glottal pulse, and the impulse response resulting from the transfer function of the vocal tract. This model of voice production is often referred to in literature as a source-filter model, where the source represents the flow of the air leaving the lungs and passing through the glottis (space between the vocal folds), and the filter representing the resonances of the vocal tract and lip radiation.

The estimation of the shape of the glottal pulse from the speech signal is of significant importance in many fields and applications, since all the features of speech related to voice quality, vocal effort and speech disorders, for example, are mainly due to the voice source, although the glottal flow waveform is a very difficult signal to measure directly and non-invasively.

Several methods for estimating the glottal pulse have been proposed over the last decades, but there is not yet a complete and automatic algorithm. Most of the developed methods are based on a process called inverse filtering. The inverse filtering represents the deconvolution, i.e., it seeks to obtain the input signal by applying the inverse of the vocal tract transfer function to the output signal. Despite the simplicity of the concept, the inverse filtering procedure is not simple because the output signal may include noise and it is not linear to accurately model the characteristics of the vocal tract filter.

This work presents an anatomical and physiological study of the human speech production system and an analysis of some glottal flow models and different approaches for estimating the glottal pulse. Some of the processes of estimation of the glottal pulse are applied to speech signals (real and synthetic), and the respective results are presented, compared and discussed.

Key words: glottal pulse; glottal model; estimation of the glottal pulse; inverse filtering.
Resumo

O processo de produção humana de voz é, resumidamente, o resultado da convolução entre o sinal de excitação, o impulso glótico, com a resposta impulsiva resultante a função de transferência do tracto vocal. Este modelo de produção de voz é frequentemente referido na literatura como um modelo fonte-filtro, em que a fonte representa o fluxo de ar que sai dos pulmões e passa pela glote, fazendo vibrar as pregas vocais, e o filtro representando as ressonâncias do tracto vocal e a radiação labial.

Estimar a forma do impulso glótico a partir do sinal de voz é de importância significativa em diversas áreas e aplicações, uma vez que as características de voz relacionadas, por exemplo, com a qualidade da voz, esforço vocal e distúrbios da voz, devem-se, principalmente, ao fluxo glotal. No entanto, este fluxo é um sinal difícil de determinar de forma directa e não invasiva.

Ao longo das últimas décadas foram desenvolvidos vários métodos para estimar o impulso glótico mas ainda não existe um algoritmo completo e automático. A maioria dos métodos desenvolvidos baseia-se num processo designado por filtragem inversa. A filtragem inversa representa, portanto, a desconvolução, ou seja, procura obter o sinal de entrada aplicando o inverso da função de transferência do tracto vocal ao sinal de saída. Apesar da simplicidade do conceito, o processo de filtragem inversa não é simples uma vez que o sinal de saída pode incluir ruído e não é linear modelar com precisão as características do filtro do tracto vocal.

Este trabalho apresenta um estudo anatômico e fisiológico do processo humano de produção de voz e alguns modelos de impulso glótico, assim como de diferentes abordagens na estimação deste. Alguns dos processos de estimação do impulso glótico são aplicados em sinais de voz (reais e sintético), fazendo-se uma análise, comparação e discussão dos resultados obtidos.

Palavras chave: *impulso glótico; modelos de impulso glótico; estimação do impulso glótico; filtragem inversa.*
# Contents

Abstract .......................................................................................................................... 5  
List of Figures .................................................................................................................... 9  
Notations ........................................................................................................................... 12  
Acronyms .......................................................................................................................... 13  
1. **INTRODUCTION** ................................................................................................... 14  
   1.1. **OVERVIEW** .................................................................................................. 14  
   1.2. **OBJECTIVES** ............................................................................................ 15  
   1.3. **STRUCTURE OF THE DOCUMENT** .............................................................. 15  
2. **HUMAN SPEECH PRODUCTION SYSTEM** .......................................................... 17  
   2.1. **ANATOMY AND PHYSIOLOGY OF THE VOICE ORGAN** ............................ 17  
   2.2. **GLOTTAL FLOW** ....................................................................................... 20  
      2.2.1. Electroglottography .................................................................................... 21  
      2.2.2. Imaging Techniques .................................................................................. 24  
3. **MODELS OF VOICE PRODUCTION** ..................................................................... 27  
   3.1. **SOURCE-FILTER MODEL** ............................................................................ 27  
   3.2. **EXTRACTION OF CHARACTERISTICS FROM A SPEECH SIGNAL** ....... 33  
      3.2.1. Linear Predictive Coding (LPC) Method .................................................... 34  
      3.2.2. Discrete All-Pole Modelling (DAP) ........................................................... 38  
      3.2.3. Mel-Frequency Cepstral Coefficients (MFCC) Method .............................. 40  
         3.2.3.1. Cepstral Analysis ................................................................................. 40  
         3.2.3.2. Mel scale .............................................................................................. 41  
         3.2.3.3. MFCC calculation process ................................................................... 42  
   3.3. **GLOTTAL FLOW MODELS** ......................................................................... 43  
      3.3.1. Time-domain methods .............................................................................. 43  
      3.3.2. Frequency-domain methods ..................................................................... 45  
      3.3.3. Model-based methods ............................................................................... 46  
         3.3.3.1. Fant Glottal Model .............................................................................. 48  
         3.3.3.2. Liljencrats-Fant Glottal Model (LF Model) .......................................... 49  
         3.3.3.3. Rosenberg Glottal Model .................................................................... 53  
         3.3.3.4. Rosenberg++ Glottal Model ................................................................. 54  
         3.3.3.5. Klett Glottal Model ............................................................................ 55  
         3.3.3.6. Causal-Anticausal Linear Model (CALM) ............................................. 56  
4. **ESTIMATION OF THE GLOTTAL FLOW** ......................................................... 58
4.1. **INVERSE FILTERING TECHNIQUES** ................................................................. 59  
4.1.1. Iterative Adaptive Inverse Filtering (IAIF/PSIAIF) .............................. 60  
4.1.2. Javkin *et al.* Method ........................................................................... 65  
4.2. **ZEROS OF THE Z-TRANSFORM (ZZT) AND COMPLEX CEPSTRUM (CC)** ..................................................... 67  
4.3. **CAUSALITY/ANTICAUSALITY DOMINATED REGIONS** .................... 71  
4.4. **COMPARING RESULTS** ........................................................................... 73  

5. **CONCLUSION** .............................................................................................. 83  
5.1. **OVERVIEW** .............................................................................................. 83  
5.2. **FUTURE WORK** .......................................................................................... 84  

**APPENDIX – FUNDAMENTALS OF DIGITAL SIGNAL PROCESSING** ...................... 85  
- **WINDOW FUNCTIONS** .................................................................................. 85  
- **DISCRETE TIME FOURIER TRANSFORM (DTFT)** ........................................... 86  
- **DISCRETE FOURIER TRANSFORM (DFT)** ..................................................... 86  
- **FAST FOURIER TRANSFORM (FFT)** .............................................................. 86  
- **DISCRETE COSINE TRANSFORM (DCT)** ....................................................... 87  
- **Z-TRANSFORM** ............................................................................................ 87  
- **ZEROS OF THE Z-TRANSFORM** ..................................................................... 88  

**BIBLIOGRAPHY** ............................................................................................. 89
List of Figures

Figure 2.1  Human speech production mechanism. .......................................................... 18
Figure 2.2  Top view of the larynx, showing the positions of the vocal folds. ............... 19
Figure 2.3  Glottal phases. ............................................................................................... 20
Figure 2.4  EGG measurement setting. ............................................................................ 22
Figure 2.5  Schematic representation of a single cycle of vocal fold vibration (above) viewed coronally (left) and superiorly (right), and an EGG of a normal phonation (below). .................................................................................. 22
Figure 2.6  Electroglottogram of the normal phonation of a male subject. ..................... 23
Figure 2.7  Video laryngoscopy. ....................................................................................... 24
Figure 2.8  Kimography. ................................................................................................. 25
Figure 2.9  Videostrobscopy. ........................................................................................... 25
Figure 2.10  Fundamental principle of videostrobscopy. .................................................. 26
Figure 3.1  Schematic illustration of the generation of voice sounds. .............................. 28
Figure 3.2  Block diagram of the source-filter model. .................................................... 29
Figure 3.3  Schematic view of voice production models. .................................................. 30
Figure 3.4  General discrete-time model of speech production. ...................................... 31
Figure 3.5  Source-filter model: the speech signal as a result of the convolution between the excitation source signal, the impulse response of the vocal tract and the lip radiation. ............................................................................. 31
Figure 3.6  Speech signals. .............................................................................................. 33
Figure 3.7  FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the LP method (thick line). ................................................................. 36
Figure 3.8  FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the DAP method (thick line) together with the LP method (dashed line). ................................................................................... 38
Figure 3.9  The Mel scale. ............................................................................................... 42
Figure 3.10  Block diagram of the MFCC calculation process. ........................................ 43
Figure 3.11  Three periods of a sound-pressure waveform and the respective glottal flow and its derivative. .......................................................... 44

Figure 3.12  Flow spectrum. ................................................................. 45

Figure 3.13  Main scheme of the glottal pulse used by most of glottal models. .......... 47

Figure 3.14  Glottal flow model proposed by Fant [Fan79]. ................................. 48

Figure 3.15  The LF model of the glottal flow and the derivative of the glottal flow. ........................................................................................................ 50

Figure 3.16  Linear stylization of the LF-Model spectrum. ..................................... 51

Figure 3.17  Examples of the LF$^{rd}$ model for different $Rd$ values. ...................... 52

Figure 3.18  The Rosenberg model of the glottal flow. ........................................ 53

Figure 3.19  LF and R++ glottal-pulse time derivatives. ....................................... 55

Figure 3.20  The derivative of the Klatt glottal flow model. ................................... 56

Figure 4.1  Schematic spectra of the represented signals and filters. ....................... 58

Figure 4.2  Speech production model used in IAIF. ............................................. 60

Figure 4.3  The bock diagram of the IAIF method for estimation of the glottal excitation from the speech signal. ......................................................... 61

Figure 4.4  The graphical user interface and the signal view of Aparat. ................. 63

Figure 4.5  The spectra view in Aparat of a speech signal, the calculated glottal flow and the used vocal tract filter (left) and parameters computed from the estimated glottal flow. ........................................ 63

Figure 4.6  Block diagram of the main steps of the PSIAIF method. ....................... 63

Figure 4.7  Inverse filtering model by Murphy [Mur08]. ...................................... 64

Figure 4.8  Inverse Filter interface....................................................................... 66

Figure 4.9  Synthesizer interface. ........................................................................ 67

Figure 4.10  Sky interface. .................................................................................. 67

Figure 4.11  Illustration of the source-filter modelling for one period. The Glottal Closing Instant (GCI) allows the separation of glottal open and glottal closed phases, corresponding to the anticausal and causal signals [DDM’08]. ......................................................... 68

Figure 4.12  The model of speech production used on ZZT .................................... 68

Figure 4.13  The ZZT of a speech signal. ............................................................... 69
Figure 4.14  Block diagram of the ZZT decomposition algorithm. ........................................ 70
Figure 4.15  Anticausality-based decomposition of speech using CC. ................................. 71
Figure 4.16  Effect of a sharp CGI-centered windowing on a two-period long speech frame. . 72
Figure 4.17  Glottal flow estimation of a synthetic vowel /a/ using the LF model. .......................... 75
Figure 4.18  Glottal flow estimation of a synthetic vowel /a/. ............................................... 76
Figure 4.19  Glottal flow estimation of a female vowel /a/. .................................................... 77
Figure 4.20  Glottal flow estimation of a female vowel /a/. .................................................... 78
Figure 4.21  Glottal flow estimation of a male vowel /a/. ....................................................... 79
Figure 4.22  Glottal flow estimation of a male vowel /a/. ....................................................... 80
Figure 4.23  Glottal flow estimation of a female vowel /i/. ...................................................... 81
Figure 4.24  Glottal flow estimation of a singing signal of a male voice. ............................... 82
Figure A.1  ZZT representation of a signal in cartesian and polar coordinates. ..................... 88
Notations

\( f_s \) : Sampling frequency of a discrete signal.

\( f_0 \) : Fundamental frequency in Hz of a periodic signal.

\( T_0 = 1 / f_0 \) : Fundamental period.

\( x(t) \) : Continuous signal with respect to time \( t \).

\( x[n] \) : Discrete signal with respect to sample \( n \).

\( x(t) \ast y(t) \) : Convolution between \( x(t) \) and \( y(t) \).

\( A_v \) : Peak amplitude of the glottal pulse.

\( \alpha_m \) : Asymmetry coefficient of the glottal pulse.

\( T_p \) : Opening glottal phase length.

\( T_l \) : Closing glottal phase length.

\( T \) : Total length of the glottal cycle.

\( t_c \) : Duration of the period of the glottal flow waveform \( (t_c = T_0 = 1 / f_0) \).

\( t_p \) : Time of the maximum of the glottal pulse.

\( t_e \) : Time of the minimum of the time-derivative of the glottal pulse.

\( t_a \) : Return phase duration.

\( F_s \) : Glottal formant.
### Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQ</td>
<td>Amplitude Quotient</td>
</tr>
<tr>
<td>AR</td>
<td>AutoRegressive</td>
</tr>
<tr>
<td>CALM</td>
<td>Causal-Anticausal Linear Model</td>
</tr>
<tr>
<td>CC</td>
<td>Complex Cepstrum</td>
</tr>
<tr>
<td>CIQ</td>
<td>Closing Quotient</td>
</tr>
<tr>
<td>DAP</td>
<td>Discrete All-Pole Modelling</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DTFT</td>
<td>The Discrete Time Fourier Transform</td>
</tr>
<tr>
<td>EGG</td>
<td>ElectroGlottoGraphy</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GCI</td>
<td>Glottal Closing Instant</td>
</tr>
<tr>
<td>GOI</td>
<td>Glottal Opening Instant</td>
</tr>
<tr>
<td>H₁-H₂</td>
<td>Difference of the first two harmonics on the decibel scale</td>
</tr>
<tr>
<td>HRF</td>
<td>Harmonic Richness Factor</td>
</tr>
<tr>
<td>IAIIF</td>
<td>Iterative Adaptive Inverse Filtering</td>
</tr>
<tr>
<td>IDFT</td>
<td>Inverse of the Discrete Fourier Transform</td>
</tr>
<tr>
<td>LF</td>
<td>Liljencrants-Fant glottal model</td>
</tr>
<tr>
<td>LFₚd</td>
<td>Transformed-LF glottal model</td>
</tr>
<tr>
<td>LP</td>
<td>Linear Prediction</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstrum Coefficients</td>
</tr>
<tr>
<td>NAQ</td>
<td>Normalized Amplitude Quotient</td>
</tr>
<tr>
<td>OQ</td>
<td>Open Quotient</td>
</tr>
<tr>
<td>PSIAIF</td>
<td>Pitch Synchronous Iterative Adaptive Inverse Filtering</td>
</tr>
<tr>
<td>SQ</td>
<td>Speed Quotient</td>
</tr>
<tr>
<td>TL</td>
<td>Spectral Tilt</td>
</tr>
<tr>
<td>VTF</td>
<td>Vocal Tract Filter</td>
</tr>
<tr>
<td>ZZT</td>
<td>Zeros of the Z-Transform</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

"Nothing so surely reveals the character of a person as his voice."
(Benjamin Disraeli)

1.1. OVERVIEW

Voice is one of the most important instruments of human communication and it is such a complex phenomenon that, despite being investigated over the years in different areas as engineering, medicine and singing, not all of its attributes seem to be known.

As sound identity of human beings, the voice reflects individual characteristics such as age, sex, race, social status, personal characteristics and even emotional state. Claudius Galen, a Greek second century physician, physiologist and philosopher, studied the voice production and believed that voice was the mirror of the soul.

Voice is the unique signal generated by the human vocal apparatus and it is perceived as the sounds originated from a flow of air from the lungs, which causes the vocal folds to vibrate, and that are subsequently modified by the vocal tract. Very briefly, voice is the result of a balance between two forces: the force of the air leaving the lungs and the muscle strength of the larynx, where the vocal folds are located. As a physical phenomenon, voice is defined as a complex sound, which voiced regions, i.e., those resulting from a vibration of the vocal folds, consist of a fundamental frequency and a large number of harmonics [Sun87].

Several mathematical models have been proposed over the years both to model the voice production system or to estimate the flow of air passing through the glottis (i.e., the space between the vocal folds) (e.g. [JBM87], [BDA‘05], [Alk95]), called the glottal flow or glottal pulse. Most models of the voice production system presume that voice is the result of an excitation signal, consisting in the voice source, and that is modulated by a transfer (filter) function determined by the shape of the vocal tract. This model is often referred to as the "source-filter model of speech production" (e.g. [Fan60], [JBM87], [Air08b], [Mag05]).

1 In http://www.acsu.buffalo.edu/~duchan/new_history/ancient_history/galen.html
According to Fant’s source-filter theory, the glottal flow and the transfer function of the vocal tract are linearly separable from the speech signal [Fan60]. Specifically, using a technique called inverse filtering, it is possible to cancel the spectral effects of the vocal tract and lip and nostrils radiation on a speech signal and, then, to estimate the glottal source. Separation between source and filter is one of the most difficult challenges in speech processing.

The importance of the estimation of the glottal source is well established in speech science, providing insight into the voice signal, which is of potential benefit in many application areas such as speech coding, synthesis or re-synthesis, speaker identification, the non-invasive assessment of laryngeal aspects of voice quality and the study of pathological voices (since perturbations on the glottal flow component are considered to be one of the main sources of speech disorders), the vocal perception of emotions, and the assessment of the vocal performance of singers.

Although this topic has been extensively studied over the last decades, it is very likely that it will continue to be an open topic over the next few years, which shows its importance and complexity.

1.2. **Objectives**

Although many inverse filtering methods and processes of estimation of the glottal pulse were developed over the last years, a fully automatic procedure is not yet available.

The main purpose of this work is to study the state of art in estimation of the glottal flow and to compare the results delivered by the representative methods.

It is expected that this study will be useful for future research into the estimation of the glottal pulse.

1.3. **Structure of the Document**

A first requirement in our study is the analysis of the anatomy and physiology of the organ of voice. In chapter 2, the human speech production system is discussed and the three systems that integrate the voice organ (the breathing apparatus, the vocal folds and the vocal tract) are briefly described. In chapter 3, the fundamental source-filter theory of speech production is presented. Different methods of extraction of characteristics from speech signals and several
glottal waveform models are summarized and their respective mathematical details are presented. In chapter 4, several techniques of estimation of the glottal flow that are representative of the state of art, namely inverse filtering methods, are described, analysed and compared.

The closing chapter sets out the concluding remarks for our study and presents objectives for future research.
Chapter 2

HUMAN SPEECH PRODUCTION SYSTEM

An understanding of the human speech production system is essential in the context of our study. This chapter begins with a brief anatomic and physiologic study of the voice organ and a description of the speech production process. A particular emphasis is placed on the glottal source signal and associated phases.

2.1. ANATOMY AND PHYSIOLOGY OF THE VOICE ORGAN

The voice organ, also called phonetic system, consists of three different systems: the breathing apparatus, the vocal folds and the vocal tract. Figure 2.1 illustrates the human voice production mechanism. The lungs or respiratory organs, are spongy structures with numerous wells that provide a large surface area for gas exchange with the blood. They are located in the chest, which is separated from the abdominal cavity by the diaphragm. The latter, together with the intercostal muscles, promote respiratory movements [KG00].

On expiration, the diaphragm and intercostal muscles relax causing a decrease in the volume of the thorax and hence the increase of pressure in the chest pushes the air out of the lungs. This causes an increase in subglottal pressure that forces the opening of the vocal folds, situated at the midpoint of the larynx. As air rushes through the vocal folds, these may start to vibrate, opening and closing, in alternation, the passage of air flow. Thus, the air flow causes a series of short pulses of air, which increases the supraglottal pressure, and then, the suction phenomenon known as the Bernoulli effect is observed. This effect, due to the decrease of the pressure across the constriction aperture (i.e., the glottis), sucks the folds back together, and the subglottal pressure increases again, so that the vocal folds open giving rise to a new pulse of air [Sun87]. This phonation process has a fundamental frequency directly related to the frequency of the vibration of the vocal folds, as it will be explained below. The phonation from the larynx then enters the various chambers of the vocal tract: the pharynx, the nasal cavity and the oral cavity. The pharynx is the chamber stemming the length of the throat from the larynx to the oral cavity. The position of the velum, a piece of tissue that makes up the back of
the roof of the mouth, determines the access to the nasal cavity [Cin08]. For the production of certain phonemes, the velum can be raised or lowered to prevent or to allow acoustic coupling between the nasal and oral cavities. The tongue and lips, in conjunction with the lower jaw, are called the articulators and act to provide varying degrees of constriction at different locations, helping to change the “filter” and, therefore, the produced sound [Gol00].

Thus, according to this theory, the sustained vibration of the vocal folds is described as the balance between three aspects: the lung pressure, the Bernoulli principle and the elastic restoring force of the vocal fold tissue. However, according to recent studies, together with the Bernoulli forces, there must be also an asymmetrical driving force that is exerted on the folds and that changes with the direction of their velocity, supplying the vocal fold tissue with more energy – without which the vibrations would dissipate too readily. The Bernoulli force, along with the asymmetrical force at the glottis due to the mucosal wave as well as the closed and open phase of the folds, is now considered to be the sustaining model for vocal fold vibration [Mur08].

The vocal folds are the most important functional components of the voice organ, because they function as a generator of voiced sounds. They are covered by a mucous membrane and
the space between them is given the name of glottis, an end-point of which is found at the site of the Adam’s Apple. Images of the human glottis are shown in Figure 2.2.

The length of the vocal folds varies: in a newborn it is approximately 3 mm, and increases to 9-13 mm and 15-20 mm in adult female and male, respectively [Sun87]. This length is what defines the frequency of vibration of the vocal folds, called the fundamental frequency (or pitch). Because frequency is inversely proportional to length, the values of the fundamental frequency of female voices are higher than those of male voices. For female voices, the values of the fundamental frequency are close to 220 Hz while for male voices are around 110 Hz [Per09]. So, when a tenor sings a note with a fundamental frequency of 330 Hz, this means that his vocal folds open and close 330 times per second.

Arytenoid cartilages control the movement of the vocal folds, separating them, in the case of breathing, and joining them and tightening them to produce a voiced sound emission. The action of the cartilage joining the vocal folds is known as adduction and the opposite action (i.e., separating the vocal folds) is abduction. It is the combination of adduction and abduction actions, when performed at certain frequencies, that cause the production of a sound wave and that is then propagated towards the lip opening.

The sounds produced as a result of the vibration of the vocal folds are called voiced sounds, and the sounds produced without vibration of the latter (the vocal folds remain open only) are designated unvoiced [Per09].

The tube formed by the larynx, the pharynx, and the oral and nasal cavities is called the vocal tract, so it can be defined as the space downstream the glottis that ends with the mouth cavity or the nostrils [Kob02]. The individual morphology determines it length but in adult males, the vocal tract length is about 17 – 20 cm and 3 cm of diameter. Children and adult females have shorter vocal tracts [Sun87].
Chapter 2: Human Speech Production System

It is known that the longer the vocal tract, the lower the formant frequencies. This knowledge is very useful to singers: if they want to sing a lower note, they have to increase the vocal tract, for example, projecting the lips or lowering the larynx.

When the vocal folds vibrate and form pressure pulses near the glottis (which, in turn, are propagated towards the vocal and nasal openings), the energy of the frequencies of the excitation is altered as these travel through the vocal tract [Gol00]. Consequently, the sound produced is shaped by the resonant cavities above the glottal source and, therefore, the vocal tract is responsible for changing acoustically the voice source.

2.2. Glottal Flow

According to the anatomy and physiology of speech production, the glottal flow is the airflow velocity waveform that comes out of the glottis and enters the vocal tract. As the vocal folds open and close the glottis at identical intervals, the frequency of the sound generated is equal to the frequency of vibration of the vocal folds [Sun87]. Each cycle consists of four glottal phases, as it can be seen in Figure 2.3: closed, opening, opened and closing (or return).

![Figure 2.3. Glottal phases (adapted from [San09]).](image)

Typically, when the folds are in a closed position, the flow begins slowly, builds up to a maximum, and then quickly decreases to zero when the vocal folds abruptly shut [Kaf08]. However, studies have indicated that total closure of the glottis is an idealistic assumption and that the majority of the individuals exhibit some sort of glottal leakage during the assumed closed phase [Cin08]. Still, most of the glottal models assumed that the source has zero flow during the closed phase of the glottal cycle.
Chapter 2: Human Speech Production System

The time interval during which the vocal folds are closed and no flow occurs is referred to as the glottal closed phase. The next phase, during which there is nonzero flow and up to the maximum of the airflow velocity is called the glottal opening phase, and the time interval from the maximum to the time of glottal closure is referred to as the closing or return phase. Many factors can influence the rate at which the vocal folds oscillate through a closed, open and return cycle, such as the vocal folds muscle tension, the vocal fold mass and the air pressure below the glottis [Kaf08].

Due to the location of the larynx, the glottal flow cannot be measured directly, but there are some medical procedures that allow the observation of the larynx and the vocal fold vibration. These techniques can be divided into two categories. First, electrical and electromagnetical glottography extract specific features of the vocal fold vibration related to the changing electrical properties of the human tissue. Second, imaging techniques are based on visual analysis of larynx by observing the vocal folds using a mirror [Air08b].

Glottal inverse filtering is also a technique used to estimate the airflow through the glottis, which will be analysed in the next chapter.

2.2.1. Electroglottography

Electroglottography (EGG), a very common technique used both in voice research and clinical work, is a non-invasive method for examination of the vocal fold vibration [Air08b]. However, EGG involves contact with the skin and even physical pressure, and some specialists consider this to be an invasive technique.

EGG is based on the fact that human tissues are conductors of electric current [Hen09], giving a variable resistance to electric current, whereas air is a particularly poor conductor. It measures the contact area of the vocal folds by placing one electrode on each side of the thyroid cartilage, as it can be seen in Figure 2.4.

The conductivity of the vocal fold tissue is much larger than that of the air within the laryngeal cavity and the glottis, which makes that the impedance between the electrodes varies in step with the vocal fold vibration [Air08b]. A high-pass filter (with cut-off frequency between 5 and 40 Hz) removes the low frequency noise components, mainly due to the movement of the larynx during phonation, the blood flow in arteries and veins of the neck, and the contractions of laryngeal muscles [Hen01].

The resulting electroglottographic signal, the electroglottogram, allows to analyze the vocal folds movement and to identify the glottal phases (see Figure 2.5).
Figure 2.4. EGG measurement setting. Electrodes have been placed on the subject’s skin and a band has been adjusted around the neck to hold the electrodes in place. The electroglottograph (Glottal Enterprises MC2-1) is on the right on top of the oscilloscope [Pul05].

Figure 2.5. Schematic representation of a single cycle of vocal fold vibration (above) viewed coronally (left) and superiorly (right), and an EGG of a normal phonation (below). The numbered points on the trace correspond approximately to the points of the cycle depicted above (adapted from [Ken04]).
Figure 2.6 shows an example of an electroglottogram from a male subject during normal phonation.

![Electroglottogram of the normal phonation of a male subject.](image)

In the opening phase (1-3), the vocal folds are separating from lower margins towards upper margins and then upper margins start opening. Then the open phase (3-4) begins, where the vocal folds are maximally opened. The closing phase (4-6) follows, where the vocal folds are closing from lower margins towards upper margins. Finally, the closed phase (6-1) takes place, where the vocal folds are fully closed.

Despite its relative simplicity, the EGG allows the investigation of the vocal fold vibration during phonation and a measurement of the glottal activity, independently of the supraglottic system. However, many authors argue that the EGG signal does not allow an exact determination of the instants of closure and glottal opening, and some prefer to analyze the derivative of the signal EGG [Pul05]. This signal is often studied because it allows to visualize the changes in the tilt of the signal: if the derivative is negative in a given instant, it means that the EGG signal is decreasing at that instant, which corresponds to the closing phase. If the derivative is positive, than the EGG signal is increasing, which denotes the opening phase; otherwise, the derivative is equal to zero, meaning either the open phase (maximum) or the closed phase.

Nathalie Henrich [Hen01] regarded the peaks of the derivative of the EGG signal as reliable indicators of glottal opening and closing instants defined by reference to the glottal air flow. But this kind of approach is unreliable because, as it will be possible to see in the next chapter, often such peaks are imprecise or absent, or double peaks may occur [Pul05]. Also, the EGG
signals denote the area of contact of the vocal folds and thus do not represent directly the glottal airflow pulse shape.

### 2.2.2. Imaging Techniques

Video laryngoscopy (Figure 2.7) consists of a video camera attached to a laryngoscope so that images and sounds of the larynx and vocal folds can be simultaneously recorded and later analysed [Gui08].

From this technique a test, called kymography, is performed what makes a quantitative analysis of the vocal fold vibration, by joining a sequence of lines obtained from the captured video frames. This is illustrated in Figure 2.8.

From the examination of the kimography, it is possible to measure the duration of each phase of the glottal cycle and the opening amplitude of the glottis.


Videostroboscopy (Figure 2.9) is other procedure used to assess the structure and movement of the vocal folds. It uses a video camera attached to a stroboscopic light source, which illuminates the vocal folds quasi-synchronized with vocal fold vibration to provide what appears to be a slow-motion view of vocal fold movement and vibration [Gui08].
Illustration of the fundamental principle of videostroboscopy is shown in Figure 2.10.

By enabling the vocal folds to be viewed both in slow motion and at standstill, assessment of amplitude and glottic closure is enhanced using the videostroboscopy procedure [WB87].

Despite the fact that these techniques allow the analysis of the glottal source, they can interfere with normal phonation behaviour. Also, the logistic requirements of these techniques (equipment and health professionals for video laryngoscopy and videostroboscopy), the
invasive nature of the procedures and the time required make these techniques not practical and, therefore, not unattractive.

The estimation of the glottal pulse directly from the speech signal seems to be much more attractive, due to the relative simplicity and non-invasiveness of the process, which explains, somehow, the numerous studies done in this area in recent decades.

Figure 2.10. Fundamental principle of videostroboscopy.
Flashes of light are fired one time each frame of video at a given moment (above). The images captured from each frame are combined to create an artificial cycle (below) [Gui08].
Chapter 3

MODELS OF VOICE PRODUCTION

Any sound, given the physical nature of human sound waves, requires an energy source, an oscillator and a medium to travel through. In the body system, these are represented by the lungs, the vocal folds and the vocal tract, respectively. From the action of the lungs, the vocal folds vibrate as an oscillating force and, together with the resonant cavities of the vocal tract, mouth and nose, this force creates the sound waves required for voice [Mur08]. Thus, this is the basis of any model of voice production, as the source-filter model, that will be presented and analysed below. Also, different methods of extraction of characteristics from speech signals and glottal waveform models are described and their respective mathematical details are presented.

3.1. SOURCE-FILTER MODEL

The voice organ, as a generator of sounds, has three major units: a power supply (the lungs), an oscillator (the vocal folds) and a resonator (the vocal tract).

As it was explained in the previous chapter, when the vocal folds are closed and an airstream arises from the lungs, the pressure below the glottis forces the vocal folds apart: the air passing between the folds generates a Bernoulli force that, along with the mechanical properties of the folds (and the asymmetrical force), almost immediately closes the glottis. The pressure differential builds up again, forcing the vocal folds apart again. This cycle of opening and closing the glottis feeds a train of air pulses into the vocal tract and produces a rapidly oscillating air pressure in the vocal tract: in other words, a sound [Sun77]. During this process, an entire family of spectrum tones is generated, called partials, where the lowest tone is known as the fundamental and the others as overtones.

The vocal tract has four or five important resonances, called formants, that shape the initial sound wave, setting frequency amplitudes and formant features, which define the quality and vowel type when the wave is perceived audibly [Mur08].

The glottal source spectrum is filtered by the vocal tract and since the partials have different frequencies, the vocal tract treats them in different ways: the partials closest to a formant
frequency reach higher amplitudes than neighboring partials [Sun87]. This is illustrated in Figure 3.1.

Many models for voice production system are based on Fant’s source-filter theory: the voice is the result of the convolution between the excitation source and the filter system, i.e., the source represents the air flow at the vocal folds and the filter represents the resonances of the vocal tract which change over time [Fan60]. For voiced speech, the excitation is a periodic series of pulses, whereas for unvoiced speech, the excitation has the properties for random noise [Kaf10]. Thus, the source is the creation of the puffs of air at the glottis (glottal pulses) generating the sound wave (glottal source), which propagates through the vocal tract, and that is then filtered by varying shapes and cavities encountered therein and radiated by the lips [Mur08]. This model is, then, a simplification of the intricate relationship between the glottal source, the vocal tract and the lip radiation.

Figure 3.2 illustrates this simple model.
Chapter 3: Models of Voice Production

This model has two strong assumptions:

1. The source and the filter are independent from each other (i.e., the glottal source is equal to the glottal flow).
2. In time-domain, voice production can be represented by means of a convolution of its elements (i.e., the glottal source, the vocal tract filter and lip radiation) [Deg10].

The first assumption, in reality, is not perfectly valid, because the glottal flow is actually influenced to some degree by the vocal tract configuration. Nevertheless, the validity of the theory can be considered sufficient for most cases of interest and the assumption is very common in speech processing systems [Pul05].

This model is a simplification of the physiology and acoustic model of voice production and the scheme in Figure 3.3 emphasizes the links between these elements. According to Gilles Degottex ([Deg10]), the author of the scheme, the articulators are in blue, the passive structures are in grey and the glottis, which is acoustically active, is in orange like the vocal folds.

On the left of Figure 3.3 is a synthesis of the physiology of the voice production system, described in the previous chapter.

In the center, an acoustical model is presented, in which the impedance of the vocal apparatus is represented by area sections and their physical properties all along the structures. The impedance of the larynx is mainly defined by the glottal area, which is an implicit variable influenced by the imposed mechanical properties of the vocal folds [Deg10].

On the right of Figure 3.3, the source-filter model is depicted: the speech signal is the result of a glottal flow filtered by the vocal tract cavities and radiated by the lips and nostrils.
Figure 3.3. Schematic view of voice production models [Deg10].
The source-filter model is, as it was explained, a simplification of the discrete-time model of speech production, represented in Figure 3.4.

The mathematical framework of the classic source-filter model of speech production model can be calculated as follows:

\[ s[n] = g[n] * v[n] * l[n] \]  

where \( s[n] \) is the output signal, i.e., the speech signal, \( g[n] \) is the excitation source signal, \( v[n] \) is the impulse response of the vocal tract and \( l[n] \) is the lip radiation. This is illustrated in Figure 3.5.
In Z-domain, equation (1) can be written as:

\[ S(z) = G(z)V(z)L(z) \]  

(2)

where \( G(z) \) is the spectrum of the acoustic excitation at the glottis level. The resonances and anti-resonances of the vocal tract are combined into a single filter \( V(z) \), termed Vocal Tract Filter (VTF) and the lip and nostrils radiation are combined into a single filter \( L(z) \), termed radiation. Therefore, the glottal inverse filtering requires solving the equation:

\[ G(z) = \frac{S(z)}{V(z)L(z)} \]  

(3)

that is, to determine the glottal waveform, the influence of the vocal tract and the lip radiation must be removed. In the case of a voiced speech signal, the glottal waveform presents a typical period shape, shown in Figure 2.3.

The VTF can be written as an \( p \)-order all-pole filter:

\[ V(z) = \frac{1}{1 - \sum_{i=1}^{p} b_i z^{-i}} \]  

(4)

The lip radiation can be modelled as a derivative operator that applies to the produced acoustic signal, meaning that the derivative of the glottal flow is the effective excitation of the vocal tract. Therefore:

\[ L(z) = 1 - \alpha z^{-1} \]  

(5)

where \( \alpha \) is the lip radiation coefficient, which value is close to (but less than) 1.

This equation can be written as:

\[ L(z) \approx \frac{1}{\sum_{k=0}^{N} \alpha^k z^{-k}} \]  

(6)

where \( N \) is theoretically infinite but in practice finite because \( \alpha < 1 \).

Although in the literature, most often, these three processing stages are implemented in the discrete-time-domain, the spectral approach to voice source modelling has a number of advantages, as it will be seen in this study.


3.2. **EXTRACTION OF CHARACTERISTICS FROM A SPEECH SIGNAL**

The speech signals, as already mentioned, can be classified into voiced and unvoiced. This classification is fundamental in signal analysis, since each of them has a different kind of excitation in the synthesis of the signal.

Voiced speech signals are those that are generated as a result of the vibration of vocal folds, for example, all the vowels that we pronounce. These signals have an important feature: a well defined periodicity.

Unvoiced speech signals are generated by the passage of air at high speed through the vocal tract while the glottis is partially open. These kind of signals have almost no periodicity. Sounds such as “s”, “p” or “ch” are unvoiced sounds.

In Figure 3.6, one can see a voiced segment speech signal and another unvoiced segment speech signal.

![Figure 3.6. Speech signals: A – voiced speech signal; B – unvoiced speech signal.](image)

Speech signals are non stationary signals, i.e., over time their main attributes and, in particular, their waveform, are constantly being changed. The mathematical tools used in signal processing typically require that these signals remain time invariant so that its characteristics can be conveniently analysed. If a speech signal is divided into short enough segments (approximately between 10 and 30 ms), these "new" signals of short duration can be considered almost stationary since, throughout this duration, the phonatory articulators movements are sufficiently reduced and slow, and, thus, the transfer function associated to the vocal tract shape remains fixed (or nearly fixed) and the acoustic characteristics of the "new" signal can be considered virtually time invariant.

Thus, to extract the features of a voice signal it is first necessary to perform a segmentation of the signal into segments of sufficiently short duration. This segmentation is achieved by
applying a sliding window (e.g., Hanning, Hamming, Sine or Blackman\(^2\)) to the complete voice signal, i.e., the segmentation of the speech signal is made by multiplying the window with the voice segment. Each of these resulting segments is called a frame and can be defined as:

\[
x_m[n] = x[n]w_m[n]
\]

where \(x[n]\) is the speech signal and \(w_m[n]\) is the window function, which is zero everywhere except in a small region.

The short-time Fourier (DTFT) representation for frame \(m\) is defined as:

\[
X_m(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x_m[n]e^{-j\omega n} = \sum_{n=-\infty}^{\infty} w_m[n]x[n]e^{-j\omega n}.
\]

There are several methods to extract features from a voice signal, and some will be analysed: Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) and Discrete All-Pole Modelling (DAP).

### 3.2.1. Linear Predictive Coding (LPC) Method

Linear Predictive Coding (LPC) is a mathematical technique introduced in the sixties that may be applied to time series data and was used, primarily, as a direct proposal to modelling the voice spectral envelope in digital form [Mur08]. Given a speech signal, LPC assumes that a current sample can be determined by a weighted linear combination of a certain number of past previous samples. This method is widely used because it is fast and simple, yet an effective way of estimating the main parameters [HAH01].

The most common representation of linear prediction (LP) is:

\[
y[n] = \sum_{i=1}^{N} a_i x[n-i]
\]

where \(y[n]\) is the linear prediction of a signal \(x[n]\), \(a_i\), for \(i=1\) to \(N\), are the prediction coefficients and \(N\) is the predictor order, which is usually in the range of 8 to 14, with 10 being the most common for coding applications. The choice of the prediction order determines the number of poles that result in the designed filter and is essentially determined by the sampling

\(^2\) See mathematical details in the Appendix.
rate. This equation (9) is sometimes referred to as an autoregressive (AR) model. The difference between the predicted sample and the actual sample is called the residual error or the prediction error, and is given as:

\[ e[n] = x[n] - y[n] \]  

(10)

and from equation (9), yields:

\[ e[n] = x[n] - y[n] = x[n] - \sum_{i=1}^{N} a_i x[n-i]. \]  

(11)

The aim of LPC analysis is getting the most suited coefficients such that the residue is as small as possible. Their estimation is termed linear predictive analysis, and the most common choice for the optimization of these coefficients is the criterion of minimizing the square error function given by:

\[ E = \sum_{n=0}^{L-1} [e[n]]^2 = \sum_{n=0}^{L-1} \left[ x[n] - \sum_{i=1}^{N} a_i x[n-i] \right]^2. \]  

(12)

This function can be minimized by imposing:

\[ \frac{\partial E}{\partial a_i} = 0, \quad i = 1, ..., N. \]  

(13)

From this equation, one obtains:

\[ 2 \sum_{n=0}^{L-1} \left[ e[n] \cdot \frac{\partial e[n]}{\partial a_i} \right] = 0 \iff \]  

(14)

\[ \iff 2 \sum_{n=0}^{L-1} \left[ e[n] \cdot (-x[n-i]) \right] = 0 \]  

(15)

\[ \iff -2 \sum_{n=0}^{L-1} \left[ x[n] - \sum_{k=1}^{N} a_k x[n-k] \right] x[n-i] = 0 \]  

(16)

\[ \iff -2 \left( \sum_{n=0}^{L-1} [x[n] \cdot x[n-i]] - \sum_{n=0}^{L-1} \sum_{k=1}^{N} a_k x[n-k] \cdot x[n-i] \right) = 0 \]  

(17)
This last equation leads the normal equations for linear prediction [Cin08].

Thus, the result of LPC analysis is a set of coefficients $a_i$, for $i = 1$ to $N$, that can be computed using the autocorrelation or covariance methods, and an error signal $e[n]$, which will be as small as possible and represents the difference between the predicted signal and the original signal.

In order to understand the LP method is convenient to use a frequency-domain approach [Mag05]. From a spectral standpoint, linear prediction attempts to match the power spectrum of the signal $x[n]$ to the predicted filter given by the coefficients $a_i$, as it can be seen in Figure 3.7.

In Z-Transform domain, the equation (11) is equivalent to:

$$E(z) = X(z) \left[ 1 - \sum_{i=1}^{N} a_i z^{-i} \right] = X(z)A(z)$$

(19)

where $A(z)$ is the model Z-Transform, $E(z)$ is the Z-Transform of the prediction error and $X(z)$ is the Z-Transform of the speech signal.

Usually, $A(z)$ is referred to as the LPC analysis filter and $\frac{1}{A(z)}$, an all-pole filter, as the synthesis filter.

**Figure 3.7.** FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the LP method (thick line). For the sake of clarity, the magnitude level of the prediction model has been lifted 10 dB [Mag05].
Defining the power spectrum \( P(\omega) \) of the signal \( x[n] \) as

\[
P(\omega) = |X(e^{j\omega})|^2
\]

then, from equation (19), one has:

\[
P(\omega) = \frac{|E(e^{j\omega})|^2}{|A(e^{j\omega})|^2}.
\]

(21)

If the prediction filter is effective, the prediction error is white noise, which means its power spectral density is flat. As a consequence, the signal spectrum \( P(\omega) \) is approximated by the all-pole model spectrum \( \hat{P}(\omega) \) of the estimated filter. Assuming that the noise is white, then \( |E(z)|^2 = \sigma^2 \) and

\[
\hat{P}(\omega) = \frac{\sigma^2}{|A(e^{j\omega})|^2}
\]

(22)

where \( \sigma^2 \) is the error energy. Comparing equations (21) and (22), one can conclude that the more “flat” the residual power spectrum is, the better approximation is obtained \( \left( |E(e^{j\omega})|^2 \approx \sigma^2 \right) \).

Defining the total error as the infinitive sum

\[
\varepsilon(a) = \sum_{n=-\infty}^{\infty} e_n(a)^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 \, d\omega
\]

(23)

where \( e_n \) is the prediction error, defined in equation (10). Combining equations (21) and (22), equation (23) reduces to:

\[
\varepsilon(a) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 \, d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) |A(e^{j\omega})|^2 \, d\omega = \frac{\sigma^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} \, d\omega.
\]

(24)

Analysing equation (24), one can conclude that, if a small region of the spectrum is considered, the error can be minimized and a better fit is obtained when \( \hat{P}(\omega) > P(\omega) \), because \( P(\omega)/\hat{P}(\omega) \) is small [Mag05]. This implies that the resulting estimated \( \hat{P}(\omega) \) is above the original spectrum, which can be seen in Figure 3.7.
One disadvantage of this method is that for discrete spectrum, the LPC error measure possesses an error cancellation property, which means that the contributions to the error when \( \hat{P}(\omega) > P(\omega) \) cancel those when \( \hat{P}(\omega) < P(\omega) \). Thus, an envelope can be selected rather the only one which passes through all the spectral points [JM91].

The LPC has also some limitations as: it assumes an all-pole spectrum, but nasalized vowels, involving the lowering of the velum within the vocal tract and the radiation of speech sounds through the nose, can create zeros in the spectrum. On the other hand, consonants produced with a continuous airflow through a vocal tract constriction, called fricative sounds, also introduce anti-resonances into the spectrum.

### 3.2.2. Discrete All-Pole Modelling (DAP)

In 1991, El-Jaroudi and Makhoul [JM91] proposed a new method for parametric modelling of spectral envelopes, called the Discrete All-Pole (DAP) modelling. The main purpose of the DAP method is to fit the all-pole model using only the finite set of spectral locations that are related to the harmonic positions of the fundamental. Figure 3.8 shows an example of a spectral envelope given by DAP and LP.

![Figure 3.8.](image)

**Figure 3.8.** FFT spectrum of a male vowel /a/ (thin line) and the all-pole spectrum of the DAP method (thick line) together with the LP method (dashed line). For the sake of clarity, the magnitude levels of the prediction model have been lifted 10 dB [Mag05].
This method uses the discrete Itakura-Saito error measure and the optimization criterion is derived in the frequency-domain.

The Itakura-Saito error measure is given by:

\[
E_{IS} = \frac{1}{N} \sum_{m=1}^{N} \left[ \frac{P(\omega_m)}{\hat{P}(\omega_m)} \ln \frac{P(\omega_m)}{\hat{P}(\omega_m)} - 1 \right]
\]  

where \( P(\omega_m) \) and \( \hat{P}(\omega_m) \) are the given discrete spectrum and the all-pole spectrum, respectively, defined at \( N \) frequency points. This error reaches the minimum only when \( P(\omega_m) = \hat{P}(\omega_m) \), i.e., in the DAP modelling, the error function is equal to zero only when the model spectrum coincide on all discrete points.

The minimum error is obtained when [JM91]

\[
\frac{1}{N} \sum_{m=1}^{N} P(\omega_m) = 1.
\]

Thus, from equation (25):

\[
E_{IS\text{ min}} = \frac{1}{N} \sum_{m=1}^{N} \left[ -\ln \frac{P(\omega_m)}{\hat{P}(\omega_m)} \right] = \ln \left[ \prod_{m=1}^{N} \frac{\hat{P}(\omega_m)}{P(\omega_m)} \right]^{1/N}
\]

and one can conclude that the minimum error is equal to the logarithm of the ratio of the geometric means of the model spectrum and the original spectrum.

According to Alku et al. [AVV02], the DAP allows a more accurate estimation of the formants of the vocal tract, particularly the \( F_1 \), and decreases the amount of formant ripple in the estimated glottal flows.
3.2.3. Mel-Frequency Cepstral Coefficients (MFCC) Method

Mel-Frequency Cepstral Coefficients (MFCC) method is based on feature extraction of speech signal from its cepstrum, changed according to the Mel scale, a scale that seeks to replicate the human auditory perception.

3.2.3.1. Cepstral Analysis

In the frequency-domain, the convolution between the two components of the signal, the excitation source signal $g[n]$, and the impulse response of the vocal tract $v[n]$, is transformed into a product:

$$g[n] * v[n] \leftrightarrow g(w)v(w).$$  \hspace{1cm} (28)

Applying the logarithm operator to the signal, it yields:

$$\log(g(w)v(w)) = \log(g(w)) + \log(v(w)).$$  \hspace{1cm} (29)

Thus, the resulting signal is a linear combination between the two components from equation (29).

Applying the inverse of the Fourier transform to the resulting signal from equation (29), the signal will be represented in a new domain, called the cepstral domain. In this domain, it is possible that the two components of the signal, the excitation and the impulse response of the vocal tract, are linearly combined and separated from each other. Thus, representing the speech signal in the cepstrum, it is possible to select specific components of the latter, applying a linear filter to remove unwanted parts. According to the components not removed, it is possible to apply a reverse transformation to the production of the cepstrum.

The cepstral domain can be divided into real cepstrum and complex cepstrum. The difference between them is that, in the first, the phase information of the speech signal is lost, i.e., any signal is minimum phase. In the latter, the cepstral coefficients have real part and imaginary part and, thus, the phase information is maintained.

The real cepstrum and the complex cepstrum of a signal $x[n]$ are given, respectively, by [GoI00]:
\[ c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln|X(e^{j\omega})|e^{j\omega}d\omega \]  
(30)

and

\[ \hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln X(e^{j\omega})e^{j\omega}d\omega. \]  
(31)

In the latter, the complex logarithm is used and can be written as:

\[ \hat{X}(e^{j\omega}) = \ln X(e^{j\omega}) = \ln|X(e^{j\omega})| + j\theta(\omega) \]  
(32)

and the phase \( \theta(\omega) \) is given by:

\[ \theta(\omega) = \arg[X(e^{j\omega})]. \]  
(33)

From equations (30) and (31) it is simple to conclude that if the signal \( x[n] \) is real, both the real cepstrum and the complex cepstrum are real signals and homomorphic transformations, i.e., they convert a convolution into a sum. Thus, if \( x[n] \) is a speech signal, then the real cepstrum and the complex cepstrum can be written as the sum of the excitation and the vocal tract filter.

### 3.2.3.2. Mel scale

Several studies state that the correlation between the human perception of the fundamental frequency of a sound and its real value is not linear (e.g. [CL09; Cua07]). The perceived frequency by human people is known as pitch. In 1940, Stevens and Volkman developed a scale, called the Mel scale, that seeks to approach the characteristics of sensitivity of human hearing [Gol00].

“Mel” is a unit of measure of perceived pitch or frequency of a tone, defined as:

\[ f_{mel} = 1127 \ln \left(1 + \frac{f}{700}\right) \]  
(34)

where \( f \) is the actual frequency in Hz.

Figure 3.9 shows the Mel scale.
As it can be seen, this scale has almost a linear frequency behaviour below 1000 Hz and a logarithmic behaviour above 1000 Hz.

Some experiments in human perception have shown that frequencies of a complex sound within a certain bandwidth of some nominal frequency cannot be individually identified unless one of the components of this sound falls outside the bandwidth [Cua07]. This bandwidth is known as critical bandwidth and its width varies with the frequency.

### 3.2.3.3. MFCC calculation process

In the MFCC analysis, the first steps are the pre-emphasis and the segmentation of the speech signal, typically using the Hamming window with 50% overlapping. The next processing step is the Fast Fourier Transform (FFT), which converts each frame of $N$ samples from the time-domain into the frequency-domain. Then, the amplitude of the FFT obtained is filtered through triangular windows, and then a cepstral analysis follows from a minimum frequency of 0 Hz to the maximum frequency of half of the sampling frequency. As a result the scale of frequency is converted from linear to Mel scale. In the final step, the cepstral coefficients are extracted, applying the logarithm function to the different frames of the spectrum and then using the Discrete Cosine Transform (DCT)\(^3\).

A block diagram of the MFCC calculation process is represented in Figure 3.10.

---

\(^3\) See Appendix for mathematical details.
According to Jankowski et al. [JVL95], the MFCC provide superior noise robustness in comparison with linear prediction based feature extraction techniques.

### 3.3. GLOTTAL FLOW MODELS

There are three main categories of parameterization methods of the glottal flow: time-domain, frequency-domain and model-based methods [Air08a].

#### 3.3.1. Time-domain methods

The glottal cycle, as it was referred, can be divided to a few phases. Usually, in time-domain methods, the critical time instants, as the instant of the glottal opening and the glottal closure, are marked in the glottal flow pulse and the durations of the phases are measured. This is illustrated in Figure 3.11. From these values some time-domain parameters can be obtained, as the open quotient ($OQ$), the closing quotient ($CIQ$) and the speed quotient ($SQ$).

Open quotient, also sometimes referred to in the literature as the glottal pulse width [Cin08], is defined as the ratio of the open phase length to the total length of the glottal cycle. Following this definition, the open quotient is 0 for a sealed larynx, and when turbulent air is travelling unhindered through the glottis, the open quotient is 1. According to some authors, the open quotient of a normal phonation is a value between 0.4 and 0.8 [Cin08].

Closing quotient is the ratio of the closing phase duration to the glottal cycle length and due to abrupt closure of the vocal folds during phonation, this value is, usually, less than half of the open quotient.
Speed quotient, also called the glottal pulse skewness, is the ratio of the opening phase duration to the closing phase duration. This last value indicates the speed of glottal opening to glottal closing and, if it is equal to 1, it means that the opening phase has the same duration as the closing phase.

Denoting $T$ the duration of the glottal cycle, $T_p$ the duration of the opening phase and $T_i$ the duration of the closing phase, $OQ$, $CIQ$ and $SQ$ result as:

$$OQ = \frac{T_p + T_i}{T},$$  \hspace{1cm} (35)

$$CIQ = \frac{T_i}{T},$$ \hspace{1cm} (36)

$$SQ = \frac{T_p}{T_i}.$$ \hspace{1cm} (37)

Despite the simplicity of these definitions, it is problematic to determine their exact values because it is difficult to find the exact locations of the glottal opening and closing instants. This reduces the precision and robustness of these parameters [Air08b].
Some other time-based parameters can be defined, combining the amplitude-based time instants to express properties related to the time-domain of the signals, such as the amplitude quotient \((AQ)\), defined as:

\[
AQ = \frac{A_{\text{ac}}}{d_{\text{min}}}.
\]  

(38)

where \(A_{\text{ac}}\) is the difference between the maximum and the minimum value within one period and \(d_{\text{min}}\) is the minimum value of the flow derivative. This parameter has been shown to correlate with the phonation type [Air08b].

Normalized amplitude quotient \((NAQ)\) is other time-based parameter, defined as the ratio of the amplitude quotient to the total period duration:

\[
NAQ = \frac{AQ}{T}.
\]  

(39)

This parameter seems to correlate with the expression of the phonation type in intensity changes [Air08b].

The measurements of amplitude levels are straightforward to obtain and the absolute scale of the glottal pulses is not required to be known, which makes the \(AQ\) and \(NAQ\) more robust than their time-based counterpart, \(C1Q\) [Air08a]. However, time-domain methods are not very robust to noisy data since the time-domain waveforms and landmarks vary a lot with noise [Boz05].

### 3.3.2. Frequency-domain methods

Glottal flow parameterization is achieved in the frequency-domain by taking measurements from the power spectrum of the flow signal, as shown in Figure 3.12.

![Figure 3.12. Flow spectrum. The levels of the first five harmonics are depicted as \(H_1\) to \(H_5\) [Air08a].](image)

Figure 3.12. Flow spectrum. The levels of the first five harmonics are depicted as \(H_1\) to \(H_5\) [Air08a].
Some parameters have been proposed to facilitate parameterization of the spectra of the glottal flow pulses, as the difference of the first and second harmonics decibels, denoted by $H_1 - H_2$, or $\Delta H_{12}$, and the harmonic richness factor ($HRF$), defined as:

$$HRF = \sum_{k=2}^{\infty} \frac{H_k}{H_1}$$  \hspace{1cm} (40)

where $H_k$ is the $k$th harmonic. This parameter is related to $H_1 - H_2$ but tries to approximate the spectral energy distribution from more than one higher harmonic [Air08b].

A trivial observation on these parameters is that they co-vary with the fundamental frequency. When the fundamental frequency increases, the distance between the harmonics grows and, thus, the value of $H_1 - H_2$ increases while the value of $HRF$ decreases.

Parabolic spectrum parameter ($PSP$) is another frequency-domain parameter used to improve spectral parameterization, which fits a second-order polynomial to the flow spectrum to gain an estimate of the spectral slope [Air08a].

There are several advantages on the spectral approach to voice source modelling, such as the better description of the voice quality spectral parameters, the more efficient voice quality modification and voice source parameter estimation [DAH03]. Frequency-domain methods seem to be also more robust in handling both noisy and phase distorted data [Boz05]. However, there are few of frequency-domain glottal source models.

### 3.3.3. Model-based methods

The model-based parameterization methods attempt to capture to general flow shape of the non interactive glottal source with the finer subtleties of the vocal fold motion, using some mathematical formula that yields artificial waveforms similar to glottal flow pulses [Cin08].

Several glottal models have been proposed to define one period of the glottal flow analytically. In this section, some of these models will be described.

Although the glottal models do not use the same number of parameters or the same name for similar parameters, which makes it rather difficult to understand the differences and similarities among models, all share some common features, because:

- the glottal flow is always positive or null;
- the glottal flow is quasi-periodic;
- the glottal flow is a continuous function of time;
Chapter 3: Models of Voice Production

- on a single period, the glottal flow is bell-shaped: it starts increasing, then decreasing and finally null;
- the glottal flow is a differentiable function of time, except in some instants as the glottal closing instant (GCI) and the glottal opening instant (GOI);
- the glottal opening phase is longer than the glottal closure phase [DA99].

According to the general properties of the glottal flow expressed above, the existing models use mainly a set of time instants, as it is shown in Figure 3.13:

- $t_p$: duration of the period ($t_p = T_0 = 1/f_0$);
- $t_p$: time of the maximum of the pulse. This maximum is termed the voicing amplitude $\lambda_v$;
- $t_e$: time of the minimum of the time-derivative;
- $t_a$: the return phase duration.

![Figure 3.13. Main scheme of the glottal pulse used by most of glottal models [Hen01].](image)

Most glottal flow models, if not all, show an asymmetry that gives more importance to the right part of the flow [DAH03]. Some of these models have a time-based parameter, the asymmetry coefficient, denoted by $\alpha_m$, that regulates this asymmetry. This is a dimensionless glottal flow parameter defined as the ratio between the flow rise time and the open time. It is equivalent to the speed quotient, as shown by the following relationship:

$$\alpha_m = \frac{SQ}{1 + SQ}.$$  \hspace{1cm} (41)

This parameter was introduced by Doval and d’Alessandro to simplify the equations of glottal flow models [HSA’03]. Another advantage of using the asymmetry coefficient instead of the
speed quotient is that the values of this parameter are more easily understandable: $\alpha_m$ ranges between 0 and 1 (which corresponds to $0 < SQ < \infty$), with typical values between 0.5 ($SQ = 2$) and 0.8 ($SQ = 4$). For $\alpha_m < 0.5$ ($SQ < 1$), the glottal pulse is skewed to the left, for $\alpha_m = 0.5$, the glottal pulse is symmetric, and for $\alpha_m > 0.5$, the glottal pulse is skewed to the right.

### 3.3.3.1. Fant Glottal Model

In 1979, Fant [Fan79] proposed a glottal model which is described by:

$$
g_p(t) = \begin{cases} 
\frac{1}{2}(1-\cos(\omega_g t)), & 0 \leq t \leq t_p \\
K \cos(\omega_g (t - t_p)) - K + 1, & t_p < t \leq t_c = t_p + \frac{1}{\omega_g} \arccos \frac{K-1}{K} \\
0, & t_c < t \leq T_0
\end{cases}
$$

(42)

where $\omega_g = \pi / t_p$ and $K$ is a parameter that controls the slope of the descending branch.

This model of glottal flow is illustrated in Figure 3.14.

![Figure 3.14. Glottal flow model proposed by Fant [Fan79].](image)

According to this model, if $K = 0.5$ the pulse is symmetric.
3.3.3.2. Liljencrants-Fant Glottal Model (LF Model)

Liljencrants and Fant [FLL85] suggested a model for the derivative of the glottal flow, called the LF model. This is a four parameter model that, together with the length of the glottal cycle, determines uniquely the pulse shape.

The LF model is described by the following equations:

\[
g'_{LF}(t) = \begin{cases} 
  E_0 e^{\alpha t} \sin \left( w_g t \right), & 0 \leq t \leq t_e \\
  -\frac{E_e}{\beta t_e} \left( e^{-\beta (t-t_e)} - e^{-\beta (t_{e} - t_{c})} \right), & t_e < t \leq T_{0}
\end{cases}
\]  \tag{43}

where,

- \( E_0 \) is a scale factor necessary to achieve area balance;
- \( E_e \) is the amplitude of the negative maximum;
- \( \alpha = C \pi \), where \( C \) controls the exponentially growing sinusoid;
- \( w_g = 2\pi F_g \) is the frequency of the sinusoid, where \( F_g = \frac{1}{2T_p} \);
- \( \beta \) is a decay constant for the recovery phase of the exponential.

The parameters \( \alpha \) and \( \beta \) can be calculated from equation (43) by imposing:

\[
g'_{LF}(t_e) = E_e \tag{44}
\]

and the energy balance,

\[
\int_{0}^{T_{0}} g'_{LF}(t) \, dt = 0. \tag{45}
\]

The LF model is, thus, a piecewise function, consists of two parts. The first part models the glottal flow derivative from the instant of glottal opening to the instant of the maximum negative extreme and corresponds to the opening phase. The second segment characterizes the closure phase. In addition to these two equations, the model is governed by the principle of area balance, meaning that the integral of the function over the entire period must be equal to zero [Cin08]. In Figure 3.15 the LF-model is represented as well as the derivative of the glottal flow.

The timing instants of the LF model have a correspondence with the behaviour of the vocal folds, as it is illustrated in Figure 2.3.
The LF model can also be described by other parameters. The most relevant parameters are the open quotient \((OQ)\), speed quotient \((SQ)\), and the return quotient \((RQ)\). The latter is defined by the ratio of the return phase duration \((t_r)\) to the glottal cycle length.

In the spectral domain, the LF model can be stylized by three asymptotic lines with \(+6 \text{ dB/ octave}\), \(-6 \text{ dB/ octave}\) and \(-12 \text{ dB/ octave}\) slopes. Figure 3.16 shows this spectral representation.

The crossing point of the first two asymptotes corresponds to a peak, called glottal spectral peak, at the frequency \(F_g\), called the glottal formant. Thus, \(F_g\) is the frequency of the maximum on the spectrum of the time-derivative of the glottal pulse. The frequency of the crossing point of the two last asymptotes is denoted by \(F_c\).

The crossing point of the first two asymptotes corresponds to a peak, called glottal spectral peak, at the frequency \(F_g\), called the glottal formant. Thus, \(F_g\) is the frequency of the maximum on the spectrum of the time-derivative of the glottal pulse. The frequency of the crossing point of the two last asymptotes is denoted by \(F_c\).
The last line is due to the spectral tilt (\(TL\)), also known as spectral slope, which is an important parameter of voice quality, especially for female voices [Kaf08]. \(TL\) is related to the spectrum behaviour when the frequency tends towards infinite and can be defined as a “measure of how the amplitudes of successive components decrease with increasing harmonic number” [Tit94], i.e. the amount of decreasing of the spectral amplitude and it is related to the parameters of the return phase [Hen01]. \(TL\) contributes with an additional \(-6\ dB/octave\) above the frequency \(F_c\), and this latter can be obtained as:

\[
F_c = \frac{1}{2\pi O_T} \sqrt{\frac{E_{LF}(\alpha_m)}{J(\alpha_m)}}
\]

where \(E_{LF}(\alpha_m)\) and \(J(\alpha_m)\) are functions of the asymmetry coefficient \(\alpha_m\). Thus, one can deduce that, for given \(E_{LF}(\alpha_m)\), \(J(\alpha_m)\) and \(T_0\), the glottal formant frequency is inversely proportional to the open quotient.

\(F_c\) depends on several glottal parameters, but it mostly depends on the glottal parameter \(t_a\) and it can be approximated by a simpler expression [CRR’07]:

\[
F_c = \frac{1}{2\pi t_a}.
\]

This model seems to be the preferred glottal model of many researchers (e.g. [Boz05], [Hen01], [CRR’07]), because of its ability to accommodate a wide range of natural variation.
Also, several studies have shown that the LF model is superior to other models when the objective is to model natural speech [Cin08]. However, its use in speech synthesizers is limited because of its computational complexity, since it involves solving a nonlinear equation (45) [Vel98].

An alternative model also exists, called the LF$^{Rd}$ model, proposed by Fant [Fan95], which is a particular parameterization of the LF model. In this model, the curve is parameterized by only one shape parameter $Rd$. Fant has shown that this parameter allows to better describing voice qualities into a single value.

The $t$ parameters can be expressed in a normalized form, known as $R$ parameters:

* $R_0 = t_e / T_0$;
* $R_\pi = T_0 / (2t_p)$;
* $R_\kappa = (t_e - t_p) / t_p$;
* $R_a = t_a / T_0$.

From several measurements of the $R$ parameters on various speakers having different types of phonation, the following statistical regression has been proposed:

$$Rd = \left(1 / 0.11\right)\left(0.5 + 1.2R_\pi\right)\left(R_\kappa / 4R_\pi + R_a\right).$$ \hspace{1cm} (48)

Figure 3.17 illustrates some examples of the LF$^{Rd}$ model for different $Rd$ values.

![Figure 3.17. Examples of the LF$^{Rd}$ model for different $Rd$ values [Deg10].](image)
From this parameter, the parameters can be predicted as follows:

\[
R_{sp} = \left( -1 + 4.8Rd \right) / 100
\]
\[
R_{kp} = (22.4 + 11.8Rd) / 100
\]
\[
R_{kp} = 1/ \left[ 4 \left( \left( \frac{0.11Rd}{0.5 + 1.2R_{sp}} \right) - R_{sp} \right) / R_{kp} \right].
\]

(49)

According to the author [Fan95], \(R_d\) is a value between 0.3 and 2.7.

### 3.3.3.3. Rosenberg Glottal Model

Rosenberg proposed six different models to model a glottal pulse [Deg10]. One of the most known model is:

\[
g_{R_1}(t) = \begin{cases} 
\frac{A_1}{2} \left[ 1 - \cos \left( \frac{\pi t}{t_c} \right) \right], & 0 \leq t < t_c \\
A_1 \cos \left( \frac{\pi (t-t_c)}{2t_c} \right), & t_c \leq t < t_c \\
0, & t \leq t < T_0
\end{cases}
\]

(50)

According to Degottex [Deg10], in 1971, Rosenberg made a test to select the model which sounds as the best source and selected one model that shapes the glottal volume velocity with two polynomial parts:

\[
g_{R_2}(t) = \begin{cases} 
A t^2 (t_c - t), & 0 < t < t_c \\
0, & t_c < t < T_0
\end{cases}
\]

(51)

These models are illustrated in Figure 3.18.

![Figure 3.18](image-url)  

**Figure 3.18.** The Rosenberg model of the glottal flow. In blue is represented the glottal flow defined by the model (50) and in red glottal flow defined by the model (51).
This latter model has only one shape parameter, $t_e$, the instant of closure and it is easy to determine the instant of maximum flow: $t_p = \frac{2}{3} t_e$.

### 3.3.3.4. Rosenberg++ Glottal Model

Veldhuis [Vel98] proposed a glottal source model which has two extensions of the Rosenberg model (thus the name R++) described by the expression in (51). As is was mentioned above, in the Rosenberg model the glottal source has always the same instant of maximum flow $\left( t_p = \frac{2}{3} t_e \right)$, which, according to Veldhuis, limits is flexibility, and does not have a return phase.

The R++ model is given by [Vel98]:

$$g_{R^{++}}(t) = \begin{cases} 
  f(t), & 0 < t < t_e \\
  f(t_e) + t_a K_a \left(1 - e^{-\frac{t - t_e}{t_a}} - \frac{t - t_e}{t_a} e^{-\frac{t - t_e}{t_a}}\right) \frac{1}{1 - e^{-\frac{t - t_e}{t_a}}}, & t_e < t < T_0 
\end{cases} \tag{52}$$

with

$$f(t) = At^2 \left(t^2 - \frac{4}{3} t (t_p + t_s) + 2t_p t_s\right) \tag{53}$$

$$K = 4 A t_e (t_p - t_e) (t_s - t_e). \tag{54}$$

The parameter $t_s$ can be found solving equation:

$$t_s = t_e \left(1 - \frac{t_e^2 - t_e t_p}{2t_e^2 - 3t_e t_p + 6t_a (t_e - t_p) D(t_e,t_e,t_a)}\right) \tag{55}$$

with

$$D(t_e,t_e,t_a) = 1 - \frac{t_e - t_e}{t_a} \frac{1}{e^{-\frac{t_e}{t_a}} - 1}. \tag{56}$$

Figure 3.19 illustrates LF and R++ glottal-pulse time derivatives for two sets of parameters. The top panel shows glottal-pulse time derivatives for a modal voice with a distinct closed phase and the bottom panel for an abducted voice without a distinct closed phase.
According to the author of the model, R++ model is computationally more efficient than LF model and requires less processing time.

3.3.3.5. Klatt Glottal Model

Klatt and Klatt [KK90] proposed two models of the glottal source, in which the characteristics of the waveform are described by conventional parameters as the fundamental frequency of voicing \( F_0 \), the peak amplitude of the glottal pulse \( A_p \), the open quotient \( OQ \) and the spectral tilt \( TL \) or the spectral change associated with “corner rounding” in which closure is nonsimultaneous along the length of the vocal folds.

One of the models is a slightly modified version of the LF model to allow consider turbulence noise generation at the glottis. Thus, a cutoff frequency is specified below which the source consists of harmonics, and above which the source is flat-spectrum noise.

The other proposed model, known as the KLGLOTT88 model, or the Klatt model, synthesizes the glottal pulse in the same way as the Rosenberg model described by the analytical expression (51).

Figure 3.20 illustrates the derivative glottal wave of the Klatt model.
The derivative of the glottal flow model is given by:

\[
G_k'(t) = \begin{cases} 
2a \frac{t - T_0/Q}{f_s} - 3b \left( \frac{t}{f_s} \right)^2, & 0 \leq t \leq T_0/Q \cdot f_s \\
0, & T_0/Q \cdot f_s < t \leq T_0 \cdot f_s 
\end{cases}
\]  

(57)

where

\[
a = \frac{27A_v}{4Q^3T_0}
\]

(58)

\[
b = \frac{27A_v}{4Q^3T_0^2}.
\]

(59)

In this model the closed phase of the glottal cycle is clearly visible, as it can be seen in Figure 3.20.

3.3.3.6. Causal-Anticausal Linear Model (CALM)

This model, proposed by Doval, d’Alessandro and Henrich [DAH03] and referred to as CALM model, is totally described in the spectral domain, based on the assumption that the glottal flow can be considered as the impulse response of a linear filter.

The authors showed that glottal flow characteristics contribute to the speech signal spectrum with two components: the glottal formant \(F_g\) and spectral tilt \(T_L\) [Boz05]. \(T_L\) is linked, in this model, to the causal part of the model glottal closure and it is related, in terms of spectra,
to the position of the single causal pole of the model. $F_g$ corresponds to the anticausal part of the glottal model and to the pair of poles outside the unit circle.

Observing the LF model (43), one can recognize that the open phase is the truncated impulse response of an anti-causal stable pole and the return phase is the truncated impulse response of a first order continuous time causal filter. Therefore, this model can be defined by two filters. The first anti-causal filter is given by:

$$H_A(z) = \frac{b_1 z}{1 + a_1 z + a_2 z^2}$$  \hfill (60)$$

with

$$a_1 = -2e^{-a_p/f_s} \cos\left(b_p / f_s\right)$$  \hfill (61)$$

$$a_2 = e^{-2a_p/f_s}$$  \hfill (62)$$

$$b_1 = E \frac{\pi^2}{b_p} e^{-a_p/f_s} \sin\left(b_p / f_s\right)$$  \hfill (63)$$

$$a_p = -\frac{\pi}{O_q T_0 \tan\left(\pi \alpha_m\right)}$$  \hfill (64)$$

$$b_p = \frac{\pi}{O_q T_0}.$$  \hfill (65)$$

The second causal filter is equivalent to the low-pass filter of the KLGLOTT88 model, used to control the spectral tilt in high frequencies [Deg10].

The anti-causal pole pair, as it can be observed by the equations above, does not depend on the time-parameters and the causal real pole is independent of the parameters $OQ$ and $\alpha_m$.

This approach implies that it is possible to obtain an estimation of the voice source parameters without any inverse filtering procedure, only requiring a process to separate the causal and the anticausal parts of the speech signal.
Chapter 4

**ESTIMATION OF THE GLOTTAL FLOW**

In the last decades, several methods and techniques have been developed for the estimation of the glottal waveform during voiced speech. Many of these approaches are based on Fant’s source-filter theory: the glottal flow and the transfer function of the vocal tract are linearly separable from the speech signal [Air08].

The source-filter theory of speech production states that speech can be described as a sound source being modulated by a dynamically changing filter. This is a simplification of the relationship between the glottal source and the vocal tract and implies that speech signals are produced by exciting the vocal tract system with periodic source (glottal flow) signals [Boz05]. If the transfer function of the vocal tract filter is known, an inverse filter can be constructed in order to estimate the voice source.

Inverse filtering is a procedure that tries to estimate the glottal pulse by cancelling the spectral effects of the vocal tract and lip radiation on a speech signal (Figure 4.1.).

![Figure 4.1. Schematic spectra of the represented signals and filters.](image)

The upper row represents the separated speech production model. The lower row represents the corresponding inverse filtering process, in which the lip radiation and vocal tract are inverted to produce an estimate of the glottal flow waveform [Air08].

This procedure has, usually, three different stages: first, modelling the vocal tract filter, i.e., the transfer function of the vocal tract is estimated; second, cancellation of the effects of formants in the voice signal, filtering the voice signal through the inverse filter of the vocal
tract; and finally, estimation of the glottal pulse. So, basically, the procedure involves extracting two signals, the volume velocity waveform at the glottis, and the effect of the vocal tract filter, from a single source signal [Pul05]. If voice production is described as a convolution between the glottal excitation and the vocal tract filter, than inverse filtering can be understood as the deconvolution.

In the last years other glottal flow estimation techniques have been developed that are not based on inverse filtering procedures.

In this chapter different glottal flow estimation processes, including inverse filtering techniques, are described and analysed.

### 4.1. Inverse Filtering Techniques

Inverse filtering has many applications in both research and clinical examination of voice production, as it was stated above, but few voice inverse filtering software packages exist and it is still a challenge to develop a complete and automatic inverse filtering method. Some of the existing tools implement manual inverse filtering techniques, as DeCap, developed by Svante Granqvist, and Waveview, developed by Glottal Enterprises. This latter software allows to analyze both CV-mask (airflow) and microphone (radiated pressure) waveforms during speech and singing. Also, Paul Milenkovich [Mil01] developed a time-frequency analysis software, TF32, that is able of linear predictive inverse filtering, and Kreiman et al. [KGB06], based on the inverse filtering method proposed by Javkin et al. [JBM87], developed an open-source inverse filtering and analysis software, Inverse Filter and Sky. Another open-source licence software package that implements glottal inverse filtering and several time-based parameters of the voice source in a graphical user interface is HUT Voice Source Analysis and Parametrisation Toolkit (Aparat), develop by Airas et al. [APB’05]. This software is used in Matlab environment and implements the inverse filtering algorithms Iterative Adaptive Inverse Filtering (IAIF) and Pitch Synchronous Iterative Adaptive Inverse Filtering (PSIAIF). Also, in Matlab, there is VOICEBOX⁴, a speech processing toolbox maintained by and mostly written by Mike Brookes, which includes inverse filtering routines but with no graphical user interface.

---

⁴ Available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
In this section the above mentioned techniques of inverse filtering and others will be presented and analysed and some of the inverse filtering softwares will be used in a next section to estimate the glottal pulse from speech signals.

4.1.1. Iterative Adaptive Inverse Filtering (IAIF/PSIAIF)

The Iterative Adaptive Inverse Filtering (IAIF) is a semi-automatic inverse filtering method proposed by Alku [Alk92]. The method uses a speech pressure signal as input and generates an estimate of the corresponding glottal flow signal. This procedure has three fundamental parts: analysis, inverse filtering and integration. The glottal contribution to the speech spectrum is initially estimated using an iterative structure. This contribution is cancelled and, then, the transfer function of the vocal tract is modelled. Finally, the glottal excitation is estimated by cancelling the effects of the vocal tract (using inverse filtering) and lip radiation (by integration). A scheme of speech production model used in this approach is presented in Figure 4.2.

![Figure 4.2. Speech production model used in IAIF.](image)

The algorithm has been changed from that described by Alku [Alk92] by replacing the conventional linear predictive analysis (LPC) with the discrete all-pole modelling (DAP) method. These modifications, by Alku et al. [APB°05], allow to reduce the bias due to the harmonic structure of the speech spectrum in the formant frequency estimates. The block diagram of the IAIF procedure is shown in Figure 4.3.

The method operates in two iterations: the first phase, which consists on the stages 1 to 6, makes an estimation of the vocal tract function and applies inverse filtering to the signal with that estimate, and generates an estimate of the glottal source which is used as input of the second phase (stages 7 to 12) to achieve a more accurate estimate. A more detailed description of each step is provided below.[Pul05].

1. The input signal (speech signal) is first high-pass filtered using a linear-phase finite impulse response (FIR) filter to remove disturbing low-frequency fluctuations. To avoid filtering out relevant information, the cut-off frequency should be lower than the
fundamental frequency of the speech signal. The high-pass filtered signal is used as the input of the next stages.

2. The output of the previous step is analysed by a first-order DAP and there is a first estimate of the combined effect of the glottal flow and the lip radiation effect on the speech spectrum.

3. The input signal is inverse filtered using the filter obtained in step 2. The spectral tilt caused by the spectrum of the excitation signal and the lip radiation effect is removed.

**Figure 4.3.** The bock diagram of the IAIF method for estimation of the glottal excitation $g(n)$ from the speech signal $s(n)$ [APB’05].
4. A new analysis by DAP is calculated to obtain a model of the vocal tract transfer function. The order $p$ of the DAP analysis can be adjusted by the operator of the IAIF method and it is related to the number of formants modelled in the relevant frequency band.

5. The input signal is inverse filtered using the inverse of the $p$th-order model from the previous step.

6. Lip radiation is cancelled by integrating the output of the previous step. Then, a first estimation of the glottal flow is obtained and completes the first phase of the procedure.

7. A $g$th-order analysis of the obtained glottal flow estimate is calculated and then starts the second phase of the IAIF method. This gives a spectral model of the effect of glottal excitation on the speech spectrum. The value of $g$ is usually between 2 and 4.

8. The input signal is inverse filtered using the model of the excitation signal to eliminate the glottal contribution.

9. The output of the previous step is integrated in order to cancel the lip radiation effect.

10. A new model of the vocal tract filter is formed by an $r$th order DAP analysis. The value of $r$ can be adjusted and it is usually set equal to the value of $p$ in step 4.

11. The input signal is inverse filtered with the vocal tract model obtained in the step 10, in order to remove the effect of the vocal tract.

12. The final estimate of the glottal flow is obtained, removing the lip radiation effect by integrating the signal. This yields, finally, to the output of the IAIF method.

A toolkit for voice inverse filtering, Aparat, which is a Matlab implementation of the IAIF method, was developed at the Laboratory of Acoustics and Audio Signal Processing at Helsinki University of Technology. This software (Figure 4.4.), available as open-source\textsuperscript{5}, is an interface for the semi-automatic IAIF inverse filtering algorithm. This software also allows to determine the time and amplitude based parameters and to visualize the spectra of the speech signal, the estimated glottal flow and the used vocal tract filter (Figure 4.5).

According to the authors, Aparat has been used successfully in several research projects. However, the application of automatic inverse filtering has been problematic in high pitched signals [APB’05].

Pitch Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) is an inverse filtering method based on IAIF. The glottal pulse is obtained by applying the IAIF method twice to the speech signal. The output of the first IAIF procedure is used to calculate the fundamental period which

\textsuperscript{5} Available at http://www.acoustics.hut.fi/software/aparat/
is important to calculate the new windowing, before applying IAIF again. The block diagram of PSIAF procedure is shown in Figure 4.6.

**Figure 4.4.** The graphical user interface and the signal view of Aparat.

**Figure 4.5.** The spectra view in Aparat of a speech signal, the calculated glottal flow and the used vocal tract filter (left) and parameters computed from the estimated glottal flow.

**Figure 4.6.** Block diagram of the main steps of the PSIAF method.
Also based on IAIF is the method used by Murphy [Mur08], represented in Figure 4.7. The process uses repeated paired lattices to eliminate the effects of the vocal tract and lip radiation effects on the sound wave produced at the glottis.

The model operates as follows.

1. The input voice, $s(n)$, is filtered using an inverse radiation model filter to eliminate the effect of the lip radiation and produce a signal for the radiation compensated voice, $s_l(n)$.

2. A first estimation of the simple glottal pulse inverse function is obtained, which is used to eliminate the behaviour of the glottal pulse on the radiation compensated voice. Then, a trace for the deglottalised voice, $s_v(n)$, is produced.

3. A model for the vocal tract is derived by inverse filtering $s_v(n)$ with lattice filters and extracting the model of the vocal tract.

4. The vocal tract inverse model is applied to the radiation compensated voice, $s_l(n)$, and generates a residual trace that contains information on the glottal pulse second derivative, $u_g(n)$, which is related to the relative speed between each fold’s centre of mass.

5. The glottal pulse is extracted and, by repeating step 2 and, then, steps 3 to 5, according to the author, reliable estimations for the glottal pulse second derivative and vocal tract function are made.

The author advocates that this method generates robust estimates for the voice signal decompositions, which have been used for determining any unusual vibration patterns that may be caused by pathological masses on the vocal folds or in their immediate environment.
4.1.2. Javkin et al. Method

The first assumption of Javkin et al. [JBM87] is that speech waveforms are the product of both the phonatory setting and the shape of the vocal tract, and if the effect of the vocal tract can be subtracted from the speech waveform, then the glottal waveform can be examined without requiring any invasive procedure.

The algorithm proposed was developed in the frequency-domain because, according to the authors, the formants introduced by the vocal tract, as well the effect of the lip radiation, are best understood in this domain rather than time-domain. Since the estimated glottal flow is a function of time, the Z-Transform is used converting between these two domains.

Because of interactions between the vocal tract and the source, formant frequencies and bandwidths modulate during the open phase of the glottal cycle. Then, reliable estimate of the vocal tract parameters should be obtained during the glottal closed phase, which can be detected from the LPC residual signal.

To remove the formants and the effects of the lip radiation, it is necessary to construct a filter that has the inverse response. A digital filter that models the vocal tract is proposed and a model for each formant is obtained.

\[
VT(z) = \prod_{k=1}^{M} F_k(z) \tag{66}
\]

and

\[
F_k(z) = \frac{1 - e^{-b_kT} \cdot 2 \cos(f_kT) + e^{-2b_kT}}{1 - e^{-b_kT} \cdot 2 \cos(f_kT) \cdot z^{-1} + e^{-2b_kT} \cdot z^{-2}} \tag{67}
\]

where,

\(M\) is the number of formants,

\(b_k\) is the (one-sided) bandwith in radians of formant \(F_k\),

\(f_k\) is the formant frequency in radians of formant \(F_k\),

\(T\) is the sampling period.

To invert the effect of the vocal tract it is necessary to invert the effect of all the formants, and to invert a formant, the numerator and denominator of equation (67) are simply reversed, yielding:

\[
IF_k(z) = \frac{1 - e^{-b_kT} \cdot 2 \cos(f_kT) \cdot z^{-1} + e^{-2b_kT} \cdot z^{-2}}{1 - e^{-b_kT} \cdot 2 \cos(f_kT) + e^{-2b_kT}} \tag{68}
\]

The lip radiation is modelled by:
\[ L(z) = 1 - z^{-1}. \] (69)

Inverting the effect of lip radiation is less straightforward, because that effect amounts to a differentiation. Although, at zero frequency, \( z \) will be equal 1, and the value of the inverse of the expression (65) would be infinity, which means that low frequencies would be greatly amplified, provoking an unstable response. To avoid this, \( z \) is multiplied by a constant \( (k) \) that is less than 1 and multiplying the resulting expression by \( 1 - k \) will make the amplification (or gain) equal to 1 at zero frequency. Then, the expression that will invert lip radiation is:

\[ IL(z) = \frac{1-k}{1-kz^{-1}}. \] (70)

Finally, inverting the effects of both vocal tract and lip radiation yields to the glottal pulse. The proposed inverse filtering method implements the inversion of the nine formants in cascade.

Based on this inverse filtering method, Kreiman et al. [KGB06] developed an interactive software\(^6\) for inverse filtering, Inverse Filter, voice synthesis, Synthesizer, and voice analysis, Sky. Although the latter allows to estimate the glottal pulse, in an automatic process, the inverse filter version is a simplified version of Inverse Filter.

The interfaces of each one of these computer programs are shown, respectively, in Figures 4.8, 4.9 and 4.10.

---

\(^6\) Available at http://www.surgery.medsch.ucla.edu/glottalaffairs/download.htm.
4.2. **Zeros of the Z-Transform (ZZT) and Complex Cepstrum (CC)**

Zeros of the Z-Transform (ZZT) is a spectral decomposition method that considers the “source-filter model” as an “excitation-filter model” [SAD07]. This approach, proposed by Bozkurt et al. [BDA05], relies on the observation that speech is a mixed-phase signal, where the anticausal component corresponds to the glottal source open phase and the causal component comprises both the glottis closure and the vocal tract contributions. Thus, the Glottal Closing Instant (GCI), also known as Glottal Closure Instant, allows the separation of glottal open and closed phases, corresponding, respectively, to the anticausal and causal signals, as it is illustrated in Figure 4.11.
The authors stated that, while the contribution of the glottal flow in the vocal tract dominated spectrum is hardly observed, the vocal tract contribution in the glottal flow dominated spectrum is observed as ripples of low amplitude ([Boz05], [BDA’05]).

The block diagram of the model of speech production used on ZZT is shown in Figure 4.12.

ZZT is a representation of the z-transform polynomial through its zeros (roots) and, according to Bozkurt [Boz05], the set of ZZT of a speech signal is just the union of ZZT sets of the three components: the impulse, the glottal flow and the vocal tract filter.

The ZZT pattern for the impulse train is a set of zeros on the unit circle (i.e., with modulus 1), with a gap at each multiple of the fundamental frequency. For the differential glottal flow, the ZZT pattern is a union of two sets of zeros: one inside the unit circle (i.e., with modulus lower than 1), in which can be seen a gap, corresponding to the spectral tilt; other outside the unit circle (i.e., with modulus greater than 1), showing a gap between the zero on the real axis and the others, which correspond to the glottal formant. Also, the glottal formant and the spectral tilt can be seen on the spectrum representation, respectively, as a local maximum in the low-

\[ \text{See Appendix for mathematical details.} \]
frequency region and a global slope. The ZZT pattern for the vocal tract is a line of zeros inside the unit circle, in which can be seen zero gaps, each one corresponding to a formant [BDA’04]. This is illustrated in Figure 4.13.

Before beginning the ZZT process, the glottal closure instants (GCI) of the speech signal have to be detected and for each GCI synchronously windowed speech frame, the roots of the $z$-transform are computed and separated in two subsets based on their modulus: the roots with modulus lower than 1 (i.e., inside the unit circle) and the roots with modulus greater than 1 (i.e., outside the unit circle). According to this representation, the first subset of roots corresponds to the anticausal part of the voice source and the other to the causal part of

**Figure 4.13.** The ZZT of a speech signal [SAD07].
source and vocal tract. Then, computing DFT for each of these groups, the corresponding spectrum is obtained. Finally, using IDFT, the estimation of the glottal source and the vocal tract filter are obtained.

A block diagram of the ZZT decomposition algorithm is shown in Figure 4.14.

---

**Figure 4.14.** Block diagram of the ZZT decomposition algorithm.

Despite the simplicity of the ZZT decomposition algorithm, the need of finding roots of high degree polynomials makes it computationally heavy [BDA’05]. Also, windowing the speech signal is very critical, because the exact determination of the GCI instants of a speech signal is still an open problem and this seems to quite influence the zeros computation.
In 2009, Drugman et al. [DBD09] proposed a method, called Complex Cepstrum-based Decomposition (CC) based on the same principles of the ZZT decomposition, i.e., the speech signal is a mixed-phase signal where the maximum-phase contribution is related to the glottal open phase and the minimum-phase to the glottis closure and the vocal tract component (Figure 4.15). Although, this approach has an advantage: computationally with much higher speed than ZZT.

This decomposition is based on the fact that the complex cepstrum (equation (32)) of an anticausal signal and causal signal is zero for, respectively, all positive and negative. Thus, if one considers only the negative part of the CC, it will be possible to estimate the glottal contribution.

According to the authors, a difficulty of CC implementation is the estimation of the phase (equation (33)) because requires an efficient phase unwrapping algorithm and the windowing is still a critical issue in this approach.

4.3. _Causality/Anticausality Dominated Regions_

In 2008, Thomas Drugman et al. [DDM’08] proposed a new approach for glottal source estimation directly from the speech signal. This method uses the knowledge mentioned above, that GCI marks the separation between the glottal open and closed phases, which correspond, respectively, to the anticausal and causal signals. Thus, the causality and anticausality dominated regions are delimited by GCI and, according to the authors, the Anticausality Dominated Region makes a good approximation of the glottal open phase, since the causal contribution, i.e., the glottal source return phase and the vocal tract filter have almost no
Chapter 4: Estimation of the glottal flow

contribution before GCI. This approach requires a GCI-centered and sharp window (typically a Hanning-Poisson or Blackman window\(^8\)), as it is illustrated in Figure 4.16.

![Figure 4.16](image)

Figure 4.16. Effect of a sharp CGI-centered windowing on a two-period long speech frame. The Anticausality Dominated Region (ACDR) approximates the glottal source open phase and the Causality Dominated Region (CDR) the source return phase and the vocal tract filter (DDM08).

\(^8\) See Appendix for mathematic details.
4.4. COMPARING RESULTS

This section reports the results obtained using some of the previously described methods in order to estimate the glottal pulse of selected speech signals: the IAIF method, which was applied using Aparat software; the Javkin et al. method, using Inverse Filter; LPC inverse filtering, using TF32 software, and the ZZT and CC decompositions, using Matlab.

Before the analysis of the performances, it is important to mention that, as it was stated by Walker and Murphy [WM07], it is difficult to evaluate the quality of the estimated glottal airflow signal, because it is still unknown how the real shape of the glottal flow looks like without using special equipment and invasive procedures.

There are researchers that analyse the quality of the estimated glottal flow waveform using some parameters, as the glottal formant, the spectral distortion, time-based parameters and the harmonic richness factor (e.g. [DDM+08], [Kaf10]), or evaluate the quality of the algorithm used calculating the error between the estimated glottal flow waveform and the input glottal waveform, when this latter is known. It is also possible to compare the estimated glottal flow waveform with the one obtained using other techniques but, in this case, it is difficult to determine which one is a better estimation. Another alternative is to compare the resulting waveform of synthetic speech with the original source and, thus, evaluate the quality of each procedure.

In this study eight signals samples were selected with normal phonation, in which two are synthetic signals:

- two signals of a synthetic vowel /a/ (Figures 4.17 and 4.18);
- two real signals of a female vowel /a/ (Figures 4.19 and 4.20);
- two real signals of a male vowel /a/ (Figures 4.21 and 4.22);
- a real signal of a female vowel /i/ (Figure 4.23);
- a real singing male signal (Figure 4.24).

In each figure (from left to right) we show first the glottal flow waveform, for the synthetic signals, or the speech signal, for the real speech signals. Then estimates follow using the Aparat, the Inverse Filter, the TF32, and finally the ZZT and CC decompositions, in which, as it was mentioned above, the maximum-phase component corresponds to the glottal open phase and the minimum-phase both to the glottis closure and the vocal tract component.

Analysing the Figure 4.17, since we have the glottal airflow one can conclude that the better estimations are the ones obtained using Aparat and the ZZT and CC decompositions, and the estimations using these two latter methods (ZZT and CC) are very similar. Although, only in the
estimations obtained using Aparat, Inverse Filter and TF32, it is possible to denote the abrupt closure of the glottal impulse, often referred in literature. In ZZT and CC decompositions, the glottal flow derivative waveforms do not denote what corresponds to that, i.e., there is no abrupt decreasing on the negative amplitudes. This may be caused by the GCI windowing issue or, simply, by the procedures of the decompositions.

It is also possible to visualize in the estimation by Aparat that the glottal impulse has almost no closed phase, contrarily to the real input glottal flow waveform.

The estimations obtained in Figure 4.18 are worst, since they are very distant from the real input waveform, which shows that any of these approaches gives a robust estimation of the glottal pulse.

Analysing the estimation of the glottal flow waveforms from real speech signals (Figures 4.19 to 4.24), the abrupt closure it is visible as well as the asymmetric shape of the glottal pulse in any of the estimations. In these cases, as it was mentioned above, it is difficult to evaluate the quality of each estimation since the real glottal pulse waveform is not known.

What seems to be common in the approaches using Aparat, Inverse Filter and TF32 is the ripple in the closed phase of the glottal flow. This is often assumed to illustrate the non-zero air flow in the closed phase or the effect of the noise when the signal was captured. Although, Walker and Murphy [WM05] state that it may just mean that the estimation of the glottal flow is not exactly achieved. However, this phase is clearly visible on the estimations using ZZT and CC decompositions, because the derivative is equal, or approximately, zero in some intervals.
Chapter 4: Estimation of the glottal flow

Figure 4.17. Glottal flow estimation of a synthetic vowel /a/ using the LF model.

a) Glottal flow waveform (top) and glottal flow derivative (bottom) ($Fs = 20000\text{ Hz}$);

b) The glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;

c) The glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;

d) The glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;

e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Chapter 4: Estimation of the glottal flow

Figure 4.18. Glottal flow estimation of a synthetic vowel /a/.

a) Glottal flow waveform (blue) and glottal flow derivative (red) \((F_s = 32000 \, Hz)\);
b) The glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;
c) The glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;
d) The glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;
e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Figure 4.19. Glottal flow estimation of a female vowel /a/.

a) Waveform (Fs = 32000 Hz);

b) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;

c) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;

d) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;

e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Figure 4.20. Glottal flow estimation of a female vowel /a/.

a) Waveform (\( F_s = 22050 \text{Hz} \));
b) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;
c) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;
d) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;
e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Chapter 4: Estimation of the glottal flow

Figure 4.21. Glottal flow estimation of a male vowel /a/.

a) Waveform ($F_s = 44100\, Hz$);
b) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;
c) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;
d) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;
e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZTZ and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Figure 4.22. Glottal flow estimation of a male vowel /a/.

a) Waveform (\( F_s = 22050 \text{ Hz} \));
b) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;
c) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;
d) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;
e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Figure 4.23. Glottal flow estimation of a female vowel /i/.

a) Waveform (Fs = 32000 Hz);
b) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;
c) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;
d) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;
e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CCD decompositions. On the top of the panel is the speech waveform and the applied window.
Chapter 4: Estimation of the glottal flow

Figure 4.24. Glottal flow estimation of a singing signal of a male voice.

a) Waveform ($F_s = 24000 \text{ Hz}$);

b) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by IAIF;

c) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by TF32;

d) The corresponding glottal flow (upper row) and derivative (under row) waveforms estimated by Inverse Filter;

e) The estimation of the derivative of the glottal flow (maximum-phase component) and vocal tract contribution (minimum-phase component) using the ZZT and CC decompositions. On the top of the panel is the speech waveform and the applied window.
Chapter 5

CONCLUSION

5.1. OVERVIEW

The goal of this work was to study the state of art of the estimation of the glottal flow from a speech signal, which required knowledge, namely, about:

- the human voice production in different perspectives, anatomic, physic and acoustic;
- the glottal cycle;
- the source-filter model;
- methods of extraction of characteristics from a speech signal;
- glottal flow models;
- inverse filtering.

The focus was laid on the mathematical description of different procedures of estimation of the glottal pulse which included different inverse filtering techniques. Some of these procedures and inverse filtering techniques were applied into eight signals (two synthetic and six real signals) in order to estimate the glottal flow and to compare the results. The selected methods were: IAIF, using the Aparat software, Javkin et al. method, using Inverse Filter, LCP inverse filtering, using TF32, and ZZT and CC decompositions.

The results have shown that neither of these approaches is completely robust since the estimations of the glottal flow for the synthetic signals were not very close to the original waveform. However, in these samples, the Aparat and the ZZT and CC decompositions seem to give the most accurate estimations.

On the real signals, some features of the glottal flow were mostly common: the asymmetric shape and the abrupt closure of the glottal impulse, and the quasi no-existing glottal closed phase. Only in the ZZT and CC decompositions this last feature is clearly visible.

These results denote that the accurate estimation of the glottal flow waveform is still a challenge.
5.2. **Future Work**

The importance of an accurate estimation of the glottal flow waveform is well established in many different areas. Despite the several developed methods in the last decades, the results presented in this work demonstrate that this issue is still open.

The goal of our following work is to develop a new glottal inverse filtering technique. Our approach will be performed on the frequency-domain and based on the Normalized Relative Delays (NRDs) of the harmonics representing the quasi-periodic component of the glottal source, in order to decouple the cumulative group delay effects due to the vocal tract filter and the quasi-periodic glottal excitation [DSF11].

The next step is to capture, at the same time, the acoustic signal with a microphone as close as possible to the glottis and the speech signal outside, using a microphone with the same characteristics, in order to estimate accurately the vocal tract transfer function. Also, since the lip radiation corresponds to a first-order discrete-time differentiation, we believe that the integration should be performed in the frequency-domain, and we have some recently promising results in this area.
**APPENDIX – FUNDAMENTALS OF DIGITAL SIGNAL PROCESSING**

- **WINDOW FUNCTIONS**

  **Blackman**

  \[ w_{b}(n) = 0.42 + 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1 \]

  **Hamming**

  \[ w_{hm}(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1 \]

  **Hanning**

  \[ w_{hn}(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), 0 \leq n \leq N-1 \]

  **Hanning-Poisson**

  \[ w_{hp}(n) = 0.5\left[1 + \cos\left(\frac{2\pi n}{N}\right)e^{-\frac{2|n|}{N}}\right], |n| \leq \frac{N}{2} \]

  **Sin**

  \[ w_{s}(n) = \sin\left(\frac{\pi}{N}\times\left(n + \frac{1}{2}\right)\right), 0 \leq n \leq N-1 \]

  **Triangular**

  \[ w_{t}(n) = \frac{2}{N-1}\left(\frac{N-1}{2} - \left|\frac{n - N-1}{2}\right|\right), |n| \leq \frac{N-1}{2} \]
• **Discrete Time Fourier Transform (DTFT)**

The Fourier transform represents a signal in terms of complex exponentials (or sinusoids, because \( e^{-j\omega n} = \cos(\omega n) - j\sin(\omega n) \)).

The discrete time Fourier transform (DTFT) of a discrete signal \( x[n] \) is defined as:

\[
X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}.
\]

The inverse of the discrete time Fourier transform (IDTFT) is given by:

\[
x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega.
\]

• **Discrete Fourier Transform (DFT)**

The discrete Fourier transform (DFT) of a periodic signal \( x[n] \), with period \( N \), is defined as:

\[
X_N(k) = \sum_{n=0}^{N-1} x_N[n] e^{-j\frac{2\pi nk}{N}}
\]

and the corresponding inverse (IDFT) is given by:

\[
x_N[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_N[k] e^{j\frac{2\pi nk}{N}}.
\]

• **Fast Fourier Transform (FFT)**

The fast Fourier transform (FFT) of a signal produces exactly the same result as evaluating the DFT definition directly but is much faster.

There are several algorithms of FFT, as the function `fft.m` in Matlab.
• **Discrete Cosine Transform (DCT)**

The discrete cosine transform (DCT) of a signal \(x[n]\), with length \(N\), is defined as:

\[
X[n] = c(n) \sum_{m=0}^{N-1} x[m] \cos \left( \frac{(2m+1)n\pi}{2N} \right)
\]

where

\[
c(n) = \begin{cases} 
\sqrt{1/N}, & n = 0 \\
\sqrt{2/N}, & n \neq 0.
\end{cases}
\]

The inverse of a discrete cosine transform (IDCT) is:

\[
x[m] = \sum_{n=0}^{N-1} c(n) X[n] \cos \left( \frac{(2m+1)n\pi}{2N} \right)
\]

• **Z-Transform**

The z-transform is a generalization of the Fourier transform.

The z-transform of a discrete signal \(x[n]\) is defined as

\[
X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n}
\]

and a sufficient condition for convergence of the infinite series is:

\[
\sum_{n=-\infty}^{\infty} |x[n]| |z|^{-n} < \infty
\]

which is true only for a region of converge in the complex z-plane.

The inverse z-transform is defined as:

\[
x[n] = \frac{1}{2\pi j} \oint X(z) z^{n+1} dz
\]

where the contour integral is evaluated on a closed contour, within the region of convergence for \(z\) and enclosing the origin.
• **ZEROS OF THE Z-TRANSFORM**

For a series of \( N \) samples \((x(0), x(1), ..., x(N-1))\) taken from a discrete time signal \(x(n)\), the zeros of z-transform (ZZT) representation is defined as set of roots (zeros), \(\{Z_1, Z_2, ..., Z_{N-1}\}\) of the corresponding z-transform polynomial \(X(z)\):

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

considering that \(x(0)\) is non-zero.

Figure A1 shows an example of a ZZZT representation of a signal.

![ZZT representation of a signal in (a) cartesian coordinates and (b) polar coordinates](image)

**Figure A.1.** ZZZT representation of a signal in (a) cartesian coordinates and (b) polar coordinates [BDA’05].

*Adapted from [Ort05], [HAH01], [Gol00] and [Boz05].*
BIBLIOGRAPHY


