Towards a Voice Analysis Toolbox for the Extraction, Parametrization and Analysis of the Glottal Source Waveform and its Application for Senescence Voice

Itay Ben-Dom

A thesis submitted in fulfilment of the requirements for the degree of Master of Engineering in Electrical Engineering, The University of Auckland, 2017.

[This page intentionally left blank]

Abstract

Interest in examination of senescence voice and its phonetic traits are at the cornerstone of this thesis work. While age-induced changes had been shown to have a detrimental affect on the phonetic system, the analysis methods were most often invasive in nature (i.e. Electroglottography). An alternative approach involves the derivation of the glottal flow waveform, which is complementary to the behaviour of the vocal folds during cyclic vibration. In the past, the advancement of glottal analysis was hindered due to its computational complexity. Over the past two decades, however, advancements in speech processing operations allow for reliable and efficient extraction of the glottal source. In turn, quantitative analysis of descriptive features of the glottal source provide a looking glass into physiological changes with advanced age. The contribution of this thesis work is two-fold. First, the design and implementation procedures for a voice analysis toolbox in R is provided. The R environment was chosen as it links to the emuR framework, enabling efficient analysis of large volumes of speech data. This toolbox implements popular speech processing algorithms in a computationally efficient manner for the extraction and parametrization of the glottal source. Using the popular iterative adaptive inverse filtering algorithm, six time- and amplitude-domain parameters were investigated, including the newly proposed open quotient criterion, OQsub50. Second, the impact of age, vowel, and age-vowel interactions are discussed. The findings show correlation between senescence voice and the shape of the glottal pulse. Furthermore, vowel types were found to alter the glottal signal, implying coupling of the glottal source and the vocal tract filter. Finally, ageing had an affect on the production of vowels, implying results cannot be generalized across different vowels. This thesis work propels glottal signal analysis to new heights, highlighting the robustness of the extraction process, its correlation to speech production, and its usefulness for voice analysis.

[This page intentionally left blank]

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Catherine Watson, for taking me on for this ride. I'd like to thank her for her guidance, support, interest in my work, pushing me towards deadlines (especially that Wellington conference), talking to me about the work place and her experiences, and broadening my horizons about everything speech.

I definitely wouldn't be at this point without the weekly speech-group meetings, so I'd like to extend a warm gratitude to Tian, Jesin, and Saima, for listening to me talk about my progress and sharing theirs. Having chosen to do a PhD dissertation, I wish them all the best. I'd also like to thank Stephen Bier for helping me get established when I was only starting to find my footing in the world of speech. Also, I must thank Justine for lending a hand with conquering LaTeX, and answering emails at a frightening speed. I'd like to extend my gratitude towards Raphael Winkelmann for his guidance and help with everything EMU. A special thanks goes to Laura Thompson, Stephen Bier (again), and the MAONZE project for allowing me to use their speech corpora for this study.

Finally, I'd like to thank people outside of university (who will most likely never read this). I'd like to thank my parents and my sister for reminding me time and time again, rather eloquently, that the decision to undertake a Master thesis project was my choice and not theirs and therefore it is not their problem. I'd like to thank my friends for not understanding why I could not "hang out" in February. A special thank you goes out to my flatmates for throwing a huge party two days before this thesis was due, making sleep into luxury. I'd also like to thank Drake, Kanye, Chance, and Led Zeppelin, for making music that put me in writing mode.

[This page intentionally left blank]

Contents

List of Figures

List of Tables

Acronyms

ANOVA Analysis of Variance
AR AutoRegressive
ARMA AutoRegressive and Moving Average

ClQ Closing Quotient

DAP Direct All-Pole Modelling

EGG Electroglottography, Electroglottogram

F0 Fundamental Frequency
FFT Fast Fourier Transform
FIR Finite Impulse Response

GCI Glottal Closing Instance
GOI Glottal Opening Instance
GVV Glottal Volume Velocity waveform

IAIF Iterative Adaptive Inverse Filtering
IIR Infinite Impulse Response
IPA International Phonetic Alphabet

LF Liljencrants-Fant glottal model
LP Linear Prediction
LPC Linear Predictive Coding

MAONZE The Māori and New Zealand English project
MRI Magnetic Resonance Imaging

MRPA Machine Readable Phonetic Alphabet

NAQ Normalized Amplitude Quotient
NZE New Zealand English

OQ Open Quotient

PGG Photoglottography

SAMPA Speech Assessment Methods Phonetic Alphabet
SQ Speed Quotient
SRH Summation of Residual Harmonics

VTF Vocal Tract Filter

[This page intentionally left blank]

Chapter 1

Introduction

1.1 Project Motivation

How is speech generated? This relatively simple question took a considerable amount of time to fully understand. In large, the voice production system consists of three main parts: the lungs, the voice box, and the vocal tract. Those vocal organs work together in concert to produce audible and perceptually-comprehensible speech. A partial aim of this thesis document is to bridge the gap between the medical and the practical aspects of speech production, by providing a detailed analysis of the vocal organs and the vocal mechanism. In short, speech is produced when air expels from the lungs, rises up the windpipe, through the voice box, and propagates outwards through the oral or nasal cavities. During the production of voiced speech, the *vocal folds* open and close quasi-periodically, effectively modulating the airflow into a series of pulses. This modulated airflow, known as *glottal flow*, is the source signal for voiced speech. Ignoring the rigorous theory involved, we can simply say that voiced speech is the result of a *source* signal that passes through a *filter*, which then shapes the signal into what can be auditorially perceived as speech. This simplification is known as the *source-filter* model theory and is the cornerstone of most modern speech processing techniques. Notwithstanding advancements in speech processing, derivation of the source signal is still prone to errors. If anything, this is a testament to the inherent complexity of the voice mechanism. With this work, we created a voice analysis toolbox for the extraction and subsequent parametrization of the glottal source signal.

Despite the seemingly-simple theory of voice production, the human speech exhibits incredible variability. For over half-a-century, studies have attempt-ed to deduce information about a speaker using information solely from his voice. Such information may include the speaker's age, gender, habits, and physiological condition. One of the most widely-researched topics, which is at the core of this thesis work, is the *glottal flow* signal. It has been used in speech analysis, synthesis, and recognition. The glottal flow signal is especially important, as it also conveys information about the behaviour of the vocal folds during vibration. Presently, it can be extracted from the speech signal, which means the analysis method is non-invasive. However, this has not always been the case. Prior to the development of speech processing algorithms for the extraction of the source signal, the behaviour of the vocal folds during speech production was observed in a rather invasive manner. Many methods had been devised to obtain this information, such as laryngeal scoping, high-speed imaging photography, videostroboscopy, and electroglottography, to name a few. In this work, the glottal source is extracted through an inverse filtering operation using the speech signal. However, at its core, this project was conceived following prior research which centred around electroglottography.

At the signal processing speech research laboratory in the University of Auckland, the speech signal had been researched via electroglottography. One of the pitfalls of electroglottography is its requirement for an invasive apparatus. We considered this to be a limiting factor, with a non-invasive method being highly preferable. Hence, we chose to focus on the glottal signal. We devised the development and implementation of a voice analysis toolbox in the R environment. R was chosen for three main reasons. First, it's a free software. Second, it was developed by the university of Auckland. Third, it includes the emuR library, which enables simplistic speech data management. Subsequently, we focused our efforts on the examination of senescence voice. Although senescence voice had been analysed through electroglottography, its analysis through examination of the glottal flow signal is scarce in literature. It was our aim to establish glottal analysis as a valid approach for voice analysis. We set out to answer three key questions: How does ageing affect the glottal signal? What impact do different vowels have on the production of the glottal signal? Are there age-vowel interactions? With regards to ageing effect, it was our aim to show the robustness of the glottal source extraction procedure and its usefulness in discriminating aged voice. With regards to vowel effect, we questioned the linearity properties of the source-filter model, as we set out to examine source-filter coupling. With regards to age-vowel effect, we believe this to be the first study to examine this phenomenon. It was our aim to show the results cannot be generalized between vowels, implicating the combined influence of ageing and vowels on the glottal signal. As part of the investigation, we wanted to contribute to the progress of New Zealand English investigation. Hence, aged voice investigation was carried out across four speech corpora comprised entirely of native New Zealand speakers (English and Māori). Finally, upon completion of this thesis work, it was decided to release the toolbox under a free open-source license to be used by researchers everywhere.

On a personal note, it is our hope that this document will enable newcomers in this area (future post-graduate students) an easy integration into the world of speech. Great care has been put towards creating a literature review that is extensive yet not exhaustive. We hope that this toolbox will grow with time and exceed its original purpose.

1.2 Thesis Structure Breakdown

Chapter 2 presents a collection of theoretical knowledge gathered from literature. The aim of the chapter is to provide sufficient information with regards to the speech production mechanism. Anatomically, the lungs, larynx, and vocal tract structures are discussed. Physiologically, the role of each vocal organ in speech production is highlighted. The vocal folds vibration cycle is explored in detail, with various physical principles discussed. The chapter concludes with an overview of the existing glottal analysis methods. This launches us into the next chapter.

Chapter 3 introduces the glottal flow waveform. Description of the glottal pulse temporal instances, as well as phases, is presented. Building upon this, we introduce the time-domain and amplitude-domain parameters utilized in this thesis: open quotient, speed quotient, normalized amplitude quotient, fundamental frequency, jitter, and shimmer. In addition, a new open quotient criterion, OQsub50, is presented. The OQsub50 parameter is a new open quotient with computationally simple specifications. It OQsub50 was devised during the development of this present work. This chapter concludes with a discussion of vocal quality, particularly with regards to ageing; what is it? what affects it? and how is it measured? This chapter rounds off the introduction of the glottal flow and catapults us into the realm of speech signal processing.

Chapter 4 delves into the mathematical and signal processing theory behind the popular inverse filtering technique. First, the source-filter model is explained. We show how a sample of speech can be described in both the time-domain and frequency-domain. The vocal tract is modelled as an all-pole filter, which in turn introduces the notion of an inverse filter, followed by linear predictive coding. This allows us to relate the vocal tract filter coefficients to linear prediction coefficients. Through least squares minimization we obtain the *normal equations.* This set of equation is solved using the auto-correlation method. The Yule-Walker equations and the Levinson-Durbin recursive algorithm are described in-depth. An overview of the inverse filtering method as a whole, and the iterative adaptive inverse filtering algorithm in particular, follows. The chapter concludes with a summary of the available polarity detection algorithms, including the RESKEW method which was utilized for this toolbox. The next chapter is where we cut-our-teeth on the implementation of the theory put in place.

Chapter 5 presents the methodology employed in the creation of the toolbox, which fuses the theory presented in the preceding chapters. By this point, the reader is expected to have a substantial understanding of the methods put in place. The R environment is presented, followed by its emuR library. Next, the blueprint for the toolbox is provided. An overview of the existing speech analysis packages is provided. A detailed description of the glottal flow, from extraction to analysis, is provided. The implementation procedure for the signal processing algorithms is included. This chapter concludes with an introduction of the four speech corpora used in this study. All four corpora are related to vocal ageing. The speech data shows variety, including speech data in both New Zealand English and Māori, male and female speakers, continuous- and citation-form speech, archival as well as recorded data, and a longitudinal study. This chapter prepares the ground for the presentation of our results.

Chapter 6 presents the outcome of our senescence voice investigation through glottal analysis. For each speech corpus, the mean and standard deviation measurements for each glottal parameter are presented. Statistical analysis, through ANOVA and *post hoc t-test* is performed to identify age, vowel, and age-vowel interaction between the speaker groups for each corpus. Visual representation in the form of box-plots and scatter-plots is included when deemed appropriate. The next chapter discusses our findings are relates them to concepts, theory, and results presented in the literature.

In the penultimate Chapter 7 we reflect on our results and discuss their implications. The effect of ageing on the voice is discussed. We discuss the anatomical and physiological implications of the results for each glottal parameter. The results show the impact of ageing on the shape of the glottal flow waveform and vocal quality. The effect of vowels in speech production is also discussed. The results are indicative of a vowel effect on the glottal source signal. This implies vowel production is not solely governed by changes in formants, but is rather inclusive of the shape of the glottal source. In addition, age-vowel interactions are presented. The findings suggest that failing to perform vowel analysis in voice analysis studies may result in incorrect conclusion. A significant age-vowel effect indicates that the glottal pulse shape is not only dependent on the vowel type, but is it also codependent on the age of the speaker. Finally, we discuss the limitations inherent in the speech corpora in terms of recording procedure, phonetic environment, and quality of the speech recordings. Formant ripples and the importance of detecting polarity inversion are also highlighted.

We conclude this thesis document in Chapter 8, in which we provide a final overview of the ideas, methodology, and outcomes discussed in this work. Future work is discussed with regards to the expansion of the toolbox, desired follow-up studies, and integration of speech processing algorithms.

Chapter 2

Physiology of Voice Production

The physiological process of human speech production is most clearly explained via an anatomical description of the organs forming the phonetic system (also referred to as the vocal organs). The following section focuses on the organs that have a functional role in speech production. However, before an intricate description of the vocal organ is given, a more simplistic and idealistic overview of the voiced speech production mechanism is presented. The production of speech can be conveniently regarded as a system of three components in ascending anatomical order. Those are: the *lungs* or the sub-glottal respiratory system; the *larynx* or the *vocal box*; and the *vocal tract* or the supra-glottal area. Air is key to the phonation process. Lungs act as the source of energy which supplies airflow to the larynx. Air expelled from the lungs rises through the lower respiratory airways consisting of the bronchi and the trachea, or *windpipe*. The airflow reaches the larynx, where the free airflow is modulated into a series of pulses. The larynx houses the *vocal folds*, often referred to as *vocal cords*. The vocal folds operate as a valve. When air stream reaches the vocal folds it causes them to vibrate, open and close, periodically. The opening between the vocal folds is referred to as the *glottis*. The modulated airflow, also known as the *glottal flow* or *glottal volume velocity* waveform, provides a source to the vocal tract. The vocal tract consists of a series of the cavity airways of the *nose*, *mouth*, and *pharynx*. The vocal tract acts as a variable acoustic filter, which alters in shape while a person speaks. The vocal tract gives the sound its timbre, effectively shaping the spectral characteristics of the source. This sound wave is radiated by the lips and nasal cavity. The radiated speech signal then propagates through a medium (air) and can be received by the listener's auditory system and interpreted as speech. The human voice production system is illustrated in Figure 2.1 (lungs excluded). We note that the content of this chapter was assembled from a comprehensive collection of literature. Readers interested in an expanded, all-inclusive theoretical information are encouraged to explore the following books by Titze *et al* [1], Stevens *et al* [2], and West *et al* [3].

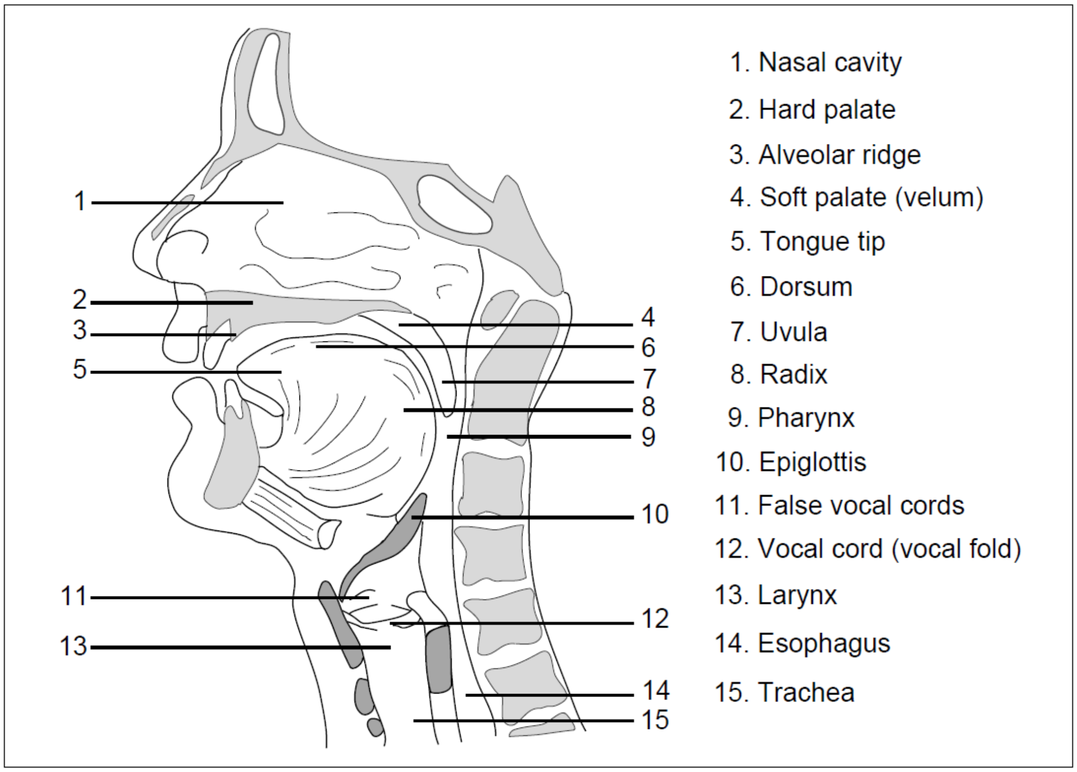

Figure 2.1: *The human speech production mechanism (as in [4] adapted from [5])*

2.1 The Lungs

Respiration, or the action of breathing, is performed by the respiratory system. The respiratory system consists of three main parts: the lungs, the airways, and the respiratory muscles. The lungs are the main respiratory organs with their primary function being the exchange of gases. During the respiration process, oxygen is absorbed into the bloodstream during inhalation/inspiration, while exhalation/expiration allows for the expulsion of carbon dioxide toxins [6]. The airways form a path for the air to flow from the lungs to the body's exterior. The airways include the trachea, bronchi, and the glottal and supra-glottal organs. The respiratory muscles consist of the diaphragm and the intercostal muscles (rib muscles). The lungs, along with the trachea and bronchi, are anatomically located below the glottis, thus are collectively referred to as the *sub-glottal respiratory system*. In speech, lungs are the power source for phonation.

2.1.1 Anatomy of the Airways

The lungs are a pair of spongy-textured, elastic, air-filled organs located in the chest cavity on either side of the heart. The lungs are a complex branching network of air passages. These passages reduce in size and increase in number as they descend into the lungs [6, 2]. Although anatomically complex, the anatomy of the lungs can be simplified and greatly clarified using the idealization of the human airways according to *Weibels model* [7]. Weibel airways model details the branching of the airways. It allows for the visualization of the airflow's path into the lungs in the form of a respiratory tree diagram. It begins at the trachea and terminates at a cluster of the respiratory cells of the lungs, called the *alveolar sacs*. The model can be divided into two regions: the *conducting zone*, or anatomic dead space, as it does not involve any gas exchange, followed by the *respiratory zone*, where blood-gas exchange takes place. Blood-gas exchange occurs at the alveolar sacs. Alveolar sacs are formed by a cluster of *alveoli*. Alveoli are tiny air sacs forming the respiratory surface of the lung, where gas exchange takes place.

In Weibel's model, each subdivision of airways is referred to as a *generation* (Z), with 23 generations observed from the trachea until the alveolar sacs. The model begins at the trachea. The trachea is a single tube of 10-12 cm in length, with a cross sectional area of about 2.5 cm^2 [2]. Trachea is referred to as the first generation of airways. After the trachea, all the airways diverge dichotomously, meaning a parent airway divides into two daughter airways. Thus, the trachea diverges into the right and left bronchi. The bronchi then subdivides into smaller branches, called bronchioles. There are 16 generations of bronchioles branching, terminating at the terminal bronchioles, before alveoli cells appear. This region is referred to as the conducting zone. By this point the air should be sterile, warm, and wet. This is followed by the respiratory zone, consisting of respiratory bronchioles (three generations), leading to the alveolar ducts (three generations), and terminating at the alveolar sacs formed by a cluster of alveoli, where the blood-gas barrier is located [3, 8].

The volume of the conducting zone is about 150 mL, while the volume of the respiratory zone is about 3 L [3]. The total air-volume capacity of the adult human lung is between six and seven litres. The lungs have a two litres residual volume capacity, referring to the minimum volume capacity which is always present and cannot be expelled unless the lung collapses. Therefore, the operational (respiration or voice) volume of the lungs ranges between four and five litres, with only 10 to 15 percent of the operational volume capacity, or tidal, being used during normal breathing. However, a higher percentage may be used in case of increased physical activity [1, 2].

2.1.2 Respiratory Movements

The lungs are located inside the chest cavity, enclosed by the *chest wall* (ribcage), and rest on the dome of the *diaphragm*. The lungs are sealed inside the chest cavity by a linking tissue called pleurae. Pleurae is a smooth membrane which covers the lungs and lines the chest cavity. These two layers of pleurae are separated by a thin film layer of fluid, which prevents the adhesion of pleurae layers, allowing for frictionless sliding. Between the two layers of pleurae exists a vacuum. Therefore, when the ribcage and/or diaphragm move, the lungs follow [6, 8].

During the respiratory process, the lungs expand and contract. Since the lungs do not contain muscles of their own, it is the increase and decrease in the volume of the chest cavity, by the movement of the ribcage and diaphragm, which causes them to expand and contract, respectively [9, 1, 2, 10]. The volume of the lungs varies by the contraction of the *muscles of respiration*. The muscles of respiration consist of the diaphragm and the *intercostal muscles*, which allow for expansion and contraction in the chest cavity volume. The diaphragm is a dome-shaped, thin sheet of muscle which forms the floor of the chest and the roof of the abdomen [8]. It is connected to the ribs at the side and the spine at the back [3]. The intercostal muscles are a group of muscles found between the ribs, with each rib connected to its neighboring ribs by the intercostal muscles. There are two types of intercostal muscles: *external* intercostal and *internal* intercostal. External intercostal muscles run obliquely between the ribs, from the bottom of one rib to the top of the rib below it. The internal intercostal run at right angles to the externals, from the bottom of each rib to the top of the next rib above [6, 8]. Since the lungs are housed inside the chest wall, their expansion requires the chest wall to be a moveable structure. The chest wall consists of 12 pairs of ribs, with the top 10 attached to the breast bone to form a closed cage. The two lowest pairs of ribs do not attach and are called free ribs. This allows the ribs to move upon intercostal muscle contraction, allowing the ribcage to expand and contract, thus change in volume [6, 2].

2.1.3 Respiration

Breathing is a muscular effort, in which inspiration is active and expiration is generally passive. Inhalation results in the expansion of the chest cavity. Under quiet breathing conditions, inhalation is the result of the contraction of the diaphragm and the contraction of the external intercostal muscles. The most important muscle in inspiration is the diaphragm. It contracts and descends, effectively increasing the vertical dimension of the chest cavity [11, 1]. The diaphragm can move vertically down up to 10 centimetres. When the diaphragm contracts, it increases the pressure within the abdominal cavity, which causes the ribs to elevate and twist slightly ("move out"), with a motion that resembles the upwards swing of a bucket handle. The secondary muscles in the inspiration process are the external intercostal muscles. The external intercostal muscles contract, causing the ribs to move upward and outward. When the ribs are raised they also rotate and the diameter of the chest wall increases [3].

Exhalation follows inhalation and vice versa. During inhalation, the chest walls have expanded to allow for great chest cavity volume. As a result, during exhalation, it will tend to return to its rest position. This mechanism is known as *elastic recoil*. Elastic recoil occurs due to the contraction of the internal intercostal muscles, which are the primary expiratory muscles. Contraction of the outer parts of these muscles lowers the ribs, effectively reducing the size of the chest cavity [3]. During expiration, the compression of the ribs decreases the chest cavity volume, resulting in elastic recoil of the lungs, which can provide the force necessary to expel air out of the lungs [9, 1].

2.1.3.1 Respiration Pressure Principles

Thus far, a detailed explanation of the physiological mechanics of respiration was given. The contraction of the respiratory muscles, altering the volume of the chest cavity, results in expansion/contraction of the lungs. Those breathing mechanics are governed by the laws of physics. In order to better understand the mechanics of breathing, we need to understand the pressure-volume relationship.

Airflow is generated by pressure exerted on the lungs by the respiratory muscles. The unit of measure for pressure is Pascal (Pa). Pascal's Law states: "*Pressure is transmitted rapidly and uniformly throughout an enclosed fluid at rest*". The human lungs are made of a network of airways and contain millions of alveolar sacs. According to Pascals Law, there exists a uniform pressure along all the alveolar sacs, called *alveolar pressure*, P_{al} [1]. Another key principal is Boyle's Law. Boyle's Law, also referred to as the gas law, states: "*In closed space under constant temperature, pressure (P) and volume (V) are inversely proportional*". This law applies to the lungs, as the lungs are a soft-walled enclosure, maintained at constant temperature (body temperature). Therefore, considering the lungs, the product of $P \cdot V$ remains constant, meaning an increase in volume results in a proportional decrease in pressure and vice versa [6, 9, 1].

It can be readily observed that the mechanism of respiration is governed by both Pascal's and Boyle's laws. In order for air to flow there must exist a pressure gradient to form its path [8]. Air flows from a region of high pressure to a region of low pressure, until the pressures are equalized. Thus, for air to flow in and out of the lungs, there must be a difference between the air pressures of the lungs, P_{al}, and the outside atmospheric pressure, $\mathrm{P}_{\mathrm{atm}}$. During quiet inhalation, the ribcage expands and the diaphragm contracts, leading to an increase in chest cavity volume. As a result, lungs volume increases and the alveolar pressure drops. When the alveolar pressure reduces, air flows into the lungs, until the pressures are equalized. In exhalation, the chest walls contract and the diaphragm relaxes. This compression of the lungs causes pressure to rise. When $\mathrm{P_{al}}$ exceeds $\mathrm{P_{atm}}$ air is expelled from the lungs [6, 1, 8].

2.2 The Larynx

The larynx, sometimes colloquially referred to as the voice box, is a hollow muscular organ situated at the neck, forming an air passage to the lungs and housing the vocal folds. The larynx and the vocal folds form the vibratory system of the voice mechanism. Hence, they are of major importance for the scope of this thesis. The larynx is a key organ for both the respiratory and phonatory mechanisms. It forms a protective layer of muscle around the airways and it enables respiratory control, as well as voice production. Anatomically, the larynx is suspended from the *hyoid bone* by ligaments and muscles, located in front of the esophagus, and located above the trachea [12]. The larynx framework consists of four main components: cartilages, muscles, nerves, and the vocal folds. Information about the laryngeal nerves is excluded as it exceeds the scope of this thesis. In large, the larynx consists of four main cartilages, two pairs of joints, two sets of muscles, and fibro-elastic tissue. An illustration of the larynx is given in Figure 2.2. The four main laryngeal cartilages are: Thyroid, Cricoid, Arytenoid, and Epiglottis. These are inter-articulated by muscular, fibrous and elastic tissue. There are two sets of laryngeal muscles: intrinsic and extrinsic. The intrinsic muscles connect the cartilages of the larynx. There are both abductor and adductor intrinsic muscles, which are responsible for altering the length, tension, shape, and spatial position of the vocal folds. The extrinsic muscles connect the laryngeal framework to the body and maintain its stability. A detailed description of these organs follows.

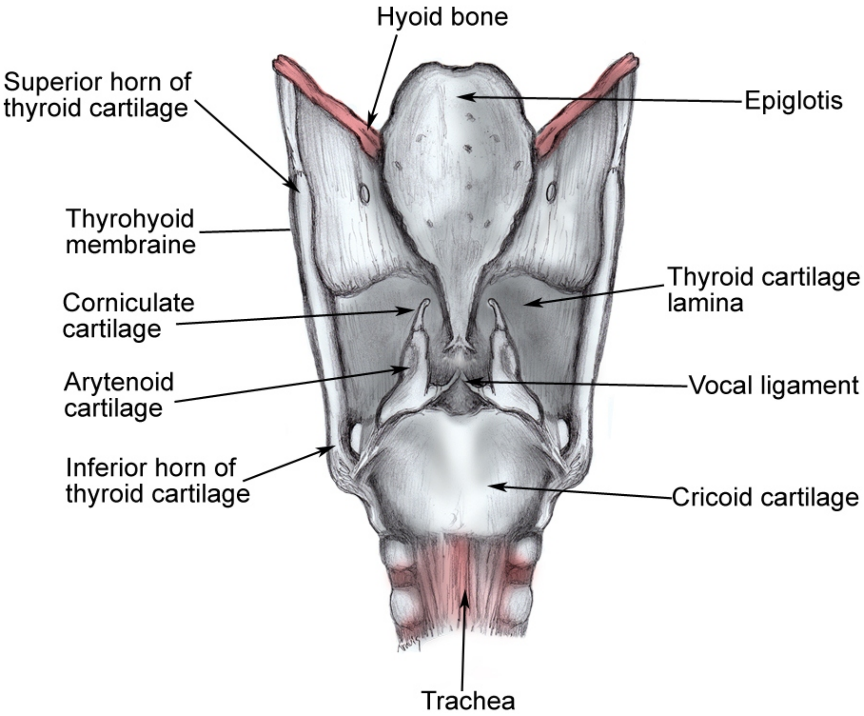

Figure 2.2: *Posterior view of the laryngeal cartilages (as in [13])*

2.2.1 Laryngeal Functions

The larynx has two main functions in the human body: airway protection, and phonation [14]. The larynx has evolved to allow phonation, but its main purpose is airway protection, especially during swallowing, by sealing the inlet to the larynx, thus preventing solids and fluids from descending to the lungs [15]. In phonation, the larynx houses the vocal folds which modulate the airflow through vibration, providing the input source signal to the speech production system. The activity of the larynx can be classified in terms of four basic gesture types:

1. Account for glottal stops, fricatives, and phonation types.
2. Account for fundamental frequency variations.
3. Raising-lowering gestures, which shift the larynx vertically (e.g. swallowing)

4. Supplementary constricting gestures which constrict the supraglottal part of the larynx (e.g. swallowing)

In voice phonation, with regards to the operation of the vocal folds, the larynx has two important roles: opening and closing gestures, which regulate the degree of opening between the vocal folds; stiffening and slacking gestures, which regulate the length, thickness and stiffness of the vocal folds [12].

The larynx has other key functions, all involving the opening and/or closing of the glottis: respiration, coughing, yawning, and swallowing. In respiration, the larynx channels the air to the respiratory organs for gas exchange. The laryngeal muscles contract and open the vocal folds' orifice, allowing passage of air. In coughing, the laryngeal muscles contract and close the glottis for pressure to build-up, then open the glottis for a burst of expelled air. In yawning, the larynx descends to enlarge the cavity above it. The backward motion of the tongue forces the epiglottis over the opening of the glottis to cover the laryngeal inlet and prevent aspiration of swallowed material into the lungs. In swallowing, the larynx also moves superiorly and anteriorly, which opens the esophagus for the passage of the swallowed material [16].

2.2.2 Laryngeal Cartilages & Joints

Cartilage is a firm, flexible connective tissue. The main laryngeal cartilages, in ascending anatomical order, are: *Cricoid*, *Arytenoid*, *Thyroid*, and *Epiglottis* cartilages (see Figure 2.2). The *thyroid cartilage* is the largest cartilaginous structure in the larynx. It is composed of two flattened rectangular plates of hyaline cartilage, the lamina, that are fused in the midline, forming the *thyroid notch* [17]. The fused plates diverge as they extend backwards at an angle. The angle between the two plates is around 90 degrees in men and 120 degrees in women. The sharp angle in males causes the fused plates to form an anterior projection, also known as *Adam's apple* [6, 18]. Each thyroid lamina posteriorly projects two *cornua*, or horns; the *superior* and *inferior*. The two superior cornua project upwards and attach to the thyrohyoid ligaments, which link the thyroid cartilage to the hyoid bone. The two inferior cornua articulate the thyroid cartilage to the cricoid cartilage, forming the *cricothyroid joint*. (Note that the names of laryngeal ligaments and joints indicate the names of the laryngeal cartilages they link). The cricothyroid joints enable anteroposterior (back-and-forth) sliding of the thyroid inferior cornua against the cricoid cartilage. Moreover, it allows it to rotate, which in turn can lead to the lengthening and tensing of the vocal folds, allowing for pitch modification [17].

The *cricoid cartilage* is the lowermost laryngeal cartilage. It is a signet ring-shaped cartilage that lies below the thyroid cartilage and is the only laryngeal cartilage to encircle the airway completely. It is articulated with the thyroid cartilages inferior cornua, connected to the upper ring of the trachea via the cricotracheal ligament, and is also articulated with the arytenoid cartilage to form the *cricoarytenoid joint.* The cricoarytenoid joints are multiaxial joints and are the primary moving structure of the larynx. It allows the arytenoid cartilage to have multiaxial freedom of motion (sliding, rocking, and twisting rotation; excluding forward movement [18]), which, when combined with the contraction of the laryngeal muscles, can alter the position and shape of the vocal folds [17].

The *Arytenoid cartilage* are a pair of pyramidal cartilages, located at the posterosuperior border of the cricoid cartilage and articulated at the base with the posterior plate of the cricoid cartilage.

The *epiglottis cartilage* is the uppermost cartilage in the larynx, and is made of fibro-elastic cartilage. It is inferiorly attached to the thyroid cartilage via the thyro-epiglottic ligament. It is connected to the hyoid bone via the hyo-epiglottic ligament. The major function of the epiglottis is to help prevent aspiration during swallowing. During swallowing, the hyoid bone is elevated, causing the larynx to ascend, in turn causing the epiglottis to descend and its superior free-edge forms a lid-like protection over the larynx, preventing food from going down the airway [17, 18].

2.2.3 Laryngeal Muscles

The muscles of the larynx are traditionally divided into two groups: *intrinsic* and *extrinsic* muscles. In overview, the intrinsic muscles interconnect the laryngeal internal structure and control the mechanics of the vocal folds; the extrinsic muscles are attached to the larynx at one extremity and provide laryngeal stabilization

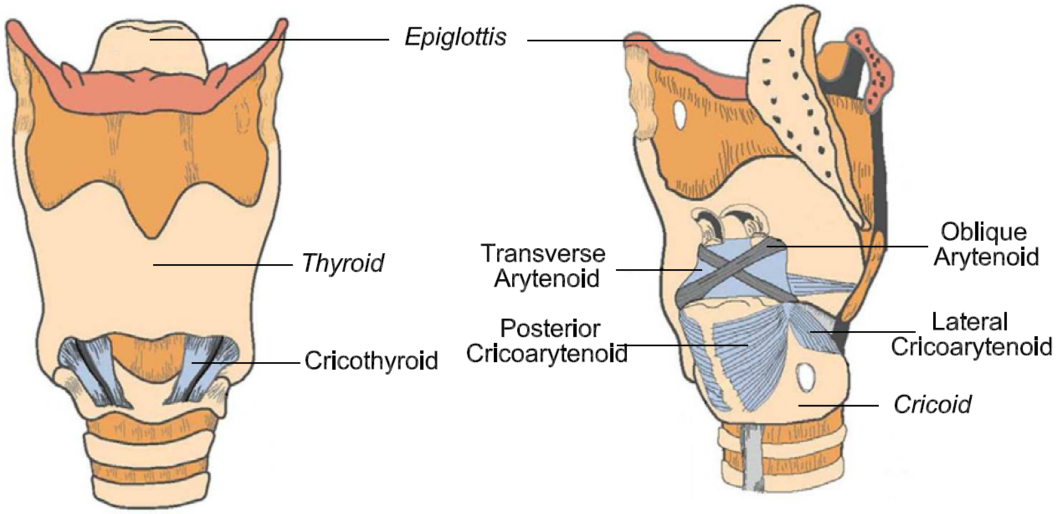

Figure 2.3: *Anterior (left) and oblique (right) view of the laryngeal muscles & cartilages (as in [19] adapted from [20])*

and displacement [17, 18]. The extrinsic muscles allow vertical movement, e.g. the larynx ascends during swallowing [18]. Since we are only interested in the voice production mechanism of the larynx, the extrinsic muscle structure is omitted from this chapter. The intrinsic muscles are responsible for altering the length, tension, shape, and spatial position of the vocal folds [17]. The three intrinsic muscles responsible for vocal fold adduction (closing of the glottis) are: *lateral cricoarytenoid*, *thyroarytenoid*, and *interarytenoid* muscles. An illustration of the laryngeal muscles is given in Figure 2.3.

The pair of *lateral cricoarytenoid muscles* govern the closing of the glottis. When contracted, the vocal folds move downwards and medially, leading to lengthening of the vocal folds, resulting in adduction (closure).

The pair of *thyroarytenoid muscles* control the tension of the vocal folds. The thyroarytenoid muscles cause the approximation of the arytenoid cartilages, drawing them closer to the thyroid, thus relax and shorted the vocal ligaments. This muscle structure consists of two muscle bellies (fiber bundles): the *internus* and *externus*. The thyroarytenoid internus muscles, also known as *vocalis* muscles, are the upper portion of the thyroarytenoid muscle and play a major role in the production of voice. The vocalis muscles form the main bulk of the vocal folds' body and are adherent to the vocal ligaments (this will be expanded upon below). The contraction of the vocalis muscles alter the tension of the vocal folds during speech production. The thyroarytenoid externus muscles cause the vocal folds to shorten and adduct.

The *interarytenoid muscle* is a singular muscle inserted transverse and obliqu-e between the right and left arytenoid cartilages. Contraction of this muscle causes the approximation of the arytenoid cartilages, closure of the posterior glottis, and narrowing of the inlet to the larynx. The intrinsic muscle responsible for vocal fold abduction (opening of the glottis) is the *posterior cricoarytenoid muscle.* The posterior cricoarytenoid facilitates vocal folds abduction via separation of the arytenoid cartilages, and has a key role in respiration by control of the glottic airway.

The intrinsic muscle responsible for vocal fold tension is the *cricothyroid muscle*, also referred to as a tensor muscle. The cricothyroid muscle connects the cricoid cartilage with the thyroid cartilage. Upon contraction, it rotates the cricoid cartilage about the horizontal axis via the cricothyroid joint. This causes the narrowing of the cricothyroid space, tilting the thyroid cartilage anteriorly, resulting in tension, elongation and thinning of the vocal folds. Therefore, the tensor muscle plays a primary role in voice pitch control, e.g. increasing the vocal folds resonant frequency. The thyroarytenoid and cricothyroid muscles are effectively an opposition pair and their relative degree of operation accounts for the variation in length of the vocal folds [16].

2.3 The Vocal Folds/Cords

The vocal folds are situated inside the larynx. They are made of mucous membrane, ligaments, and muscle tissue, with the latter made of the intrinsic thyroarytenoid muscle. The vocal folds are anteriorly attached to the thyroid cartilage and posteriorly attached to the arytenoid cartilages of the larynx. The orifice between the vocal folds is known as the glottis. When air passes through the vocal folds they vibrate. An illustration of the glottis is given in Figure 2.4.

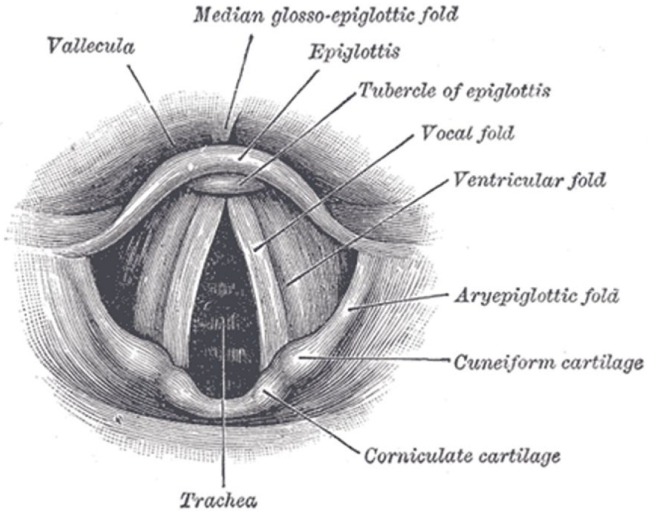

Figure 2.4: *The glottis is defined as the orifice between the vocal folds. From Gray's Anatomy of the Human Body, 20th edition.*

2.3.1 Anatomic Structure & Body-Cover Model

The vocal folds run across the narrowest point of the laryngeal airway, stretching between the thyroid notch and arytenoid cartilages. They are a complex layered structure with flutter-valve-like bahaviour; allowing airflow inwards and outwards of the lungs in respiration, effectively modulating airflow in the phonation process. The vocal folds are composed of layers of mucosa membrane and muscle fibers, forming a flexible structure that can stretch between three and four millimeters. The orifice of the vocal folds is known as the *glottis*, and it is the opening which enables the passage of airflow. The glottis can be divided into two sections: the membranous and cartilaginous. The membranous glottis lies between the vocal folds and comprises two-thirds of the glottis structure. The remaining third of the glottis lies between the arytenoid cartilages and is referred to as the cartilaginous glottis [6].

Although the vocal folds are complex in structure, their structure can be greatly simplified using the *body-cover model* [22]. The body-cover model divides the vocal folds into cover and body, and is useful for the purposes of understanding the biomedical structure of the folds and their vibratory cycle. Before elaborating

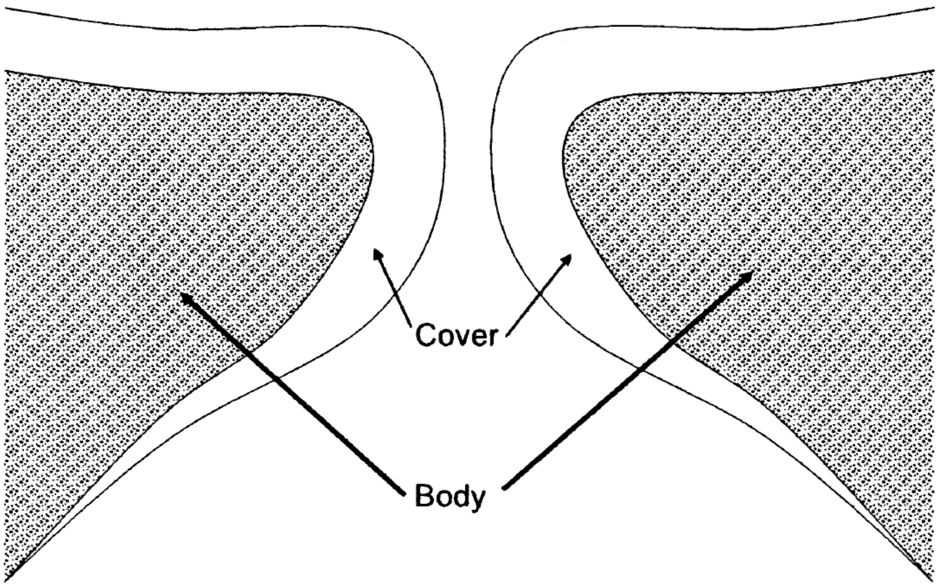

Figure 2.5: *The body-cover model (as in [21] based on [22])*

on this model, it is important to better understand the different layers in the vocal folds. The vocal folds consist of three primary layers: the *epithelium*, the *lamina propria*, and the *thyroarytenoid muscle*. The epithelium is the outermost layer of the vocal folds. It is a thin layer that protects the folds from degradation due to stress and friction. The lamina propria consists of three layers: *superficial*, *intermediate*, and *deep*. The superficial layer consists of loose connective tissue [14]. The intermediate layer consists primarily of elastic fibers. The deep layer consists primarily of collagenous fibers. The intermediate and deep layers form the *vocal ligaments* [14]. The thyroarytenoid muscle forms the bulk of the vocal folds and are adherent to the vocal ligaments [2]. Upon contraction, it causes vocal folds adduction at the midline. Moreover, it moderates vocal loudness and can alter vocal quality [6].

The body-cover model groups these layers into two levels. The *cover* is the top layer of the folds and it consists of the epithelium and the superficial layer of the lamina propria [6]. It is a non-contractile mucosal tissue, often referred to as the vocal folds mucosa, which serves as a sheath around the body [23]. The *body* consists of the thyroarytenoid muscle and the vocal ligaments. We note that the vocal ligaments, comprised of the intermediate and deep layers of the lamina propria, are sometimes referred to as the transition region between the cover and the body [14, 2]. In literature, the transition region is often grouped into the body layer for added simplification. The thickness of the cover-body layers varies with position along the vocal folds, with greatest thickness at the midpoint of the membranous vocal folds and lesser thickness at the cartilaginous edges. The different layers differ in properties, with relative stiffness between the layers, enabling the body to stiffen while the cover remains flexible, allowing it to move freely around the body, which is essential during voice production [6].

2.3.2 Vibratory Mechanism

The vibration of the vocal folds is essential in the process of speech production. If the vocal folds are abducted, the glottis is open and no vibration occurs. This would allow the glottal flow to be a continuation of the steady, outward flow of air expelled from the lungs. During vocal folds vibration, however, the glottal opens and closes cyclically, which modulates the airflow travelling through into a train of airflow pulses, providing an excitation source for the speech signal. The vibration of the vocal folds is related to the fundamental frequency of the speech, and its inverse is known as the *pitch period*.

In the 1950s, two vocal fold vibration theories were proposed: the *myoel-astic-aerodynamic* theory and the *neurochronaxic* theory. The former was conceived by Muller (1848) and formulated by Van der Berg (1958) and states that vocal fold vibration is the product of aerodynamic forces and the elastic recoil of the vocal folds tissue [6, 1]. The latter was suggested by Raoul Husson in 1950, proposing a neuromuscular theory which contrasted the myoelastic-aerodynamic theory. The neurochronaxic theory suggests that vocal folds vibration was the result of neural impulses and not airflow. This theory has been since refuted and discarded [24].

The myoelastic-aerodynamic theory has been updated through the years using computational models, such as lumped models [25, 26, 22, 27, 28] and finite-element models [29]. For this, we will give an overview of three lumped models. We will first look into the definition of the myoelastic-aerodynamic theory. We will then elaborate on the physical properties behind three lumped models, showing the development in the vocal folds oscillation theory. This will be followed by a detailed description of the kinematics principles of the vocal folds vibration cycle with regards to the structure of the folds, leading to the definition of the mucosa wave and the asymmetric forces driving the vocal folds self-sustained oscillation.

The myoelastic-aerodynamic theory is based on the elastic properties of the vocal folds and the Bernoulli Effect. At the beginning of the vibration cycle, the vocal folds are approximated medially and are adducted as a result of laryngeal muscles contraction. During phonation, the lungs act as bellows, expelling air up the trachea into the larynx. Since the glottis is closed, the subglottal air pressure builds up. When the subglottal pressure is sufficient to overcome the elastic forces of the vocal folds, the vocal folds are blown open, allowing a burst of air to pass through the now-open glottis. After the burst of air through the glottis, the vocal folds are adducted once again. As air flows through the orifice in the vocal folds, the subglottal pressure drops. As the pressure drops, the vocal folds return to their original closed configuration by the elastic recoil of the tissue structure of the glottis. This results in narrowing of the orifice. The Bernoulli Effect, also referred to as Bernoulli's Energy Law in fluids, states that if energy is conserved, pressure and velocity of gas/fluids are inversely related. Therefore, an increase in airflow velocity results in decrease in pressure, and vice versa. As the opening of the glottis narrows, the higher air velocity results in reduction of pressure, adducting the folds back towards the midline [6, 23]. This process then repeats cyclically. This mechanism is known as *self-sustained oscillation*. An illustration of the vibratory mechanism is given in Figure 2.6.

Subsequent studies have deemed the myoelastic-aerodynamic theory to be overly simplified, not providing an adequate representation for the vocal folds oscillating system. In turn, this led to the development of models for vocal folds self-oscillation [23, 30, 31]. Those models are an extension of the myoelastic-aerodynamic theory of phonation. The three lumped parameter models are: one-mass model [25], two-mass model [26], and the three-mass model [22] (see Figure 2.8). The one-mass model of the vocal folds is a single mass-spring oscillator, driven by airflow from the lungs. The one mass model determined that the fundamental frequency was more correlated with the aerodynamic forces. Moreover, it introduced the contribution of vocal tract inertia to vocal folds vibration. They found that the vocal tract will sustain oscillation only when coupled with an inertive vocal tract load [32, 33]. The model suggests that as the elastic recoil narrows the glottis opening, an area of negative air pressure forms above the

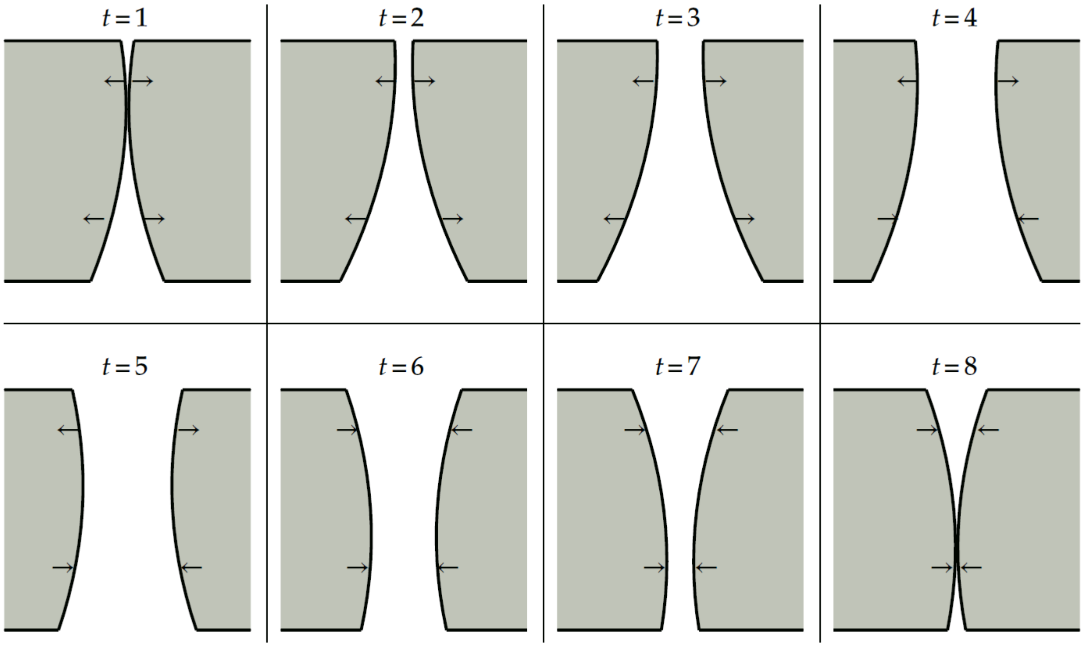

Figure 2.6: *The idealized cycically vibration pattern of the vocal folds. The wave-like propogation about the vertical axis illustrates the propogation of the mucosal wave (as in [23])*

glottis. This negative pressure contributes to the adducting of the vocal folds to end the vibratory cycle. Although the model can be viewed as an improvement, it models the vocal folds as a single unit of mass, discarding the effect of the glottis. Moreover, the one mass model only allows for lateral movements of the mass in oscillation, not producing the vertical phase difference required for oscillation [30, 25, 32].

The one-mass model was improved with the two-mass model, where the vocal folds were approximated as a self-oscillating source of two couple masses [26]. An illustration of the two-mass model is given in Figure 2.7. This model was further enhanced into the three-mass model. An illustration of the three-mass model is given in Figure 2.8. The three-mass model gives a better approximation of the vocal folds cover-body structure. It is a lumped element model similar to the two-mass model, where two masses coupled to each other, m_1 and m_2, are used to approximate the cover of the vocal folds, with a third, greater mass, m_b, added to approximate the body (thyroarytenoid muscle and vocal ligaments) of

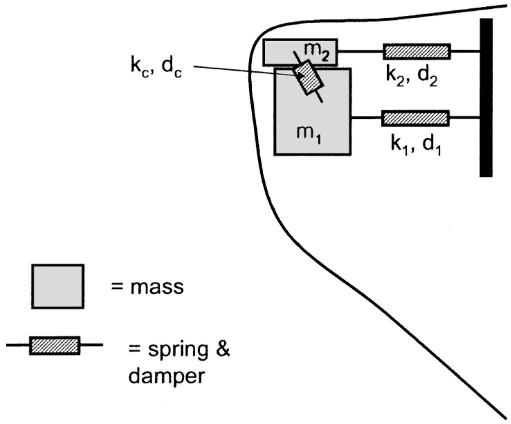

Figure 2.7: *The two-mass model based on the cover-body structure [21].*

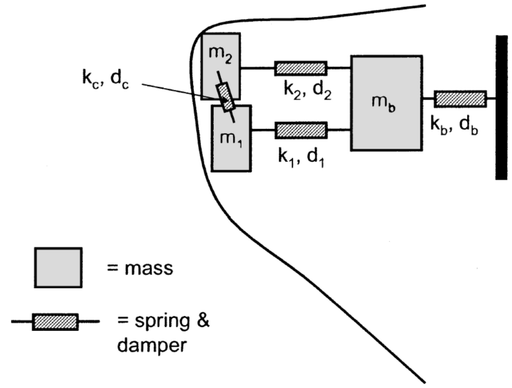

Figure 2.8: *The three-mass model based on the cover-body structure [21].*

the vocal folds. The masses are coupled through non-linear springs and damping elements, k and d [34]. The two cover masses are coupled to one another through a linear spring, which represents the vertical phase difference in glottal contact during vibration, also known as the *mucosal wave* (see Figure 2.6). This model shows a glottal vertical phase difference, with altering convergence and divergence of the glottis as it adducts and abducts. The multiple mass-models have been further developed, introducing models with varying number of masses, with upwards of 16 masses per model [27, 28]. The variations in models led to the current theory that suggests vocal folds oscillation occurs when an asymmetry exists between the inter-glottal aerodynamic driving forces during the vibratory cycle [23, 30, 31, 34].

There are two mechanisms which account for self-sustained oscillation of the vocal folds: *non-uniform glottis deformation* (glottis convergence/divergenc-e) and *opposing supraglottal and subglottal pressures* (inertive vocal tract loading) [34].
The driving force for vocal fold oscillation is the intra-glottal pressure inside the glottis over the medial surfaces of the vocal folds, P_g. The mean intra-glottal pressure equation is shown in Equation 2.1 below:

$$P_g = (1 - \frac{a_2}{a_1})(P_s - P_i) + P_i \tag{2.1}$$

Where P_s is the subglottal pressure (Pa), P_i is the input pressure to the vocal tract, and a_1 and a_2 are the cross-sectional areas at the glottal entry and exit, respectively [1, 23]. Non-uniform deformation of glottis geometry is anchored on the vocal folds body-cover model. During oscillation, the cover varies in shape over the stiffer body. In the process, the flexible tissue of the glottis alters in shape throughout the vibration cycle (see Figure 2.6). It is known that the vocal folds vibrate with a vertical phase difference [35]. This lateral phase difference is known as the mucosal wave. As a result, the glottis undergoes convergence and divergence throughout the vibratory cycle. The mucosal wave propagation is the result of intra-glottal pressure differences. The mean intra-glottal pressure, P_g, over the medial surfaces of the vocal folds can be approximated as per Equation 2.2 (assuming no vocal tract loading in the system) [23]:

$$P_g = (1 - \frac{a_2}{a_1})P_s \tag{2.2}$$

By utilizing Bernoulli's energy conservation law and the flow continuity principle of fluid dynamics, duct area is proportional to duct pressure. As the mucosal wave propagates upwards, altering the shape of the glottis, it produces different glottal pressure profiles. Therefore, as the glottis converges a_1 is greater than a_2, and P_g is greater than zero. As the folds abduct outwards, the elastic recoil force resists the motion, resulting in glottis divergence, in which a_2 is greater than a_1 and P_g becomes negative. This establishes the driving force (intra-glottal pressure) asymmetry [34].

The inertive vocal tract loading can be modelled using the aforementioned Equation 2.1. When modelling the non-uniform deformation of the glottal geometry, the equation is simplified by assuming the glottal source is open to the atmosphere, where $P_i = 0$. This assumption eliminates the contribution of the vocal tract load. In this case, however, we are interested in modelling the inertive vocal tract driving force, thus we eliminate the deformation of the glottis, effectively equating the cross-sectional areas and removing the mucosal wave, resulting in Equation 2.3 [23]:

$$P_g = P_i \tag{2.3}$$

Now, the only driving force for vocal folds oscillation is the supraglottal, input pressure to the vocal tract, P_i [34]. The pressure at the input of the vocal tract is correlated to time variations in the glottal flow. This concept is more easily understood by considering the vocal tract to be made of a series of concatenated tubes. The tubes are interconnected at their boundaries, with the edge of the first tube connected posteriorly to the glottis and the edge of the last tube connected superiorly to the output of the system. The cross-sectional area, A, varies between the linked tubes. Therefore, as the glottal flow propagates upwards through the vocal tract, it is reflected at the tube boundaries, returning to the base of the vocal tract and resulting in input pressure perturbation. Therefore, at frequencies below the first formant frequency (concentration of acoustic energy about a frequency value), the pressure at the input of the vocal tract is proportional to time-variations in the glottal flow, U, through an inertance coefficient, I [23, 36]. This function is given in Equation 2.4 below:

$$P_i = I\frac{dU}{dt} \tag{2.4}$$

This equation is essentially Newton's 2$^{\text{nd}}$ Law of Motion, where force is analogous to P_i and mass is inertance. Inertance is defined as a function of the vocal tract length and area as per Equation 2.5 given below [23, 36]:

$$I = \frac{\rho L}{A} \tag{2.5}$$

Where ρ is air density, L is the vocal tract length, and A is the vocal tract cross-sectional area. Since I is inversely proportional to A, it can be readily seen that changes in the cross-sectional areas of the tubes lead to varying I throughout the vocal tract. As a result, as the pressure wave propagates through the vocal tract it passes through a medium of varying inertance [36]. These variations cause the pressure wave to partially-reflect and partially-transmit with each change in inertance at the boundary of neighboring tubes. Since P_i is proportional to the rate of change in glottal flow, P_i is positive during the glottis opening phase, where the rate of change is positive, and vice versa [23, 34]. During the closing phase, the flow rate of change is negative, resulting in negative pressure at the input of the vocal tract, which assist in closing the glottis. Thus, self-sustained oscillation is the result of an inertive vocal tract loading.

In speech phonation, both mechanisms act in concert to produce speech. Vocal folds self-oscillation models utilize these asymmetric driving forces to model the vibratory cycle of the vocal folds, e.g. the three-mass model. However, we note that in theory those are not mutually exclusive; the presence of just one of those asymmetric forces is sufficient to produce sustained vocal folds oscillation [23, 34].

2.4 The Vocal Tract

The vocal tract acts as a resonant cavity and an acoustic filter, shaping the spectral properties of the glottal airflow source. The vocal tract is a respiratory airway extending from above the glottis to the lips. It consists of the larynx, pharynx (throat), oral cavity, and nasal cavity. The vocal tract length is about 17 cm and 15 cm for adult males and females, respectively [2]. The length and cross-sectional area of the vocal tract affect the nature of the speech signal. Its parameters can be varied by altering the shape and placement of the tongue, teeth, jaws and lips. We first present breakdown of the anatomical structure of the vocal tract cavities, followed by their role in speech sound production. This is followed by an overview of the vocal tract role as an acoustic filter, presenting some of the popular models used to approximate the effect of the vocal tract, which is of primary importance in speech processing and speech synthesizing.

2.4.1 Vocal Tract Cavities

The anatomical structure, location, and functions of the larynx have been previously presented in this chapter and therefore are omitted from this section. The larynx is connected to the pharynx at the epiglottis (uppermost cartilage) boundary. The *pharynx* is an air passage which extends from the back of the mouth and nose to the larynx and esophagus. The pharynx belongs to both the respiratory and digestive systems. It is divided into three sections: nasopharynx, oropharynx, and laryngopharynx. The *nasopharynx* is the superior part of the pharynx, extending vertically into the nasal cavity. The *oropharynx* refers to the middle part of the pharynx. It is situated behind the mouth, extending from below the tongue to the anterior end of the soft palate, and forms a pass way for both food and air. The *laryngopharynx* is the inferior part of the pharynx. It extends from the oropharynx and diverges into the larynx (respiration pathway) and the esophagus (digestive pathway).

The *nasal cavity* refers to the air space above and behind the nose. It is divided into two cavity chambers by a part-cartilage, part-bone vertical partition known as the nasal septum. Each one of the cavity chambers terminates anteriorly by a nostril and posteriorly by the nasopharynx. The nasopharynx extends to the nasal passage through the velopharyngeal opening. This pathway can be constricted using the soft palate, also known as *velum* (rear of the roof of the mouth). The velum can be elevated to form the floor of the nasal cavity, thus restricting airflow into the nasal passage. Alternatively, the velum can be lowered, opening an air passage into the nasal cavity. When the velum is lowered and the oral cavity is sealed by the lips, the sound wave can propagate upwards through the nasal passage and outwards through the nostrils. The nasal cavity is coupled with the oral cavity via the velum.

The *oral cavity*, also referred to as the *mouth cavity*, consists of the lips, tongue, and teeth. The oral cavity is considered the space bounded by the lips and the oropharynx wall longitudinally, the cheeks laterally, and the floor of the mouth and the palate (roof of the mouth; hard and soft palates) vertically. The primary function of the oral cavity is to allow consumption of food and begin the digestive process. Its secondary functions are respiration and phonation. The lips are the anterior opening of the vocal tract and allow for sound propagation. The tongue is a muscular organ located above the mandible (lower jaw) and connected to it via muscle fibre. Changing the shape and position of the tongue varies the cross-sectional area of the vocal tract.

2.4.2 The Vocal Tract as an Acoustic Filter

The vocal tract is an acoustic chamber. As an acoustic system, the vocal tract has resonance. Resonance frequencies are the system's natural frequencies of vibration. As a sound wave propagates in the vocal tract, it will be amplified at frequencies matching the vocal tract resonances. In speech production, resonance frequencies are often referred to as *formant frequencies*, or *formants*. Formants are represented by capital **F** with a number indicating their order, i.e. F1 is the first formant frequency. The vocal tract formants are correlated to the length and shape of the vocal tract. During speech production, the length and cross-section of the vocal tract (area function) change due to movement of its articulators (tongue, teeth, velum, jaw, lips), corresponding to change in speech sounds. Different speech sounds have different vocal tract area functions and formants. This subsection presents a simplified model of the vocal tract known as the *single tube model*. In turn, it leads to the understanding of the *concatenated tube model*. Theory regarding the concatenated tube model and techniques used to collect area function data from the vocal tract follow. Understanding the correlation between vocal tract shape and formats is of key importance, as it merges with the previous section and provides the background for the next section which details the correlation between vocal tract articulators and speech sounds.

Different speech sounds have different vocal tract area functions and formants. To better understand this concept, the structure of the vocal tract is stripped of its articulators, and its shape is simplified to a cylinder. Hence, for simplicity, first consider the vocal tract as a single, uniform, lossless cylinder or tube. Consider a cylinder closed at one end (accounting for the glottis during the closing phase) and open at the other (lips). This is known as the single tube model.
Consider a standing wave in the tube. The frequency of the sound wave is calculated using Equation 2.6:

$$f = \frac{c}{L} \tag{2.6}$$

Where c is the speed of light, L is the length of the tube, and f is the frequency of the sound wave. As f is correlated to L, the sound wave will partially radiate and partially reflect at certain frequencies. Therefore, the sound wave exhibits constructive interference (anti-node) and destructive interference (node) patters. Resonance frequencies are those for which the sound wave is amplified (anti-nodes). Consider a tube 17 centimeters in length (average vocal tract length in male speakers). Utilizing the above equations for tube lengths of $4L$, $1.33L$ and $0.8L$ its formants are $F1 = 500$ Hz, $F2 = 1500$ Hz, $F3 = 2500$ Hz. The formant frequencies can be calculated via Equation 2.7; where F_n is the n$^{\text{th}}$ formant frequency of the vocal tract [9]:

$$F_{n+1} = \frac{(2n+1)c}{4L} \tag{2.7}$$

The above model for the vocal tract is conceptually simple. However, it negates the effects of the articulators and the mechanical properties of the vocal tract walls, as well as potential energy losses. An improvement was made on the single tube model in the form of the *concatenated tube model.* An illustration of the concatenated tube model is given in Figure 2.9, where $\mathrm{A_n}$, represents the variation in cross-sectional area along the vocal tract.

This model depicts the vocal tract as a series of linked, short, lossless cylindrical units with varying cross-sectional area ($\mathrm{A_1}$, $\mathrm{A_2}$, $\mathrm{A_3}$, $\mathrm{A_4}$). Increasing the number of

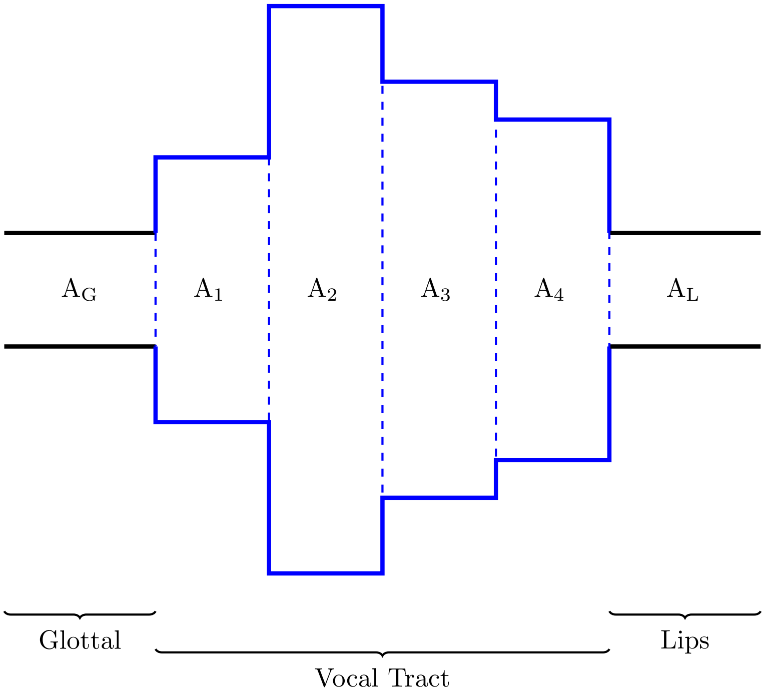

Figure 2.9: *Vocal tract concatenated tube model*

cylinders allows for greater precision when modelling the vocal tract. Therefore, this model is more robust and is widely used due to its linearity and computational simplicity. The concept of vocal tract representation as an acoustic tube comprised of multiple cylinders varying in length and cross-sectional areas has been in place for over half a century [37]. The combination of the cylindrical lengths and cross-sectional areas is known as the vocal tract *area function*. Variations in the vocal tract length and area function affect the formant frequencies. The area function models are more complex and require area function data inventory (an overview of the methods used for the collection of such data is given in the following paragraph). The correlation between the area functions and formants is given via the acoustic sensitivity functions. The sensitivity functions describe how small perturbations in area function regions cause format frequencies to become higher or lower [38, 39, 40, 41]. Therefore, sensitivity functions give indication to which sections of the vocal tract should be spatially altered in order to change the formants values. The sensitivity functions were used in the Distinctive Region and Mode theory [42]. This theory proposed the *Distinctive*

Region model. The Distinctive Region model divided the uniform vocal tract tube into eight regions of different length, or distinctive regions. The model showed how small perturbation in area function, caused by altering the dimensions of either one of the eight regions, induced an alteration in formant frequency [42, 40]. Thus, different articulatory modes for different speech sounds could be quantified by the length and cross-sectional area of the distinctive regions.

Vocal tract models are used in articulatory synthesizers to create human speech. The closer the vocal tract approximation is to the area function of a speaker, the more realistic the approximation. Therefore, an area function inventory is required [43]. There has been great progress in area function data collection techniques. Fant *et al* [37] collected data from vowel production of a single speaker to model the vocal tract as concatenated tubes. He then combined the vocal tract area function data to create an inventory which was used as input for articulatory synthesizers [37, 12]. The formants of the models were calculated and compared to the formants of human speech signal for validity. Fant collected the data using radiographic film (X-ray) [37, 44, 45]. Advances led to the three-dimensional modelling of the area function by interpolation of data collected via the following methods: CT scan [46]; Magnetic Resonance Imaging (MRI) [43, 47]; Articulators ultrasound sound [45], and Acoustic Reflectometery (AR) [48, 49]. These data collection techniques allow for the modelling of a more accurate and realistic area function, in turn allowing for better understanding of speech sounds articulation. The role of the vocal cavity in speech sound production is further elaborated on in the next section.

2.4.3 Elements of Language

Spoken language can be represented by a set of discrete symbols, i.e. alphabet. In language, the smallest unit of sound in a word is known as *phoneme*, and for minimal word pairs it distinguishes one word from another, e.g. 'pet' and 'pit'. Phonemes provide a link between alphabetical written language and spoken language. Phonemes can be sub-divided into classes of sounds: vowels, diphthongs, semi-vowels, and consonants. Each of these classes can be further divided into sub-classes. Phonemes can be classified, in terms of the vocal tract behaviour during their phonation, as *continuant* and *non-continuant*. Continuant phonemes are produced with a time-invariant vocal tract and include the vowels, nasals, and fricatives classes. In contrast, non-continuant phonemes are produced with a time-variant vocal tract and include diphthongs, semi-vowels, stops, and affricates. In general, sounds are classed as either *consonants* or *vowels*. For the scope of this thesis, vowels are of primary interest and their properties will be discussed further. However, a brief overview of the other phoneme classes is given as well.

Aside from vowels, consonants are the other class of phonemes. Consonants are produced with a time-varying vocal tract. The sub-classes of consonants include nasals, fricatives, plosives, and affricates. A brief description of several of those sub-classes follows. Nasal are produced with the velum lowered and the vocal tract constricted. Thus, air flows upwards through the *velopharyngeal* opening into the nasal cavity and is radiated through the nostrils. The nasal consonants are /m/ and /n/ and they differ by the location of the vocal tract constriction. Fricatives can be either voiced or unvoiced. They are produced by a noisy airflow source into the vocal tract, where two articulators combine to form a constriction, i.e. /f/. Plosives, also known as stops, are similar to fricatives in that they can be either voiced or unvoiced. Plosives are produced by building up pressure inside a constricted vocal tract and then unblocking it, resulting in a burst of air, i.e. /p, k, t/ [50, 51].

Vowels result from a time-invariant configuration of the vocal tract, where the glottal flow is unrestricted except for at the glottis. As mentioned in the previous section, the shape of the vocal tract can be quantified via its *area function*. The area function varies between different sounds due to the change in spatial position of the vocal tract articulators (i.e. tongue). The changes in the cross-sectional areas of the vocal tract determine the formants. A vowel can be categorized by the vocal tract area function and/or by the resonance frequencies of the vocal tract. In terms of area function, the vowels can be plotted in vowel space, which is related to the openness of the jaw in the y-axis and the backness of the tongue in the x-axis. Through proper choice of sections lengths and cross sectional areas, a vocal tract can be modelled to exhibit formant frequency distributions of a desired vowel [50, 43, 37]. This can be readily seen in the International Phonetic Alphabet (IPA) vowel chart, which depicts how the location of the tongue affects the vowel produced, as shown in Figure 2.10.

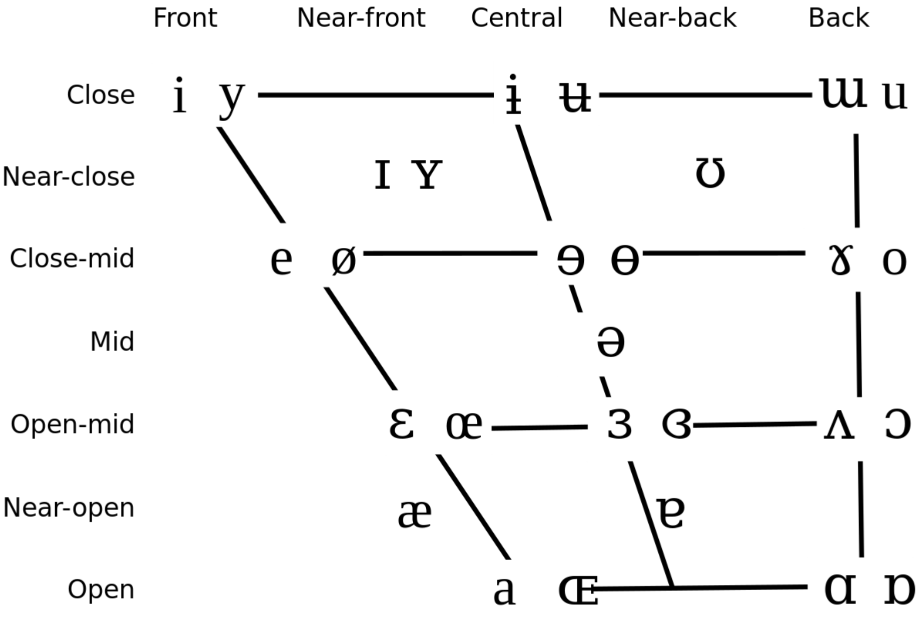

Figure 2.10: *IPA vowel chart* [52]

Alternatively, formant frequencies convey information about vowels. The first and second formants, F1 and F2 respectively, convey most of the information with regards to the vowel quality, whereas F3 and the fundamental frequency convey information about the speaker [37]. The energy of the formants can be plotted in the form of *spectrograms*. Spectrograms are a graphical display of energy density of a frequency at a time instance, computed using the short time Fourier Transform [50, 51]. This led to the F1 × F2 plane, where the x-axis and y-axis correspond to the F1 and F2 formants, respectively. The F1 × F2 plane is a visual representation of the variation in formant frequencies for each vowel separately, and across all vowels collectively [53]. The first two formants are associated with the tongue location in the vocal tract, with tongue height inverse to F1 and tongue backness related to F2 [37, 38]. Vowels are often described in terms of articulatory features: back/front, high/low, rounded/unrounded, tense/lax. Tense and lax vowels are also referred to as long and short vowels, respectively, with the former denoted by the vowel's phonetic symbol followed by a column, i.e. /i:, a:, o:/. Note that tense vowel, unlike lax vowels, can be word final. Low and high vowels refer to the location of the jaw i.e. in high vowels the jaw is relatively high, implying the mouth is closed. Therefore, low/high vowels are also known as open/closed vowels, respectively [50, 51, 53, 38]. For the purposes of this thesis, 11 New Zealand English (NZE) vowels were chosen for analysis. The 11 NZE vowels in International Phonetic Alphabet (IPA), Machine Readable Phonetic Alphabet (MRPA), and Speech Assessment Methods Phonetic Alphabet (SAMPA) is provided in the Table 2.1 below. Mapping between the 11 NZE monophthongs and their respective hVd words is provided in Table 2.2.

Table 2.1: *Phonetic Symbols of 11 New Zealand English monophthongs*

	short vowels						*long vowels*				
IPA	I	ʊ	ɛ	ɒ	ʌ	æ	i:	u:	ɔ:	a:	ɜ:
MRPA	I	U	E	O	V	A	i:	u:	o:	a:	@:
SAMPA	I	U	e	O	6	{	i:	}:	o:	6:	3:

A single vowel is referred to as a monophthong. If the vowel quality changes within the syllable, it is referred to as a diphthong. Diphthongs are a two-part vowel sounds made of two vowels transitioning from into the other under the same syllable. In diphthongs articulation, the vowel sound changes as the tongue and lips transient between the two different vowel positions. Another type of vowels are the nasal vowels. Nasal vowels are produced with the velum lowered and the vocal tract constricted. Thus, air flows upwards through the *velopharyngeal opening* into the nasal cavity and is radiated through the nostrils. The English language, unlike French, does not have nasalized vowels [50, 51].

Phonemes may exhibit *coarticulation*. Coarticulation refers to changes in the articulation and acoustic characteristics of a phoneme due to effects of preceding/following phonemes (known as phonetic context). In continuous speech, the acoustic characteristics of vowels will change depending on the context in which they are spoken [53]. Therefore, coarticulation must be considered when performing vowel analysis. In addition, when performing vowel analysis, we extract the *vowel target*. The vowel target refers to the time instance that is in the temporal midpoint of the monophthong. It is at this point that the vowel typically exhibits the smallest spectral variations, and it is the part of the vowel that is least influenced by coarticulation effect [53].

Table 2.2: *hVd Words for 11 New Zealand English monophthongs*

	vowels		
	IPA	*MRPA*	*SAMPA*
HID	I	I	I
HOOD	ʊ	U	U
HEAD	ɛ	E	e
HOD	ɒ	O	O
HUD	ʌ	V	6
HAD	æ	A	{
HEED	i:	i:	i:
WHO'D	u:	u:	}:
HOARD	ɔ:	o:	o:
HARD	a:	a:	6:
HERD	ɜ	@:	3:

2.5 Glottal Analysis Methods

Analysis of the glottal and vocal folds behaviour during speech production is of primary importance in this thesis. As the folds are located in the larynx they are not easily accessible. The extraction of information is therefore either invasive or non-invasive, with the latter being naturally favourable. Throughout the years, there have been various procedures to extract information about the behaviour of the vocal folds during phonation. The information extracted was mainly used for research studies and/or clinical observations. In this section, a description of the following analysis techniques is given: high speech cinematography, high speed imaging, Photoglottography (PGG), and Electroglottography (EGG). We note that the inverse filtering method is the one utilised in this thesis project for extraction of the glottal flow waveform. This topic is further discussed in Chapter 4.

2.5.1 High Speed Imaging

The vibratory pattern of the vocal folds motion can be captured using a high-speed imaging system. High-speed imaging system comprises of a film camera with high shutter speed. This imaging system, which uses high speed cinematography, was utilized at the forefront of vocal folds behaviour research [54, 55]. When observing the behaviour of the vocal folds, it is favourable to record the speech signal simultaneously for subsequent analysis [56, 57]. Since high speed cinematography utilized film cameras, the equipment produced high levels of mechanical noise. As a result, the recorded speech signal was distorted by noise and deemed unusable. This equipment was plagued with other technical issues which increased its complexity and made it disadvantageous to researchers. A notable problem was the use of film stock. The images were captured on film and required development before analysis, thus restricting the flexibility of the data-collection process, not allowing for interactive analysis of results.

Dissatisfied with the existing equipment at the time, a new device for high-speed recordings of vocal folds motion was developed in the form of an *endoscope*. The endoscope is similar in structure to scopes used at the time for clinical laryngeal observations. It is a solid-state device, consisting of an image sensor and an image processor. It allowed for high speed digital imaging, where the video output of the sensor was fed through an A/D converter to the processor. Thus, the images collected were displayed on an external screen simultaneous to the data collection process, allowing for interactive investigation of the data captured. The endoscope was superior to the high speed cinematography procedure, as the device is small, compact, and does not produce mechanical noise. However, the endoscope is invasive in nature, as it required oral insertion in order to capture data during speech production. Due to its solid state, it could only be used for sustained phonation. The endscope was improved by introducing the *fiberscope* [56, 57]. The fiberscope is an improved flexible, fiber-optic endoscope. Unlike the endoscope, the fiberscope is inserted nasally and lowered through the respiratory airways to the glottis. It captures the vocal folds behaviour via video at 2000-5000 frames per second. The fiberscope's flexible and slimmer frame allowed for greater flexibility in speech data recording compared to the endoscope. Whereas the endoscope is limited to sustained phonation, the fiberscope could capture the behaviour of the vocal folds during continuous speech [55, 56, 57].

2.5.2 Electroglottography

Electroglottography (EGG) is a simple, non-invasive, innocuous electric method for investigation of vocal folds behaviour during phonation. Electroglottography is used to measure the varying impedance between the vocal folds, when an alternating current is applied. EGG builds upon the electrical conductivity of the human tissue. During the vibratory cycle, the vocal folds open and close. The opening of the glottis introduces an air gap. Air is a poor conductor, with smaller conductivity compared to human tissue. Thus, the impedance is higher for air than it is for human tissue. This change in impedance is recorded and is known as the Electroglottographic signal, or *electroglottogram*.

Electroglottography measures the impedance about the larynx using an external apparatus. The apparatus consists of two electrodes on the middle of each thyroid lamina [58]. Each electrode has a surface are of a few centimetres. A steady oscillator generates a high-frequency AC current, with frequency ranging between 300 kHz and 5 MHz, and current no more than 10 mA [59, 60]. The current converges from one electrode to another, forming a current field. The path of the current field does not form a straight line between the electrodes. Instead, current is radiated in all directions and is coupled into the neck which acts as a volume conductor [59]. Electroglottography utilizes Ohm's Law and the biological conductivity of the human tissue. According to Ohm's Law, resistance is inversely proportional to current (V = I·R). Since the electrodes radiate an AC current, EGG deals with impedance. Therefore, impedance is inversely proportional to current, with changes in current flow resulting in changes in impedance. This is known as the amplitude modulation of an AC carrier [59]. The changes in impedance is plotted against time to form the electroglottogram. In turn, the electroglottogram provides information about the vocal fold contact area as a time-varying signal. The vocal fold contact area is determined via the changes in impedance as a result of different conductive properties of human tissue and air. Human tissue has superior conductivity to air. During the vocal folds vibration, the vocal folds abduct and the glottis opens. This introduces an air gap, which adheres the current flow. Therefore, during the vibratory cycle, impedance rises and falls as the glottis opens and closes, respectively. Impedance is inversely related to vocal fold contact area; vocal fold contact increase results in impedance decrease, and vice versa.

The EGG signal is invaluable for research studies and clinical observations. The signal is advantageous to vocal folds research as it is simple, non-invasive, and free of supraglottal influences. The electroglottogram is used to examine glottal behaviour, speech disorders, and respiratory behaviour of the larynx. Moreover, the signal is used to derive quantitative parameters that provide information about the speaker's age, gender, and voice quality [61, 62]. In addition, the signal is perfectly in phase with the recorded speech signal. Its periodicity allows for extraction of vocal fundamental frequency (F0) and fundamental frequency cycle-to-cycle perturbations (jitter) [59, 62]. The electroglottogram provides a two-dimensional plot of impedance changes (y-axis) with time (x-axis). There are two conventions regarding the behaviour of the waveform as one moves upwards on the y-axis. The conventions are derived from two popular ideal EGG waveforms. Childers [63] model suggests that vocal folds contact decreases up the y-axis, with the peak of the wave signifying minimum contact. Rothenberg [64] model suggests the opposite, with vocal folds contact increasing up the y-axis. These models are inverted versions of one another and both suggest that electroglottogram provides a representation of the relative contact area between vocal folds during phonation.

The original, unfiltered electroglottogram signal is referred to as G_x [65]. Most

EGG devices include built-in signal processing applications which account and compensate for signal perturbations. The signal is often high filtered to remove low frequency perturbation. High pass filtering eliminates the slow impedance changes due to laryngeal movements, thus amplifying the movement of the vocal folds vibration. The high-pass filtered signal is referred to as L_x [59, 58, 65]. The device also includes electric circuit to compensate for other errors: An Automatic Gain Control circuit which eliminates the potential DC-bias which may arise due to larynx movement or changes in electrode-to-skin contact; a circuit to minimize the effects of uneven current field distribution about the neck; a circuit to account for slow changes in impedance due to head movement of skin moisture [59].

The impedance measurements are affected by the conductivity of the human tissue. As the current is coupled into the neck, the electric field is unevenly distributed between the two electrodes due to changes in tissue structures in the larynx [59]. This current field spread about the different tissue structures of the neck results in multiple paths for current flow. This leads to only one or two percent of the total generated current reaching the vocal folds [58]. Moreover, the impedance is affected by electrode-to-skin contact. Skin moisture or oil have different conductivity than skin tissue. Therefore, the electrodes should be placed on a smooth and clean skin [58]. In addition, the location of the electrodes is of major importance. The electrodes should be placed at their ”ideal” place; middle of the thyroid lamina, where the EGG signal is most prominent. Incorrect placement of the electrodes can be detected by observing the electroglottogram, which may result in a weak signal or reversal in signal polarity [58]. Displacement of the electrodes can lead to perturbation in the impedance signal. During phonation, the electrodes could shift from their ideal position due to laryngeal movements; raised, lowered, bent. In turn this results in slow impedance changes in the electroglottogram [58, 59].

One of the more significant shortcomings of the EGG waveform is its inability to convey information with regards to non-contact vocal folds physiology. During the open phase, after the abduction of the vocal folds, the folds may continue to depart laterally. This leads to an increase in glottal area, which in turn results in an increased glottal flow. However, such an event will not be registered in the EGG waveform. Similarly, during the vibratory cycle the vocal folds exhibit changes in the amount and region of contact. Furthermore, the folds exhibit changes in folds contours that do not involve contact. Since the EGG waveform provides a record of change in contact and not movement, non-contact vibratory events will go unregistered and unnoticed [63]. It is important to note the EGG signal is sensitive to the inter-glottal changes caused by the vertical propagation of the mucosal wave [63, 66]. Those changes may register as spikes in impedance, due to the anterior-posterior changes in vocal folds contact area, even though there was no change in glottal area and thus no change in glottal airflow.

The EGG waveform is also susceptible to *mucous bridge* effect. A mucous bridge forms a current path between the vocal folds. As its conductivity is higher than air, it registers as a contact region between the vocal folds, and observed as downward deflection in the EGG signal. This indicates the adduction of the already partially-abducted vocal folds, suggesting an opening-closing-opening behaviour mid-cycle [58]. Moreover, when the folds abduct to break the mucous bridge, a large spike will be registered in the EGG time derivative signal. This positive peak approximates the instance of glottal opening, the instance at which the glottal increases in size from its minimum value. In the occurrence of the mucous bridge, however, this instance will not correspond to the moment of glottal opening. Thus, a mucous bridge provides false temporal information about the abduction of vocal folds [63]. This could lead researchers to misguided conclusions about the vocal folds behaviour during phonation.

Electroglottography has been extensively used and featured in literature. Its inherent invasive nature, however, makes it for a nuisance. In this project, we strive to examine the behaviour of the vocal folds through examination of the glottal flow waveform. The glottal flow can be extracted from the speech signal, and thus it is a non-invasive procedure. The topic of glottal flow is heavily featured in this thesis work and is the topic of the next chapter.

Chapter 3

Glottal Source Signal

3.1 The Glottal Pulse

Obtaining the glottal flow is the first step in the process to parametrize and analyse its temporal characteristics. As shown in the next chapter (Chapter 4), the glottal source can be separated from the speech signal through *source-filter deconvolution*. The extracted glottal flow signal for voiced speech can then be fitted with a parametric model. The parametric model relates the glottal flow to the behaviour of the vocal folds during the vibratory cycle. Thus, we can use information derived from the glottal flow to analyse the speaker's condition. Although the glottal flow can be fitted with various time-domain and frequency-domain models, several properties hold true for all glottal waveforms: the glottal waveform is a quasi-periodic positive function that is both continuous and differentiable in time, except for at instances of glottal closure (GCI) [67, 68]. Consider the Liljencrants-Fant (LF) glottal model, illustrated in Figure 3.1 (see Appendix A).

The Liljencrants-Fant parametric model is fitted to the glottal flow and provides a tool for temporal parametrization of the glottal source signal. As illustrated in Figure 3.1, the glottal pulse can be segmented to temporal phases. These temporal phases are: *open, closed, opening, closing* and *return* phases. The acronyms GOI and GCI refer to the *glottal opening instance* and *glottal closing instance*, respectively. The different phases and instances correspond to vocal

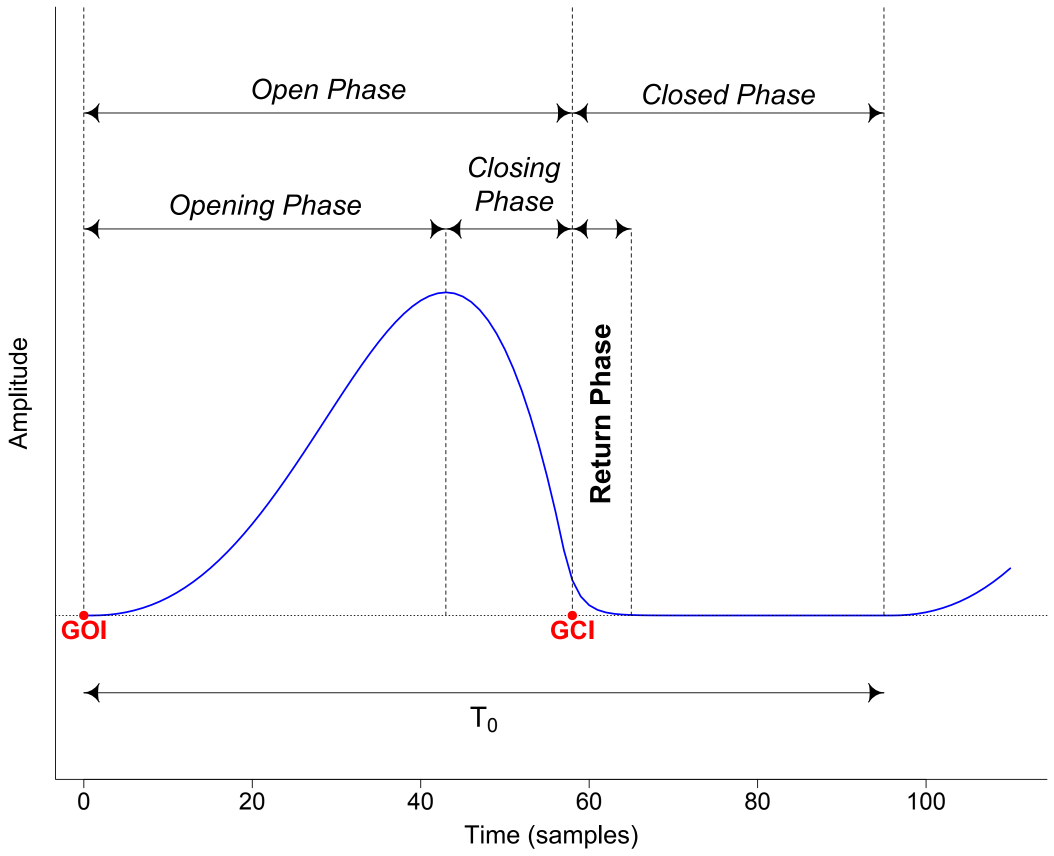

Figure 3.1: *Liljencrants-Fant glottal pulse model with temporal phases*

folds behaviour. A description of each temporal information segment ensues.

- T_0: fundamental period of the glottal cycle; inversely related to the fundamental frequency of vocal folds vibration; $T_0 = 1/f_0$. Determines the perceptual parameter pitch.

- *Open Phase*: the time duration where the vocal folds are open; extends between neighbouring glottal opening instance and the following glottal closing instance temporal points; embodies of the opening and closing phases, which are temporally divided by the instance of maximum glottal flow. The open phase determines the width of the glottal pulse, and is quantified using the Open Quotient parameter. It is different than the Opening Phase (see below).

- *Opening Phase*: extends between moments of glottal opening and maximum glottal flow; corresponds to increasing glottal flow. The ratio between the opening phase and the closing phase determines the skewness of the glottal pulse, and is quantified using the Speed Quotient.

- *Closing Phase*: extends between moment of maximum glottal flow glottal closing instance; corresponds to decreasing glottal flow. The ratio between the closing phase and the fundamental period is quantified using the Closing Quotient (ClQ). This phase is different than the Closed phase, which is given below.

- *Closed Phase*: the time duration where the vocal folds are closed; extends between neighbouring glottal closing instance and the following glottal opening instance temporal points; embodies of the return phase. The ratio between the closed phase and the fundamental period is quantified using the Closed Quotient (CQ).

- *Return Phase*: the vocal folds return to their neutral medial state; researchers may define the return phase as independent of the closed phase. Earlier parametric models omit the return phase.

- *Glottal Opening Instance*: initial time instance for open phase; time instance at which the glottis is defined as open.

- *Glottal Closing Instance*: time instant at which the glottis is defined as closed; instant of minimum glottal waveform derivative value; large excitation to the vocal tract system.

3.2 Glottal Pulse Parametrization

The segmentation of the glottal pulse into phases (refer to Figure 3.1) was made possible using the Liljencrants-Fant (LF) model. This section presents six glottal descriptive features/parameters, which are used for vocal characteristics quantification and subsequent analysis. Those parameters are: speed quotient, open quotient, normalized amplitude quotient, pitch, jitter, and shimmer. A description of each parameter follows. Also included in this section is a description of a new open quotient criteria, the OQsub50, which was devised during the development of this present work. The next section in this chapter relates the aforementioned parameters to voice quality. Figure 3.2 and Figure 3.3 illustrate the glottal flow and the derivative glottal flow, respectively.

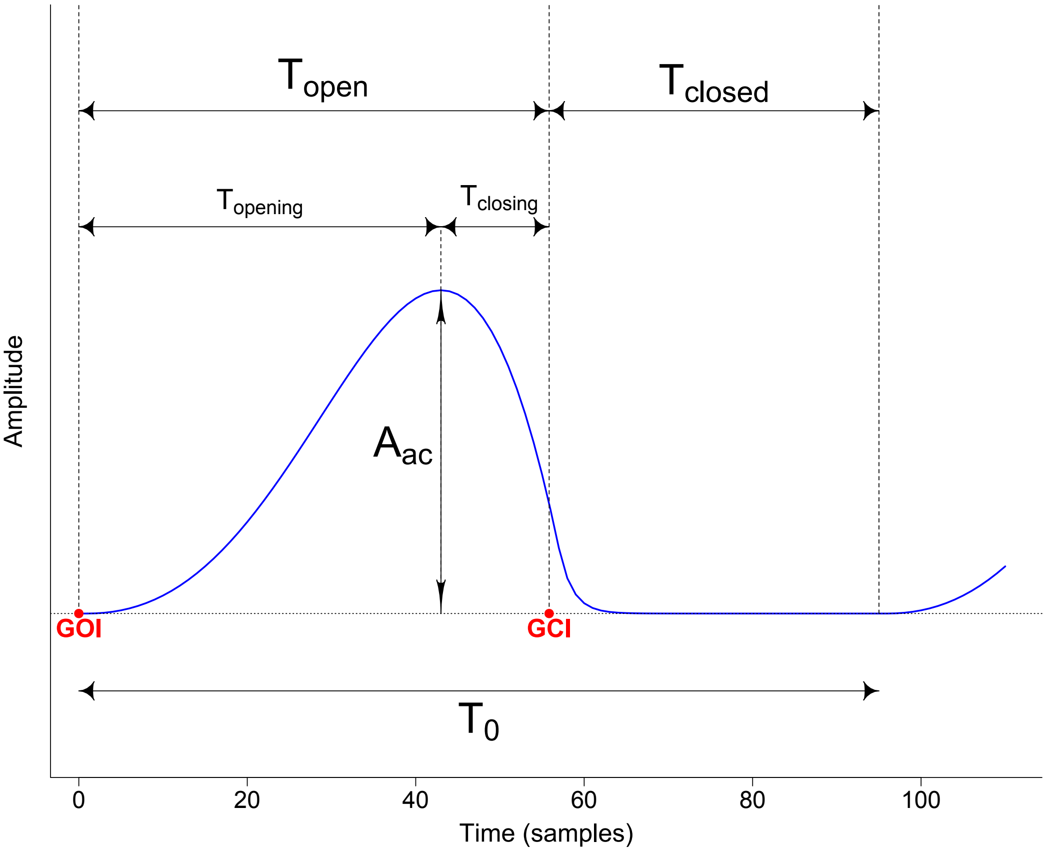

Figure 3.2: *Liljencrants-Fant glottal pulse model with temporal phases*

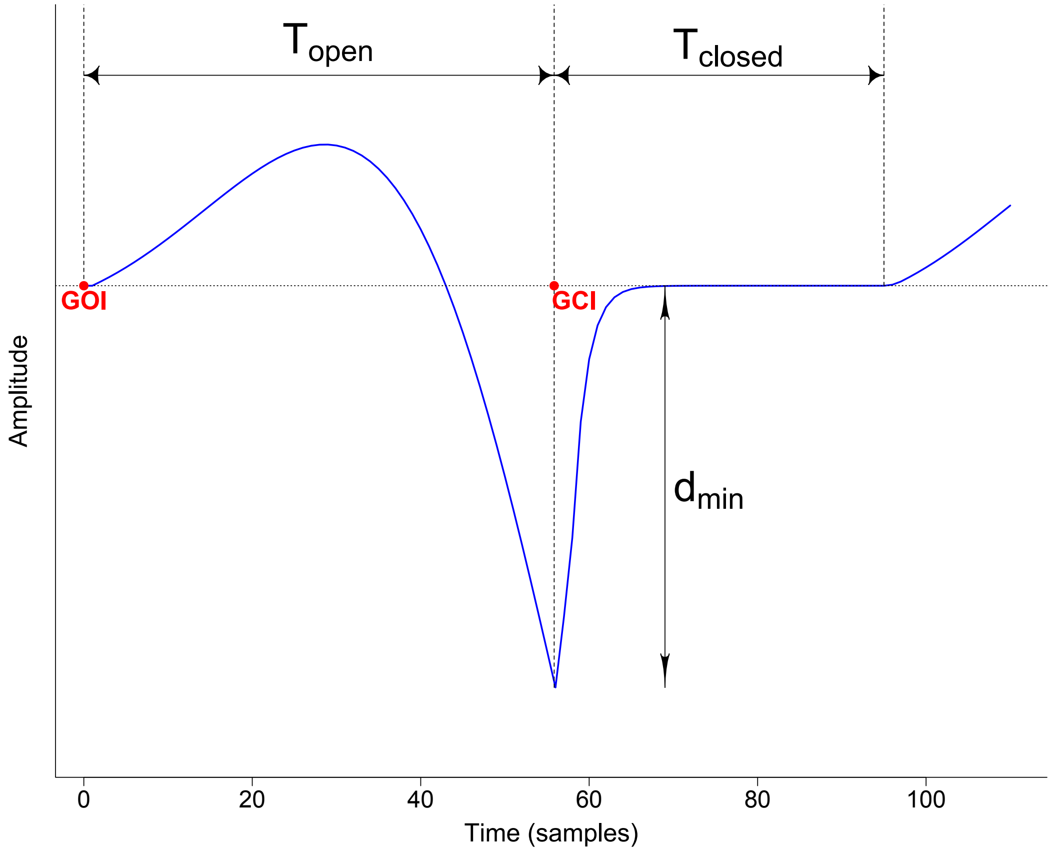

Figure 3.3: *Liljencrants-Fant derivative glottal pulse model with temporal phases*

3.2.1 Speed Quotient (SQ)

The speed quotient is a time-domain parameter derived from the glottal waveform. The speed quotient is expressed as the ratio of the opening phase duration ($T_{opening}$) to the closing phase duration ($T_{closing}$) (see Equation 3.1) [69, 70]. A visual depiction of the phases is shown in Figure 3.2. It is a measurement of the glottal pulse skewness, which quantifies the asymmetry of the glottal pulse [68]. The speed quotient allows to monitor the behaviour of the vocal folds during speech production. For $SQ = 1$, the glottal pulse is symmetric about the instance of maximum glottal flow. Due to the mechanical properties of the vocal folds, abduction is longer than adduction. Since the lateral separation occurs at slower rate than the elastic recoil to the medial position, the speed quotient is expected to be greater than one, $SQ > 1$ [71]. Hence, the speed quotient proves helpful in detection of voice pathologies (i.e. vocal fry). In addition, the speed quotient provides a measure for glottal efficiency. An asymmetrical glottal pulse with a high SQ value corresponds to greater efficiency [72].

$$SQ = \frac{T_{opening}}{T_{closing}} \tag{3.1}$$

3.2.2 Open Quotient (OQ)

The open quotient is a time-domain parameter which quantifies the width of the glottal pulse in relation to the duration of the glottal cycle. It is the ratio between the glottis open phase duration ($\mathrm{T_{open}}$) and the glottal pulse period duration ($\mathrm{T_0}$) (see Equation 3.2) [69]. A visual depiction of the phases is shown in Figure 3.2. The open quotient has become a reference time-domain parameter in most studies [73, 74, 75]. Moreover, it can be used as a parameter to model the glottal flow [68]. The open quotient is computationally straight forward as its extraction does not require prior knowledge of the absolute flow values of the glottal waveform, and its measurements can be directly measured from the time-domain glottal waveform [76]. Physiologically, the open quotient reflects the behaviour of the vocal folds, providing a numerical representation for the duration for which the vocal folds are abducted during the glottal cycle. The open quotient is related to two other quotients: the closed quotient (CQ) and the contact quotient (Qx). OQ and CQ account for the entire glottal cycle; $1 - CQ = OQ$. The contact quotient is used in EGG studies and it is a time-domain parameter for the duration of vocal folds contact in the glottal cycle. Although the open quotient is not directly proportional to the contact quotient, the two parameters allow EGG/speech-signal data comparison.

$$OQ = \frac{T_{open}}{T_0} \tag{3.2}$$

3.2.2.1 Existing Open Quotient Criterion

As previously mentioned, the open phase is defined as the duration between a GOI point and a successive GCI point. Accurate detection of GOI and GCI points is of significant importance for glottal parametrization. As per the Liljencrants-Fant model, the instance of glottal closure is defined to occur at the instance of maximum negative flow in the derivative glottal flow. An abundance of glottal closing instance detection algorithms is available in the literature [77, 78, 79]. A review of the existing glottal closure instance detection algorithms is available in the publication by Naylor *et al* [80]. The location of the glottal opening instances is less defined. In theory, the GOI instances correspond to a second, weaker positive peak in the derivative glottal flow. However, since their energy of excitation is much weaker than that of the GCI instances, their detection poses greater difficulty [77]. Another issue facing GOI and GCI detection is computational complexity. It could be argued that researchers are willing to trade accuracy for reduced CPU runtime. Although GOI and GCI instances are paramount in glottal parametrization, publications often fail to mention how the GOI and GCI instances were obtained. This is especially true for the GOI instances. As a result, there currently is no unified consensus on the extraction method for glottal opening instances. This division in opinions has led to various definitions of the open phase. In turn, this led to the definition of several open quotient criteria. There are two widely used OQ criteria: the *OQ50* and *OQsub*. These criteria have different definitions for the GOI and GCI points, as well as definition of the pulse period [81].

- **OQ50**
  OQ50 uses a 50% criterion level, where the open phase is defined as the time where the glottal flow is greater than a nominal threshold amplitude level. This is the predominantly used criterion. These criteria was proposed due to the difficulty in determining GOI and GCI points. Instead of extracting the GOI and GCI points, a computationally complicated task with inherent difficulties and susceptible to errors, the true GOI and GCI points are replaced by the time instances for which the glottal flow exceeds an arbitrary amplitude level, i.e. 50% [81, 76]. This threshold is defined as 50% of the ac component (ac equals peak flow minus minimum flow). The GOI and GCI points correspond to the temporal points in the glottal waveform at the amplitude threshold level. The glottal pulse period is defined as the duration between two successive GOI

points [81]. We note that other threshold criterion levels are often used to OQ20, OQ30, OQ80 [82].

- **OQsub**
  OQsub (subjective airflow open quotient) is the ratio between the open phase and the period of vibration [82]. The OQsub is less prominent in research due to its imprecise definition of GOI and GCI points. The GOI point is defined as the point of initial increase from minimum flow. Although easy to implement, this definition is prone to errors if minima flow points exhibit small ripples due to noise coupling. The GCI point is defined to occur at the end of a sharp decline in glottal flow, followed by a sudden increase in flow. This definition is case-specific, as not all waveforms exhibit an increase in flow during the closing phase. Even though the OQsub measurements are performed pitch synchronously, no mention of results interpolation had been noted in any other research papers. This criterion is seldom used as it is heavily reliant on the shape of the glottal pulse. As the glottal flow may not exhibit ripples at the opening phase, the detection of the glottal opening instance becomes difficult [73].

3.2.2.2 A New Open Quotient Criterion: OQsub50

Although the open quotient parameter is cited in literature, the variations in open quotient criteria and glottal flow parametrization methods have made it difficult to compare results across studies. Furthermore, the lack of criteria implementation procedure make it impossible to draw comparison between studies utilizing the same speech corpus [82]. The three main open quotient criteria specifications that are often failed to be mentioned are: the author's definition of the open and closed phases, whether interpolation of glottal opening and glottal closing instances was performed, and whether the analysis was performed pitch synchronously [81]. In addition, studies fail to specify whether the open quotient values were gathered for the entire duration of the utterance or only at the most stable point (i.e. mid-point for vowels). The lack of detail compromises the integrity of the mean and standard deviation statistical analysis of the open quotient results.

We propose a new open quotient criterion, the *OQsub50.* This criterion was first presented in a publication by the author [83]. The OQsub50 offers a computationally simple method for quantifying the open phase of the glottal flow. As the name suggest, this criterion is a hybrid of OQ50 and OQsub. In accordance with OQ50, the glottal opening instance is defined as the time instances for which the glottal flow exceeds an arbitrary amplitude level (in this case 50%). The extraction of the glottal closing instance has been extensively researched in speech processing [77]. Our definition of the glottal closing instance is based on the LF-model, where the time instance of glottal closure corresponds to the point of minimum flow in the glottal waveform derivative [84, 68, 85]. This GOI and GCI extraction methods make the OQsub50 a computationally simplistic parameter. The OQsub50 measurements are performed pitch synchronously, where the glottal pulse period is defined as the duration between two consecutive glottal flow amplitude minima instances, as in accordance with OQsub [81].

3.2.3 Normalized-Amplitude Quotient

When extracting time-domain parameters from the glottal waveform, it is critical to identify the instances of glottal opening and glottal closure using a robust and reliable method. The detection of the GOI and GCI points is susceptible to error due to glottal ripples arising from formants coupling. Alku *et al* [76] suggested an amplitude-domain parameter to replace the time-domain parameter, thus excluding the requirement for accurate detection of GOI and GCI points. This parameter is the normalized amplitude quotient (NAQ). The NAQ is an amplitude-based parameter set to provide an alternative representation for the closing quotient (ClQ), where ClQ is defined as the ratio between the closing phase duration and the fundamental period of the pulse. Hence, NAQ quotient is a time-domain parameter for the characterization of the glottal closing phase. The NAQ quotient is more reliable and robust compared to the ClQ quotient, resulting in smaller variance across different sound pressure levels [86]. It was proposed as a new iteration for the amplitude quotient (AQ). The amplitude quotient is defined as the ratio of the glottal pulse intensity ($\mathrm{A_{ac}}$)to the negative peak of the differentiated glottal pulse ($\mathrm{d_{min}}$), as per Equation 3.3 [73]. Visual depiction of the amplitudes for both the glottal flow and its derivative is shown in Figure 3.2 and Figure 3.3, respectively.

$$\mathrm{AQ} = \frac{A_{ac}}{d_{min}} = \frac{A_{max} - A_{min}}{d_{min}} \tag{3.3}$$

Where A_{ac} is the peak-to-peak amplitude of the glottal waveform (ac flow) and d_{min} is the minimum negative amplitude of the glottal waveform derivative. The amplitude quotient does not require knowledge of the glottal pulse scale, which makes it straightforward and reliable [73]. However, the AQ quotient is susceptible to fundamental frequency variations, as the closing phase is correlated with the period of the glottal pulse. This hinders the attempts at comparison of results across phonation types [76, 86]. This problem was mitigated by accounting for the fundamental period dependency via normalization. The AQ parameter is normalized to the fundamental period of the pulse (T_0), yielding the NAQ parameter. The NAQ quotient is computed per Equation 3.4 [76].

$$\mathrm{NAQ} = \frac{AQ}{T_0} = \frac{A_{ac}}{d_{min}T_0} \tag{3.4}$$

3.2.4 Pitch Detection

The importance of robust pitch detection is embedded in the history of speech processing, coding, analysis, and recognition. Correct estimation of pitch is notoriously difficult due to several factors. Its transient nature makes pitch determination difficult. Pitch variability, in terms of both periodicity (jitter) and amplitude (shimmer) is observed across all phonation types, even modal [87, 88]. Another issue is the duration of voiced speech segments. Pitch detection algorithms require the speech segments to exhibit multiple cycles for a reliable pitch detection. A comprehensive discussion of additional factors which complicate the task of pitch detection can be found in the following publication by Talkin *et al* [89] (p. 500). What follows is a review of the existing methods for pitch detection from the speech signal.

One of the first pitch estimation methods utilized the *auto-correlation* method. The auto-correlation function exhibits a peak about the fundamental frequency. Markel *et al* [90] devised the simplified inverse filter tracking (SIFT) algorithm.

The SIFT algorithm utilizes the auto-correlation function for an excitation signal obtained via inverse filtering. The signal exhibits a sharp peak which is used to estimate the pitch of the voiced segment. This technique is best used for band-limited speech signals, as it is highly susceptible to the influence of the second and third formants [88]. This method is based on the source-filter model, assuming the vocal-tract to be time invariant for short durations (about 20 ms). One of the most popular pitch estimation methods is the YIN algorithm, developed by Kawahara *et al* [91]. The algorithm implements the auto-correlation function alongside a series of preventive steps to minimize pitch estimation error.

Another pitch detection methodology which is based on the source-filter model is the *cepstrum* method. This method was first employed by Noll *et al* [92]. Cepstrum pitch detection is carried out by separating the excitation signal from the vocal tract in the cepstral domain. The cepstrum signal then exhibits peaks at multiples corresponding to the fundamental period. A discussion of the algorithm by Noll, describing its pitfalls for a periodic signal, is given by Camacho *et al* [93]. Pitch detection can also be performed using the *normalized cross-correlation function* (NCCF). Talkin *et al* [89] developed the RAPT algorithm based on this method. First, a down-sampled speech frame is used to obtain the location of local maxima via NCCF. Next, NCCF is computed for a high-sampled speech signal, in order to refine the location of the local maxima. Each local maxima is a potential candidate for the pitch period. Finally, the local maxima corresponding to the "true" pitch period is observed to be the largest (close to 1).

The aforementioned pitch detection techniques rely on robust peak-picking methods. Moreover, peak-picking is dependent on the resolution of the signal: sampling rate in the time-domain or FFT size in the frequency domain. As a result, interpolation of the signal is often required prior to the peak-picking process. The *multi-band harmonic* methods do not require those steps. A multi-band pitch detection algorithm was proposed by Saul *et al* [94, 95]. The algorithm models a low frequency spectrum of a voiced speech signal as the sum of sinusoids. First, the speech signal is low-pass filtered (reduce noise components) and transformed by a half-wave rectifier (concentrate the spectral energy about the fundamental frequency). The transformed signal then passes through a band-pass filter bank. The output signals are either sinusoids at multiples of the fundamental frequency, or signals that are low frequency noise. Sinusoid detection via least squares method is then applied to determine the pitch frequency of the voiced segments.

Another pitch detection method is based on the *harmonics* of the speech signal. Sun *et al* [96] proposed a new descriptive parameter; the Subharmonic-to-Harmonic Ratio (SHR). The parameter is a measurement of the amplitude ratio between the sub-harmonics and harmonics. Through a perceptual test, it is observed that a SHR in the range of 0.2 to 0.4 affects pitch perception. Subsequently, a pitch detection algorithm based on SHR is proposed. The SHR pitch detection algorithm was modified by Sun *et al* [87]. Drugman & Alwan *et al* [97] proposed a pitch tracking algorithm based on the residual harmonics. The pitch tracking method computes the fundamental frequency for voiced speech using the residual signal ,$e[n]$. This algorithm exploits the harmonic structure of the residual signal's spectrum. The linear prediction residual signal is derived through inverse filtering, where the spectral contribution of the vocal tract is removed from the speech signal, $s[n]$, to obtain the residual signal. Next, the spectral amplitude of the residual signal, $E(f)$, is computed using the Fast Fourier Transform (FFT). Finally, a Summation of Residual Harmonics (SRH) method is employed to detect the fundamental frequency of the speech signal. The algorithm is also used to compute the voiced/unvoiced speech segments. This method was shown to be particularly robust in adverse conditions compared to other state-of-the-art pitch detection techniques.

3.2.5 Jitter

Jitter is a percentile parameter for measurements of frequency modulation. It allows for quantification of cycle-to-cycle perturbations in fundamental frequency. Jitter is correlated to the behaviour of the vocal folds during the vibration cycle, accounting for the quasi-periodicity of the glottal waveform. Consider a glottal pulse train with controlled amplitude as illustrated in Figure 3.4. The five glottal pulses have fundamental frequencies denoted as $\{f_1, f_2, f_3, f_4, f_5\}$. For this example, $f_{1,2,4,5} = 80$ Hz, while $f_3 = 120$ Hz. Thus, the glottal waveform exhibits fundamental frequency modulation between successive glottal cycles. Jitter is computed per Equation 3.5, where N is the number of cycles in the waveform and T_i is the fundamental period at cycle i. This is the measure for *relative jitter*, also referred to as *local jitter*, which is the absolute jitter normalized to the average fundamental period and expressed as a percentage [98]. Other jitter parameters are available: *absolute*, *rap*, *ppq5* [99, 100, 101, 102].

$$Jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \tag{3.5}$$

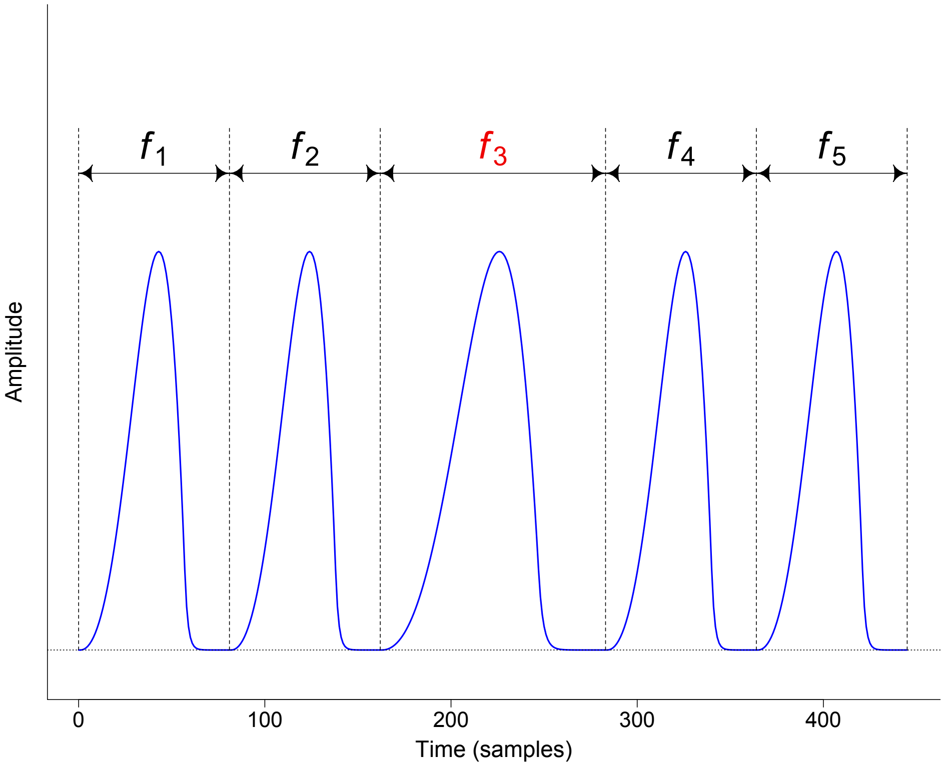

Figure 3.4: *Jitter (adapted from [103])*

3.2.6 Shimmer

Shimmer is a parameter for measurements of peak-to-peak amplitude modulation. It quantifies cycle-to-cycle perturbations in glottal pulse intensity. It is known that natural speech is susceptible to amplitude modulation regardless of the speaker's voice pathology [104]. Consider a glottal pulse train with controlled fundamental frequency as illustrated in Figure 3.5. It is evident that there exists significant variations of peak-to-peak amplitudes, $\{A_1, A_2, A_3, A_4, A_5\}$, between consecutive periods. The amplitude modulation across a glottal waveform is quantified using Equation 3.6, where N is the number of cycles in the waveform and A_i is the peak amplitude at cycle i. This is the measure for *relative shimmer*, which is the mean absolute difference between consecutive intensity levels normalized to the average intensity and expressed as a percentage [98]. Other shimmer parameters are available: *dB*, *apq3*, *apq5* [98, 100, 101, 102].

$$\mathit{Shimmer} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{3.6}$$

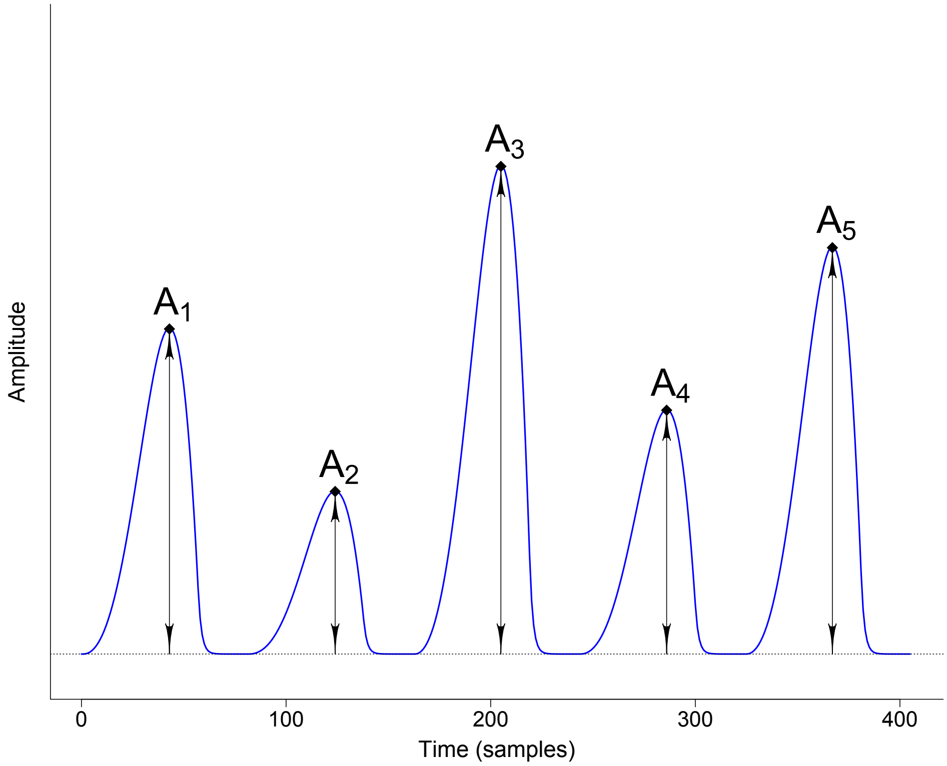

Figure 3.5: *Shimmer (adapted from [103])*

3.3 Vocal Quality

The vocal organs form a layered and complex biological system, which is physiologically utilized for respiration and phonation. Consequently, the vocal system alters and deteriorates from childhood to adulthood and into old age. Interest in examination of senescence voice and its phonetic traits are at the cornerstone of this thesis work. It doesn't take a speech scientist to be able to aurally discern between speech samples of young and old speakers. With advanced age, the vocal organs exhibit deterioration which manifests as voice that is often described as creaky, hoarse, breathy, slow, and weak. Listeners are capable of making a distinction between the aforementioned voices through a set of acoustic cues, such as pitch level and variability, loudness, and speech rate. Despite this set of acoustic cues, labelling a speaker as "old" based solely on a speech sample may prove difficult due to the notion of *physiological age* (subjective acoustic perception of a speaker's age). Prior to the consideration of physiological age, voice deterioration has been attributed to neurological and muscular inefficiencies due to advanced chronological age. In the 1970s, however, the extent of the influence of chronological age on a speaker's voice was being questioned. Inter-subject variability in data gathered from groups of similar chronological age suggested the existence of other variables aside of chronological age. Ramig & Ringel *et al* [105] proposed that speakers of same chronological age will exhibit different speech characteristics based on their physiological condition ("poor" or "good" health) [106, 107]. The introduction of physiological age has made speaker's age classification into a laborious task. We can no longer consider a speaker's age as the sole variable in determining vocal quality. Instead, we rather need to consider his/hers physiological condition, as well as other external factors that may have influence on the vocal system, e.g. strees, nutrition, smoking. It can be concluded that with ageing, the phonatory mechanism loses its efficiency, which leads to voice quality pathology. First, we present the features of four vocal registers. The age-related changes in the vocal apparatus system affect the lungs, larynx and vocal folds. The affect of ageing on the vocal organs is described, followed by a review of the acoustic cues used to identify and quantify vocal ageing. This discussion relates back to topics described in previous sections of this document.

3.3.1 Four Voice Registers

Voice pathology is often the result of abnormal vocal folds vibration patterns. The motion of the vocal folds is reflected in the glottal pulse. Thus, the behaviour of the vocal folds induces changes in the shape of the glottal pulse in terms of length of glottic closure, amplitude, and skewness, as well as cycle-to-cycle changes in amplitude and periodicity. Thus, it is apparent that abnormal vibratory patterns will render pathological voice. In linguistic and paralinguistic studies, the quality of the voice is assessed through unified descriptions. Six phonation types (registers) were suggested by Laver *et al* [108] to to describe voice quality: modal/normal, breathy, creaky, whispery, tense, and lax. The role of these registers is to allow comparison of results across studies by using a single term to identify physiological, perceptual, and acoustic aspects of voice quality. The phonation types observed in normal speech are illustrated in Figure 3.6. For the purpose of study, the voice qualities examined were modal, breathy, creaky and hoarse voice, all associated with aged voice. An overview of the selected registers follows.

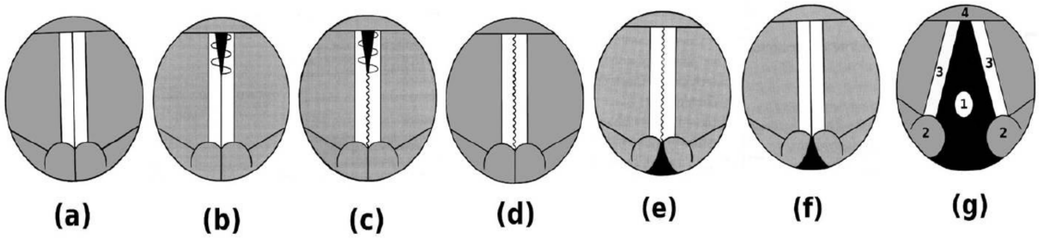

Figure 3.6: *Vocal folds configuration for various phonation types: (a) Glottal Stop, (b) Creak, (c) Creaky voice, (d) Modal voice, (e) Breathy voice, (f) Whisper, (g) Voicelessness. In (g): 1. Glottis, 2. Arytenoid cartilage, 3. Vocal fold and 4. Epiglottis (as in [85]*

3.3.1.1 Modal (normal) Voice

Modal voice is the most common vocal register in the English language. The modal register describes the optimal vocal outcome from a normal and healthy vocal folds vibratory cycle. In theory, the glottal pulse shape for modal voice has symmetric pulse width, with about 50 to 60 percent open phase (open quotient). The pulse is observed to exhibit slight right-skewness, indicating a rapid closing phase compared to opening phase [74]. Modal voice is conveniently characterized as having moderate vocal folds tension, length, thickness, glottal pulse width, and glottal opening. The fundamental frequency range for modal voice is between 90 to 290 Hz for men and 140 to 540 Hz for women [109]. Compared to other vocal registers, modal voice is reported to have low jitter and shimmer values. Modal voice is the phonetic register against which all other registers are evaluated, especially in perceptual studies done to compare voice quality among speakers.

3.3.1.2 Hoarse Voice

Hoarseness in a common voice quality symptom in older speakers. The term *hoarse voice* is used to describe vocal abnormalities such as voice that is breathy, rasp and/or strained. Physiological changes to the vocal folds as a result of senescence are attributed to hoarseness. Among those pathologies, vocal fold bowing, unilateral vocal folds paralysis, and benign vocal fold lesions have been attributed as the main cause for hoarseness in individuals [110, 111]. Hoarseness had also been attributed to pathological larynx [112]. Other impacting factors may include smoking, disease, and functional usage (i.e. singing). Hoarseness is often associated with low pitch and low vocal intensity [113].

3.3.1.3 Creaky (fry) Voice

Creaky voice is another type of vocal quality. Creaky voice is used to characterize speech quality with irregular vocal folds vibratory patterns. Vocal creak is common in English speech (referred to as glottalization) as a linguistic marker of sentence boundaries [114, 115]. Creaky voice is also known as vocal fry, glottal fry, pulse register, and stiff voice. In literature, the interchangeability of those names is subjectively defined by the author. The perception of creakiness has changed throughout the years as the cause for its symptoms became more evident. Initially, creaky voice had been classified as a voice disorder due to abnormal vocal folds vibration pattern. However, subsequent studies have suggested it to be a phonatic register for low-pitch voice [116]. Recent studies have shown vocal fry to be a selective vocal register in speakers with no vocal pathology. Wolk *et al* [115] reports selective use of vocal fry at the end of sentences among college students without vocal pathology. The authors postulate the influence of pop culture on selective use of vocal fry, showing speakers to modify their vocal register to mimic popular figures. In essence, creaky voice is described by three main acoustic features: low pitch compared to modal phonation [117], pitch irregularity (jitter or F0(SD)) [115], and constricted glottis [109]. Variations in the degree of influence for each one of these variables yielded sub-categories [118]. Vocal fry does not exhibit irregular vocal folds vibration. Aperiodic voice has no perceived pitch due to extremely irregular vocal folds patterns. Non-constricted creak has an open glottis (also known as breathy creak). Physiologically, creaky voice is the result of tightly adducted vocal folds with increased thickness and reduced elasticity. It is characterized by short glottal pulse open phase followed by a long closed phase, resulting in small glottal pulse width (lowest Open Quotient values compared to other voice registers) [109]. As a result, creaky voice is the lowest voice register on the frequency continuum, with pitch range between 5 and 50 Hz [119]. Pitch irregularity of vocal creakiness is quantified by large jitter and shimmer values compared to modal phonation [115]. Childers *et al* [109] cites the occurrence of multiple opening/closing phases in a single glottal pulse, indicating a highly irregular vocal folds vibration pattern, termed *dicrotic vibration*. Moreover, the authors cite significantly high Speed Quotient values for creaky voice, indicating significant glottal skewness/asymmetry during the vibratory cycle. Chen *et al* [120] reported high SQ values for both male and female speakers, noting higher SQ values for female speakers. This phenomenon was attributed to the increased stiffness in vocal folds, resulting in longer abduction period. The authors also observed dicrotic vibration pattern for both male and females, noting a higher frequency of occurrence in male speakers. This phenomenon was attributed to the increase in vocal folds thickness. We note that vocal creak could also develop due to smoking [121].

3.3.1.4 Breathy Voice

Breathy voice, also known as murmur, is a voice quality pathology caused due to incomplete medial closure of the vocal folds (open glottis). In production of breathy voice, the vocal folds do not approximate simultaneously; First, the anterior parts of the vocal folds are closed, followed by the posterior parts. In breathy phonation, the vocal folds approximation does not occur simultaneously for both vocal folds. As a result, there is a considerable coupling of aspiration noise into the speech signal. This aspiration noise component is the key acoustic cue for perceptual classification of breathy voice. Breathy voice is classified as a voice pathology and is symptomic to laryngeal disorders and trauma [122], as well as abductor spasmodic dysphonia [123], vocal fold bowing (a permanent gap between the vocal folds due to muscle atrophy), and vocal folds paralysis. Similar to other vocal pathologies, breathiness may occur due to improper functional usage (i.e. singing). Breathiness in voice involves vocal folds vibration mode which is considered inefficient compared to modal voice. Incomplete closure allows for turbulent air (noise) leakage through the posterior glottis during the closed phase of the vibratory cycle. This is characterized by a large open period relative to the pitch period. Childers *et al* [109] observed rapid glottal closure in EGG waveforms for breathy voice. Moreover, the authors indicate the absence of glottal closure in speakers. As a result, breathy voice is quantified by a large open quotient and was found to have the highest open quotient values compared to hoarse, creaky and modal voice registers (large glottal pulse width). Aside of OQ, the normalized amplitude quotient was also found to be correlated with breathiness [124]. Furthremore, breathiness limits the vocal intensity and is usually accompanied by a low pitch [125]. Large pitch and amplitude perturbations are also observed for breathy voice [109]. In literature, breathy voice has been observed for both male and female speakers [126]. In women, however, incomplete glottal closure of the posterior parts of glottis is not considered a pathology. Moreover, female speakers were reported to exhibit less rapid vocal folds closure compared to their male counterparts [127]. Mendoza *et al* [113] remarks that with regards to auditory perception of voice quality, breathiness is more associated with female voices. This notion of higher perceived breathiness in female speakers compared to male speakers was established in other publications [126, 127]. Breathiness also encompasses paralinguistic information. Similar to the hypothesis by Wolk *et al* [115] with regards to creaky voice, Henton *et al* [125] suggests that breathy voice may be selectively exploited by speakers for paralinguistic communication.

3.3.2 Ageing Effect on the Respiratory System

Air expelled from the lungs drives the speech production mechanism. Age-related changes affect the muscle and tissue structure of the respiratory system. From birth to old age, the skeletal framework of the ribcage, in particularly the angle of the ribs, is correlated to respiratory efficiency. In adults, the ribs are angled downwards, allowing for chest-wall movements upon inspiration. The ageing process causes loss of elasticity and atrophy of the muscles of respiration impedes the mobility of the ribs during the respiratory function [128]. With advancing age those changes manifest as loss of operational capacity (about 1 litre). A loss of lung capacity, coupled with loss of diaphragm muscle strength, leads to reduction in expiratory air volume, which impacts vocal loudness. Those factors impair respiration and voice production efficiency.

3.3.3 Ageing Effect on the Larynx

Age-related degradation of anatomical structure of the larynx impedes voice quality. The extent of laryngeal structure changes is considered to be greater in males than females, and is attributed to changes in vocal pitch and variability [129]. Intrinsic muscles atrophy and cartilage ossification (new bone formation) had been observed to occur as part of the ageing process. Muscle atrophy and reduced density, particularly for intrinsic laryngeal muscles, impair the motor control for vocal folds abduction and adduction, which prevents complete closure of the glottis, leading to voice pathologies (breathy voice) [130, 131]. Cartilage ossification had been studied in literature for over half a century. Jurik *et al* [132] reported intensified laryngeal ossification, conjoined with calcification, with advancing age. Ossification of the thyroid, artenoid and cricoid cartilages was observed for both male and female participants. Greater degeneration was found in males compared to females, with ossification occurring later in life for females [128]. Changes to the artenoid and cricoid cartilages lead to changes in the cricoarytenoid joints, which are the primary moving structure of the larynx. Subsequent lowering of the larynx is correlated to pitch variability. Pitch variability might also be related to the degree of mucous membrane, with greater vulnerability attributed to females. [103]. Moreover, reduced laryngeal cartilages and joints elasticity (stiffening) alter the mechanical characteristics of the vocal ligaments' insertion zones [133]. Consequently, reduced elasticity of the cartilages is attributed to reduced vocal ligaments elasticity, which results in changes in fundamental frequency.

3.3.4 Ageing Effect on the Vocal Folds

It has been well established in literature that ageing affects the physiology of the vocal folds. The ageing process leads to histological changes in the vocal folds, which in turn affect voice quality. The vocal folds consists of the epithelium layer, three layers of lamina propria (superficial, intermediate, and deep), and the thyroarytenoid muscles. These layers change differently with regards to the speaker's age and gender. Variation in the vocal folds mass are the result of atrophy or edema (swelling due to excessive collection of fluids in tissue). Edema hinders propagation of the mucousal wave, causing irregularities. Variation in mass manifests in the form of changes in vocal folds tissue thickness and density. These changes alter the vibratory mechanism of the vocal folds. Edema was observed on the superficial layer of the lamina propria in aged speakers. There is a lowering of elastic fibres density in the intermediate layer, which subsequently leads to thinning of the layer. The deep layer exhibits increased thickness and distortion of the collagenous fibres. The collagenous fibres lose their linearity as they run in all directions, resulting in fibrosis. The aforementioned histological changes of the vocal folds occur at different prominence in male and female speakers. Hirano *et al* [134] noted an increase in vocal folds thickness for female subjects. Edema was noted for both males and females, but was more prominent in females. Edema leads to increase in the mass of the vocal folds, which could be attributed to the lowering in pitch with advanced age. Atrophy of the intermediate layer due to reduced thickness and lowering of elastic fibres density was observed in males. Atrophy of this layer is associated with vocal folds bowing, as it leads to deterioration of the contour shape of the layer. The deep layer stiffens with age due to its increase thickness and fibres distortion. These changes were observed in males but not females. Honjo & Isshiki *et al* [135] noted yellow/gray discolouration of the vocal folds for both genders with age. This discolouration was attributed to fat degeneration of the epithelium. Edema was observed in both genders and was more prominent in females. Edema in females was attributed to hormonal changes after menopause. Vocal fold bowing was also noted for both genders and was more prominent in males. Vocal folds bowing in males was attributed to age-related changes in the mucousal membrane of the folds. One of the causes is reduction in mucous glands secretions, which reduces vocal folds viscosity [121]. This behaviour is mainly observed in males and may result in increased pitch [103, 136]. Kahane *et al* [137] reported similar findings to Hirano, noting vocal folds atrophy in males due to fibre reduction in the vocal folds ligaments. The geriatric changes in the tissue structure of males causes an increased stiffness and shortening of the vocal folds, particularly after the age of 70 [103]. In females, fibrotic and thickness changes were not remarked with old age, however, the thickness of the cover increased. These changes may introduce irregularities in vocal folds vibration, leading to asynchronous behaviour. Decrease in vocal folds mass due to atrophy results in incomplete adduction of the folds (vocal folds bowing) and may be attributed to the increase of pitch with old age, as well as breathy voice quality.

3.3.5 Environmental Factors of Vocal Ageing

The physical development of a person is largely determined by two factors: genetic and environmental. The effect of either one of those factors is a much-discussed topic in literature. Although progress has been made in labelling and quantifying subset of each factor, the inability to judge the factors independently from one another have led to dispute over results. Genetic factors are physiological changes that are hereditary to a person's gene pool (family history) and influence voice quality degradation. Environmental factors are the result of the environment and habits of a person. Environmental factors that are considered to be primary factors in vocal ageing are low socio-economic status, emotional stress and disturbance, trauma, poor/imbalanced nutrition, being of a large family, smoking, alcoholism, and disease [127, 128, 138]. Exposure to these environmental factors and others can alter a person's physiological age. On the other hand, avoiding environmental factors which may lead to vocal deterioration will most likely allow a person to preserve his/her vocal quality even at an advanced age. Perceptual studies have shown that elder speakers in good physical condition tend to exhibit acoustic characteristics similar to younger subjects, and thus, were identified as having a physiological age below their years (chronological age). In addition, Longitudinal studies have shown that the effect of some factors is reversible. Leeuw *et al* [121] reported low fundamental frequency for two male subjects with smoking habits. Subsequent measurements showed increased fundamental frequency for the two subjects as they quit smoking. Hence, the effect of smoking was shown to be reversible in some cases.

3.3.6 Fundamental Frequency

The fundamental frequency of vocal fold vibration is directly correlated to the pitch of voiced speech. The pitch varies throughout adult life for both male and female speakers. However, the pitch pattern for aged adults vary between genders. Initially, age-related changes in fundamental frequency were examined for male speakers. Mysak *et al* [139] was the first to examine age-related changes (above adulthood) in pitch. Mysak observed a correlation between speakers' vocal pitch and age, with increased pitch for aged males. Hollien *et al* [140] were the first to collect data on the fundamental frequency changes throughout adult life of male speakers. They reported that age-related changes in speech cause a progressive lowering of F0 from pre-adolescent to middle-age, with the most dramatic change occurring at puberty. Furthermore, the authors reported a transient decrease of about 10 Hz from adulthood to middle-age, followed by a gradual increase in fundamental frequency, upwards of 35 Hz, with advancing age. This trend of vocal pitch for aged male speakers was indicated in other studies as well ([141, 142, 143, 144, 145, 146, 147]. Masaki *et al* [143] reports no significant change in fundamental frequency until the 7$^{\text{th}}$ decade. However, contradictory results of decrease in vocal pitch were presented in Benjamin [148] and Xue *et al* [99]. Other studies have shown different age-induced trends in pitch [149]. On average, the fundamental frequency of a male speaker with no vocal pathologies is around 120 Hz [103, 144]. Perceptual studies have shown that pitch is a salient discriminator for advancing age, with listeners associating high pitch with older speakers [150].

Throughout adult life, the fundamental frequency of female speakers varies differently to that of male speakers. Hollien *et al* [151] were the first to conduct a study examining the correlation between age and vocal pitch for aged female speakers. Their initial results suggested that female pitch exhibits a progressive decrease from young age to adulthood followed by little variability with advancing age. Additional studies have reported a drop in pitch between young adult speakers and elder speakers ([152, 141]). Subsequent studies have shown that female vocal pitch shows little variability into the age of 50-60 (until menopause), followed by a 10 to 15 Hz drop in pitch [153, 142, 103, 144, 145]. The post-manopause drop in pitch varies in range, where researchers had reported a more significant 20 to 25 Hz drop in pitch ([141, 154]), while Masaki *et al* [143] reported a drop of 45 Hz, and Xue *et al* [99] reported a more substantial decrease of 55 Hz. The average range of mean fundamental frequency of a female speaker with no vocal pathologies is between 190 and 220 Hz. Thus, female speakers tend to have a significantly higher pitch than their male counterparts. These fundamental frequency trends as a function of age for male and female speakers were presented by Hollien *et al* [119] as the *male-female coalescence* model. The model aims to establish the notion that pitch-related changes occur due to hormonal changes; puberty for men and menopause for women. Hollien hypothsizes that both sexes are hormonally similar prior to puberty and subsequent to menopause. Thus, attributing fundamental frequency changes throughout adult life to the hormonal shift between the genders. This model has yet to be confirmed, with contradictory results presented by both quantitive and longitudinal studies [143].

Longitudinal studies have been carried out, examining the variability of pitch with advanced age for same speakers. Most longitudinal studies constitute of two recording sessions roughly 30 years apart, as to allow for age-related changes in vocal properties. Significant reduction in pitch with advanced age for six female speakers was reported by Russell and colleagues *et al* [155]. A similar trend was reported in a study by Xue *et al* [154], examining the effect of age on eight female speakers recorded seperately 25 years apart. The progressive lowering of pitch for same speakers in longitudinal studies was further reinforced by Hollien *et al* [156]. Both Harrington *et al* [157] and Watson *et al* [158] have reported lowering of pitch for both male and female speakers. Decoster *et al* [159] reported that for 20 male speakers recorded 30 years apart, there was a 14 Hz drop in pitch. However, Decoster and collegue note that five of the speakers in the study exhibited an increase in pitch. In contrast to the findings by Decoster and colleagues, Harnsberger *et al* [160] reported an increase in pitch for male speakers, while Leeuw *et al* [121] reported unchanged pitch.

Fundamental frequency standard deviation, F0(SD) is a measurement for quantifying vocal instability. It is well accepted that elder speakers exhibit higher standard deviation of fundamental frequency compared to young adults. Linville *et al* [161] reported significantly higher F0(SD) compared to young and middle-aged female speakers. Thus, a decreased pitch stability was observed for elder female speakers. Regardless of the reported differences in men versus female speakers, old speakers showed higher pitch variability compared to young speakers for both genders. Furthermore, Linville highlights the significant contribution of F0(SD) to the statistical analysis, declaring F0(SD) as the most important acoustic variable for age discrimination. Gorham-Rowan *et al* [127] reported increased F0(SD) for both male and female aged speakers, indicating greater vocal instability in old speakers. This trend of a significant increase in F0(SD) in older speakers has been demonstrated in a number of other studies [141, 99]. It is hypothesized that pitch perturbation and instability with old age is due to age-related degradation and pathology of the laryngeal structure or partial glottis/vocal-tract coupling [112, 161]. The literature suggests that both average pitch and F0(SD) are salient acoustic cues for perceptual discrimination of perceived age in both male and female speakers [162, 129, 163].

3.3.7 Perturbation Measurements

Similar to F0(SD), fundamental frequency cycle-to-cycle perturbations (jitter, shimmer) are used to describe vocal instability in speakers. The parametrization of aged voice using jitter and shimmer has produced mixed results. Several studies have reported increased jitter values for aged speakers compared to younger speakers [164, 165, 166]. However, other studies reported that while jitter values were greater for older speakers, the was no significant changes in jitter values, noting the lack of significant interaction between jitter and aging [146, 147, 107, 161, 129, 138]. Linville *et al* [161] reports overlap in jitter range for young and elder women speakers. Although the range of jitter values shows greater variability, there is in fact, on average, great overlap between the jitter measurements for young and elder female speakers. In a study examining sustained phonation, Hollien *et al* [167] reported jitter values in the range of 0.5 to 1 for male speakers with no noted laryngeal pathology. Similar range was reported in [129], noting the range suggested by Hollien to be consistent with subsequent publications. Ramig *et al* [107, 165] and Brown, Morris & Michael *et al* [129] suggest that while aging affects the fundamental frequency of vocal folds vibration, it does not affect the cyclically synchronization of the vocal folds. Thus, jitter may be more greatly correlated to physiological age than chronological age. Several literature articles conclude that jitter is not an acoustic cue for perceptual identification of voice pathologies, such as hoarse or rough voice [161, 162, 129]. Shimmer quantifies speech variability via peak-to-peak amplitude, and is an additional cycle-to-cycle perturbation measurement correlated to hoarseness in voice. Increased perturbation with advanced age was reported for both male and female speakers [168, 99, 166, 138, 136]. Schaeffer *et al* [136] reported increased perturbation for both male and female speakers, with significantly higher shimmer values for men then women. In the study, the shimmer values for young males were higher than the shimmer values for female speakers, both young and old. Biever *et al* [168] compared shimmer values between young and geriatric women. The study shows significant increase in perturbation with advanced age, noting a correlation between shimmer and physiological age. Sataloff *et al* [169] suggest that perturbation measures are not reliable for female speakers, while jitter is a reliable measure for male speakers. They suggest that shimmer may exhibit too great variability even in normal speech to be used as a valid measurement for vocal quality analysis. Thus, it seems a consensus has yet to be established in regards to the reliability and robustness of those perturbation measurements. Kido *et al* [104] notes shimmer to be favourable to jitter, as it shows greater observable indication of vocal ageing compared to jitter. Linville *et al* [130] (see [103] page 98) concludes that neither jitter nor shimmer are reliable measurements for vocal ageing, as their measurements are too heavily correlated to external factors (sound pressure level (SPL), vocal pitch, physical variables). A lack of consistency among researchers, in terms of both measure specifications and data reliability, has hampered the development of jitter as a reliable acoustic cue. Finally, we note that expected perturbation values and range have yet to be defined for normative and pathological voice [170].

3.3.8 Speaking Rate

In literature, a person's speaking rate has long been examined as an acoustic cue for perceptual identification of aged speakers [139, 171, 172, 106]. With advancing age, the loss of fine-motor control impairs the articulatory mechanism. Hindered motor skills, as well as sensory feedback reduction and cognitive decline, cause imprecise articulation and slower reading and speaking rates in older speakers [138]. Thus, articulation rate is considered a function of age. Slower articulation rate is generally attributed to neuromuscular deterioration and laryngeal muscle degeneration [129]. Moreover, reduced articulatory motor skills is presumed to correlate to chronological age more than physiological age [129]. A speaker categorised as having "good" physiological age, however, could exhibit reduced speaking rate at a later time in life compared to a speaker with "poor" physiological age [173]. Perceptual studies have established the notion that slower speaking rate is one of the main acoustic cues listeners relied on to distinguish young and old speakers apart [174, 129, 163].

In this chapter, the usefulness of the glottal source signal and its descriptive features was highlighted. From the information presented above, it is evident that glottal processing is readily applicable for senescence voice analysis. However, the scope of glottal processing is not limited to senescence voice analysis. Glottal processing is applicable for a wide range of voice technologies. These include speech synthesis [175], speaker recognition [176], emotion recognition [177], and biomedical applications [178]. A comprehensive review of the applicability of glottal source to vocal applications is presented in [179]. In the next chapter, the extraction procedure of the glottal source signal is discussed.

Chapter 4

Speech Processing

This chapter begins by describing the source-filter model and its elements. Next, Linear Prediction is discussed. Following, the inverse filtering techniques are described. The chapter concludes with an overview of polarity detection methods.

4.1 Source-Filter Mode

The source-filter model depicts the human speech production mechanism as a linear, time-invariant system. The model consists of three elements in cascade. The airflow at the glottal level, most commonly referred to as *glottal flow* or *glottal volume velocity waveform*, is the input signal for the system. This signal is known as the *source*. The source signal acts as an input to the *vocal tract.* The vocal tract is modelled as an all-pole filter with resonances. The filter model consists of a limited number of two-pole resonators in cascade, where each resonance corresponds to a formant frequency. The third element is the radiation at the lips and nostrils in the form of a filter, termed *radiation.* Radiation is often coupled with the vocal tract filter model to form the *filter* [180]. The source-filter model system diagram is shown below in Figure 4.1.

As previously mentioned in Chapter 2, the speech production mechanism is complex in nature, with the multiple vocal organs operating in unison and couple

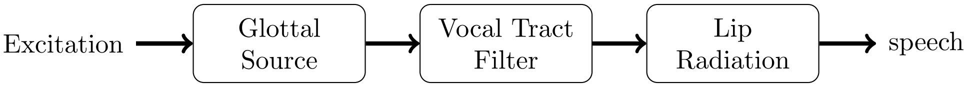

Figure 4.1: *Source-Filter Model* Diagram

with on one another. The source-filter model provides a simplification to this acoustic model via the following assumption:

The source and filter do not interact during the speech production mechanism. Thus, each element can be uniquely separated and independently acoustically modified [37, 181].

During real speech production, however, the glottal flow is dependent on the vocal tract impedance [26]. The aforementioned assumption excludes effects due to the interaction at the border between the glottal and vocal tract in favour of two independent elements. This is beneficial for analysis, as it allows to model speech as a relatively simple process. Moreover, the disunion of source and filter elements allows to examine how each element affects the output speech individually. For example, by keeping the filter function constant and varying the source function, the affect of source variations can be studied from the output speech. This model is widely used for a large range of applications due to its relative computational simplicity, such as speech analysis, speech recognition, speech synthesis, and speech coding. With this fundamental understanding of the source-filter model system diagram, we proceed to understand the physical and mathematical framework of the elements of this model.

4.1.1 Excitation Signal

There are two acoustic sources for speech production, resulting in two types of speech: *voiced* and *unvoiced*. The type of speech depends on the excitation signal, $e[n]$, to the speech production system. In voiced speech, the excitation signal is in the form of a unit impulse train, with a period corresponding to the pitch period, T, as per Equation 4.1. For unvoiced speech, the excitation signal will be a random, noise-like signal with a flat spectrum, as per Equation 4.2.

- Voiced excitation: $$e[n] = \delta[n - kT] \tag{4.1}$$

- Unvoiced excitation: $e[n] = whitenoise[n]$ (4.2)

For the scope of this thesis, only voiced speech will be discussed.

4.1.2 Glottal Filter

The excitation signal is the input to the overall speech production system. Specifically, the excitation signal is the input to a linear system, with the glottal volume velocity waveform, $u_g[n]$, as its output. The linear system is the glottal filter and it has an impulse response, $g[n]$, which results in the desired glottal pulse shape. The mathematical model used for $g[n]$ requires to have a significant roll-off at high frequencies. The model may have zeroes and poles, but an all-pole model is often more desirable [51, 50]. For the purposes of this analysis, the glottis is modelled as two 1^{st}-order low-pass filters in cascade, with an estimated cut-off frequency of 100 Hz, as per Equation 4.3 [180, 50].

$$G(z) = \frac{1}{(1 - z^{-1})^2} \tag{4.3}$$

4.1.3 Vocal Tract Filter

As mentioned in Section 2.4.2, the vocal tract area function results in a unique vocal tract shape with its own set of specific formants. The formants correspond to the spectral peaks of the magnitude response of the sound signal. These formant frequencies map to high energy frequency regions in the vocal tract. The frequencies of maximum energy in the magnitude response of the vocal tract transfer function are the resonance frequencies, which correspond to the formants [37]. It is important to note that resonance and formant frequencies are distinct of one another, as they are only approximately equal. Thus, there is inconsistency among researchers when defining the formants. Formants can refer to the peaks in the spectral envelope of the speech signal magnitude response, or they can refer to the resonances of the vocal tract, or to the poles of the transfer function of the vocal tract [182]. This transfer function was presented by Flanagan *et al*

[181] and is derived from the concatenated tube model, where all energy loses are excluded (i.e. heat, friction, movement). The vocal tract transfer function is given in Equation 4.4 below [181]:

$$V(z) = \frac{1}{\prod_{k=1}^{C_i}(1 - d_k z^{-1})(1 - d_k^* z^{-1})} \tag{4.4}$$

Where $V(z)$ is the transfer function of the vocal tract, C_i is the number of formant frequencies and, d_k and d_k^* are complex conjugate pairs. This equation can be expanded and re-arranged, as seen in Equation 4.5 and Equation 4.6, where p is the order of the model; $p = 2 \cdot C_i$.

$$\begin{align} V(z) &= \frac{1}{\prod_{k=1}^{K} 1 + b_k z^{-1} + c_k z^{-2}} \tag{4.5} \\ &= \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{4.6} \end{align}$$

Equation 4.6 provides a mathematical representation for the transfer function of an all-pole (recursive) filter. The poles occur as complex conjugates, with each pair of pole (resonance) accounts for a formant. Note that several phonemes, such as nasals and unvoiced sounds, require both poles (resonances) and zeroes (anti-resonances) in the vocal tract transfer function. Poles are more desirable than zeroes, as zeroes only contribute to the spectral balance. Therefore, since the zeroes lie within the unit circle in the z-domain, they can be approximated using multiple poles [50, 183]. Thus, the vocal tract is modelled as an all-pole, digital filter. In addition to the aforementioned assumptions, another assumption is in place for the behaviour of the vocal tract filter in this model. In continuous speech, the vocal tract articulators change positions from the production of one phoneme to the next. Thus, the vocal tract configuration is time-varying. In order to use this model, the stationary hypothesis is applied. In accordance with this hypothesis, the speech signal is analysed over a short period of time, 20-40 milliseconds, where the parameters of the model are assumed to stay constant [50, 51].

4.1.4 Radiation Load

The volume velocity waveform at the source is filtered through the vocal tract. The vocal tract terminates at the lips. The opening between the lips allows the sound wave to propagate into the speaker's surrounding and be auditorially perceived by a listener or captured by a microphone. Consider the head as a sphere and the mouth as an opening in this sphere. Assuming the sound wave is radiated from a spherical source, the sound pressure at the microphone can be approximated as the derivative of the volume velocity signal at the lips [50, 51, 182]. Thus, the pressure/volume-velocity ratio at the lips is a differentiator. The lip radiation is modelled as a finite-impulse response filter with one zero on the unit circle, as per Equation 4.7. This equation indicated that pressure is related to volume velocity via a high-pass effect; introducing a +6 dB/octave rise at high frequencies and attenuation at low frequencies [50, 51].

$$R(z) = 1 - z^{-1} \tag{4.7}$$

In this case, the radiation is modelled with a zero lying on the unit circle. Lip radiation can also be modelled with parameter ρ as to provide a more realistic model for sound propagation, as per Equation 4.8.

$$R(z) = 1 - \rho z^{-1} \tag{4.8}$$

Where ρ is an adjustable parameter smaller than 1, usually between $0.95 - 0.99$. Thus, the zero lies within the unit circle. Lip radiation effect can be included in either the *source* or the *filter* function. When included with the *source*, the input signal to the vocal tract is the derivative of the glottal volume velocity waveform, known as the *glottal flow derivative* [51]. In speech analysis, the radiation load is often removed using *pre-emphasis.*

4.1.5 Time- and Spectral-Domain Equations

To summarize the preceding sections, the elements of the source-filter model were discussed and their z-transform equations provided. Consider those elements in the time-domain. In the time-domain, speech production is the result of convolution of those elements. This relationship is illustrated in Equation 4.9.

$$s[n] = A_v(e[n] * g[n] * v[n] * r[n]) \tag{4.9}$$

Where A_v is a gain factor which controls the intensity of the excitation signal [50, 51]. This time-domain convolution translates to multiplication in the spectral domain, as per Equation 4.10.

$$S(z) = A_v E(z)G(z)V(z)R(z) \tag{4.10}$$

In order to analyse the speech signal, it is desirable to find a mathematical representation given a sampling instant of the output speech. This mathematical representation is derived by examining the relationship between the source-filter model elements and obtaining the transfer function for the model. This relationship is better expressed visually. A discrete-time system diagram for voiced speech production is given in Figure 4.2 below.

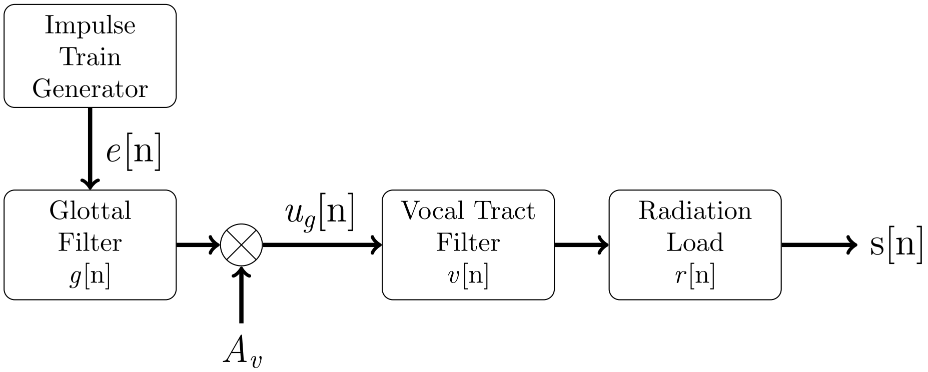

Figure 4.2: *Discrete-Time Voice Production System Diagram*

As expressed in Equation 4.9 and visualized in Figure 4.2, the output speech, $S(z)$, can be expressed as system of linear filters in cascade. The transfer function for this speech production model, $H(z)$, is the ratio between the output speech signal and the input excitation signal, as per Equation 4.11.

$$H(z) = \frac{S(z)}{E(z)} \tag{4.11}$$

$$= A_v G(z)V(z)R(z) \tag{4.12}$$

In order to formulate a mathematical representation for the speech signal sample instance, the transfer function needs to be simplified in the z-domain and transformed back to the time-domain. First, substitute into equation Equation 4.12 the z-transform filter transfer functions.

$$H(z) = \underbrace{A_v}_{\text{gain factor}} \cdot \underbrace{\frac{1}{(1-z^{-1})^2}}_{\text{G(z)}} \cdot \underbrace{\frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}}_{\text{V(z)}} \cdot \underbrace{(1-z^{-1})}_{\text{R(z)}} \tag{4.13}$$

We observe that Equation 4.13 can be readily simplified:

$$H(z) = A_v \cdot \frac{1}{(1-z^{-1})} \cdot \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{4.14}$$

Next, apply pre-emphasis on the transfer function. Pre-emphasis is in the form of differentiator; removing the remaining effects of the glottal filter, after the radiation load cancelled the other portion of it (see Equation 4.13).

$$H(z) = A_v \cdot \frac{1}{(1-z^{-1})} \cdot \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \cdot \underbrace{(1-z^{-1})}_{\text{pre-emphasis}} \tag{4.15}$$

$$= \frac{A_v}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{4.16}$$

Equation 4.15 represents the z-transform transfer function of the speech production system for voiced speech. This transfer function is advantageous for speech analysis since all the voice-production elements are modelled as one recursive filter. This allows the extraction of desired information about the individual contributions of the *source* and *filter* functions.

This equation is rearranged and transposed back to the discrete-time domain to describe a single speech sample.

$$\frac{S(z)}{E(z)} = \frac{A_v}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{4.17}$$

Then, by rearranging and expanding Equation 4.17 we get:

$$S(z) \cdot (1 - \sum_{k=1}^{p} a_k z^{-k}) = A_v \cdot E(z) \tag{4.18}$$

$$S(z) - \sum_{k=1}^{p} a_k (S(z) \cdot z^{-k}) = A_v \cdot E(z) \tag{4.19}$$

Taking the inverse z-transform of Equation 4.19.

$$\mathcal{Z}^{-1}\Big\{S(z) - \sum_{k=1}^{p} a_k (S(z) \cdot z^{-k})\Big\} = \mathcal{Z}^{-1}\Big\{A_v \cdot E(z)\Big\} \tag{4.20}$$

Use the z-transform linearity and time-shift properties [184].

$$\mathcal{Z}^{-1}\Big\{S(z)\Big\} - \sum_{k=1}^{p} a_k \mathcal{Z}^{-1}\Big\{(S(z) \cdot z^{-k})\Big\} = A_v \mathcal{Z}^{-1}\Big\{E(z)\Big\} \tag{4.21}$$

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + A_v e[n] \tag{4.22}$$

Equation 4.15 relates the excitation signal to the speech signal through the vocal tract transfer function. This all-pole model is referred to as the *auto-regressive* (AR) model, as the output is a regression of past outputs [183]. Equation 4.21 is the time-domain equivalent of this equation. It shows that a speech sample, $s[n]$, can be modelled as the sum of its previous samples, $s[n-k]$ multiplied by a constant, a_k, and added to the excitation signal, $A_v e[n]$. The a_k constants are referred to as the *linear prediction coefficients*. The linear prediction coefficient allow to model the filtering action of the vocal tract in speech [183]. Determination of these coefficients is performed via linear prediction analysis. This process is of paramount importance in speech analysis as a whole, and this thesis work in particular. By accurately computing the value of the filter coefficients, a_k, it is possible to model the speech signal parameters. Those coefficients can be approximated using Linear Predictive Coding (LPC), and is the topic of the following section.

4.2 Linear Predictive Coding (LPC)

The linear model of speech production described by the source-filter model allows the use of the advantageous linear prediction technique. In speech analysis, the input is often unknown. Consider a scenario where a speech signal is recorded and sampled. The data available contains information regarding the last $n-1$ samples. It will be highly beneficial to be able to predict the behaviour of the speech signal at the n^{th} sample, without having prior knowledge about the input to the system. Such operation can be computed using linear prediction analysis. Linear prediction is a mathematical operation where a discrete-time signal can be approximated as a linear function of weighted sum of its past output samples. Linear prediction allows for robust, efficient and reliable estimation of speech parameters. In speech coding, linear prediction is known as linear predictive coding. To understand how linear prediction can be applied to the speech production model, we examine the time-domain representation of a speech signal $s[n]$. Consider Equation 4.22 with the following added annotations:

$$\underbrace{s[n]}_{\substack{\text{speech}\\\text{signal}}} = \sum_{k=1}^{p} \underbrace{a_k}_{\substack{\text{filter}\\\text{coefficients}}} \underbrace{s[n-k]}_{\substack{\text{delayed}\\\text{samples}}} + \underbrace{A_v}_{\substack{\text{gain}\\\text{factor}}} \underbrace{e[n]}_{\substack{\text{excitation}\\\text{signal}}}$$

It is observed that a speech signal sample is given as a linear combination of its past samples and an input signal. This equation relates back to the *source-filter* model, where the gain and input are parameters of the source, and the $\{a_k, k = 1, 2 \ldots p\}$ coefficients are parameters of the vocal tract recursive filter.

A p^{th}-order linear predictor filter with p samples at equally spaced intervals can approximate the signal $s[n]$ from a linear weighted sum of its past samples. The linear predictor filter function in discrete-time is given in Equation 4.23 below.

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{4.23}$$

Where $\tilde{s}[n]$ is an approximation of the output speech signal, $s[n]$, at time n, and α_k are the prediction coefficients of the linear predictor. Since $\tilde{s}[n]$ is only an approximation of the speech signal, there exists an error between the actual value of the speech signal, $s[n]$, and the predicted value, $\tilde{s}[n]$. This prediction error is known as the *residual signal* and is denoted as $e_p[n]$. This prediction error is defined as the difference between the actual and predicted values, as per Equation 4.24.

$$\begin{aligned} e_p[n] &= s[n] - \tilde{s}[n] && (4.24) \\ &= s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] && (4.25) \end{aligned}$$

Linear predictive coding is centred around the concept of prediction coefficients. It is desirable for the LPC prediction coefficients, α_k, to match the coefficients of the recursive vocal tract filter, a_k. The optimization of the prediction coefficients is computed by minimizing the mean-squared prediction error. This is known as the *method of least squares*. Before delving into the mathematical reasoning behind this method, we digress back to explain why we strive to equate α_k and a_k. Consider the z-domain representation of the prediction error.

$$E_p(z) = S(z) - \sum_{k=1}^{p} \alpha_k \big(z^{-k} S(z)\big) \tag{4.26}$$

The transfer function of this model depicts the relationship between the output residual signal, $e_p[n]$, and the input speech signal, $s[n]$, as per Equation 4.27.

$$\frac{E_p(z)}{S(z)} = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{4.27}$$

This transfer function is denoted as $A(z)$, which is a finite impulse response (FIR) filter and is referred to as *prediction error filter*.

Consider a system where a voiced speech signal, $s[n]$, is the output of glottal excitation source, $u_g[n]$, and an all-pole system, $H(z)$, as depicted in Equation 4.16. Suppose we were interested in extracting information about the excitation source to that signal, what can we do? We'll pass the speech signal through a linear predictor and observe the residual signal. An flow diagram illustration of the process is given in Figure 4.3.

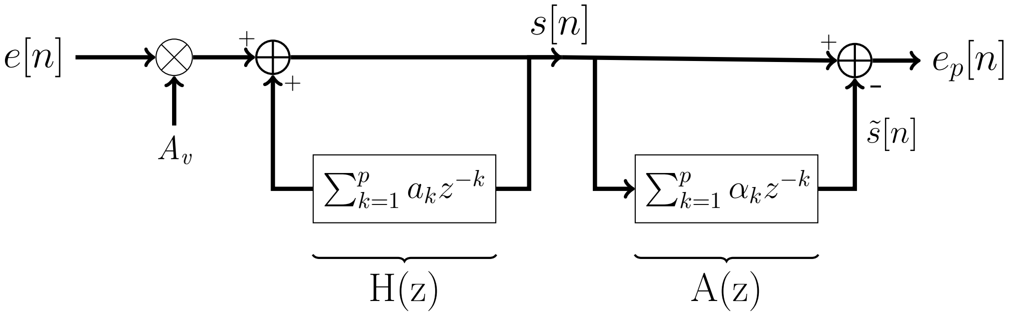

Figure 4.3: *Discrete-Time Voice Production System Diagram (adapted from [51] p. 180)*

Hence, by passing the speech signal through a linear predictor, it is possible to extract information about the original excitation source. For this reason, the error prediction filter, $A(z)$, is also known as the *inverse filter*. In order to extract an optimal representation of the source signal, the prediction coefficients need to correlate to the values of the coefficients of the FIR filter, a_k. This is done via the method of least squares which results in the *linear prediction normal equations*. These equations can be solved by various techniques, i.e. auto-correlation, covariance. For the scope of this thesis, only the auto-correlation method was implemented. Readers who are interested in broadening their horizons on this topic are encouraged to refer to the following books by Rabiner & Schaffer *et al* [50] (chapter 5) and Quatieri *et al* [51] (chapter 5). A detailed description of the auto-correlation method follows. The auto-correlation method represents the normal equations as *Yule-Walker equations*. This set of equations is solved using the *Levinson-Durbin algorithm*. These methods are elaborated upon in the following sections.

Note: The following sections include derivation of the mathematical equations used in Linear Predictive Coding. We note that understanding the mathematics is not required in order to understand the inverse filtering concept. However, we believe that in order to truly understand a concept, you have to understand the mathematics. For this reason, we chose to include the derivation in this chapter and not in the appendices. Readers who find the following sections tedious, are encouraged to skip ahead to Section 4.3.

4.2.1 Least Squares Minimization

The short-time, mean-squared prediction error, $\mathcal{E}$, is defined as the sum of all error energies in the windowed speech frame.

$$\begin{aligned}\mathcal{E} &= \sum_m e_p^2[m] && (4.28)\\ &= \sum_m \Big(s_n[m] - \tilde{s_n}[m]\Big)^2 && (4.29)\\ &= \sum_m \Big(s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k]\Big)^2 && (4.30)\end{aligned}$$

Where m indicates a short time segment; $s_n[m] = s[n+m]$. The length of the signal and the range of summation, also known as the *prediction error interval*, are discussed in the next section. At the moment, consider the speech signal to be a finite, deterministic signal tapered by a window.

The optimal α_k values which minimize the mean squared prediction error are obtained by setting the partial derivative of $\mathcal{E}$, with respect to α_k, to zero.

$$\frac{\partial \mathcal{E}}{\partial \alpha_i} = 0 \tag{4.31}$$

Where i is in the range of $\{1, 2, \ldots, p\}$, with p being the order of the linear predictor. Substituting Equation 4.25 into the above equation.

$$\frac{\partial}{\partial \alpha_i}\left[\sum_{m=1}^{N} \Big(s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k]\Big)^2\right] = 0 \tag{4.32}$$

By expanding Equation 4.32 we get:

$$-2\left[\sum_{m=1}^{N} s_n[m-i]\Big(s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k]\Big)\right] = 0 \tag{4.33}$$

Then, by rearranging Equation 4.33 we get:

$$\sum_{k=1}^{p} \alpha_k \sum_m s_n[m-i] s_n[m-k] = \sum_m s_n[m-i] s_n[m] \tag{4.34}$$

A closer examination of Equation 4.34 reveals that the right hand side is in fact the short-term convariance, i.e.

$$\phi_n[i,k] = \sum_m s_n[m-i]s_n[m-k] \tag{4.35}$$

Where $1 \leq i \leq p$ and $1 \leq k \leq p$. Reconsider Equation 4.34 and substitute new definition into it.

$$\sum_{k=1}^{p} \alpha_k \sum_m s_n[m-i]s_n[m-k] = \sum_m s_n[m-i]s_n[m-0]$$

$$\sum_{k=1}^{p} \alpha_k \phi_n[i,k] = \phi_n[i,0] \tag{4.36}$$

This set of equations is known as the *normal equations*. These equations hold true for both deterministic and random signals. It forms a set of p equations with p unknowns. Solving these equations for α_k minimizes the mean-squared prediction error, resulting in the optimal values for the prediction coefficients. In addition, the short-term covariance definition can be applied to the mean-squared prediction error, $\mathcal{E}$ [50, 51, 185].

Substituting Equation 4.35 into Equation 4.30:

$$\mathcal{E} = \phi_n[0,0] - \sum_{k=1}^{p} \alpha_k \phi_n[0,k] \tag{4.37}$$

Equation 4.36 and Equation 4.37 are invaluable in linear prediction analysis. These equations are used in combination to determine a set of prediction coefficients that provides a good estimation of the speech signal parameters. The auto-correlation method utilizes these equations for linear predictive analysis and is furthered discussed in the following section.

4.2.2 Auto-Correlation Method

Now that the *normal equations* have been defined, we can discuss the duration for signal $s_n[m]$ and the prediction error interval. The signal is tapered by a window,

$w[m]$, of finite length N. The windowed signal is therefore:

$$s_n[m] = \begin{cases} s[n+m]w[m], & 0 \leq \mathrm{m} \leq \mathrm{N} - 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.38}$$

The choice of window is correlated to the nature of the speech signal. A rectangular window is used if the signal is considered relatively stationary, whereas a Hamming or Hanning window is used for transient speech signals [185]. This tapered speech signal results in a prediction error, $e_p[m]$, which is zero outside the interval $0 \leq m \leq N - 1 + p$. As a result of the window, the prediction error is susceptible to error due to beginning and end effects. At the beginning of the interval, the predictor attempts to predict non-zero signal samples from samples that have been set to zero by the window. As a result, the error signal is large. Similarly, at the end of the interval, the predictor attempts to predict zero-value samples from non-zero samples inside the window [50, 51]. Due to this error behaviour, it is recommended to use a Hamming window and not a rectangular window.

The error prediction interval follows from the signal range. Since the error prediction is zero outside of $0 \leq m \leq N - 1 + p$, we assume that the error is minimized over all non-zero values, with the prediction error interval ranging from $-\infty$ to ∞ [185]. Hence, the mean-squared prediction error is defined over this range as:

$$\mathcal{E} = \sum_{m=0}^{N-1+p} e_p^2[m] = \sum_{-\infty}^{\infty} e_p^2[m] \tag{4.39}$$

This range is also applicable for the short-term covariance function:

$$\phi_n[i,k] = \sum_{m=0}^{N-1+p} s_n[m-i]s_n[m-k] \tag{4.40}$$

Under these conditions, the short-term covariance is determined by the overlapping region in the lag between the $s_n[m-i]$ and the $s_n[m-k]$ signals [50, 51]. Hence, it can be expressed as an auto-correlation function:

$$\phi_n[i,k] = R_n[i-k] \begin{cases} 1 \leq i \leq p \\ o \leq k \leq p \end{cases} \tag{4.41}$$

Where R_n is the auto-correlation function of the signal $s_n[m]$ and is defined as follows [185]:

$$R_n[i] = \sum_{m=0}^{N-1+p} s_n[m]s_n[m+i] \tag{4.42}$$

Since R_n is an even function ($R_n(k) = R_n(-k)$), the following definition can be made:

$$\phi_n[i,k] = R_n\big[|i-k|\big] \begin{cases} 1 \leq i \leq p \\ o \leq k \leq p \end{cases} \tag{4.43}$$

It is established that the auto-correlation function can replace the covariance function. Therefore, the *normal equations* and the mean-squared prediction error can be expressed using short-term auto-correlation functions as shown in Equation 4.44 and Equation 4.45 below:

$$\sum_{k=1}^{p} \alpha_k R_n[|i-k|] = R_n[i], \quad 1 \leq i \leq p \tag{4.44}$$

$$\mathcal{E} = R_n[0] - \sum_{k=1}^{p} \alpha_k R_n[k] \tag{4.45}$$

These equations jointly formulate the *Yule-Walker equations.*

4.2.3 Yule-Walker Equations

Thus far, both the prediction error, $\mathcal{E}$, and its partial derivative, $\partial\mathcal{E}/\partial\alpha_i$, have been expressed in terms of the short-term auto-correlation function, R_n. The result is the set of two invaluable equations; the Yule-Walker equations (see Equation 4.44 and Equation 4.45). Considering Equation 4.44, it can be expressed in matrix form as:

$$\boldsymbol{R}\vec{\alpha} = \vec{r} \tag{4.46}$$

Where $\boldsymbol{R}$ is the auto-correlation matrix, $\vec{\alpha}$ is the prediction coefficient vector, and $\vec{r}$ is a vector whose values correspond to $R_n[i]$. $\boldsymbol{R}$ is a positive-definite, Toeplitz matrix, also known as diagonal-constant matrix. To solve the predictor coefficients, we need to solve the following equation in matrix form:

$$\vec{\alpha} = \boldsymbol{R^{-1}}\vec{r} \tag{4.47}$$

To do so, we form the *Yule-Walker equations*. Combine the two equations to form the following set of matrices.

$$\begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[i] \\ R[1] & R[0] & R[1] & \cdots & R[i-1] \\ R[2] & R[1] & R[0] & \cdots & R[i-2] \\ & \vdots & & \ddots & \vdots \\ R[i] & R[i-1] & R[i-2] & \cdots & R[0] \end{bmatrix} \begin{bmatrix} 1 \\ -\alpha_1^{(i)} \\ -\alpha_2^{(i)} \\ \vdots \\ -\alpha_i^{(i)} \end{bmatrix} = \begin{bmatrix} \mathcal{E}^{(i)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.48}$$

The result is in an expanded Toeplitz matrix of dimensions $(i+1) \times (i+1)$. This can be expressed in matrix notation:

$$\boldsymbol{R^{(i)}\alpha^{(i)}} = \boldsymbol{e^{(i)}} \tag{4.49}$$

Where i is the order of the system, $1 \leq i \leq p$. This set of equations is known as the *Yule-Walker equations*. The Yule-Walker equations represent the auto-regressive model. They are satisfied by the i unknown predictor coefficients, and the mean-squared prediction error. Solving these equations results in the parameters for the auto-correlation functions - the prediction coefficients. The Yule-Walker equations can be solved using the *Levinson-Durbin algorithm*. The special properties of the Toeplitz matrix allows the system to be solved efficiently using recursion. In recursion, the Yule-Walker equations are said to have been solved after an iteration if the $\boldsymbol{e}$ matrix contains only singular non-zero value. The following section shows the recursive nature of the algorithm, depicting how to progress from one iteration to the next.

4.2.4 Levinson-Durbin Recursion Algorithm

The Levinson-Durbin (LD) algorithm is used to solve the aforementioned derived set of Yule-Walker equations. Implemented recursively, the algorithm allows to compute an equation set model of order $i + 1$ from set model of order i. Its

recursive nature is advantageous in speech processing. Since the order of the recursion may not be known to us, it is favourable to be able to see how the parameters values change between iterations. As previously mentioned, solving for prediction coefficients requires the inversion of the auto-correlation matrix; $\vec{\alpha} = \boldsymbol{R}^{-1}\vec{r}$. Such inversion requires $O(p^3)$ operations using conventional inversion techniques [51]. For a predictor of order p, $O(p^4)$ operations are required. The Levinson-Durbin recursion is computationally superior as it exploits the special form of the auto-correlation Toeplitz matrix, $\boldsymbol{R}$. The LD recursion requires $O(p)$ operations for the inversion of the auto-correlation matrix, and $O(p^2)$ operations overall for a p-order predictor [51].
Therefore, the LD algorithm can be readily considered a robust, computationally efficient, and cost-effective tool for extraction of prediction coefficients. The recursive nature of the algorithm implies that a predictor of order $i+1$ can be found using the previously found predictor coefficients values of order i. The recursion properties are discussed in this section, followed by pseudo-code implementation of the algorithm.

Once again, consider the Yule-Walker equations in matrix form.

$$\begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[i] \\ R[1] & R[0] & R[1] & \cdots & R[i-1] \\ R[2] & R[1] & R[0] & \cdots & R[i-2] \\ & \vdots & & \ddots & \vdots \\ R[i] & R[i-1] & R[i-2] & \cdots & R[0] \end{bmatrix} \begin{bmatrix} 1 \\ -\alpha_1^{(i)} \\ -\alpha_2^{(i)} \\ \vdots \\ -\alpha_i^{(i)} \end{bmatrix} = \begin{bmatrix} \mathcal{E}^{(i)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

In matrix form, this set of equations for order i is expressed as:

$$\boldsymbol{R^{(i)}\alpha^{(i)}} = \boldsymbol{e^{(i)}} \tag{4.50}$$

This set of equations is solved recursively. It is our intention to derive the next iteration, $i+1$, from this set of equations of order i, by equating the next iteration to the same form of the Yule-Walker equations. Thus, the Toeplitz matrix and the predictor vector should result in a column vector with only one non-zero entry; the mean-squared prediction error. As previously mentioned, in order to determine the predictor coefficients values, we need to solve the auto-correlation matrix. The auto-correlation Toeplitz matrix is of the following structure: $Row(1) = Row(i)$ reversed, $Row(2) = Row(i-1)$ reversed, ...etc. This structure is the basis for the recursion. Consider adding a 2nd column to the $\boldsymbol{\alpha}$ matrix, where the 2nd

column equals the 1st column reversed.
Note: *the dash line is used to provide visual cues for the changes in dimensions throughout this section.*

$$\begin{bmatrix} 1 \\ -\alpha_1^{(i)} \\ -\alpha_2^{(i)} \\ \vdots \\ -\alpha_i^{(i)} \end{bmatrix} \Rightarrow \left[\begin{array}{c:c} 1 & -\alpha_i^{(i)} \\ -\alpha_1^{(i)} & -\alpha_{i-1}^{(i)} \\ -\alpha_2^{(i)} & -\alpha_{i-2}^{(i)} \\ \vdots & \vdots \\ -\alpha_i^{(i)} & 1 \end{array}\right] \tag{4.51}$$

Substituting this new matrix into the Yule-Walker equations results in:

$$\boldsymbol{R^{(i)}} \underbrace{\left[\begin{array}{c:c} 1 & -\alpha_i^{(i)} \\ -\alpha_1^{(i)} & -\alpha_{i-1}^{(i)} \\ -\alpha_2^{(i)} & -\alpha_{i-2}^{(i)} \\ \vdots & \vdots \\ -\alpha_i^{(i)} & 1 \end{array}\right]}_{\boldsymbol{\alpha^{(i)}}} = \underbrace{\left[\begin{array}{c:c} \mathcal{E}^{(i)} & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & \mathcal{E}^{(i)} \end{array}\right]}_{\mathrm{e}^{(i)}} \tag{4.52}$$

Suppose the above equation was solved for the i^{th} iteration. For recursion to be valid, the same set of equations needs to be obtained for the next $i+1$ iteration. The $\boldsymbol{R^{(i+1)}}$ matrix will have the following form:

$$\left[\begin{array}{ccccc:c} R[0] & R[1] & R[2] & \cdots & R[i] & R[i+1] \\ R[1] & R[0] & R[1] & \cdots & R[i-1] & \\ R[1] & R[0] & R[1] & \cdots & R[i-1] & \vdots \\ & \vdots & & \ddots & \vdots & \\ R[i] & R[i-1] & R[i-2] & \cdots & R[0] & \\ \hdashline R[i+1] & & \cdots & & & R[0] \end{array}\right] \tag{4.53}$$

The expanded $\boldsymbol{\alpha}^{\boldsymbol{(i+1)}}$ matrix is of the form:

$$\begin{bmatrix} 1 & \boxed{\alpha_{i+1}^{(i)}} \\ \alpha_1^{(i)} & \alpha_i^{(i)} \\ \vdots & \vdots \\ \alpha_i^{(i)} & \alpha_1^{(i)} \\ \hdashline \boxed{\alpha_{i+1}^{(i)}} & 1 \end{bmatrix} \tag{4.54}$$

At the moment, assume that $\{\alpha_{i+1}^{(i+1)} = 0\}$ (highlighted above). Now multiply $\boldsymbol{R}^{\boldsymbol{(i+1)}}$ by $\boldsymbol{\alpha}^{\boldsymbol{(i+1)}}$ to formulate the new set of equations of order $i+1$.

$$\boldsymbol{R}^{\boldsymbol{(i+1)}} \left[\begin{array}{c:c} 1 & \alpha_{i+1}^{(i)} \\ \alpha_1^{(i)} & \alpha_i^{(i)} \\ \vdots & \vdots \\ \alpha_i^{(i)} & \alpha_1^{(i)} \\ \alpha_{i+1}^{(i)} & 1 \end{array}\right] = \left[\begin{array}{c:c} \mathcal{E}^{(i)} & \Delta_{i+1} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \Delta_{i+1} & \mathcal{E}^{(i)} \end{array}\right] \tag{4.55}$$

The equations are now of order $i+1$, where $\boldsymbol{R}^{\boldsymbol{(i+1)}}\boldsymbol{\alpha}^{\boldsymbol{(i+1)}} = \boldsymbol{e}^{\boldsymbol{(i+1)}}$. However, note that the solution for the new iteration is not in the same form as the i^{th} solution. The $\boldsymbol{e}^{\boldsymbol{(i+1)}}$ matrix now has two non-zero elements. If the only non-zero element was $\mathcal{E}$ then it can be subsequently said that the Yule-Walker equations were solved for order $i+1$. It is our intention to achieve $\{\Delta_{i+1} = 0\}$, where Δ_{i+1} is:

$$\Delta_{i+1} = 1R[i+1] + \alpha_1^{(m)}R[i] + \alpha_2^{(m)}R[i-1] + \cdots + \alpha_i^{(i)}R[1] + 0R[0] \tag{4.56}$$

$$= R[i+1] + \sum_{j=1}^{p} \alpha_j^{(i)} R[(i+1)-j] \tag{4.57}$$

Recall we've assumed that $\{\alpha_{i+1}^{(p)} = 0\}$. We need to account for this assumption. In order to solve the Yule-Walker equations for order $i+1$ we need to eliminate

$\{\Delta_{i+1}\}$. To do so, multiply the $\boldsymbol{e^{(i+1)}}$ matrix by a new, 2×2 square matrix, $\boldsymbol{k^{(i+1)}}$.

$$\boldsymbol{k^{(i+1)}} = \begin{bmatrix} 1 & -k_{i+1} \\ -k_{i+1} & 1 \end{bmatrix}, \quad k_{i+1} = \frac{\Delta_{i+1}}{\mathcal{E}^{(i)}} \tag{4.58}$$

Multiplying $\boldsymbol{e^{(i+1)}}$ by $\boldsymbol{k^{(i+1)}}$ results in the desired single non-zero element in each column.

$$\left[\begin{array}{c:c} \mathcal{E}^{(i)} & \Delta_{i+1} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \Delta_{i+1} & \mathcal{E}^{(i)} \end{array}\right] \begin{bmatrix} 1 & -k_{i+1} \\ -k_{i+1} & 1 \end{bmatrix} \tag{4.59}$$

Multiplying the first row of $\boldsymbol{e^{(i+1)}}$ by the right column of $\boldsymbol{k^{(i+1)}}$ results in:

$$\begin{aligned} -k_{i+1}\mathcal{E}^{(i)} + 1\Delta_{i+1} &= -\frac{\Delta_{i+1}}{\mathcal{E}^{(i)}}\mathcal{E}^{(i)} + \Delta_{i+1} \\ &= \Delta_{i+1} - \Delta_{i+1} \\ &= 0 \end{aligned}$$

Which eliminates the Δ_{i+1} error. Therefore, the $\boldsymbol{e^{(i+1)}}$ matrix is of the desired form of a single non-zero element in each column.

$$\begin{aligned} \boldsymbol{e^{(i+1)}} &= \begin{bmatrix} \mathcal{E}^{(i)} & \Delta_{i+1} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \Delta_{i+1} & \mathcal{E}^{(i)} \end{bmatrix} \begin{bmatrix} 1 & -k_{i+1} \\ -k_{i+1} & 1 \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{E}^{(i)}(1 - k_{i+1}) & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \mathcal{E}^{(i)}(1 - k_{i+1}) \end{bmatrix} \end{aligned} \tag{4.60}$$

Since the output is a matrix with single non-zero elements in each column, we have solved the Yule-Walker equations for this iteration. Thus, $\mathcal{E}^{(i)}(1 - k_{i+1}) = \mathcal{E}^{(i+1)}$, showing the recursive properties. Since $\mathcal{E}^{(i+1)}$ is non-negative, $\mathcal{E}^{(i+1)} \geq 0$, that means $-1 \leq k_{i+1} \leq 1$.

Since the right-side of the $\boldsymbol{R^{(i+1)}\alpha^{(i)}} = \boldsymbol{e^{(i+1)}}$ equation was multiplied by $\boldsymbol{k^{(i+1)}}$, the same multiplication needs to take place on the left-side.

$$\boldsymbol{R^{(i+1)}\alpha^{(i+1)}k^{(i+1)}} = \boldsymbol{e^{(i+1)}k^{(i+1)}} \tag{4.61}$$

Thus, the prediction coefficient matrix is multiplied by the $\boldsymbol{k^{(i+1)}}$ matrix. This shows the relationship between the predictor of order i to order $i+1$ in this recursion. The prediction coefficients are updated iteratively as shown in Equation 4.62. Hence, we've shown that the Yule-walker equations can be solved for order $i+1$ from the solution for order i.

$$\begin{aligned}\boldsymbol{\alpha^{(i+1)}k^{(i+1)}} &= \begin{bmatrix} 1 & 0 \\ \alpha_1^{(i)} & \alpha_i^{(i)} \\ \vdots & \vdots \\ \alpha_i^{(i)} & \alpha_1^{(i)} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -k_{i+1} \\ -k_{i+1} & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \alpha_{i+1}^{(i+1)} \\ \alpha_1^{(i+1)} & \alpha_i^{(i+1)} \\ \vdots & \vdots \\ \alpha_i^{(i+1)} & \alpha_1^{(i+1)} \\ \alpha_{i+1}^{(i+1)} & 1 \end{bmatrix} = \boldsymbol{\alpha^{(i+1)}}\end{aligned} \tag{4.62}$$

4.2.4.1 Levinson-Durbin Recursion Implementation

Initialization:

Initialization is for a predictor of order 0.

- $\mathcal{E}^{(0)} = R[0]$ (entire correlation of signal)
- $\alpha_j^{(0)} = 0$ (empty set, can also be set to equal 1)

Recursion:

For $1 \leq i \leq p$.

- $\Delta_i = R[i] + \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R[i-j]$
- $k_i = \frac{\Delta_i}{\mathcal{E}^{(i-1)}}$
- $\mathcal{E}^{(i)} = \mathcal{E}^{(i-1)}(1 - k_i^2)$
- $\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{j-1}^{(i-1)}, \qquad \text{for } 1 \leq j \leq m-1$
- $\alpha_i^{(i)} = -k_i$

This recursion is then performed for the desired prediction order, p. This recursion allows to compute the optimum prediction coefficients, where the optimum coefficients are:

$$\alpha_j = \alpha_j^{(i)}, \begin{cases} i = p. \\ j = 1,\ 2, \ldots, p. \end{cases} \tag{4.63}$$

In linear predictive coding, these prediction coefficients, α_k, are equated to the filter coefficients of the all-pole filter transfer function for voiced speech, a_k.

It is worth noting that the $\{k_i, 1 \leq i \leq p\}$ coefficients are an alternative parametric representation for the prediction coefficients, $\{\alpha_1,\ \alpha_2, \ldots,\ \alpha_p\}$. One of the major advantages of the k_i parameter is its range is well defined, $\{k_i = \pm 1\}$, which is useful for information compression. Moreover, if for an arbitrary order i, $k_i = \pm 1$, the mean-squared prediction error is cancelled, $\mathcal{E}^{(i)} = 0$. Such outcome indicates that we've achieved the best approximation and there is no need to increase the order of the predictor in an attempt to achieve greater resolution. In addition, if for an arbitrary order i the following is true: $k_i = 0,\ k_{m+i} = 0$, then the mean-squared prediction error remains constant, $\mathcal{E}^i = \mathcal{E}^{i+1}$.

For a perfect predictor, the mean-squared prediction error must equal zero. Referring back to the "updating equation" for $\mathcal{E}$, the mean-squared prediction error for a predictor of order i can be represented as [50]:

$$\mathcal{E}^{(i)} = R[0] \prod_{m=1}^{i} (1 - k_m^2) \tag{4.64}$$

Therefore, it can be concluded that a finite-energy procedure can be perfectly predicted if at least one of the following conditions occurs:

$$\mathcal{E} = 0, \quad \text{if} \quad \begin{cases} R[0] = 0 \\ k_i = \pm 1 \\ \sum_{i=1}^{\infty} k_i^2 = \infty \end{cases} \tag{4.65}$$

4.3 Inverse Filtering

In speech analysis, the glottal source can be parametrized and examined, thus providing information about the speaker. It is beneficial to be able to extract the glottal source from the recorded speech signal, rather than using an intrusive device or an external apparatus. This led to the development, and continued improvement, of the non-invasive inverse filtering technique. Inverse filtering is a speech analysis technique which exploits the linear time-invariant properties of the source-filter model to perform source-filter separation. At the cornerstone of inverse filtering is the estimation of the vocal tract parametric model, $V(z)$. In inverse filtering, the glottal source is estimated by removal of the vocal tract spectral contribution from the speech signal. Being one of the most popular methods for glottal source estimation, it branches into many glottal source estimation techniques. These inverse filtering methods differ by their vocal tract estimation procedure. Two of the main methods are: *Closed Phase Inverse Filtering* and *Iterative Adaptive Inverse Filtering.* For the scope of this thesis, Iterative Adaptive Inverse Filtering was used. An overview of the two methods follows.

4.3.1 Closed Phase Inverse Filtering (CPIF)

The inverse filtering method which looks towards accounting for this interaction is *Closed Phase Inverse Filtering.* As its name suggest, inverse filtering is carried out during the closed phase of the glottal cycle. During this timespan, the interaction between the sub-glottal and supra-glottal regions is minimized. This correlates to reduction in energy transfer between those regions. Hence, this method champions the notion that reduced glottal source contribution correlates to great accuracy in vocal tract estimation. In the closed phase, the vocal tract all-pole model is estimated through covariance analysis by means of linear prediction [186, 187]. This method, therefore, requires accurate determination of the closed phase. This proves troublesome, as it is notoriously difficult to obtain an accurate determination of the closed phase due to the difficulties in estimating the locations of the glottal opening and closing instances. Moreover, the closed phase may prove to be too short in relation to the speech cycle, resulting in inaccurate filter estimation for voice types (i.e. breathy) [186, 188].

4.3.2 Iterative Adaptive Inverse Filtering (IAIF)

Iterative inverse filtering techniques compute glottal flow estimation through iterative refinement of the source and filter parameters. The most widely used iterative technique is Iterative Adaptive Inverse Filtering [189]. This method computes the source signal from a pressure speech signal input via modelling and subsequent removal of the filter spectral envelope. A flow diagram of the IAIF method is given in Figure 4.4, where $s[n]$ is the speech signal and $g[n]$ is the resultant glottal source estimation. As the name suggests, IAIF extract the glottal flow signal through an iterative procedure twice repeated. Subsequent to pre-filtering of the original speech signal, $s[n]$, the glottal source is obtained in two phases, as indicated by the blue and red rectangular outlines in Figure 4.4. In the first iteration (blue), a rough estimate of the glottal flow waveform is obtained. This estimate is the input signal for the second iteration (red), which computes a greater accurate estimate of the glottal flow waveform. A more detailed description of the role of each step in the process is outlined below.

Prior to the processing and extraction of the glottal flow signal, the pressure speech signal is high-pass filtered in order to remove low-frequency noise inherent in the speech signal due to the recording procedure. Typically, the speech signal is filtered with a linear phase FIR filter with low cut-off frequency relative to the pitch [190]. Consider the first iteration of the IAIF algorithm (highlighted in blue in Figure 4.4). In **block 1**, a 2-pole model (1$^{\text{st}}$-order LPC) of the glottal flow and lip radiation contributions is obtained. The model is based on the characteristics of -12 dB/octave tilt in the spectral envelope of the glottal flow and $+6$ dB/octave contribution of lip radiation, resulting in a first-order filter of -6 dB/octave effect, $\mathrm{H_{g1}(z)}$. These glottal and lip radiation contributions are removed from the speech signal through inverse filtering in **block 2**. The result is a signal containing only the contributions of the vocal tract and impulse-train excitation. In **block 3**, the vocal tract is modelled with a p-order Linear Predictor and subsequently eliminated from the original speech signal through inverse filtering in **block 4**. In **block 5**, the lip radiation in cancelled from the $\dot{g}_1[n]$ signal and outputs the first estimate of the glottal flow. This glottal estimation drives the second iteration of the algorithm, **blocks 6-10** (highlighted in red in Figure 4.4), where a refined estimate of the glottal flow is obtained. We note that the choice of LPC order could have detrimental impact on the glottal flow. Drugman *et al* [97] suggest LPC analysis to be in the range of 10 to 18;

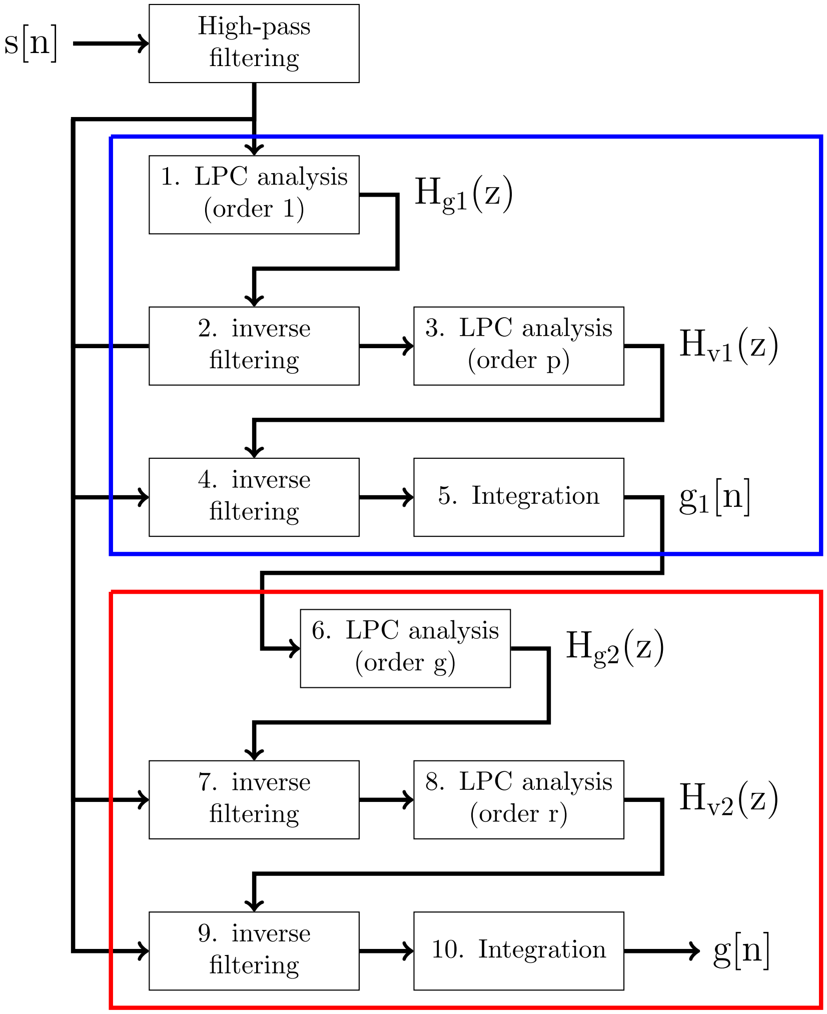

Figure 4.4: *Iterative Adaptive Inverse Filtering (IAIF) block flow diagram; 1st iteration in blue, 2nd in red.*

a higher LPC order will over-fit the vocal tract spectral envelope, while a lower LPC order will not sufficiently remove the vocal tract contributions, resulting in source-filter coupling. However, in the original pulication by Alku *et al* [189], a much lower LPC order range of 4 to 12 was applied.

The IAIF method removes the influence of the glottal pulse from the speech signal by utilising the entire pitch period, thus the method is performed pitch asynchronously. Since the method is reliant on linear prediction, it is susceptible to its deficiencies, which manifests in the incorrect formant estimation due to the underlying harmonic structure of the speech spectrum [185]. Alku *et al* [189] suggested an improvement to his IAIF method by performing linear prediction analysis pitch synchronously. By computing the IAIF analysis one fundamental period at a time, using frames that span between 2 consecutive maximal glottal openings, the influence of the pitch period is eliminated. A block system diagram of the IAIF method is given in Figure 4.5, where $s[n]$ is the original input signal, $s_{hp}[n]$ is the high-pass filtered speech signal, $g_{pa}[n]$ is the estimate of the glottal flow computed pitch asynchronously, n_i mark the time indices of the maximum glottal openings, and $g[n]$ is the glottal flow estimate computed pitch synchronously. A further improvement to this original method was suggested by replacing the LPC technique with discrete all-pole (DAP) modelling [191]. Compared to linear prediction, DAP showed greater robustness at fitting spectral envelopes to a set of points [192]. Moreover, DAP showed greater reliability for high-frequency signals (> 300 Hz), resulting in a glottal flow estimate with lower formant ripples distortion. Distortion in the glottal flow signal is attributed to glottal-vocal tract coupling [188]. DAP gives a more accurate vocal tract formant estimation, which is less influenced by pitch harmonic peaks [192, 76, 190]. Alku *et al* [190] notes that perhaps the greatest benefit of this improved technique is its implementation can be automatic, making it cost effective.

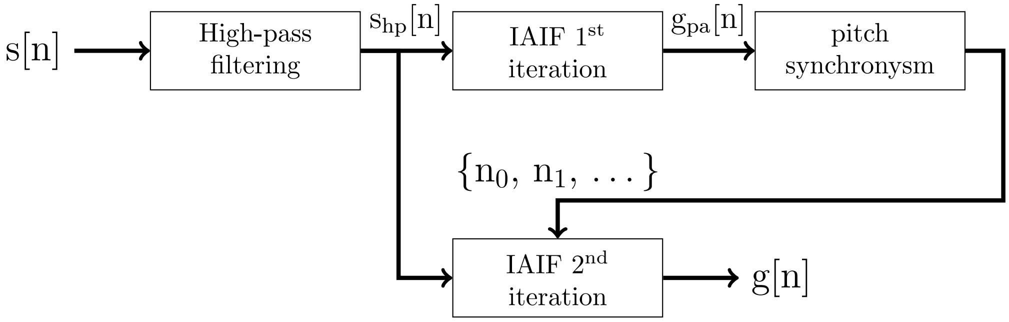

Figure 4.5: *Pitch Synchronous Iterative Adaptive Inverse Filtering (IAIF) block flow diagram.*

4.4 Polarity Detection

The polarity of the speech signal influences the results obtained by speech processing. It can be readily understood that incorrect polarity determination will impact the analysis of the speech signal. For example, polarity inversion for a speech signal will result in an inverted glottal source waveform, which in turn will yield incorrect glottal parameter measurements. In nature, speech polarity is the result of glottal pulse asymmetry. As previously discussed in this thesis, the derivative glottal waveform exhibits a discontinuity about the glottal closing instance, as depicted by the Liljencrants-Fant model [193]. This discontinuity corresponds to a negative peak and is the indicator for positive (correct) polarity. In contrast, a positive peak corresponds to negative polarity. Although polarity is paramount for speech processing, it is worth noting that the human ear is unable to discern polarity inversion, which means polarity inversion can be easily overlooked [194]. Moreover, the insensitivity of the human ear to polarity inversion allows the encription of data in the speech signal by altering the phase of the speech signal [194, 195]. As a result, robust speech polarity detection algorithms had been developed. The following discussion highlights several popular polarity detection techniques. A brief overview of the existing polarity detection algorithms was presented by Dutoit *et al* [196].

Speech polarity inversion is caused by reversal of the electrical circuitry employed in the speech data recording procedure. An input signal fed to the inverting terminal of an operational amplifier during the amplification stage will result in an output signal 180 degrees out of phase (poalrity inversion) [197]. Although the need for speech polarity detection techniques has been addressed in literature, only a handful of studies had been published, with most studies published in the last 10 years. Campbell *et al* [193] proposed a polarity detection based on the gradient of spurious glottal waveforms. The method is governed on GCI detection. Polarity detection based on the gradient of spurious glottal waveforms obtains the glottal waveform through iterative adaptive inverse filtering [189], and subsequently captures the position of the GCI instances by computing the gradient of the glottal waveform. Polarity is classified as positive if the GCI instances correspond to the negative peaks of the derivative glottal waveforms.

Drugman & Dutoit *et al* [198] proposed a polarity detection from the phase shift between two oscillating signals/statistical-moments derived from the speech signal. One signal is obtained with odd non-linearity, while the other is obtained with even non-linearity. The former is dependent on polarity while the latter is not. Those moments oscillate about the fundamental frequencies and exhibit a phase shift. This phase shift is used to determine the polarity of the speech signal. At the time, this method was shown to be the most robust compared to the available state-of-the-art techniques. However, as the moments are observed around the fundamental frequencies, this method requires an estimate of the speech signal pitch period. This method was further examined in a subsequent papter by the authors *et al* [196].

Drugman & Dutoit *et al* [199] proposed a computationally-simple, automatic speech polarity detection algorithm, referred to as Residual Excitation Skewness (RESKEW). The algorithm is based on the distribution of samples for a LPC residual signal and the derivative glottal flow signal. For this, the author notes that a residual signal will exhibit a positive skewness while the derivative glottal signal will exhibit a negative skewness. Through computation of the signals' skewness, polarity inversion is detected. The algorithm strength is apparent as it does not require an estimation of the fundamental period or the location of GCI instances prior to determining the polarity. Moreover, for clean and noisy conditions, the method was shown to be robust compared to existing polarity detection methods. The author suggests an improvement for the method in the form of "skipping" unvoiced segments. The improved suggestion was carried out by Abhiram *et al* [200]. The new embodiment of the RESKEW algorithm was termed long-term weighted skew (LT-WSKEW) algorithm. For this, the authors show the computation of LT-WSKEW to be 10 times faster than the fastest existing polarity detection algorithm. However, the algorithm does not perform as well in clean and noisy conditions as the RESKEW algorithm. Thus, it offers a more cost-effective solution at the price of reduced accuracy.

Most recently, Govind *et al* [201] proposed a polarity detection algorithm using the cosine phase of the speech signal. The method utilizes the Hilbert envelope, where the peaks of Hilbert envelope approximate the locations of epochs (instant of significant excitation of the vocal-tract, GCIs). The cosine phase is then computed about the epochs, with its slope at the zero corresponding corresponding to the polarity of the speech signal. Govind *et al* [202] proposed an improvement to this method, by comparing the amplitude variations of the Hilbert-transformed signal, $s_h[n]$, and the original speech signal, $s[n]$. The epochs are approximated as peaks in the Hilbert envelope, and the slopes of $s_h[n]$ and $s[n]$ are computed about the epochs. The polarity is determined based on the difference in average slope values; greater $s_h[n]$ slope corresponds to positive polarity.

Chapter 5

Methodology

This chapter details the techniques employed towards the creation of a toolbox for the extraction and analysis of the glottal volume velocity waveform. The outset of this chapter provides an overview of the R environment and the reasoning for its use. Subsequently, the emuR library is introduced. Next, the blueprint for the toolbox is provided in great detail. A description of available voice analysis packages and their relation to this toolbox is explored. Following, the implementation of speech processing algorithms in R is described. The implementation of the pitch tracking and open quotient parameter algorithms are included. The chapter concludes with a detailed discussion of the speech material analysed in this thesis work. This includes information regarding the speakers, the recording procedure, and related studies.

5.1 The R Environment

The voice analysis toolbox was implemented in R. Developed by Robert Gentleman and Ross Ihaka from the Department of Statistics at the University of Auckland, the *R* project is an Open-Source, Free Software environment for statistical computing and graphics [203]. *R* is a GNU project that combines ideas from the *Scheme* and *S* programming languages. Although *R* differs from *S*, it can be considered an embodiment of the language, with *S* code implemented under *R*. It is compatible with UNIX, Windows and MacOS systems. Moreover, it is compatible with other statistical systems, such as *S*, *SAS*, *Stata*, etc. Currently, the *R* project features over 2000 user-contributed statistical libraries. *R* is extended through these libraries, or packages, which are available for download via the Comprehensive *R* Archive Network, referred to as *CRAN* [204]. CRAN is a global network of ftp and web servers that store and distribute freely-available *R* libraries, code, and documentation. Although *R* is a command line interface, working with *R* is simplified using the *RStudio* integrated devvelopment environment (IDE), which includes a text-editor, debugging tools, and improved data-visualisation packages [205].

When working in *R*, data are stored as an object in the workspace. Each object has its own class, type, and dimension. Similar to *MATLAB* (matrix laboratory) and its computationally efficient matrix manipulation, *R* has its own data-types: *vector*, *array*, *matrix*, *list*, and *data-frame*. A complete disclosure of these data-types and relevant functions is available in [206].

A *vector* is the simplest object form in *R*. The most basic type of vector is the *atomic vector*. There are four types of commonly used atomic vectors: *logical*, *integer*, *numeric* (double), *character* (contains *strings*). Vectors are one-dimensional and homogeneous, meaning the contents of a vector must all be of the same type. Each vector has specific *type* and *length*, which can be computed using the *typeof()* and *length()* functions, respectively. Vectors can be created using the *c()* and *seq()* functions, with the former allowing for the combination of vector values.

An *array* is a homogeneous, n-dimensional, symmetric vector stored with additional attributes specifying its dimensions and dimensions names. The array elements are accessed using an integer index notation of the form $[i, j]$, where i and j refer to rows and columns, respectively.

A *matrix* is a two-dimensional array of atomic vectors of the same type (homogeneous). A matrix can be formed using the *matrix()* functions. Alternatively, vectors can be combined to form matrices using the *rbind()* and *cbind()* functions. Many linear algebra operations utilize the matrix data-type. In *R*, matrix operations are slower compared to vector and list operations, especially if it requires to transpose or multiply matricies. Although the {Matrix} package [207] offers enhanced matrices functionality, it is still preferable to use the aforementioned data-types.

A *list* is a type of vector, also referred to as recursive vector. However, lists differ

from vectors as they are heterogeneous, meaning they can encompass multiple data-types simultaneously. A list is a hierarchical, tree-like data-structure that can store a collection of objects (even other lists). One of the main advantages of lists is that it can contain objects with different lengths. Multiple objects can be joined into a list using the *list()* command. Alternatively, a list can be reconstructed as a singular vector using the *unlist()* command. Lists are powerful in analysis, as they can be dynamically created, altered, and named. A list subset can be accessed using either an integer index or a name: $[i]$ extracts a sub-list object as a list, $[[i]]$ extracts an element while maintaining its original type, and $\$name$ extracts named elements.
A *data-frame*, also referred to as *augmented list*, is a heterogeneous collection of equal-length variables. Data-frames are a fundamental data-structure for modelling operations in *R* and are created using the *data.table*() function. Similar to lists, data-frames have naming attributes, where its dimensions can be named for indexing. The columns/rows of the data-frame can be accessed using integer subscripts, $[i, j]$, or using the $ symbol followed by the column/row name.
Another *R* data-type that is worth mentioning is the *NULL* object. The NULL object is of class *NULL* and therefore cannot be assigned to atomic vectors, as a vector cannot combine two different data-types. When performing computation in *R*, the NULL object is returned for expressions and/or functions with undefined values. It is important to note that NULL is different than NA. While NULL is of length 0, NA is of length 1. NA is of class *logical* and indicates to a missing value. Unlike NULL, NA can be combined with atomic vectors. In addition, any computation involving a missing value will return a missing value.

An online search of for loops in *R* results in multiple threads regarding their slow computation time. For loops are notoriously slow in *R*. Since for loops are a control flow base function, the issue does not lie within the *R* base package. The main reason for the slowness of for loops is inefficient iteration through non-primitive data-types, i.e. data-frames. *R* is very specific about the way objects in lists are accessed and modified. Incorrect handling of lists in for loops results in copying of the object with every iteration and placing it into a new memory location. Such process is time consuming and inefficient. This problem can be solved by careful handling of objects, ensuring the object is not copied with each iteration by looking at its memory address using the *address()* function. This too is time consuming and may prove difficult for inexperienced users or fellow researchers who look into making small changes in the code. Other factors may affect the code running time, such as: the objects are not initialised to their full length before the loop, the loop includes operations that can be carried out of the loop [208]. To better understand the code, it is recommended to utilize the *system.time()* function, which divides the elapsed compilation time between *user* and *system*. A large *user* time indicates the code can be further optimized. However, note that even after mitigating the aforementioned issues, it is still not recommended to use for loops as they simply convey that an iteration is taking place, but do not convey the high-level operation of the code segment. Instead, another, more expressive solution is available, in the form of a *functional*. A functional is a function that in addition to the data, takes a function as an input and returns a vector as an output. A functional is set and modified to fit a task, thus better enclose information about the code. *R* has several functional which allow for computationally efficient iteration through lists and data-frames, i.e.*apply*, *lapply*.

The following question may arise: *why use R when other speech analysis packages are already available, like the VOICEBOX speech analysis toolbox in MATLAB?* [209] Although other packages are available, we chose to implement this new toolbox in *R* for several reasons. First, *R* is an open-source, free software which is readily available for researchers everywhere. MATLAB, however, requires a license and cannot be used by more than one user per license at a given time. Although tertiary institutions receive the more flexible MATLAB student-license, work places are often limited to a small number of licenses, which limits the amount of time a single user can spend on MATLAB. For this reason, we view *R* as a superior replacement for MATLAB. Second, *R* has a comprehensive collection of statistical analysis tools for both speech engineering and science disciplines. Third, there is a large number of graphics libraries in *R*, customized for the different data structures used (i.e. {ggplot2} package). Fourth, scripts can be easily adapted and edited in *R*. In a method call, input arguments can be entered in any order, which is not the case in MATLAB, where the order corresponds to the arguments [210, 211]. Finally, whereas MATLAB does not yet have a speech corpus interface for a user-friendly extraction of speech segments from a given database *R*, has speech corpus processing and analysis extension - the {emuR} package [212].

5.1.1 The emuR Library

The emuR package is a Speech Database Management System (EMU-SDMS). The package is a collection of software tools for management, preparation, extraction and analysis of speech databases [213]. Prior to its current platform embodiment, the EMU software required separate installation and needed to be run simultaneously with *R* in order to interface between *R* and the EMU speech corpora. In its current embodiment, the EMU-SDMS main component is its emuR library extension. This package is a database search engine which allows for simple extraction of speech segments from utterances based on the sequential and hierarchical structure of the database [213].

The previous EMU iteration utilizes the EMU Legacy database format. The database root directory consists of a .tpl template file, a folder with an arbitrary name to group the .wav recorded speech data, and a second folder to group to .lab label files. The template file defines the hierarchical structure of the database, specifying the phonetic level of the utterances and lists any additional attributes that might have been generated by the user (i.e. derived formants).

At the beginning of this project, the EMU Legacy database was used. The reason being the databases had already been set up in the EMU software. However, this database format was only compatible with earlier versions of the emuR package. As the project progressed, it became evident that it would be beneficial to update the emuR package, in order to take advantage of the new iteration of its commands. This update required the transition into the new database format, emuDB. In its current iteration, the Emu Legacy database format was replaced in favour of the simpler emuDB database format. NOTE: A more detailed comparison between EMU Legact and emuDB is available at (`https://mran.microsoft.com/web/packages/emuR/vignettes/EQL.html`).

One of the main advantages of the {emuR} package is its ability to handle speech corpora with user-friendly import and export of speech data. This is achieved by creating a track-data object, which contains information from a selected track in the database (e.g. speech samples). Once an EMU database is loaded into the *R* workplace environment, the database can be queried to subset the data into one of its Level definitions (i.e. word, phoneme). The query results in a segment list. A segment list is a multi-columned matrix, which contains phonetic and temporal information about the data subset. In EMU, speech data is stored as tracks. The tracks are a Simple Single File Format (SSFF) structural element of the emuDB. They may contain complementary or derived data corresponding to the original speech data (i.e. samples, formants). A specific track corresponding to entries in a segment list can be extracted using the *get_trackdata* command. In this project, this command was utilized to extract the numerical speech data corresponding to the sampled speech data stored as a .wav file; a **SAMPLES** SSFF track. We note that at the time of analysis, inclusion of samples is not automated under the iteration of the {emuR} package. Instead, a samples track is added using the *add_ssffTrackDefinition*. However, at its most current embodiment, audio samples can be accessed using the **MEDIAFILE_SAMPLES** track name, which does not require a manual definition by the user. Following the extraction of the track-data object, speech analysis methods can be applied to the speech signal. Similar to the {base} package functional in *R*, EMU has its own functional command *trapply*. Thus, analysis is performed on a segmented speech signal corresponding to the temporal location of a Level element (vowel) in the segment list.

5.2 Voice Analysis Toolbox

This section presents the methods incorporated into the vocal analysis toolbox implemented in R. First, we introduce the toolbox framework, detailing the procedures used to output glottal parameters from a speech signal input. Next, we present the parametrization algorithms implemented into this toolbox. Prior to the following discussion, we would like to address the reason for the conception of this toolbox. We are aware that an abundance of voice analysis packages are available. From the literature, it can be deduced that many research and academic facilities have implemented their own voice analysis tools for voice source extraction and parametrization.

5.2.1 Overview of Available Speech Analysis Packages

In 2008, the MATLAB-based software package *TKK Aparat*, was made freely available under an open-source license [214]. Developed at Helsinki University of Technology, TKK Aparat provides a user-interface environment for inverse filtering and glottal flow parametrization. The package includes computationally simple inverse filtering algorithms and an automatic parametrization process for parameter computation. The algorithms presented in this package were subsequently utilized as the building blocks for the COVAREP project [215]. The COVAREP project enjoyed greater popularity than the TKK Aparat environment for three main reason. First, it provided a more comprehensive package that extended beyond glottal analysis. Second, the COVAREP project was the collaborative results of five research facilities across the globe (Greece, Ireland, Belgium, Finland, and the US). Publications of new, novel speech processing techniques/algorithms by those institutions were incorporated into, and cited, the COVAREP project, which led to an increase in its reputation and popularity. Third, the source code for this project was made freely avilable on *GitHub*. Fellow academics were encouraged to offer input and improvement to the existing methods, essentially creating an online community for speech analysis.

5.2.2 The COVAREP Project

The COVAREP project is a collaborative voice analysis respiratory for speech technologies (`covarep.github.io/covarep`) [215]. Aimed towards researchers, it provides computationally fast speech processing algorithms. The COVAREP project is implemented in the Matlab language and partially compatible GNU Octave (`octave.org`). Several speech processing techniques are available to extract information required for pitch-synchronous analysis of the speech signal, such as polarity detection and pitch tracking. Subsequent techniques allows for formant tracking, glottal flow estimation, and phase processing. For the purposes of this post-graduate project, three methods were adapted from Matlab language to *R*. Given below is a description of the fundamental theory principles governing each speech processing method. This is followed by the consideration and modifications made to the Matlab code for fast computation in *R*.

Both TKK Aparat and COVAREP are considered the blueprint for this new R toolbox. indeed, several algorithms were adapted from those existing packages (IAIF, pitch tracking). Similar to the hopes of the authors of those packages, it is our hope that our R toolbox will enable fellow researchers to conduct speech research in an efficient and cost-effective manner. It is our aim to make this toolbox freely available under an open-source license upon the completion of this thesis work. Finally, although this toolbox was developed for the purposes of speech analysis, we do not consider it to be used solely for that. In our facilities, we expect this toolbox to be integrated into research areas such as the healthcare robot project [216], speech synthesis [217], and the MAONZE project [218].

5.2.3 The R Voice Analysis Toolbox Framework

This section provides a detailed exploration of the voice analysis toolbox developed in this thesis work. This is a descriptive walk-through for the various speech processing stages, from speech corpora import into the R environment, to glottal flow extraction and subsequent obtaining of glottal pulse descriptive features. A system block flow diagram for the overall structure of this toolbox is presented in Figure 5.1.

The initial stage in voice analysis is data-preparation. For the purposes of this discussion, it is assumed that a database was pre-organised to match the required emuR database format. The database is loaded into the R workspace environment. Next, the database is queried by the user, which allows access to the *levelDefinitions* of the emu database object (i.e. phoneme). The query is saved to a specialized *list* object, known as segment-list, or *seglist*. The seglist object is paramount for speech corpus analysis. Not only does it allow automatic extraction of all desired speech segment information, but it also allows access to the SSFF tracks. The segment list contains one row per segment of query (i.e. one row per vowel). Some of the most useful information presented in the seglist is the label, start & end times, file name, sample rate, and level definition. Prior to the glottal extraction stage, all speech segments with duration shorter than 100 ms are excluded from the analysis. This is done in order to ensure sufficient number of glottal pulse cycles per speech signal. Performing analysis on shorter speech signals compromises the efficiency of the pitch detection and inverse filtering algorithms. We note that this database query procedure is not automated and requires user-input. Throughout the development of this project, attempts were made to create a script for automatic segment list extraction and further segmentation. While the process can be automated for simple database segmentation, e.g. extraction of all vowel phonemes and separation of segment list into young and old speakers, the process becomes too laborious when attempting to segment a database for multiple speaker types and multiple level definitions.

Following the extraction of the segment list, an automatic glottal source extraction ensues. The segment list is used as an input to a glottal source extraction function, which accesses the SAMPLES SSFF track for each segment list entry. Prior to the extraction of the glottal flow waveform, the signal's polarity and pitch are determined, respectively. The initial stage of the processing determines the polarity of the speech signal. Polarity inversion is caused due to inversion of the electrical circuitry employed in the speech data recording procedure. Incorrect polarity would have a detrimental affect on the estimation of glottal descriptors. The polarity detection algorithm employed is the one proposed by Drugman *et al* [199]. Next, the fundamental frequency of the speech signal is determined via the pitch estimation technique based on the summation of residual harmonics (see subsubsection 5.2.4.1) [97]. Finally, in order to minimize potential errors in the detection of vowel boundaries, the first and last 5% of those extracted vowels are removed before further processing, thus reducing potential interferences from neighbouring parts of the speech signal due to coarticulation effects (i.e. consonants). Prior to the application of IAIF method, the truncated speech signal is high-pass filtered in order to remove low-frequency noise fluctuations in the speech signal due to the recording procedure [189]. This is achieved using a linear phase FIR filter with cut-off frequency of 40 Hz. At this point in the analysis the speech signal has been prepared for glottal flow extraction.

The glottal flow waveform is computed pitch synchronously using the iterative adaptive inverse filtering (IAIF) algorithm. The analysis can be performed pitch synchronously as the pitch period was already obtained in the previous stage. As described in the preceding chapter, the IAIF algorithm is a two-stage iteration process governed by the principles of linear predictive coding (LPC). We note that the LPC order was set relative to the sampling frequency of the speech signal, which resulted in a LPC order in the range of 12 to 18. The speech signal is segmented into speech frames of length 25 ms with a frame shift of 5 ms. The vocal tract filter model is estimated for each analysis frame, and glottal source is subsequently computed via inverse filtering of the vocal tract and radiation load contributions from the original speech signal. Each glottal frame is Hanning windowed. The glottal flow waveform is then computed using the overlap-and-add method. At this point in the analysis, the glottal flow waveforms for every speech segment in the aforementioned segment list was computed. The resultant glottal waveforms are stored as a list object, with every entry in the list containing the phoneme and speech file name attributions of its corresponding speech segment. This glottal list can be subsequently extracted by the user following the computation of the glottal parameters. Alternatively, the user can specify this to be the final stage of the analysis via the function call. Rigorous testing was carried out in order to minimize the computation time of this glottal derivation procedure. In its current embodiment, the runtime for a speech corpus containing 1600 vowel tokens, with an average speech segment of 200 ms (3500 samples), is 1336 seconds, which average to about 0.8 seconds per speech segment. Note that this runtime is inclusive of the SSFF SAMPLES track extraction time per speech segment.

The second stage offered in this toolbox is the extraction of descriptive glottal features via glottal pulse parametrization. In the process of time-domain parametrization, temporal and amplitude instances are acquired. This parametrization was carried out on glottal waveform frames of length corresponding to 10% of glottal waveform duration (20 ms on average). Prior to the parametrization, the glottal waveform was low-pass filtered using a Butterworth filter with a cut-off frequency relative to the pitch period. Such filtering allows to smooth the waveform and eliminate high-frequency noise effects. The acquisition of the aforementioned instances was achieved using a peak detection algorithm. The peak detection was implemented with a tolerance variable to prevent detection of neighbouring peaks at close proximity. The tolerance variable was set relative to the previously-computed pitch period. In addition, the time and amplitude instances are computed via quadratic interpolation in order reduce quantisation error. Due to the inherent fluctuations in the glottal shape (i.e. formant ripples), additional precautions were employed to prevent incorrect estimation of peaks and valleys. The mean average amplitude of a glottal waveform was computed, with peaks below 25% of this amplitude threshold excluded. Similarly, valleys above 25% of this amplitude threshold were excluded. Average sample index was used to identify neighbouring peaks/valleys with irregular spacing. Additional care was put forth towards ensuring the existence of valleys between two neighbouring peaks and vice verse. We note that such peak detection methodology greatly improved the accuracy of our results and proved robust especially for glottal waveforms extracted from speech signals distorted by noise. We also note that this peak detection method did not result in a noticeable increase in CPU runtime. Through this peak detection method, the maximum and minimum glottal flow time instances and amplitude measurements were obtained. This enables the computation of jitter and shimmer parameters.

Before the open quotient, speed quotient, and normalized amplitude quotient can be computed, information regarding the glottal opening and closing instances is required. For detailed discussion on the implementation of the GOI and GCI points, please refer to subsubsection 5.2.4.2. The glottal opening instance is determined to occur at a time-instant corresponding to an instance for which the glottal flow exceeds a user-defined threshold which corresponds to a peak-to-peak amplitude percentage (i.e. 50%). The GOI threshold is computed pitch-synchronously, on a period-to-period basis. Thus, removing the effect of a DC offset, which would have been present had the threshold been computed as an average across the entire glottal waveform. The glottal closing instance is determined through the derivative glottal flow waveform, as it corresponds to its negative peak. Since the glottal closing instance occurs between the instances of maximum and minimum glottal flow, the glottal waveform is segmented to its closing phase.We note that the sampling frequency plays an important role in the approximation of the derivative glottal flow waveform. With low sampling frequency (i.e. 16 kHz), when differentiated to produce the glottal flow derivative, the resultant wave was a crude approximation of the theoretical derivative. Thus, up-sampling interpolation was used to generate a smoother glottal flow waveform, which in turn resulted in a coherent glottal flow derivative approximation. Then, using the aforementioned pitch detection algorithm, the point of maximum negative flow is determined through quadratic interpolation.

Following the detection of GOI and GCI instances, the open quotient, speed quotient, and normalized amplitude quotient were computed. Upon completion of the method, the five glottal parameters and pitch period are stored in a list object, as well as the GOI and GCI instances. In addition, the glottal flow waveform and its derivative can be saved as well. This allows the user easy access into the data, and allows him/her to review the measurement. In addition, multiple plots can be automatically generated throughout this analysis, allowing the user to examine the shape of the waveform (i.e. glottal flow with highlighted peaks, valleys, and GOI & GCI points).

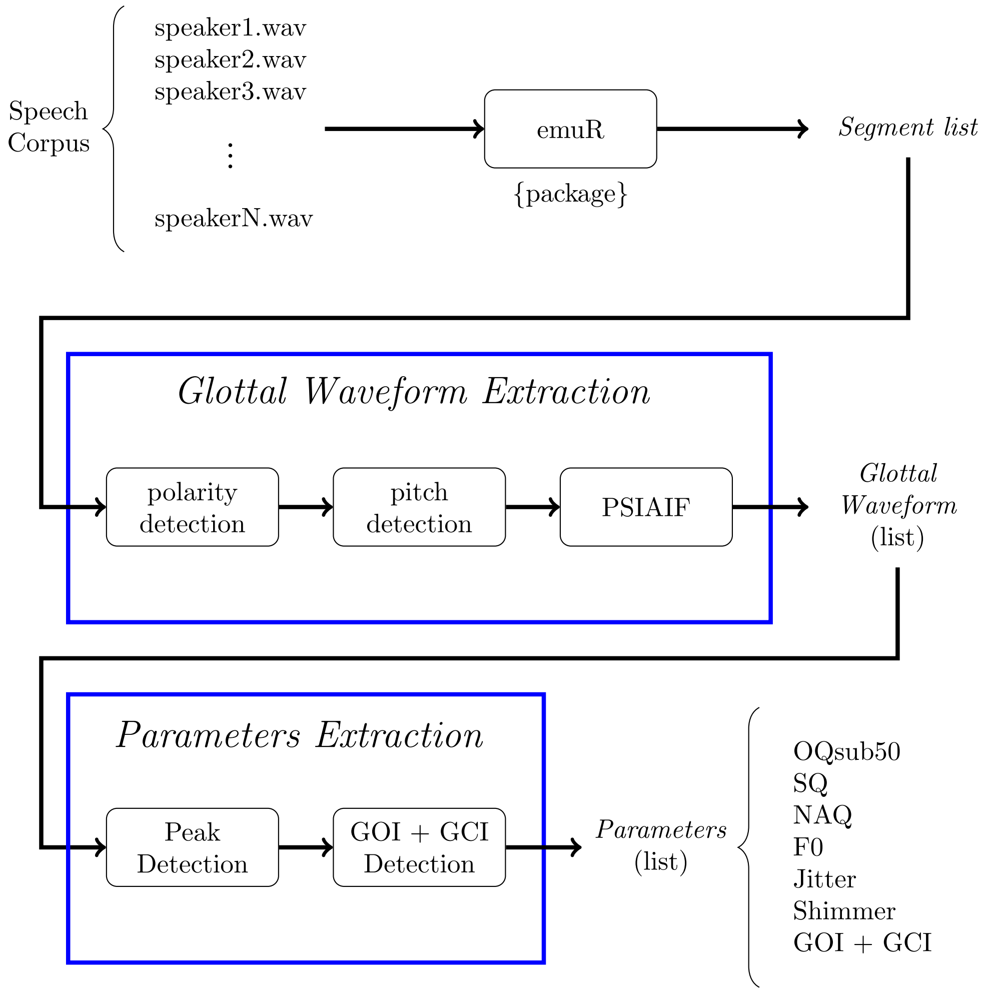

Figure 5.1: *R toolbox framework*

5.2.4 Algorithm Implementation

5.2.4.1 Fundamental Frequency Tracking

The pitch tracking algorithm developed by Drugman and Alway *et al* [97] computes the fundamental frequency for voiced speech using the residual signal ,$e[n]$. This algorithm exploits the harmonic structure of the residual signal's spectrum. The linear prediction residual signal is derived through inverse filtering, where the spectral contribution of the vocal tract is removed from the speech signal, $s[n]$, to obtain the residual signal. Following this whitening process, the spectral amplitude of the residual signal, $E(f)$, is computed using the Fast Fourier Transform (FFT). Drugman and Alwan remark that spectral peaks are present at the harmonics of the fundamental frequency for voiced speech segments. A Summation of Residual Harmonics method is employed to detect the fundamental frequency of the speech signal.

Consider a speech signal for the vowel **/a:/** uttered by a male speaker, as shown in Figure 5.2. The linear prediction residual signal is extracted via inverse filtering, with the resultant residual signal shown in Figure 5.2. The linear prediction residual signal was computed using the *lpcresidual* method as part of the Matlab® GLOAT toolbox. It was adapted to the *R* language under a free software license. The Matlab script computes the residual signal via a while loop. With each iteration, the input speech signal is segmented and windowed using a Hanning window. The windowed signal is used in an LPC method to compute the vocal tract filter coefficients. The vocal tract is removed from the speech signal to result in a residual signal corresponding to the segmented speech. The complete residual signal is computed using the overlap and add technique. A direct implementation of the code results in an average elapsed CPU time (calculated using system.time method call) of 0.07 seconds for a speech signal of average length of 3000 sample points. The code was modified as follows. The Matlab script required the input speech signal to conform into a vector. In *R*, the segmented speech signal, windowed signal, LPC filter coefficients, and the residual signal were all stored as data-frame objects. Instead of using a while loop, the start and end times of each speech frame were defined using the sequence *seq* method in the {base} package. Functions were applied using the *sapply* functional, which is preferable for data-frames and provides greater clarity of operations for the

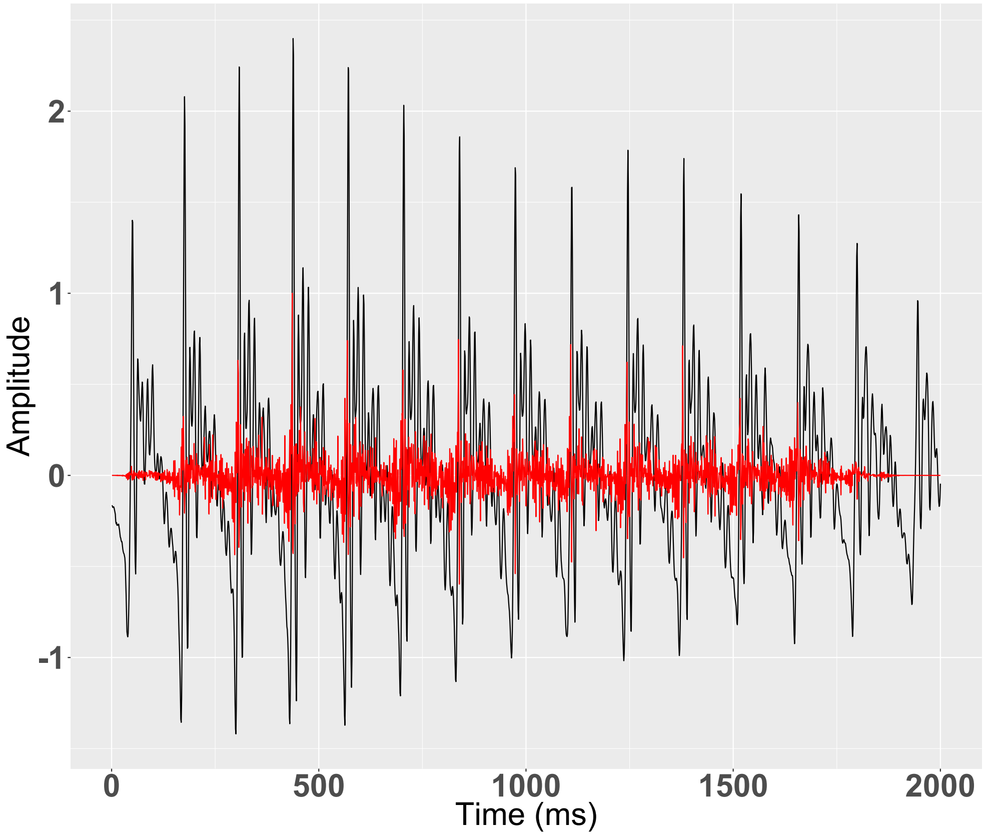

Figure 5.2: *Speech segment for vowel* ***a:*** *uttered by male speaker: speech signal (black), residual signal (red)*

user. The extraction of the vocal tract filter coefficients was carried out using the *lpc* method in the {phonTools} package [219]. This LPC method was modified to exclude pre-emphasis and normalization of the speech data. This modified implementation resulted in an average elapsed CPU time of 0.03 seconds for a speech signal of average length of 3000 sample points.

Prior to the summation of the residual harmonics, the residual signal is segmented and normalized. Analysis is performed on frames of the residual signal, with the length of each frame corresponding to the sampling frequency of the speech signal, $s[n]$. Each frame is Blackman-windowed and normalized by subtracting the mean amplitude-value of each frame from its samples (see Equation 5.1 on next page).

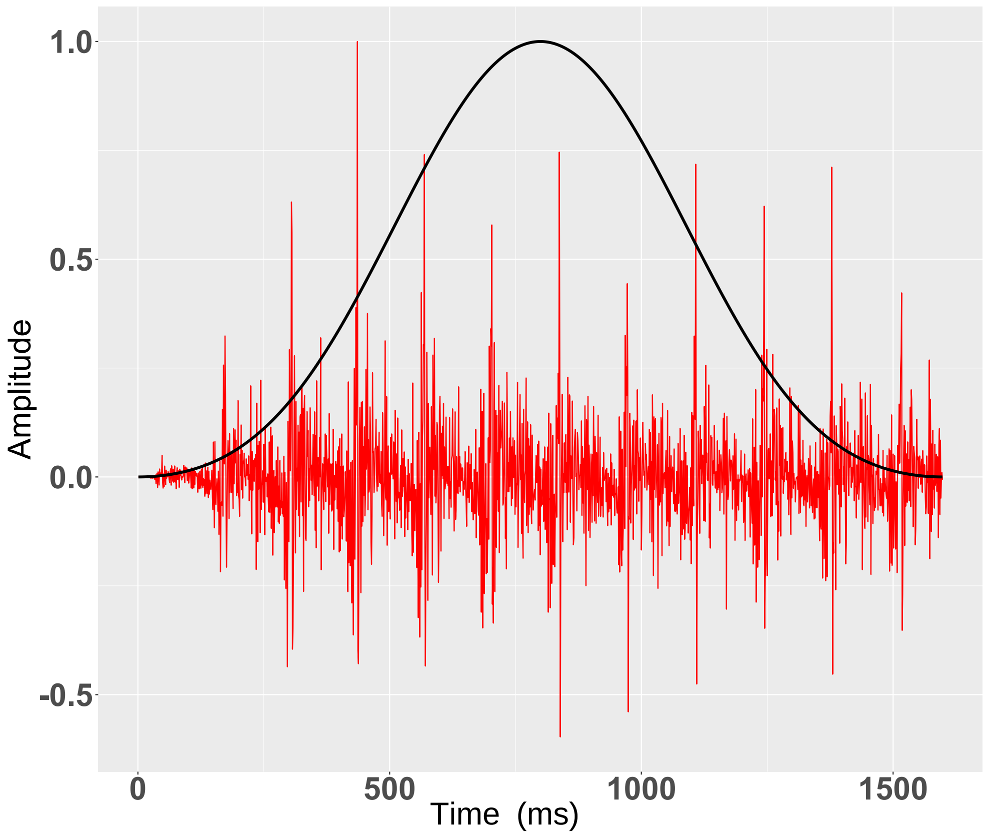

Figure 5.3: *Blackman-windowed residual signal: Blackman window (black), residual signal frame (red)*

$$Normalized\ (a_i) = a_i - \frac{1}{n}\sum_{j=1}^{n} a_j \tag{5.1}$$

Where a_i is a residual data sample at index i of a residual frame. A residual frame and its applied Blackman window are shown in Figure 5.3.

In order to perform analysis on the residual harmonics, the residual signal is transformed to the frequency domain. The amplitude spectrum, $E(f)$, is computed using a Fast Fourier Transfer function call. R provides a FFT function as part of its {stats} package. However, it does not allow for zero-padding. Thus, a $FFTN$ function was created, where N represents the length of the FFT output

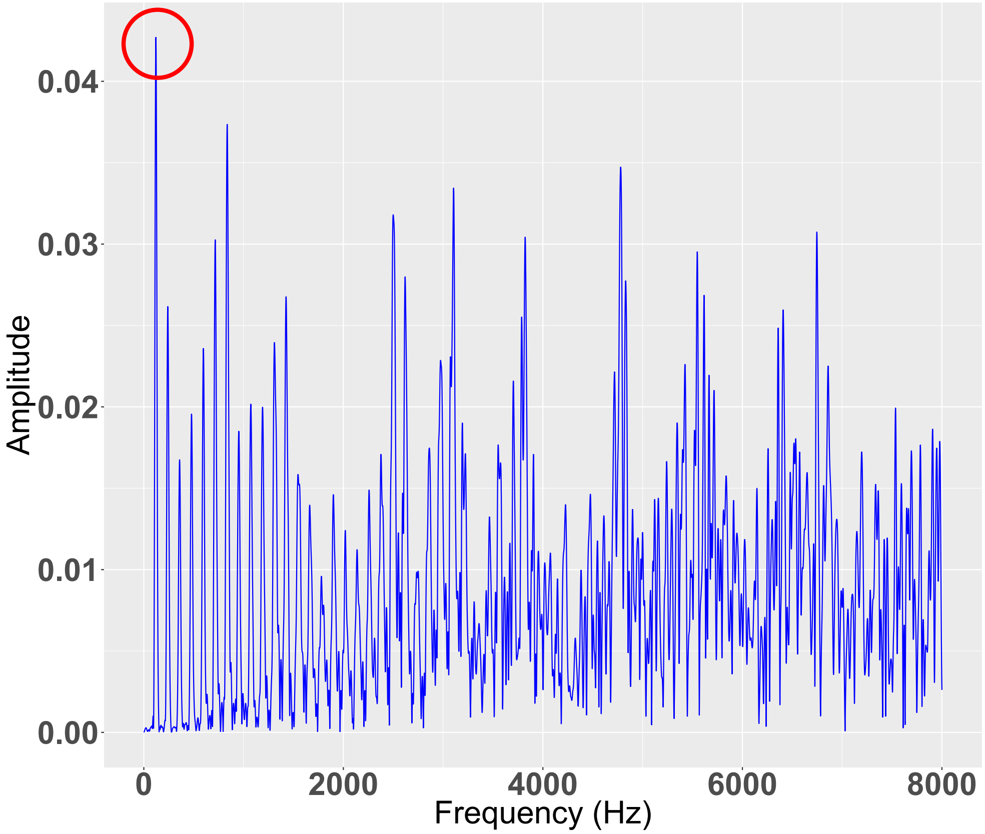

Figure 5.4: *The amplitude spectrum,* $E(f)$

signal. The symmetry property of the Fast Fourier Transform is exploited to remove redundant data. Considering a purely real input data for a Discrete Fourier Transform, there exists a symmetry about the $f_s/2$ sample point (corresponding to the Nyquist frequency). As a result, the frequency samples from $f_s/2$ to f_s are a mirror image of the first $f_s/2$ samples. The truncated spectral signal is shown in Figure 5.4. Thus, only half the samples are used for the Summation of Residual Harmonics (SRH) analysis. The spectral amplitude data is normalized through division by the mean frequency for the segment.

The Summation of Residual Harmonics is implemented as per Equation 5.2. SRH is computed for the frequency range $[F0_{min}, F0_{max}]$, where the variables $F0_{min}$ and $F0_{max}$ are inputs to the function call and correspond to the minimum and maximum possible values of the fundamental frequency, respectively.

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} \Big[E(k \cdot f) - E\big((k - \frac{1}{2}) \cdot f\big)\Big] \tag{5.2}$$

N_{harm} is the number of harmonics ($N_{harm} = 5$), and the pitch is defined as the frequency value to maximize $SRH(f)$. Upon first examination of this function, it seems logical to assume that the function could be simplified for the following relationship:

$$SRH(f) = \sum_{k=1}^{N_{harm}} E(k \cdot f)$$

Where SRH takes into account only the spectral contribution of the first harmonics. Thus, SRH is maximized for $f = F0$ for the first harmonics. However, Drugman and Alwan note that while this assumption might seem valid, SRH will be maximized for any other harmonics which are present in the frequency range $[F0_{min}, F0_{max}]$; e.g. in a speech segment with$F0 = 100$ Hz, for a frequency range $[F0_{min} = 50 \text{ Hz}, F0_{max} = 300 \text{ Hz}]$, SRH will be maximized for the first harmonic ($f = 100$ Hz which corresponds to $F0$), as well as for the second and third harmonics ($f = 200$ Hz and $f = 300$ Hz). Therefore, $E\big((k - \frac{1}{2}) \cdot f\big)$ is added to the function in order to lessen the influence of the even harmonics.

The $E(f)$ object consists of an array of N_{harm} rows and $F0_{max} - F0_{min} + 1$ columns. Each column corresponds to potential set of N residual harmonics in increasing order; e.g. $N_{harm} = 5$, $F0_{min} = 50$ and $F0_{max} = 300$ results in the $E(f)$ array shown below:

$$E(f) = \begin{bmatrix} 50 & 51 & 52 & \cdots & 300 \\ 100 & 102 & 104 & \cdots & 600 \\ 150 & 153 & 156 & \cdots & 900 \\ 200 & 204 & 208 & \cdots & 1200 \\ 250 & 255 & 260 & \cdots & 1500 \end{bmatrix}$$

In the SRH function, each column is used to index the data-frame of residual signal normalized frequency samples. This indexing operation results in N amplitude values, which are summed to a single value. This process is repeated throughout the length of the residual signal frame. As mentioned previously, the fundamental frequency is said to correspond to the peak amplitude in the spectral domain. Therefore, a function call is made to determine the maximum values and indices for each residual signal frame. The returned indices correspond to the fundamental frequency of the original speech signal, $s[n]$. The resultant pitch for the speech waveform shown in the figures above was 120 Hz, which coincides with the occurrence of the first peak of the amplitude spectrum in Figure 5.4 (marked with red circle). The SRH method is employed twice over: first, the residual harmonics are summed over the user-defined range of $[F0_{min}, F0_{max}]$ to compute the mean fundamental frequency. In the second iteration, the summation of residual harmonics is applied over the range of $[0.5 \cdot F0_{mean}, 2 \cdot F0_{mean}]$. The second iteration is applied in order to refine the pitch tracking, where it is assumed that the speaker would not exceed the new frequency range.

The SRH function was identified as the most computationally exhaustive. A direct implementation of the SRH function results in an average elapsed CPU time of 0.68 seconds for a speech signal of average length of 3000 sample points. Outliers of elapsed CPU time were recorded as high as 0.78 seconds. The main reason for the high elapsed time is the indexing of a matrix with another matrix. In addition, data was forced into a specific matrix data structure in order to allow for transposing of the data. While computationally efficient in Matlab, the operation is not recommended in *R*, as evident by the elapsed time results. Instead, data-frame and array objects were used. Using the *lapply* functional, the CPU time was reduced to 0.12 seconds for a speech signal of average length of 3000 sample points. Speech signals of various lengths were tested, yet none produced elapsed time greater than 0.2 seconds. This makes this new implementation 6 to 7 times faster than the direct implementation. We note that the CPU time will increase by changing the magnitude of the input parameters, i.e. increasing the order of linear prediction.

5.2.4.2 OQsub50

The OQsub50 is a new iteration of the open quotient parameter, as presented in Chapter 3. What follows is a description of the implementation method for the proposed criteria, including the GOI and GCI instances detection. The same GOI and GCI detection methodology was employed in order to allow computation of the SQ quotient and the NAQ quotient.

As previously discussed, the glottal opening instance is defined with accordance to the existing OQ50 criteria. The GOI instance is said to occur at a time-instant for which the glottal flow exceeds a arbitrarily-set peak-to-peak amplitude threshold. For the purpose of this discussion, we assume the amplitude threshold to be set at 50% of the peak-to-peak amplitude. This threshold is computed pitch-synchronously, on a period-to-period basis. Thus, removing the effect of a DC offset, which would have been present had the threshold been computed as an average across the entire glottal waveform. In order to compute the GOI efficiently, we refer back to our knowledge of the glottal phases. In particular, consider the open phase. It is known that the GOI instance will occur between the instance of minimum glottal flow and the subsequent instance of maximum glottal flow. Hence, the glottal waveform is segmented to about half of its duration. The GOI detection is a two-fold process. First, the GOI instance is roughly estimated. Given an amplitude threshold value, the rough GOI instance corresponds to the first glottal waveform sample which exceeds this threshold. A visual illustration of this analysis step is shown in Figure 5.5. This rough GOI instance point is the input for the second stage in the analysis. The glottal waveform is further segmented to 13 sample points, which range about the rough GOI instance (six points preceding and six points following). An 8th order polynomial linear model is then fitted to the segmented waveform and its coefficients are determined. This polynomial fitting is shown in Figure 5.6. Using simple algebraic manipulation, it is possible to compute a mathematically-robust estimation of the GOI instance. Consider a simple 2$^{\text{nd}}$ order polynomial curve with coefficients $\{a,\ b,\ c\}$, as shown in Equation 5.3.

$$ax^2 + bx + c = 0 \tag{5.3}$$

Since the threshold refers to an amplitude value, it corresponds to a y-intercept. The interception between the polynomial curve and the amplitude threshold can be expressed as follows:

$$ax^2 + bx + c = y \tag{5.4}$$

Rearrange Equation 5.4 by grouping the y-intercepts:

$$ax^2 + bx + (c - y) = 0 \tag{5.5}$$

The solution for Equation 5.5 results in root values corresponding to the x-axis (time). By finding the intersection between the polynomial curve and the amplitude threshold, the GOI temporal point is determined (see Figure 5.6).

The glottal closing instance is determined through the derivative glottal flow waveform, as it corresponds to its negative peak (minimum flow). Since the glottal closing instance occurs between the instances of maximum and minimum glottal flow, the glottal waveform is segmented to about half of its duration. We note that the sampling frequency plays an important role in the approximation of the derivative glottal flow waveform. With low sampling frequency (i.e. 16 kHz), when differentiated to produce the glottal flow derivative, the resultant wave was a crude approximation of the theoretical derivative. Thus, up-sampling interpolation was used to generate a smoother glottal flow waveform, which in turn resulted in a coherent glottal flow derivative approximation. Then, using the aforementioned pitch detection algorithm (Chapter 3), the instance of maximum negative flow is determined through quadratic interpolation. This instance corresponds to the GCI temporal point. The GCI detection process is shown in Figure 5.7.

We note that this new OQsub50 parameter was included in a study by Ben-Dom & Watson *et al* [83], which was published during this thesis work. It was subsequently used in a publication by Tian & Watson *et al* [177]. For the implementation of the algorithm in R please refer to Appendix C.

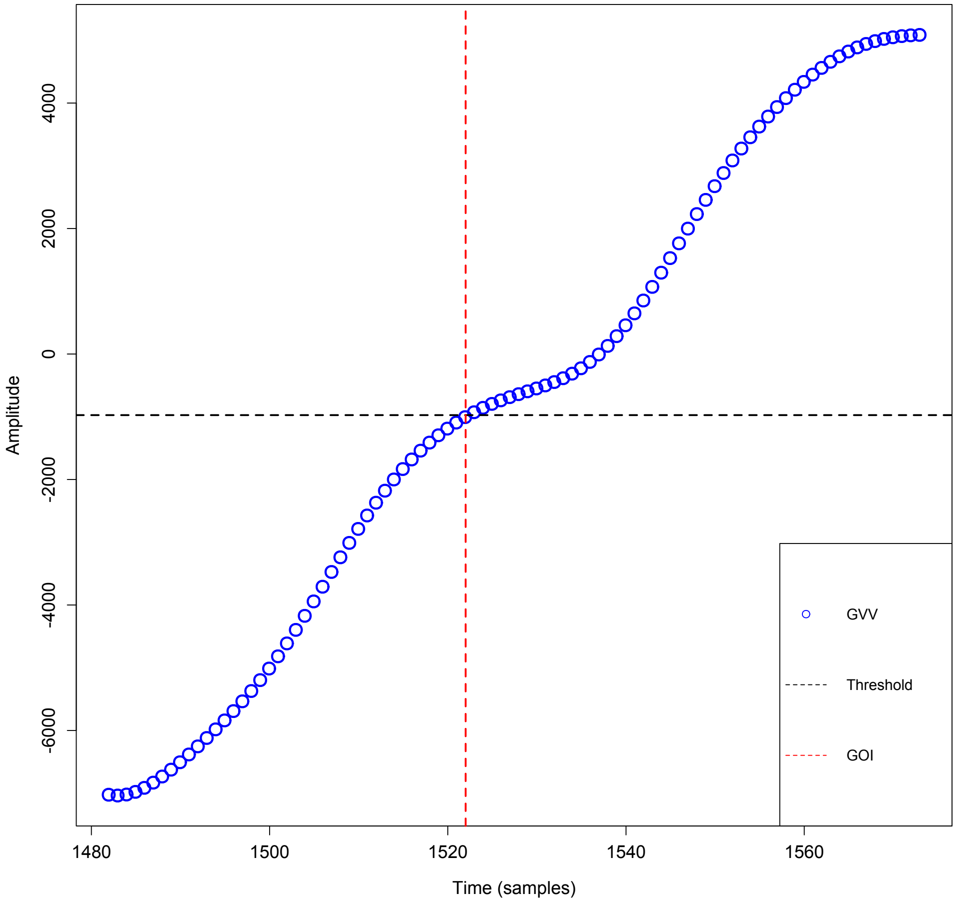

Figure 5.5: *Rough glottal opening instance detection to determine the range of the polynomial fitted curve*

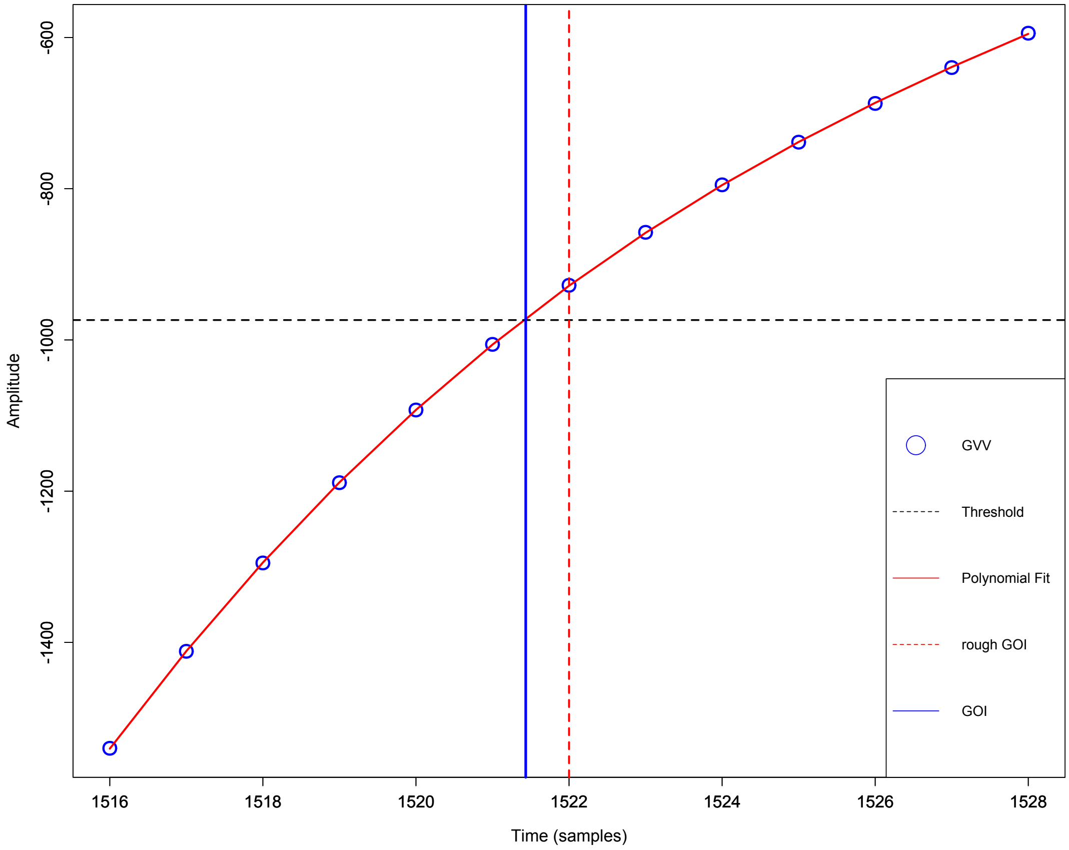

Figure 5.6: *Glottal opening instance detection from polynomial fitted curve*

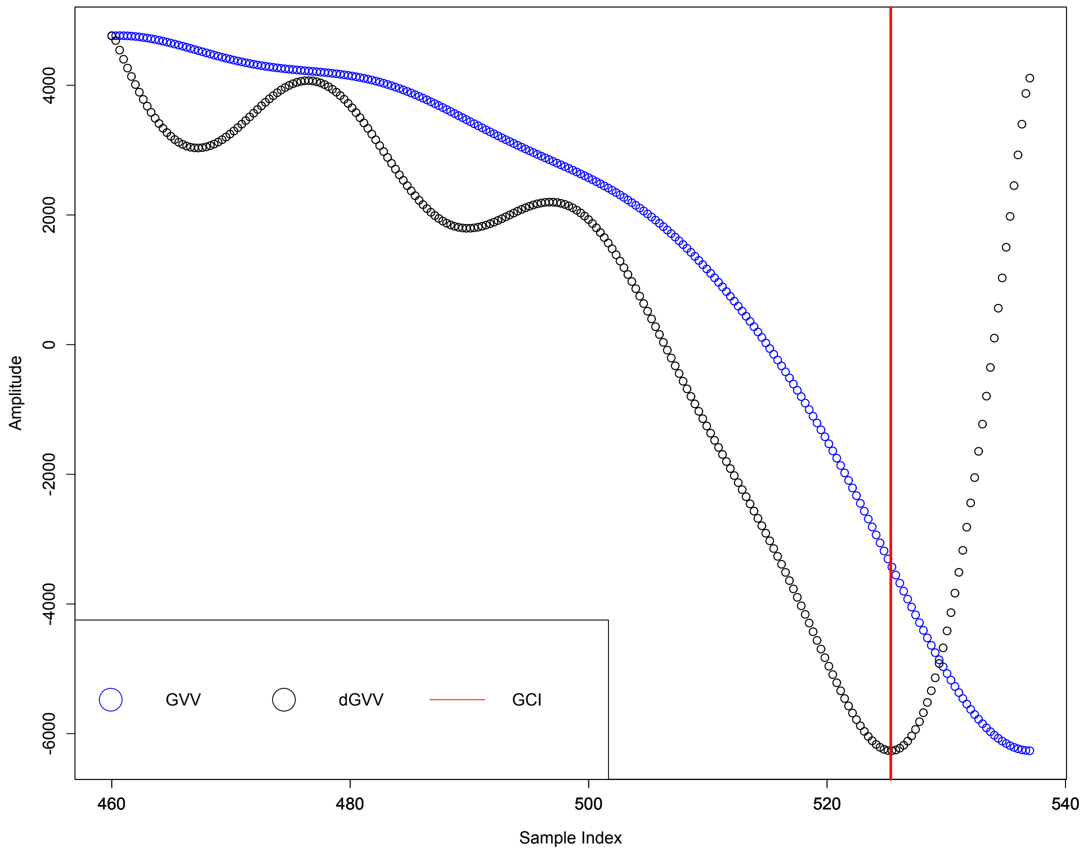

Figure 5.7: *Glottal closing instance detection from polynomial fitted curve*

5.3 Speech Corpora

This section introduces the four speech corpora investigated in this thesis work. Database A consists of New Zealand English (NZE) citation-form hVd words collected from two age groups of male speakers. Database B consisted of NZE citation-form hVd words collected from two age groups of male and female speakers. Database C consists of /a:/ (HARD) vowel tokens extracted from continuous speech in Māori, collected from archival recordings and modern-day interviews, for three age groups of male speakers. Database D is a longitudinal study for a single NZE male speaker, comprised of continuous speech archival recordings collected 40 years apart. Since different phonetic labelling sets were used, mapping of the vowels to key hVd words is provided for each database.

5.3.1 Database A: Male Ageing

This database consists of 30 male New Zealand English (NZE) native speakers [146, 147, 220]. The participants in this study were evenly divided into two age groups. The first group, referred to as the *young* group, consists of 15 speakers, 20-26 years old, with an age mean and standard deviation of 23.3 and 1.8, respectively. The second group, referred to as the *old* group, consists of 15 speakers, 56-71 years old, with an age mean and standard deviation of 62.5 and 5.6, respectively. Although the group of "old" speakers can not be classified as "elder" (typically over 80 years old), there exists a 30 year gap between the two age groups in this study, which was deemed significant enough to result in voice quality differences. The data gathering procedure was pre-empted by a questionnaire. The questionnaire was designed to obtain information about physical and habitual conditions that may affect the participants vocal health. It was gathered that none of the participants was an active smoker, with one speaker from the **young** group noting a history of smoking within five years prior to the date of data acquisition. Hearing loss was reported only for four of the speakers in the **old** group (three out of the four reported hearing loss for one ear). It is noted that "*hearing loss did not appear to be a confounding factor in the recordings*" [146]. No respiratory condition was reported for any of the speakers. Out of the 30 speakers, there is only one pair of young-old speakers that share a family relation; a father-son speaker pair.

The data acquisition procedure took place in a sound-isolation whisper room (`www.whisperroom.com`). Data was obtained using a laryngograph electroglottograph apparatus (`laryngograph.com`) with an attached lapel microphone. Two bilateral electrodes were placed at the laryngeal level and secured around the participant's neck using an elastic neck band. The lapel microphone was clipped to the electrode lead, approximately 15 cm from the participant's mouth [220]. Recordings were performed using the device's default sampling frequency of 16 kHz and were quantised to 16 bit signed number. During the recording procedure, the participants were asked to complete three tasks: read-out-loud the Rainbow Passage, produce two sustained vowels at three pitch and loudness levels, and citation form five hVd word lists of 11 NZE monophthongs (both short and long vowels) in randomized order. The recording procedure amassed to no more than 30 minutes of recordings per speaker, and a drink of water was made a available to the participants upon request. For the purposes of this study, the database comprises of a subset of the original database, with only speech data from the hVd word lists used for analysis. For this study, the speech corpus contains 1621 vowel tokens, with 803 and 813 tokens for **young** and **old** groups, respectively. This data was originally gathered for EGG analysis, which explains the use of a simple lapel microphone. Investigation of the database for the sustained vowel portion was carried out by Bier & Watson et al [146, 147, 220].

The recorded speech data had to be edited, labelled, and organized prior to its import into *emuR*. The original speech corpus contained a single, long speech file for each speaker. Using *wavesurfer* [221], each .wav file was split into five separate speech files, each file containing a sentence of 12 isolated hVd words. The 12th hVd word from each sentence was a repeat of a hVd word previously uttered in the sentence, and was removed to avoid list effects. The sentences were automatically labelled at word and phonetic level by the Bavarian Archive for Speech Signals' web services (BASWebServices) [222, 223] and exported in EMU Legacy format as a ZIP file. The label files were named after their corresponding speech data .wav files and were placed in the same root directory. The database was then converted to emuDB format. We note that the updated BASWebServices platform allows to choose emuDB as the preferred output format, thus negating the need to convert the database manually in *R*, which is highly preferable.

Note: The Speech Assessment Methods Phonetic Alphabet (SAMPA) phonetic labelling was used in this database. The vowels and their corresponding hVd *words are as follows*: /I/ "HID", /U/ "HOOD", /e/ "HEAD", /O/ "HOD", /6/ "HUD", /{/ "HAD", /i:/ "HEED", /}:/ "WHO'D", /o:/ "HOARD", /6:/ "HARD", /3:/ "HERD".

5.3.2 Database B: Acoustic Ageing

This speech corpus consists of 26 New Zealand English (NZE) native speakers [49, 224]. It comprises of both male and female speakers. Note that analysis was carried out separately for each gender group. The participants were divided into two age groups: **young** and **old**. The speakers in the **young** group were in the age range of 18 to 35 years old. The speakers in the **old** group were in the age range of 55 to 80 years old. There is an uneven distribution of the aged speakers between the two gender groups. For male speakers, the corpus comprises of seven young speakers and eight old speakers. For female speakers, the corpus comprises of size young speakers and five old speakers. The participants were not required to complete a questionnaire prior to the recording procedure. Thus, aside of their native language, there is no available information about the physical and/or psychological history of the participants.

The data acquisition procedure took place in a sound-isolation whisper room (MLD8484E) directly on to a Marantz PMD670 Solid State Recorder at a sampling rate of 16kHz using a Shure SM58 Microphone [49, 224]. The speech data was digitized and quantised to 16 bit signed number. The recorded data consists of each participants citation form speech of five hVd word lists for 11 NZE monophthongs in randomized order [224]. Overall, the speech corpus used in this study contains 1225 vowel tokens, with 695 and 530 tokens for male and female speakers, respectively. We note that while the female speakers subset included all 11 NZE monophthongs, this was not the case for the male speakers. As this corpus was originally composed for acoustic reflectometry analysis of the vocal tract, the monophthongs /V/ and /U/ were not recorded, as their vocal tract configuration mimicked that of other vowels. Hence, the male speakers subset includes only 9 of the 11 NZEmonophthongs. This limits the analysis, as noted by Watson *et al* [49].

The speech data was labelled at the phonetic level using EMU speech tools [225]. In addition, the first four formants were tracked in EMU. Similar to Database A, the recorded speech data had to be edited and organized prior to its import into *emuR*. The original speech corpus contained a single, long speech file for each speaker. Using *wavesurfer* [221], each .wav file was split into five separate speech files, each file containing a sentenced of 12 isolated hVd words. The 12$^{\text{th}}$ hVd word from each sentence was a repeat of another hVd word from the sentence. Hence, the last word in each list is a "dummy" word and was subsequently removed to avoid list effects. The label files were named after their corresponding speech data .wav files and were placed in the same root directory. The database was then converted to emuDB format, and a SSFF samples track was added to the emuDB database to allow for extraction of the speech samples. Note that two studies were performed with a partial portion of this database, to obtain vocal tract related measurements, by Watson *el at* [49, 224].

Note: The Machine Readable Phonetic Alphabet (MRPA) phonetic labelling was used in this database. The vowels and their corresponding hVd words are as follows: /I/ "HID", /U/ "HOOD", /E/ "HEAD", /O/ "HOD", /V/ "HUD", /A/ "HAD", /i:/ "HEED", /u:/ "WHO'D", /o:/ "HOARD", /a:/ "HARD", /@:/ "HERD".

5.3.3 Database C: Māori Ageing

This database consists of 30 Māori language speakers. The data used in this study is part of the Māori and New Zealand English (MAONZE) speech-corpus [218, 226, 227]. This database is a subset of the aforementioned corpus. The MAONZE project allows for examination of pronunciation changes at the segmental (vowels) level in Māori language over time. This corpus consists of three speakers groups: young male speakers, present-day male kaumātua (elders) speakers, and historical male kaumātua (elders) speakers. For the benefit of the reader, the three age groups are labelled as **Y**, **PE**, and **HE**, respectively. There is an even distribution of speakers per age group, where each group consists of 10 speakers.

The **Y** group participants were born between 1969 and 1984 and were between the ages of 18 and 30 at the time of the recording. The speech recordings were obtained between 2001 and 2006. Though recordings were made in Māori, all **Y** speakers spoke English at native competence level. Several speakers (exact number not specified) were classified as first language (L1) Māori speakers (raised in a Māori speaking environment). It was noted that most of the participants lived with a native Māori speaker (grandparent) throughout their childhood.

The **PE** group participants were born between 1925 and 1939 and were all over the age of 65 at the time of the recording. The speech recordings were obtained between 2001 and 2006. Though limited information is provided about the **PE** speakers, all participants were classified as L1 Māori speakers. It is noted that several of the participants did not read fluently in Māori. This was attributed to poor eyesight and lack of familiarity with reading in Māori.

The **HE** group participants were born between 1871 and 1885 and were between the ages of 62 and 77 at the time of the recording. The exact age of several of the participants is unknown, with the age of the other participants at the time of the recording obtained from Internet searches and from partial information collected prior to some of the interviews. Unlike the recordings for the **Y** and **PE** participants, which were collected by the founding members of the MAONZE project, the speech recordings for the **HE** group were sourced from radio and TV broadcasts by the New Zealand Broadcasting Service between the years 1946 and 1948. All participants were classified as L1 Māori speakers. The **HE** Māori pronunciation and grammar, unlike the **Y** and **PE** speakers, does not show influences from the English language.

From the information above, it is observed that the **PE** participants were born approximately half-a-century (approximately two generations) after the **Y** participants, and the **HE** participants were born approximately half a century after the **PE** participants. Thus, the participants birth dates span of 100 years. The time-span of this speech corpus reflects the differences in status, use, and exposure of the Māori language over the last century. Note that analysis of historical dialect differences was not possible due to tribal affiliation only being available for some of the **HE** participants.

The data acquisition procedure for the historical recordings consisted of formal interviews intended for radio broadcast. The recording equipment is considered limited by modern standards (equipment specifications not provided). The interviews were recorded onto a 14-inch acetate disks. Due to the limitations of the original recording equipment, the recordings were limited to a bandwidth below 5 kHz. The recordings were digitized by the MAONZE project, stored at 16 kHz as a non-compressed .wav format. Although this sampling rate is relatively low, it is sufficiently high to allow for spectral analysis of the vowels, as vowels can be adequately defined for a 0-3 kHz bandwidth [226, 228].

For the modern day recordings, the MOANZE project researchers mimicked the recording style and format of the historical recordings in order to allow for valid comparison across the age groups. For the **Y** group, the interviews were conducted as an informal conversation. For the **PE** group, the interview style was "casual", since the interviewers were familiar with the interviewees. These recordings took place in the interviewees' preferred choice of venue. Thus, recordings were not obtained in an acoustic environment and were not collected using studio-quality equipment. The recordings were made using a Sony TCD8 DAT recorder and down-sampled to 22.05 kHz [229]. Due to the choice of recorder, the modern recordings are band-limited to 10 kHz. Similar to the digitized historical recordings, the recordings were stored as non-compressed .wav format. As mentioned above, the participants were interviewed in Māori. Interviews lasted between 45 minutes and one hour, and consisted of informal conversation, citation of word lists, and passages reading. Note that for the purposes of this study, speech segments required a minimum duration of 100 ms. Analysis of the vowels duration showed most vowel segments only exhibit several cycles, resulting in short vowel duration between 20 and 50 ms. Only the vowel /a:/ (HARD) was observed to have speech segments exceeding 100 ms seconds in duration. Therefore, this is the only vowel considered in this speech corpus, for a total of 36 vowel tokens.

The recordings were transcribed using Transcriber (`www.trans.sourceforge.net/en/presentation.php`). Each phoneme was time aligned with a sound file. The transcription was then converted to *Pratt* textgrid file format for correction. Subsequen speakertly, the time alignment was modified and mistakes were identified and corrected [230]. This resulted in an EMU Legacy format-ready database. For the purposes of this study, it was updated to an emuDB format and a samples SSFF track was added. No additional preparation/correction was required.

5.3.4 Database D: One Man Ageing

This database consists of a single New Zealand English (NZE) native speaker. This is a subset of a speech corpus which was first introduced by Harrington, Palethorpe and Watson *et al* [231]. This speaker, identified as *Speaker A* in [231] and shall remain anonymous, is a native born Pakeha male speaker, whose profession was explorer and who was born between the years 1901-1916. This speech corpus was established to determine the vowel change in New Zealand English over a significant period of time. The data was compiled from broadcast and radio interviews made available by Radio New Zealand Sound Archives. Two data sets were used for this study: broadcast speech data from 1955, and radio interview data from 1992. Though the exact age of the speaker was not specified, it is assumed that the data was collected while the speaker was approximately 40 and 80 years old, respectively. This database consists of 75 recorded sentences in citation form. 27 sentences were obtained in 1955, with the rest collected in 1992. The recorded speech data obtained from the radio broadcasts was digitized to 16 bit signed number at a sampling rate of 20 kHz. The recorded data consists of continuous-form speech, from which the 11 NZE monophthongs were extracted for analysis. The database was segmented and labelled at the Word and Phonetic level using EMU. The authors *et al* [231] note that only vowels from prosodically accented words were labelled in this corpus. Overall, this speech database contains 510 vowel tokens, with 367 vowel tokens from the 1955 speech data and 143 tokens from the 1992 speech data. For the purposes of this study, it was converted from an EMU Legacy to an emuDB format.

Note: The Machine Readable Phonetic Alphabet (MRPA) phonetic labelling was used in this database. The vowels and their corresponding hVd words are as follows: /I/ "HID", /U/ "HOOD", /E/ "HEAD", /O/ "HOD", /V/ "HUD", /A/ "HAD", /i:/ "HEED", /u:/ "WHO'D", /o:/ "HOARD", /a:/ "HARD", /@:/ "HERD".

Chapter 6

Results

This chapter presents presents the results obtained from the analysis of the four speech corpora presented in the previous chapter. For each speech corpus, the mean and standard deviation (in parenthesis) were obtained for each one of the six glottal parameters: open quotient (OQsub50), speed quotient (SQ), normalized amplitude quotient (NAQ), jitter, shimmer, and pitch (F0). Statistical analysis was carried out in order to identify significant age, vowel, and age-vowel effects. For all databases, repeated measures of ANOVA were calculated with the glottal parameter as the *dependant* variable, vowels as the *within-subject* factor, and the age groups as the *between-subject* factors. For database B, male and female analysis was performed separately. For all databases, Mauchly's Sphericity test was performed to determine the sphericity of *within-subject* factors. When sphericity was violated, GreenhouseGeisser adjustments were made, with the adjusted p-values indicated as p(GGE). We note that this correction might result in parallel calculations on the same groups of speakers having different degrees of freedom [232]. In order to identify significant vowel and age-vowel interactions, *post hoc t-test* were used, with Bonferroni corrections applied for repeated tests. Please note that since different vowel symbol sets (MRPA, SAMPA) were used in this investigation, vowels will be indicated using hVd words for this section. For the mapping between the hVd words and the vowel symbols used in the various studies, refer to Table 2.1 and Table 2.2 in Chapter 2.

6.1 Database A

This speech corpus contains 1621 vowel tokens, extracted from citation-form hVd word lists. The distribution of the vowel tokens between the two speaker groups (**young** and **old**) is presented in Table 6.1. The mean and standard deviation per each glottal parameter is presented in Table 6.2. The ANOVA analysis for age interaction between the two speaker groups is also given in Table 6.2. It is observed that all glottal parameters showed significant age interaction. Box-plots for the distribution of the mean value for each glottal parameter per age group are given in Figure 6.1 to Figure 6.6. Note that the OQsub50 parameter was set to a 30% threshold, effectively making it into an OQsub30 parameter. This was done to allow for better comparison with the results obtained by Bier *et al* [220]. All parameters showed significant vowel effect via ANOVA analysis.

Significant vowel effect for the F0 parameter was indicated with ANOVA (F(10,280) = 16.11, p(GGE)<0.001). The *post hoc t-test* demonstrated multiple vowel interactions. as detailed in Table B.1. The F0 for HOOD is significantly greater than all the other vowels, WHO'D > HAD, HARD, HEAD, HEED, HOD, HOARD, and HEAD > HAD, HARD, HOD, HERD, and HERD > HAD, HARD, and HEED > HAD, this is at least for $p < 0.05$. Please see Table B.1 in Appendix B for the full t statistics. For box-plot of the mean F0 distribution between vowels, please refer to Figure 6.7.

Significant vowel effect for the OQsub30 parameter was indicated with ANOVA (F(10,280) = 14.93, p(GGE)<0.001). The *post hoc t-test* demonstrated multiple vowel interactions. The OQsub30 for HOOD is significantly greater HAD, and WHO'D > HAD, HARD, HOARD, HUD, and HOD > HAD, HUD, HARD, and HERD > HUD, HARD, and HEAD > HUD, HARD, and HID > HUD, HARD, and HOARD > HUD, this is at least for $p < 0.05$. Please see Table B.2 in Appendix B for the full t statistics. For box-plot of the mean OQsub30 distribution between vowels, please refer to Figure 6.8.

Significant vowel effect for the SQ parameter was indicated with ANOVA (F(10,280) = 1.88, $p < 0.05$). However, whilst the ANOVA suggested there was a significant interaction, *post hoc t-test* demonstrated that in fact this was not the case. For box-plot of the mean SQ distribution between vowels, please refer to Figure 6.9.

Significant vowel effect for the NAQ parameter was indicated with ANOVA (F(10,280) = 13.94, $p < 0.001$). The *post hoc t-test* demonstrated multiple vowel interactions. The NAQ for WHO'D is significantly greater than HAD, HERD, HUD, HARD, HID, HOD, HOOD, and HEED > HAD, HERD, HUD, HARD, HEAD, HID, HOD, HOARD, HOOD, and HEAD > HOD, and HOARD > HOD, this is at least for $p < 0.05$. Please see Table B.3 in Appendix B for the full t statistics. For box-plot of the mean NAQ distribution between vowels, please refer to Figure 6.10.

Significant vowel effect for the Jitter parameter was indicated with ANOVA (F(10,280) = 4.03, p<0.01). The *post hoc t-test* demonstrated multiple vowel interactions. The jitter for HUD is significantly greater than HERD, HOARD, and HOD > HERD, this is at least for $p < 0.05$. Please see Table B.4 in Appendix B for the full t statistics. For box-plot of the mean jitter distribution between vowels, please refer to Figure 6.11.

Significant vowel effect for the Shimmer parameter was indicated with ANOVA (F(10,280) = 5.85, p(GGE)<0.05). The *post hoc t-test* demonstrated multiple vowel interactions. The shimmer for HERD is significantly greater than HUD, HEAD, HID, HOD, HOARD, this is at least for $p < 0.05$. Please see Table B.5 in Appendix B for the full t statistics. For box-plot of the mean shimmer distribution between vowels, please refer to Figure 6.12.

ANOVA analysis yielded significant age-vowel effect for two glottal parameters.

Significant age-vowel effect for the OQsub30 parameter was indicated with ANOVA (F(10,280) = 3.11, p(GGE) < 0.001). The *post hoc t-test* demonstrated multiple vowel interactions, as detailed in Table 6.3. A box-plot of the mean jitter distribution for each vowel per age group is provided in Figure 6.14. Significant age-vowel effect for the Jitter parameter was indicated with ANOVA (F(10,280) = 4.11, p(GGE) < 0.01). The *post hoc t-test* demonstrated multiple vowel interactions, as detailed in Table 6.4. A box-plot of the mean jitter distribution for each vowel per age group is provided in Figure 6.13.

Note: The Speech Assessment Methods Phonetic Alphabet (SAMPA) phonetic labelling was used in this database. The vowels and their corresponding hVd words are as follows: /I/ "HID", /U/ "HOOD", /e/ "HEAD", /O/ "HOD", /6/ "HUD", /{/ "HAD", /i:/ "HEED", /}:/ "WHO'D", /o:/ "HOARD", /6:/ "HARD", /3:/ "HERD".

Table 6.1: *Phoneme distribution in Database A*

	short vowels						*long vowels*					*sum*
	I	U	e	O	6	{	i:	}:	o:	6:	3:	
Young	36	49	84	105	63	71	75	75	75	89	86	808
Old	39	49	90	105	52	74	74	75	75	90	90	813

Table 6.2: Database **A** Results

	Age		ANOVA		
	Young	Old	*F*-value	*p*-value	
F0	113.87 (18.88)	125.19 (19.41)	F(1,28) = 5.1	Old > Young	*
OQsub30	0.64 (0.09)	0.72 (0.07)	F(1,28) = 12.7	Old > Young	**
SQ	2.36 (0.76)	1.74 (0.51)	F(1,28) = 16.4	Young > Old	***
NAQ	0.18 (0.04)	0.21 (0.04)	F(1,28) = 10.4	Old > Young	**
Jitter	2.85 (2.09)	4.76 (1.88)	F(1,28) = 16.4	Old > Young	***
Shimmer	7.85 (2.88)	9.81 (2.91)	F(1,28) = 9.0	Old > Young	**

Note: *p<0.05 , **p<0.01 , ***p<0.001

Table 6.3: Database **A**: OQsub30 age:vowel interaction

		post-hoc analysis	
vowels	*age:vowel interaction*	t-test	p-value
{	Old > 6: Young	$t(27.2) = 4.2$	*
}:	Old > { Young	$t(21) = 5.38$	**
	Old > 6 Young	$t(21.4) = 5.83$	***
	Old > 6: Young	$t(23.1) = 6.38$	***
	Old > O Young	$t(23.5) = 5.0$	**
	Old > o: Young	$t(25.7) = 5.2$	**
	Old > U Young	$t(24) = 4.4$	*
	Young > { Young	$t(14) = 5.3$	*
	Young > 6 Young	$t(14) = 6.1$	**
	Young > 6: Young	$t(14) = 7.3$	***
	Young > o: Young	$t(14) = 5.8$	**
3:	Old > 6 Old	$t(14) = 5.1$	*
	Old > 6 Young	$t(25) = 4.6$	*
	Old > 6: Young	$t(27) = 5.0$	**
	Old > { Young	$t(24.9) = 4.2$	*
O	Old > 6 Old	$t(14) = 5.2$	*
	Old > { Young	$t(25.4) = 4.6$	*
	Old > 6 Young	$t(5.8) = 5.0$	**
	Old > 6: Young	$t(27.2) = 5.4$	**
	Old > O Young	$t(27.5) = 4.1$	*
	Old > o: Young	$t(28.0) = 4.2$	*
	Young > 6 Young	$t(14) = 4.7$	*
	Young > 6: Young	$t(14) = 4.7$	*
6:	Old > 6: Young	$t(23.7) = 4.3$	*
e	Old > { Young	$t(25.1) = 4.3$	*
	Old > 6 Young	$t(25.7) = 4.5$	*
	Old > 6: Young	$t(27) = 4.8$	**
	Young > 6: Young	$t(14) = 5.1$	*
I	Old > { Young	$t(21.7) = 4.6$	*

Database **A**: OQsub30 age:vowel interaction (Continued)

I	Old > 6 Young	$t(22.1) = 5.0$	**
	Old > 6: Young	$t(23.9) = 5.4$	**
	Old > o: Young	$t(26.0) = 4.1$	*
	Young > 6 Young	$t(14) = 4.9$	*
	Young > 6: Young	$t(14) = 5.6$	**
i:	Old > { Young	$t(20.8) = 5.1$	**
	Old > 6 Young	$t(21.1) = 5.5$	**
	Old > 6: Young	$t(22.9) = 6.1$	***
	Old > O Young	$t(23.4) = 4.7$	*
	Old > o: Young	$t(25.5) = 4.8$	**
	Young > { Young	$t(14) = 5.2$	*
	Young > 6 Young	$t(14) = 5.9$	**
	Young > 6: Young	$t(14) = 6.6$	**
	Young > O Young	$t(14) = 5.2$	*
	Young > o: Young	$t(14) = 5.7$	**
o:	Old > 6 Young	$t(24.8) = 4.5$	*
	Old > 6: Young	$t(26.5) = 4.8$	**
	Young > 6: Young	$t(14) = 4.6$	*
U	Old > 6: Young	$t(27.2) = 4.3$	*

Note: *p<0.05 , **p<0.01 , ***p<0.001

Table 6.4: Database **A**: Jitter age:vowel interaction

		post-hoc analysis	
vowels	*age:vowel interaction*	*t*-test	*p*-value
6	Old > }: Old	$t(14) = 5.0$	*
	Old > 3: Old	$t(14) = 5.3$	*
	Old > 6 Young	$t(27.7) = 4.1$	*
	Old > 6: Young	$t(27.9) = 4.8$	**
O	Old > }: Old	$t(14) = 6.3$	**
	Old > 3: Old	$t(14) = 5.5$	**
I	Old > I Young	$t(18.6) = 4.8$	*

Note: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

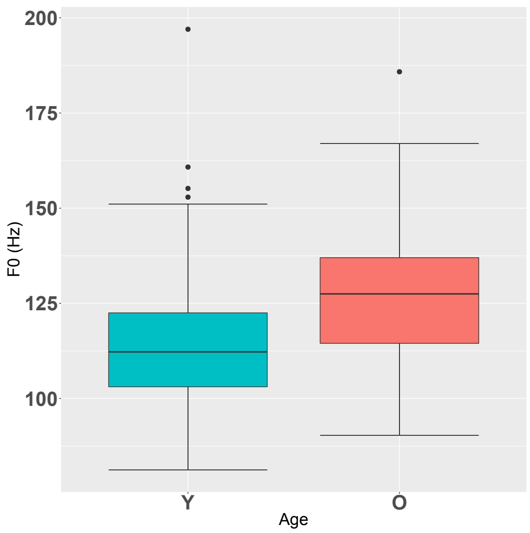

Figure 6.1: *Pitch boxplot*

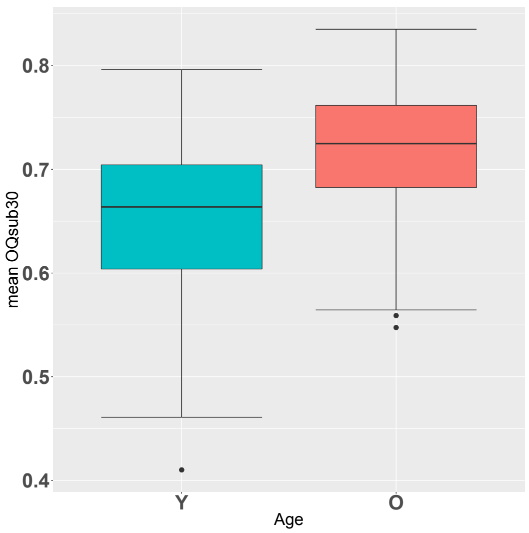

Figure 6.2: *OQsub30 boxplot*

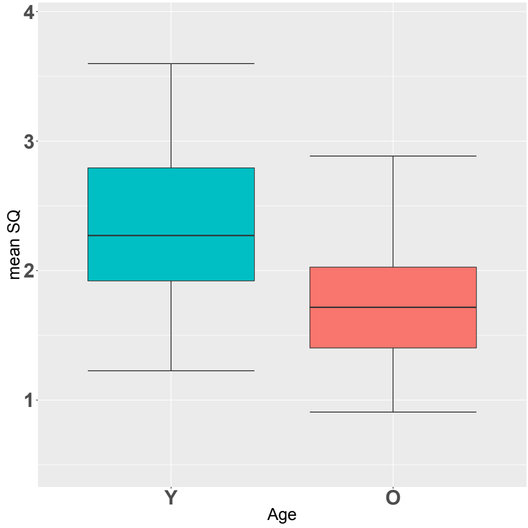

Figure 6.3: *SQ boxplot*

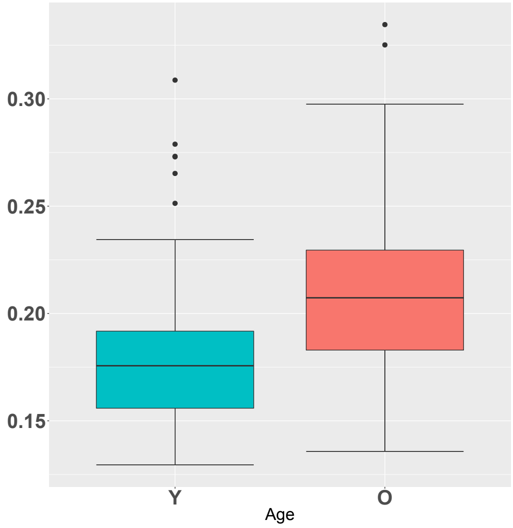

Figure 6.4: *NAQ boxplot*

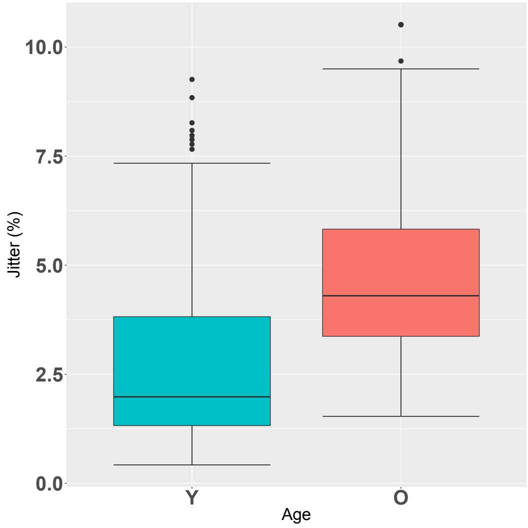

Figure 6.5: *Jitter boxplot*

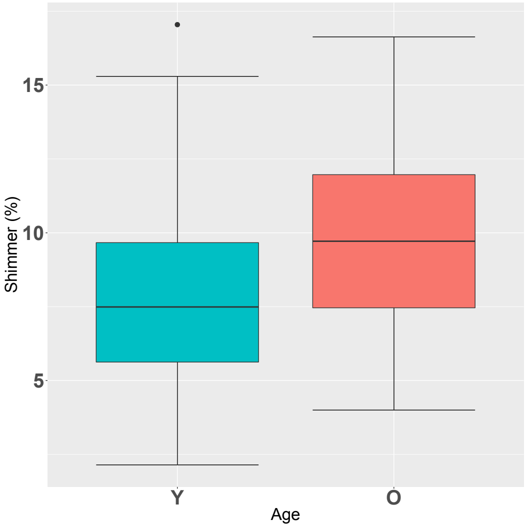

Figure 6.6: *Shimmer boxplot*

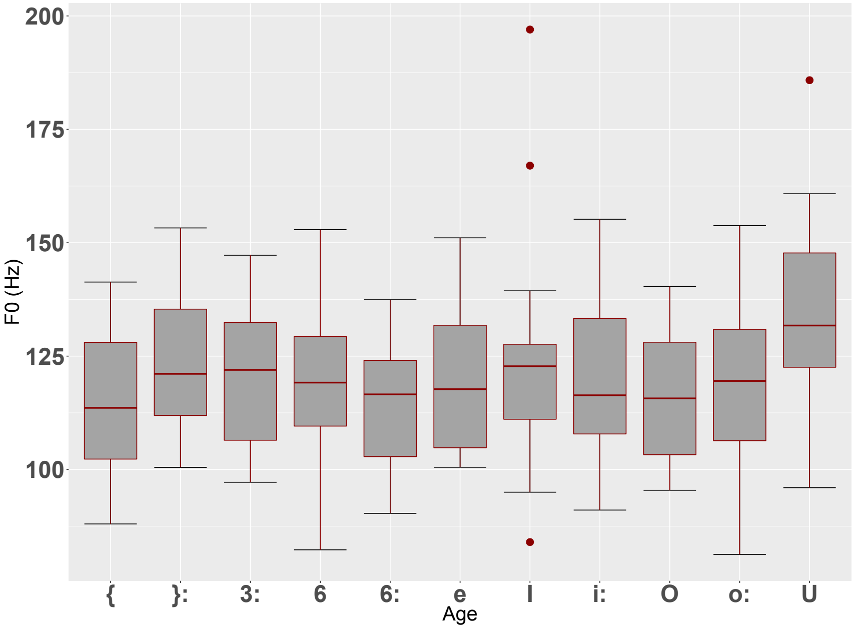

Figure 6.7: *F0 vowels interaction*

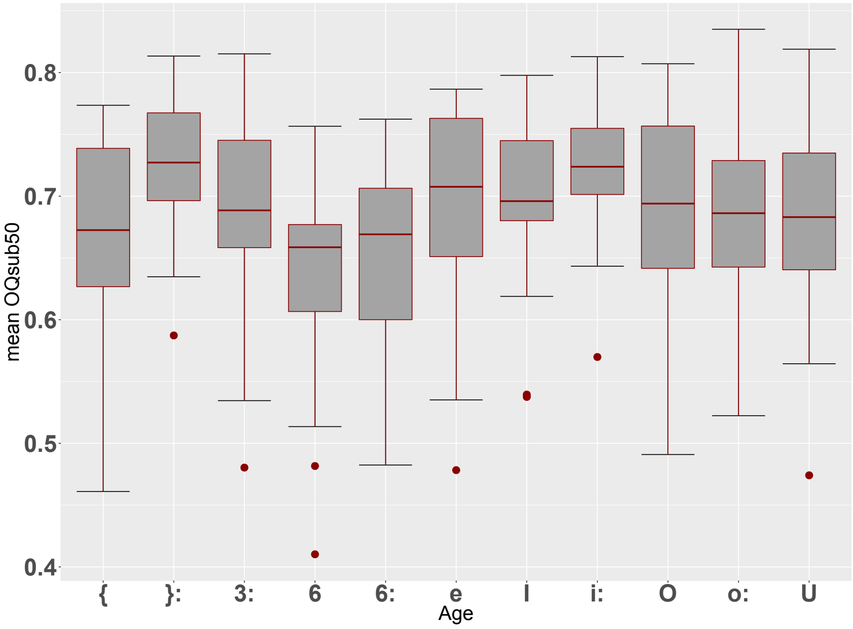

Figure 6.8: *OQsub30 vowels interaction*

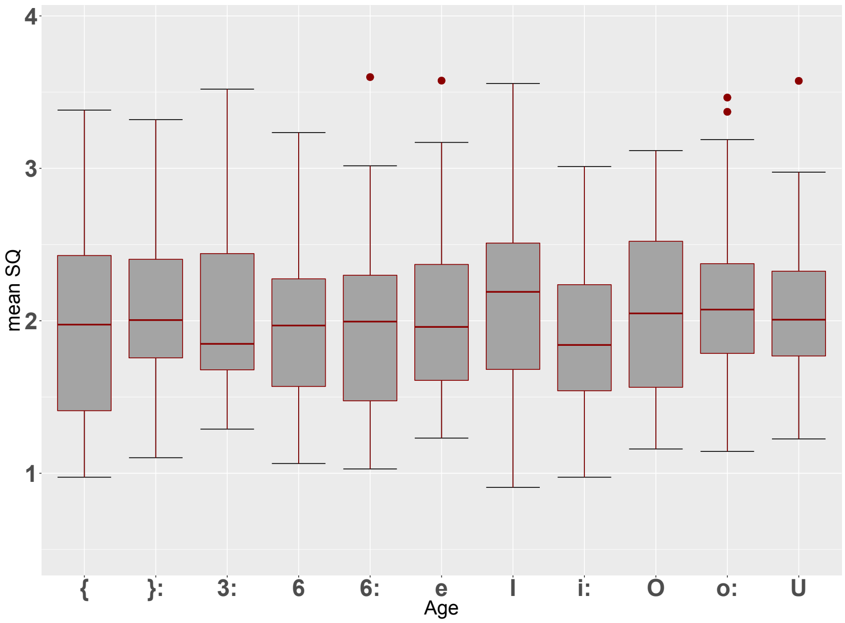

Figure 6.9: *SQ vowels interaction*

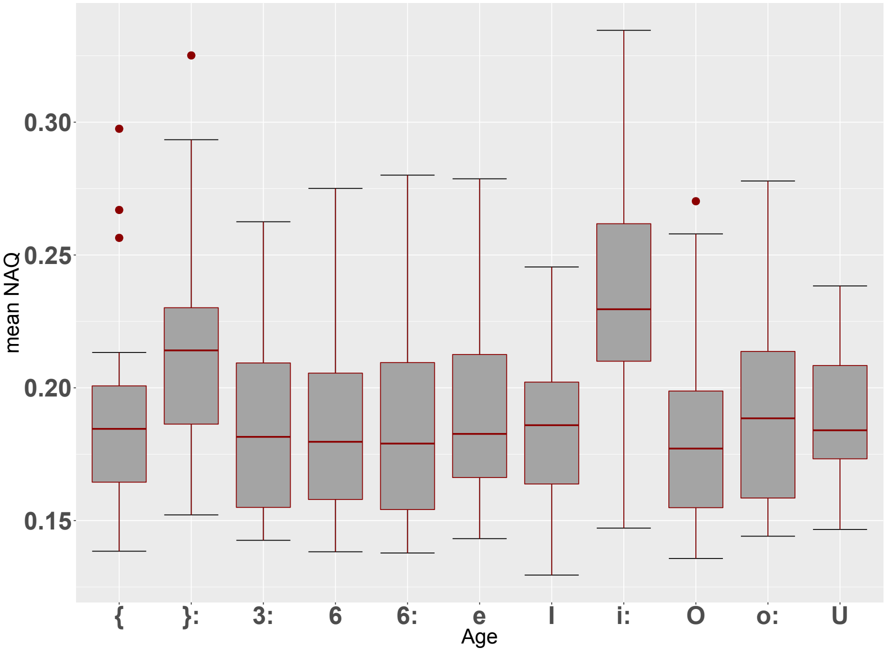

Figure 6.10: *NAQ vowels interaction*

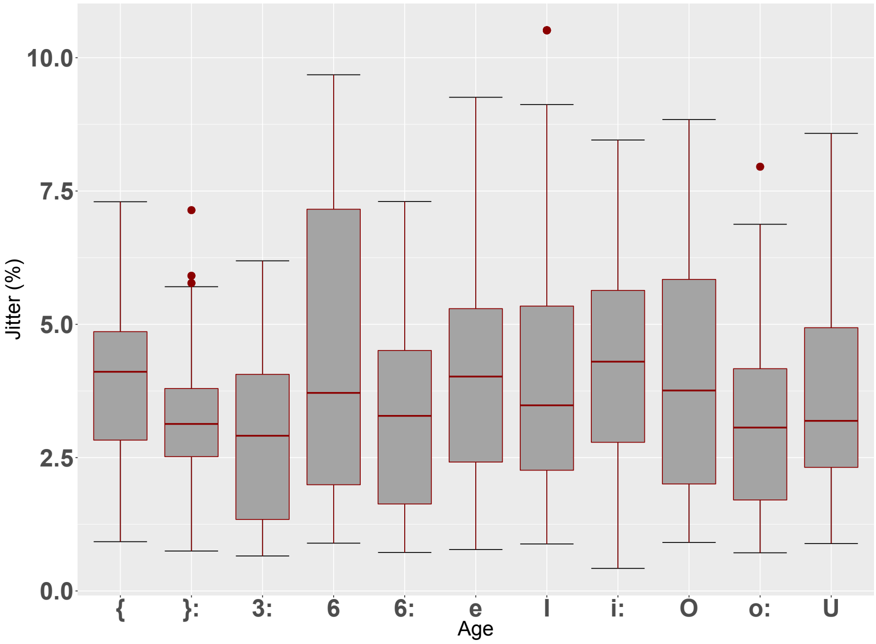

Figure 6.11: *Jitter vowels interaction*

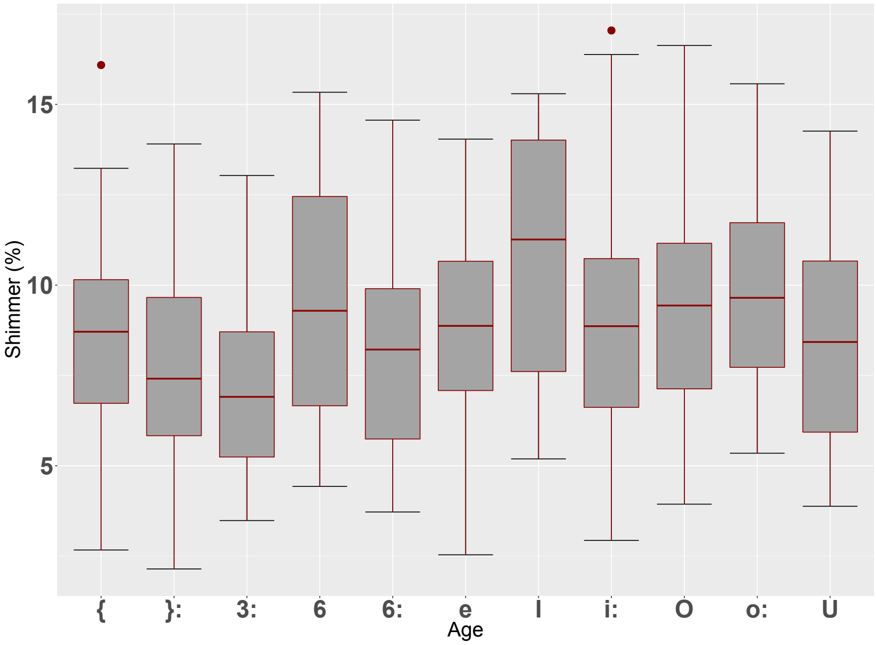

Figure 6.12: *Shimmer vowels interaction*

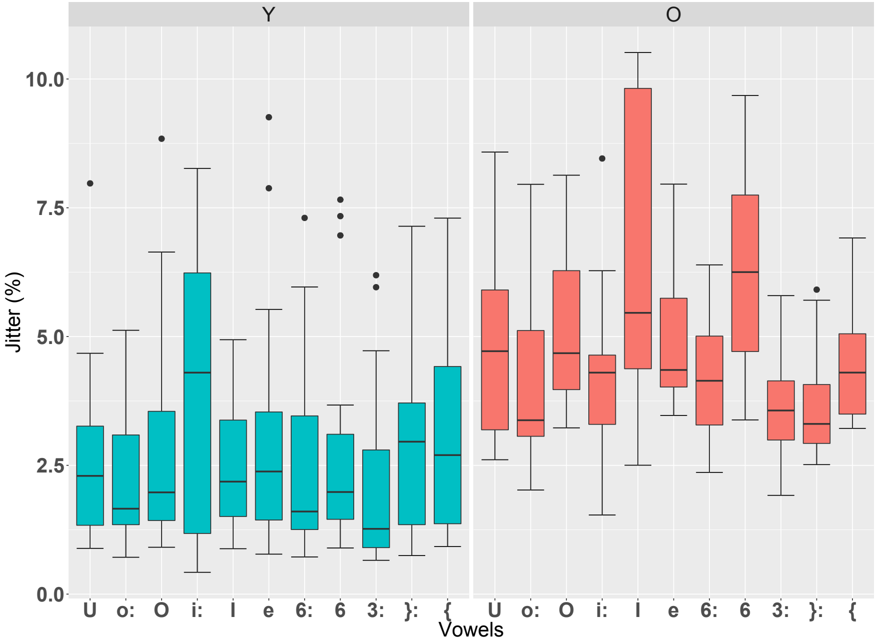

Figure 6.13: *Jitter age-vowels interaction*

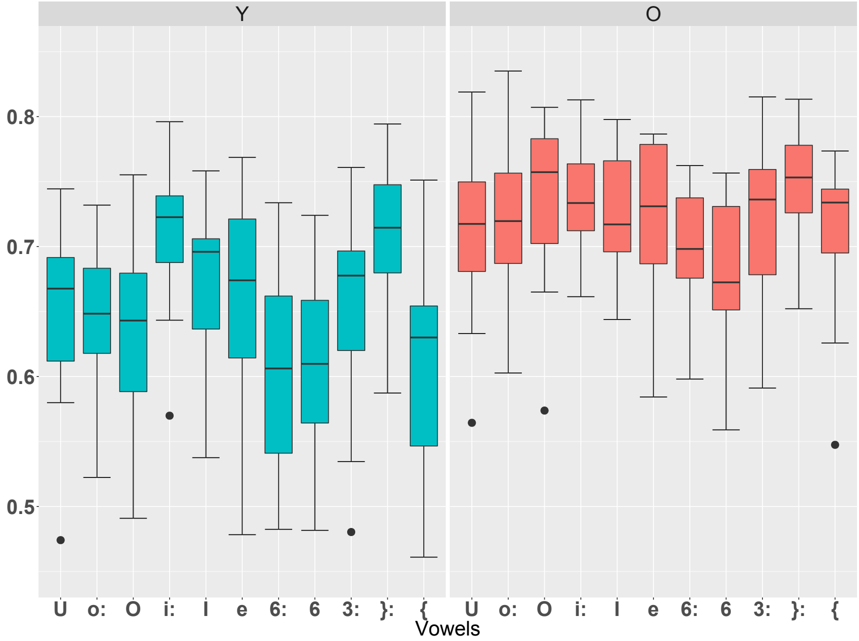

Figure 6.14: *OQsub30 age-vowels interaction*

6.2 Database B

Note: The Machine Readable Phonetic Alphabet (MRPA) phonetic labelling was used in this database. The vowels and their corresponding hVd words are as follows: /I/ “HID”, /U/ “HOOD”, /E/ “HEAD”, /O/ “HOD”, /V/ “HUD”, /A/ “HAD”, /i:/ “HEED”, /u:/ “WHO’D”, /o:/ “HOARD”, /a:/ “HARD”, /@:/ “HERD”.

6.2.1 Male Speakers

This subset of database B is comprised exclusively of male speakers. This subset contains 695 vowel tokens, extracted from citation-form hVd word lists. As noted in the preceding chapter, this data subset contains only 9 of the 11 New Zealand English monophthongs (vowels /U/ (HOOD) and /V/ (HUD) are omitted). The distribution of the vowel tokens between the two speaker groups (**young** and **old**) is presented in Table 6.5. The mean and standard deviation per each glottal parameter is presented in Table 6.6. Box-plots for the distribution of the mean value for each glottal parameter per age group are given in Figure 6.15 to Figure 6.20.

Significant age effect was observed for the OQsub50 parameter (F(1,13) = 6.06, p(GGE)<0.01). No other significant age interactions were found.

Significant vowel effect were observed for multiple parameters.

Significant vowel interaction for the F0 parameter was indicated with ANOVA (F(8,104) = 8.22, p(GGE) < 0.001). The *post hoc t-test* demonstrated multiple vowel interactions. The F0 for HAD is significantly greater than HEAD, HID, HOD, HERD, and HARD > HEAD, HOD, this is at least for $p < 0.05$. Please see Table B.6 in Appendix B for the full t statistics. For box-plot of the mean F0 distribution between vowels, please refer to Figure 6.21.

Significant vowel effect for the SQ parameter was indicated with ANOVA (F(8,104) = 3.23, p(GGE)<0.001). The *post hoc t-test* demonstrated the vowel HEAD to be significantly greater than HEED ($t(14) = 4.0$, $p < 0.05$). For box-plot of the mean SQ distribution between vowels, please refer to Figure 6.22.

Significant vowel interaction for the NAQ parameter was indicated with ANOVA (F(8,104) = 2.93, p(GGE) < 0.01). However, whilst the ANOVA suggested there was a significant interaction, *post hoc t-test* demonstrated that in fact this was not the case. For box-plot of the mean NAQ distribution between vowels, please refer to Figure 6.23.

Significant vowel interaction for the Jitter parameter was indicated with ANOVA (F(8,104) = 2.8, p(GGE)<0.01). The *post hoc t-test* demonstrated The vowel HID to be significantly greater than HERD ($t(14) = 4.19$, $p < 0.05$). For box-plot of the mean jitter distribution between vowels, please refer to Figure 6.24.

No significant age-vowel effects were observed in this subset.

Table 6.5: *Phoneme distribution in Database B - Male*

	short vowels						*long vowels*					*sum*
	I	U	E	O	V	A	i:	u:	o:	a:	@:	
Young	31	*NA*	35	37	*NA*	35	35	35	36	36	36	316
Old	31	*NA*	45	48	*NA*	41	41	41	43	45	44	379

Table 6.6: Database **B** (Male) Results

	Age			
	Young		Old	
	mean	*SD*	*mean*	*SD*
F0	116.90	15.21	117.80	21.82
OQsub50	0.68	0.05	0.73∗	0.06
SQ	1.30	0.45	1.70	0.76
NAQ	0.26	0.04	0.23	0.06
Jitter	1.25	0.71	1.71	0.63
Shimmer	9.69	3.33	14.15	7.2

Note: *p<0.05 , **p<0.01 , ***p<0.001

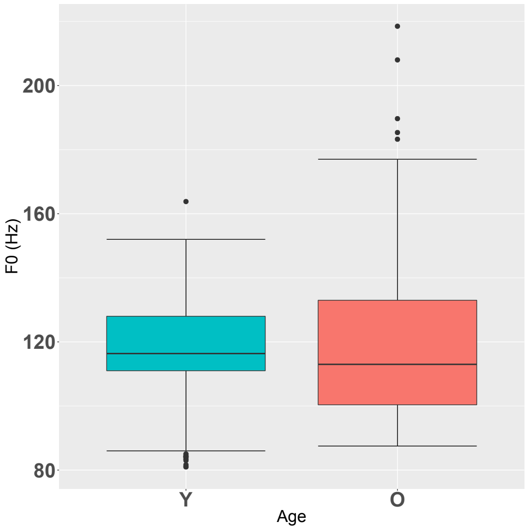

Figure 6.15: *Pitch boxplot*

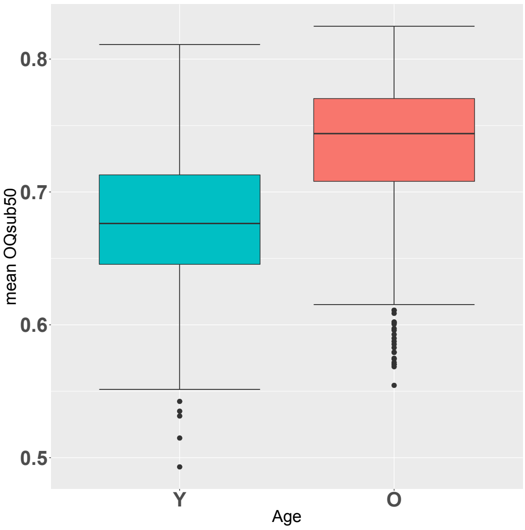

Figure 6.16: *OQsub50 boxplot*

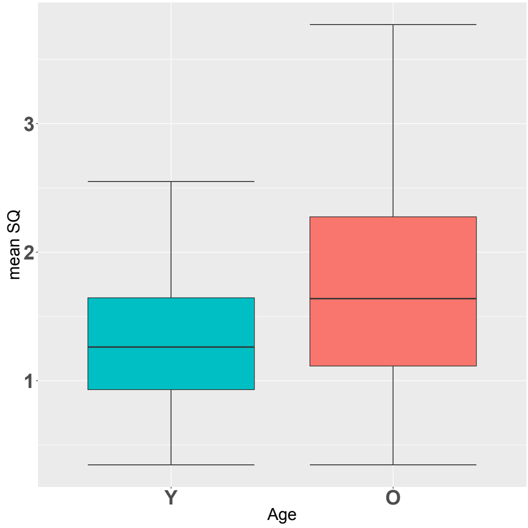

Figure 6.17: *Speed quotient boxplot*

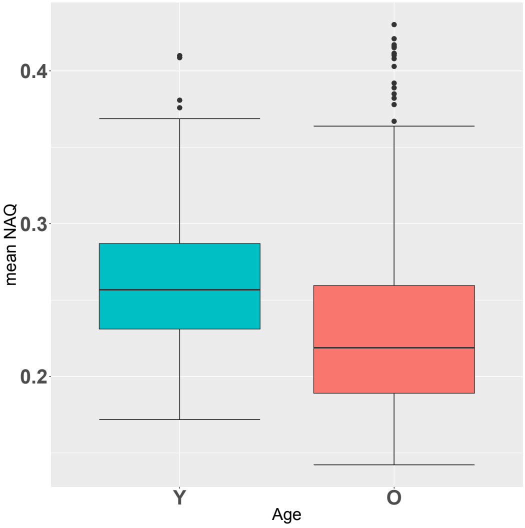

Figure 6.18: *NAQ boxplot*

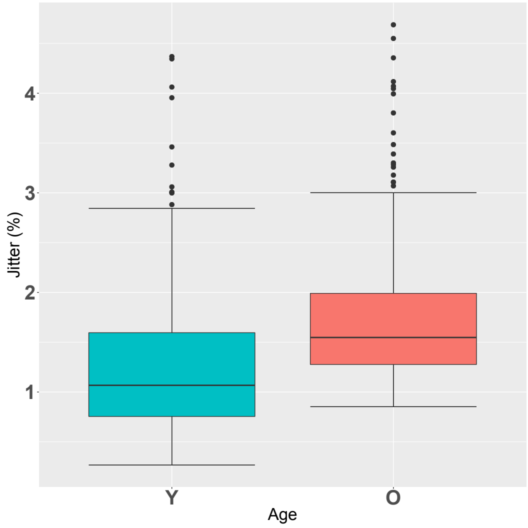

Figure 6.19: *Jitter boxplot*

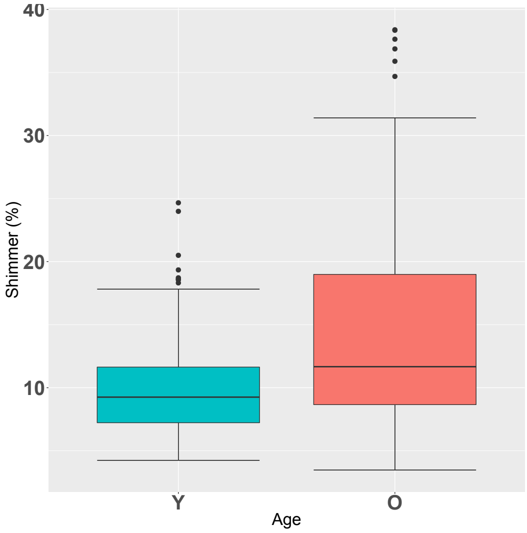

Figure 6.20: *Shimmer plot*

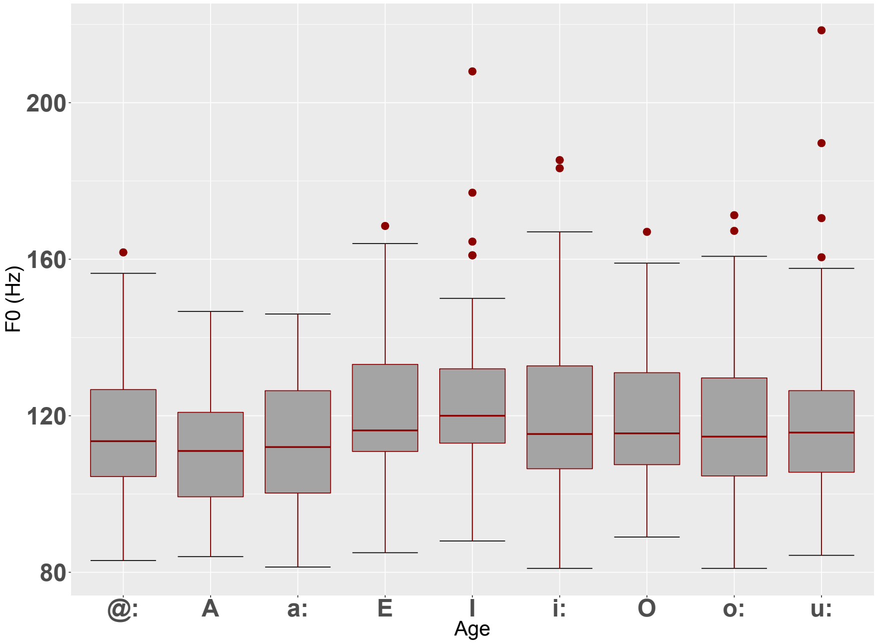

Figure 6.21: *F0 boxplots for vowels regardless of age*

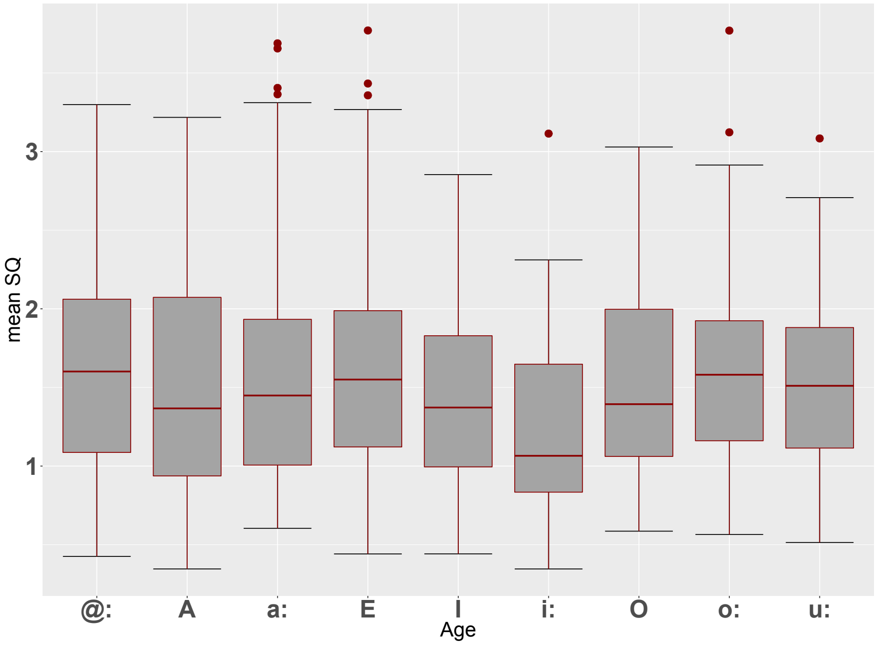

Figure 6.22: *SQ boxplots for vowels regardless of age*

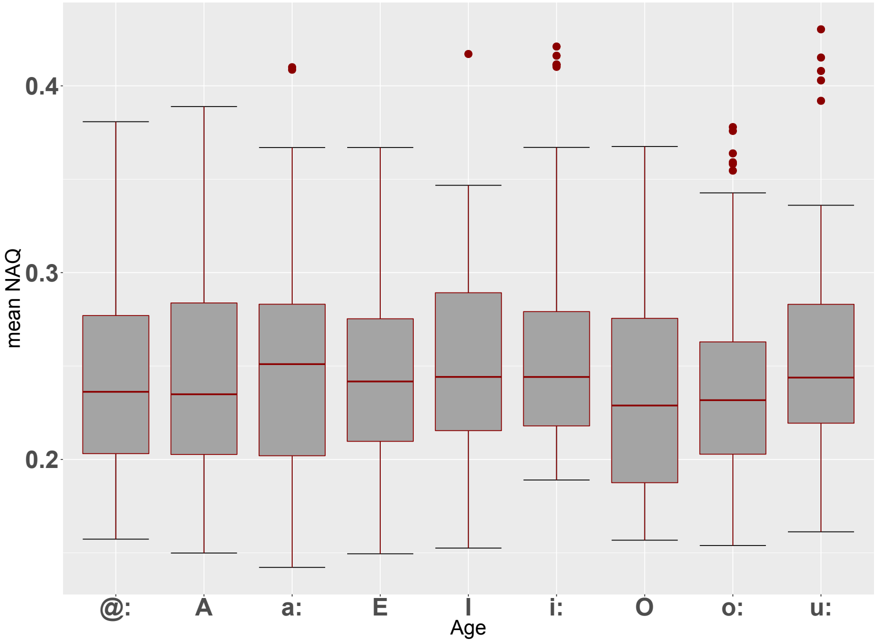

Figure 6.23: *NAQ boxplots for vowels regardless of age*

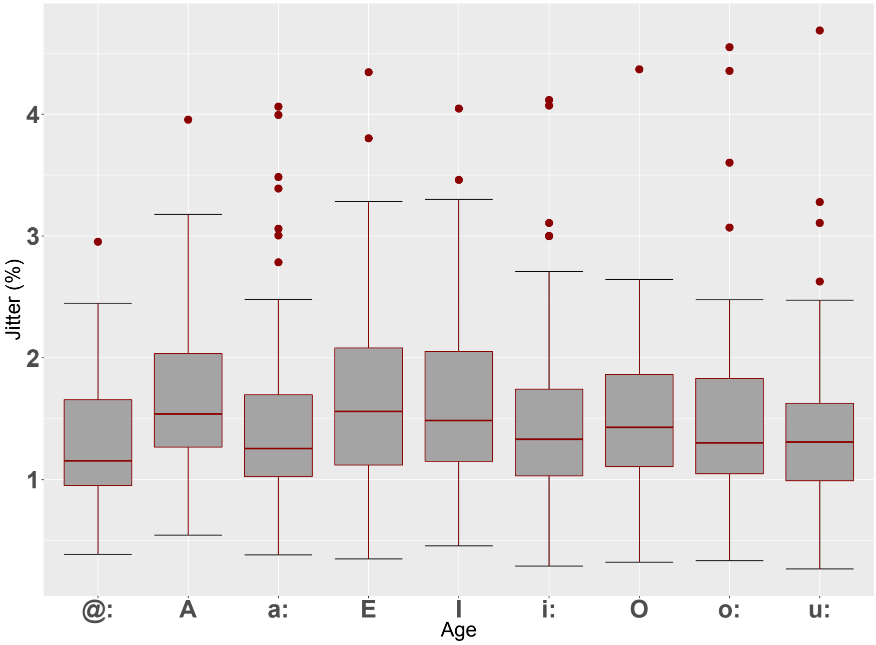

Figure 6.24: *Jitter boxplots for vowels regardless of age*

6.2.2 Female Speakers

This subset of database B is comprised exclusively of female speakers. This subset contains 530 vowel tokens, extracted from citation-form hVd word lists. The distribution of the vowel tokens between the two speaker groups (**young** and **old**) is presented in Table 6.7. The mean and standard deviation per each glottal parameter is presented in Table 6.8. Box-plots for the distribution of the mean value for each glottal parameter are given in Figure 6.25 to Figure 6.30.

No significant age effects, nor age-vowel effects, were observed for this subset.

Significant vowel effects were observed for multiple parameters.

Significant vowel interaction for the OQsub50 parameter was indicated with ANOVA (F(10,90) = 3.23, p<0.01). However, whilst the ANOVA suggested there was a significant interaction, *post hoc t-test* demonstrated that in fact this was not the case. For box-plot of the mean OQsub50 distribution between vowels, please refer to Figure 6.31.

Significant vowel interaction for the SQ parameter was indicated with ANOVA (F(10,90) = 2.33, p(GGE) < 0.001). However, whilst the ANOVA suggested there was a significant interaction, *post hoc t-test* demonstrated that in fact this was not the case. For box-plot of the mean SQ distribution between vowels, please refer to Figure 6.32.

Significant vowel interaction for the NAQ parameter was indicated with ANOVA (F(10,90) = 7.51, p(GGE) < 0.001). The *post hoc t-test* demonstrated the vowel WHO'D to be significantly greater than the vowels HOARD ($t(10) = 6.38$, $p < 0.01$), HOD ($t(10) = 6.42$, $p < 0.01$), HID ($t(10) = 4.91$, $p < 0.05$), and HERD ($t(10) = 6.47$, $p < 0.01$). In addition, the *post hoc t-test* demonstrated the vowel HEAD to be significantly greater than HOD ($t(10) = 4.78$, $p < 0.05$). For box-plot of the mean NAQ distribution between vowels, please refer to Figure 6.33.

Significant vowel interaction for the Jitter parameter was indicated with ANOVA (F(10,90) = 2.84, p(GGE)<0.01). The *post hoc t-test* demonstrated the vowel HUD to be significantly greater than HID ($t(10) = 4.74$, $p < 0.05$). For box-plot of the mean jitter distribution between vowels, please refer to Figure 6.34.

Significant vowel interaction for the Shimmer parameter was indicated with ANOVA (F(10,90) = 2.18, p<0.05). The *post hoc t-test* demonstrated the vowel

HOD to be significantly greater than HOOD ($t(10) = 5.92$, $p < 0.01$). In addition, the *post hoc t-test* demonstrated the vowel HAD to be significantly greater than HOOD ($t(10)$ = 4.97, p<0.05). For box-plot of the mean shimmer distribution between vowels, please refer to Figure 6.35.

Table 6.7: *Phoneme distribution in Database B - Female*

	short vowels						*long vowels*					*sum*
	I	U	E	O	V	A	i:	u:	o:	a:	@:	
Young	22	18	32	33	15	28	28	29	28	32	30	295
Old	20	12	25	26	14	22	21	22	22	26	25	235

Table 6.8: Database **B** (Female) Results

	Age			
	Young		Old	
	mean	*SD*	*mean*	*SD*
F0	183.72	33.90	175.31	28.70
OQsub50	0.56	0.07	0.57	0.07
SQ	1.57	0.54	1.49	0.70
NAQ	0.21	0.03	0.22	0.04
Jitter	3.55	2.25	2.69	2.16
Shimmer	4.71	2.66	5.63	2.30

Note: *p<0.1 , **p<0.05 , ***p<0.01

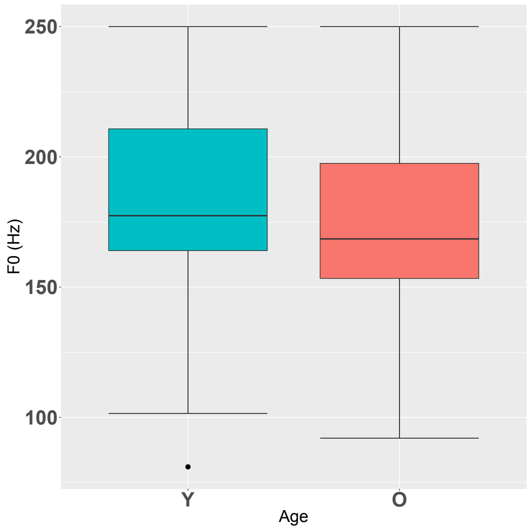

Figure 6.25: *Pitch boxplot*

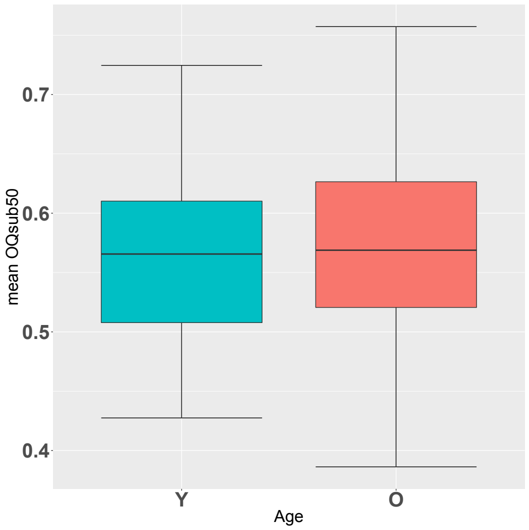

Figure 6.26: *OQsub50 boxplot*

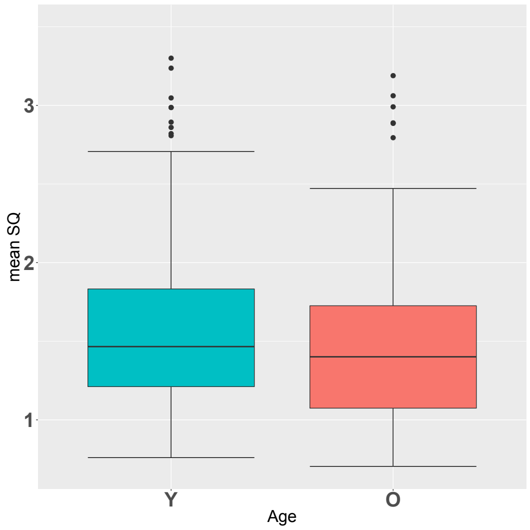

Figure 6.27: *Speed quotient boxplot*

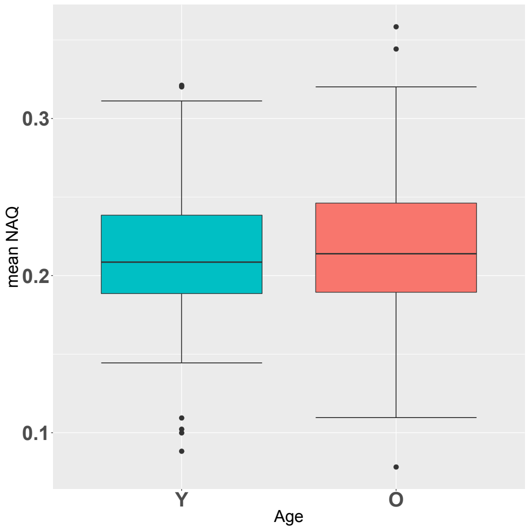

Figure 6.28: *NAQ boxplot*

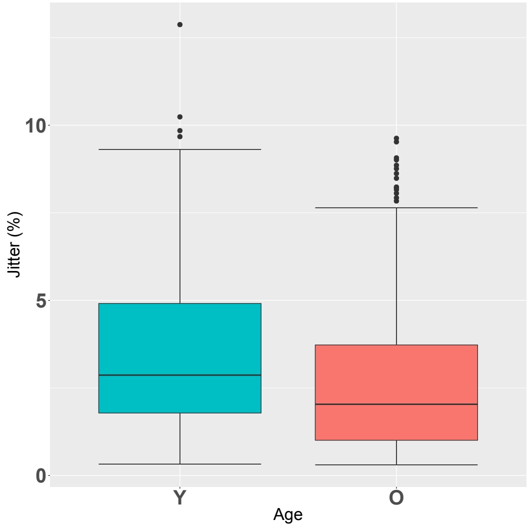

Figure 6.29: *Jitter boxplot*

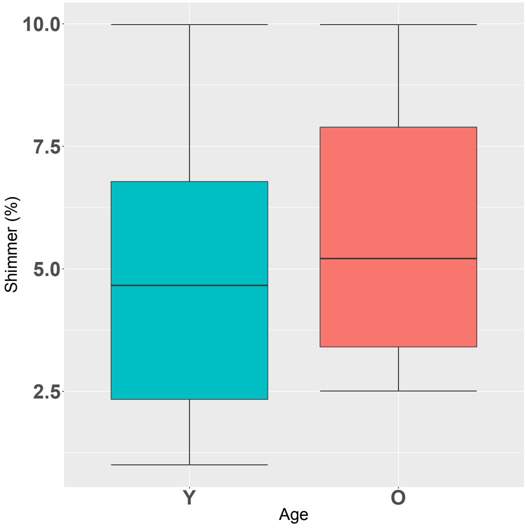

Figure 6.30: *Shimmer plot*

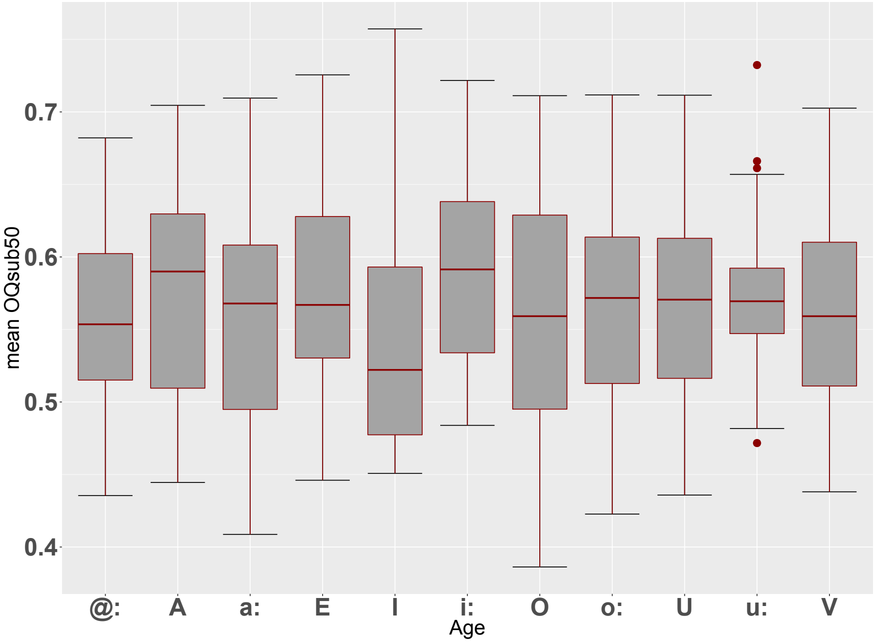

Figure 6.31: *OQsub50 boxplots for vowels regardless of age*

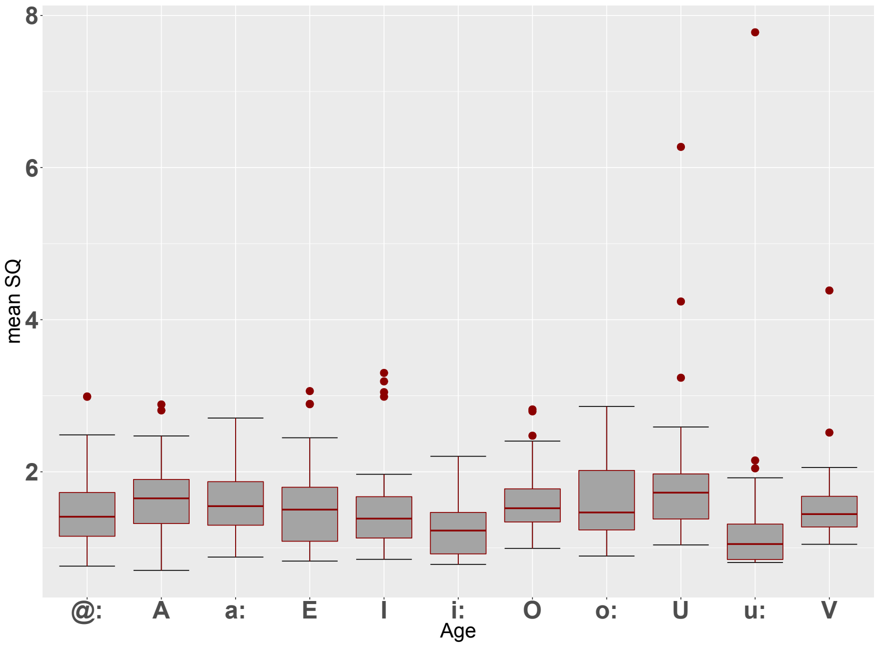

Figure 6.32: *SQ boxplots for vowels regardless of age*

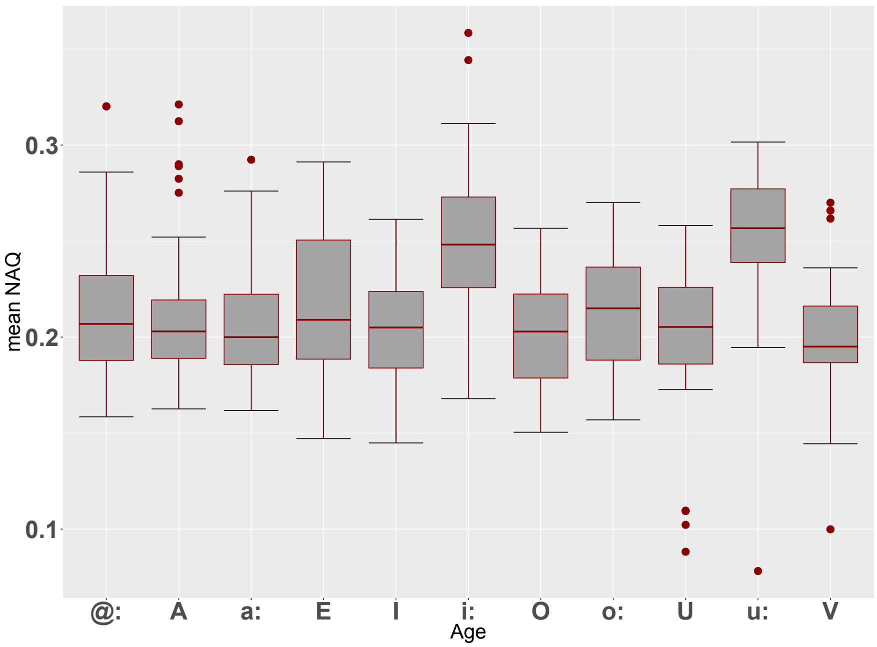

Figure 6.33: *NAQ boxplots for vowels regardless of age*

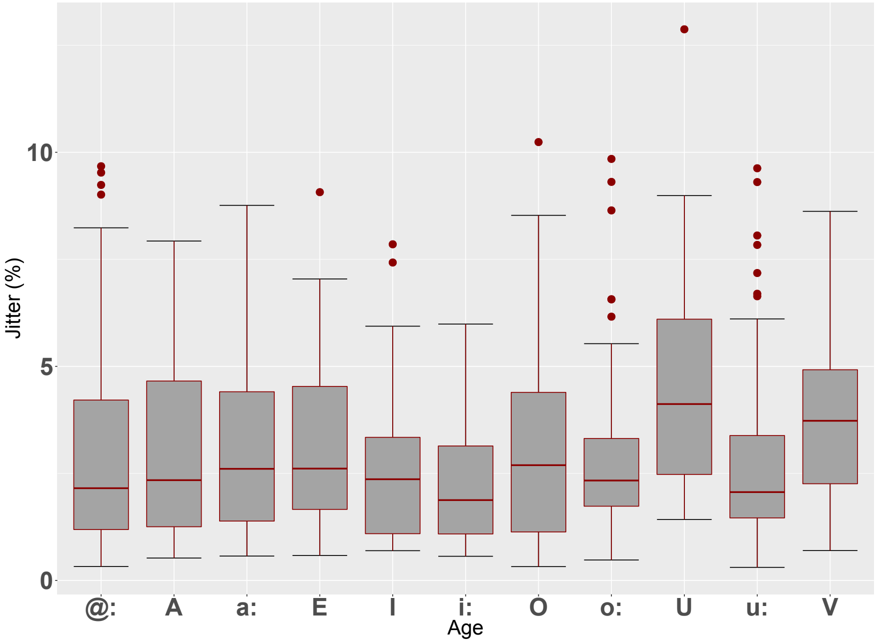

Figure 6.34: *Jitter boxplots for vowels regardless of age*

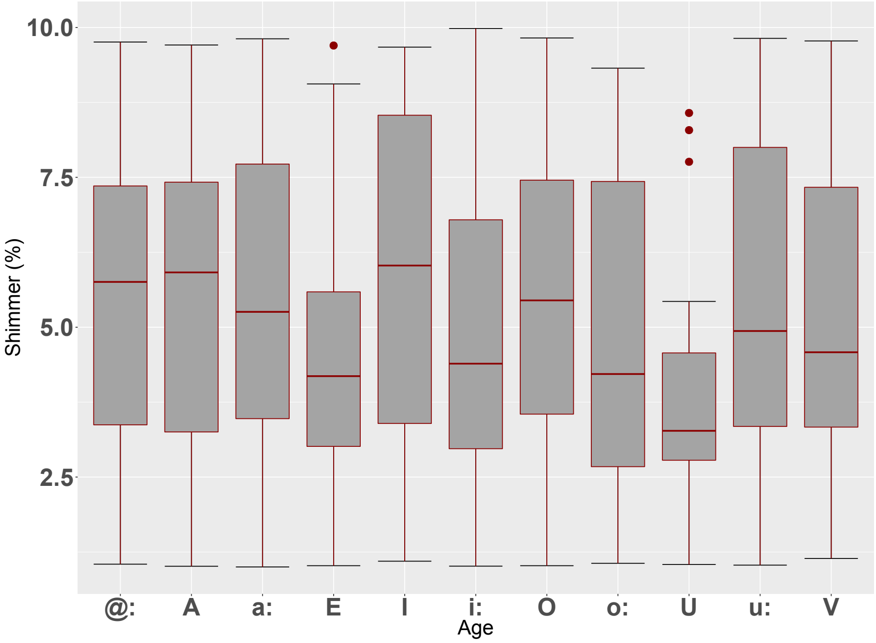

Figure 6.35: *Shimmer boxplots for vowels regardless of age*

6.3 Database C

This speech corpus contains 36 vowel tokens for the vowel /a:/ (HARD), extracted from continuous speech. The distribution of the vowel tokens between the three age groups (**Y**, **PE**, and **HE**) is as follows: there are 15 tokens for the **Y** group, 12 tokens for the **PE** group, and 9 tokens for the **HE** group. The mean and standard deviation for each glottal parameter are presented in Table 6.9. Box-plots for the the mean value of each glottal parameter per age group are shown in Figure 6.36 to Figure 6.43. Please note that Figure 6.40 & Figure 6.41 provide box-plot and scatter-plot representation for the same parameter. This is also the case for Figure 6.42 & Figure 6.43.

Significant age effects were observed for multiple parameters.

A significant age effect for the OQsub50 parameter was indicated with ANOVA (F(2,18) = 17.9, p<0.001). The *post hoc t-test* demonstrated the mean value for the **HE** group was significantly greater than that of the **Y** group (t(11.45) = 6.41, p<0.001), and the **PE** group was significantly greater than the **Y** group (t(10.41) = 4.44, p<0.01). For box-plot of the mean OQsub50 per age group, please refer to Figure 6.37.

A significant age effect for the Shimmer parameter was indicated with ANOVA (F(2,18) = 29.8, p<0.001). The *post hoc t-test* demonstrated the mean shimmer value for **HE** group was significantly greater than that of the **PE** group (t(7.75) = 5.0, p<0.01), and the **HE** group was significantly greater than the **Y** group (t(5.97) = 6.0, p<0.01). For box-plot of the mean shimmer per age group, please refer to Figure 6.39.

No vowel, nor age-vowel, effects were examined in this analysis, as all tokens are of the same vowel.

Table 6.9: Database **C** (Male) Results

	Age					
	Young		PE		HE	
	mean	*SD*	*mean*	*SD*	*mean*	*SD*
F0	124.12	31.94	114.64	31.28	147.16	33.97
OQsub50	0.37	0.09	0.63***	0.11	0.68***	0.07
SQ	1.67	0.66	2.24	0.55	1.97	0.83
NAQ	0.18	0.05	0.20	0.05	0.18	0.06
Jitter	2.55	1.99	3.65	2.30	4.38	3.21
Shimmer	11.96	6.59	14.35	8.11	41.35***	9.49

Note: *p<0.1 , **p<0.05 , ***p<0.01

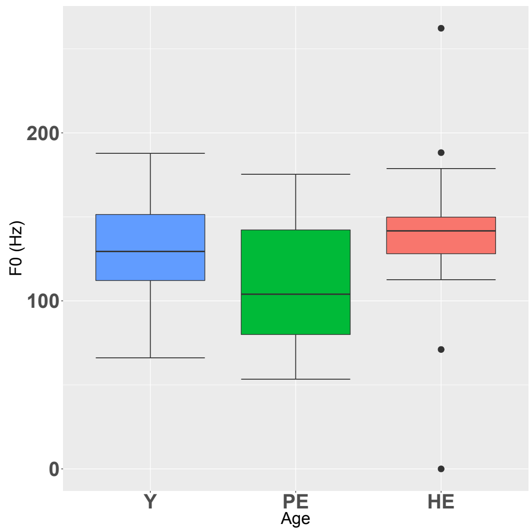

Figure 6.36: *Pitch boxplot*

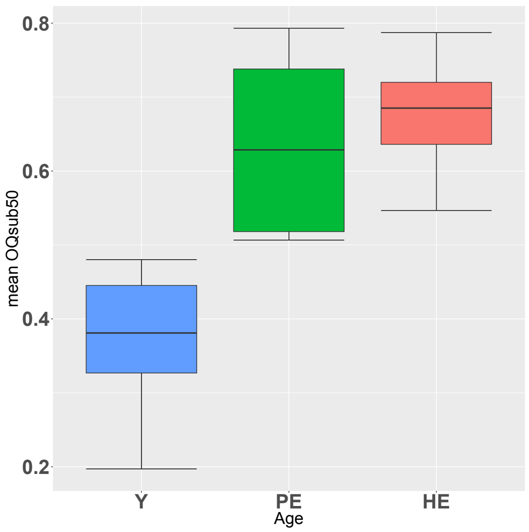

Figure 6.37: *OQsub50 boxplot*

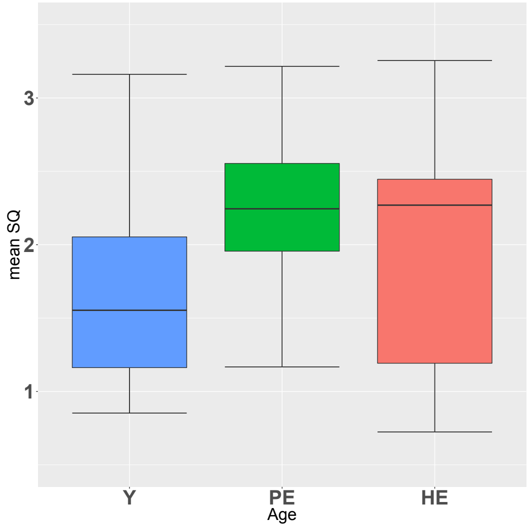

Figure 6.38: *Speed quotient boxplot*

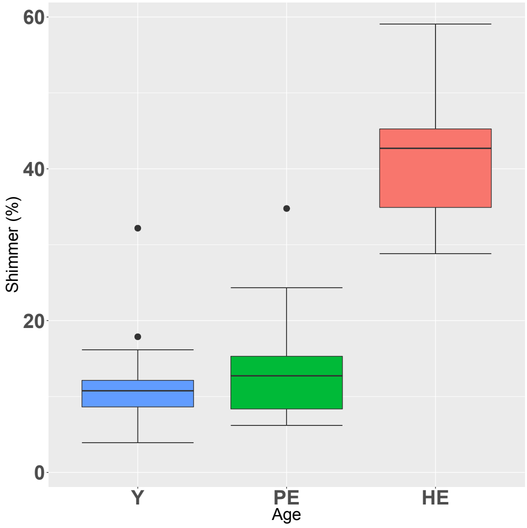

Figure 6.39: *Shimmer boxplot*

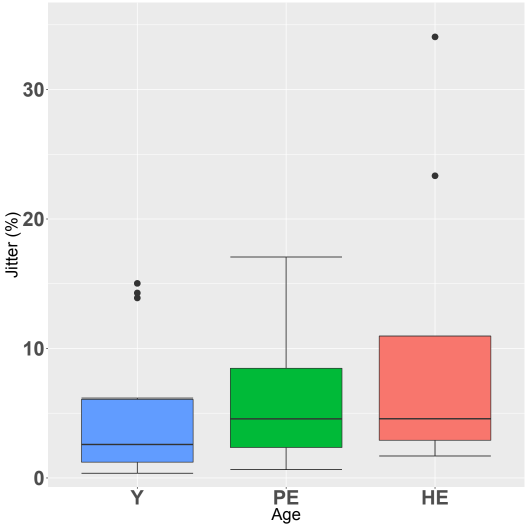

Figure 6.40: *Jitter boxplot*

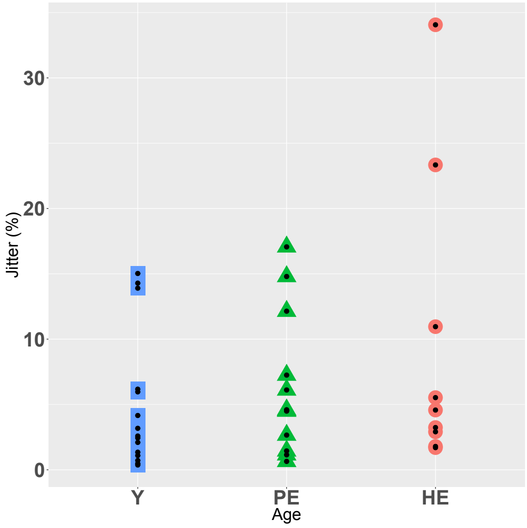

Figure 6.41: *Jitter scatter-plot*

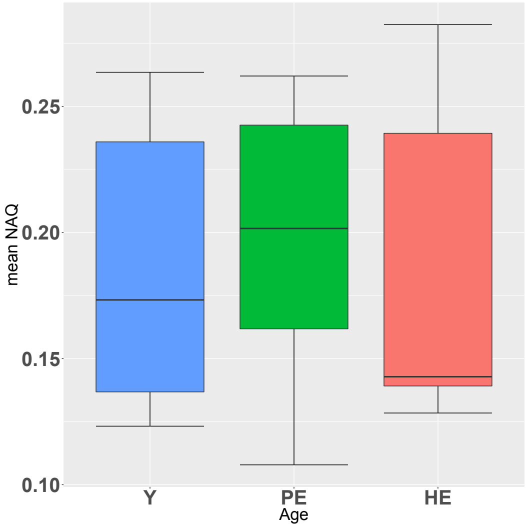

Figure 6.42: *NAQ boxplot*

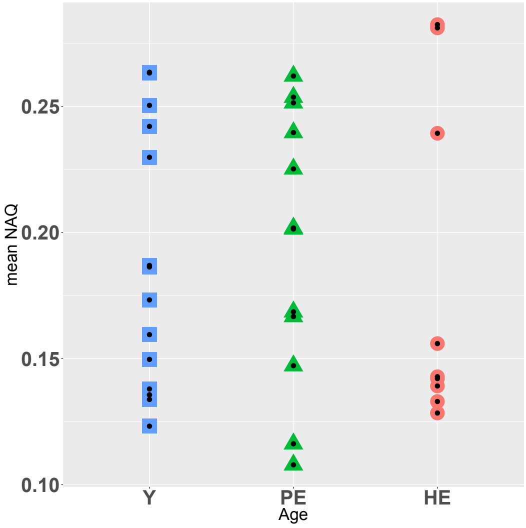

Figure 6.43: *NAQ scatter-plot*

6.4 Database D

This speech corpus contains 510 vowel tokens, extracted from continuous speech. As this is a longitudinal study, it only consists of one speaker. The speech data is divided between the years the recordings took place, effectively representing two age groups (**1955** and **1992**). The distribution of the vowels between the two recording-sessions/age groups is presented in Table 6.10. The mean and standard deviation per each glottal parameter are presented in Table 6.11. Box-plots for the distribution of the mean value for each glottal parameter are given in Figure 6.44 to Figure 6.49.

Significant age effect was observed for multiple parameters, as indicated in Table 6.11.

A significant age effect for the OQsub50 parameter was indicated with ANOVA (F(1,10) = 11.0, $p < 0.001$). For box-plot of the mean OQsub50 per age group, please refer to Figure 6.45. For box-plot of the mean OQsub50 distribution between vowels, please refer to Figure 6.50. A box-plot of the mean OQsub50 distribution for each vowel per age group is provided in Figure 6.51. For the mean and standard deviation OQsub50 per vowel for each age group, refer to Table B.7 in Appendix B.

A significant age effect for the NAQ parameter was indicated with ANOVA (F(1,10) = 6.9, $p < 0.001$). For box-plot of the mean NAQ distribution per age group, please refer to Figure 6.47. For box-plot of the mean NAQ distribution between vowels, please refer to Figure 6.52. A box-plot of the mean NAQ distribution for each vowel per age group is provided in Figure 6.53. For the mean and standard deviation NAQ per vowel for each age group, refer to Table B.7 in Appendix B.

A significant age effect for the Shimmer parameter was indicated with ANOVA (F(1,10) = 2.9, $p < 0.05$). For box-plot of the mean shimmer distribution between vowels, please refer to Figure 6.49.

No significant vowel, nor age-vowel, effects were identified.

Note: The Machine Readable Phonetic Alphabet (MRPA) phonetic labelling was used in this database. The vowels and their corresponding hVd words are as follows: /I/ “HID”, /U/ “HOOD”, /E/ “HEAD”, /O/ “HOD”, /V/ “HUD”, /A/ “HAD”, /i:/ “HEED”, /u:/ “WHO’D”, /o:/ “HOARD”, /a:/ “HARD”, /@:/

"HERD". In practice, the vowels were extracted from a wide variety of words, none of which were in fact hVd words.

Table 6.10: *Phoneme distribution in Database D*

	short vowels						*long vowels*					*sum*
	I	U	E	O	V	A	i:	u:	o:	a:	@:	
1955	7	2	27	56	14	60	61	32	40	59	9	367
1992	5	2	20	8	10	23	15	6	16	22	16	143

Table 6.11: Database **D** (Male) Results

	Age			
	1955 data		1992 data	
	mean	*SD*	*mean*	*SD*
F0	149.16	29.13	139.15	26.24
OQsub50	0.21	0.10	0.41***	0.13
SQ	1.66	0.89	1.80	1.53
NAQ	0.19	0.05	0.25***	0.05
Jitter	3.65	2.59	4.36	3.18
Shimmer	8.42	4.50	12.27*	8.72

Note: *p<0.1 , **p<0.05 , ***p<0.01

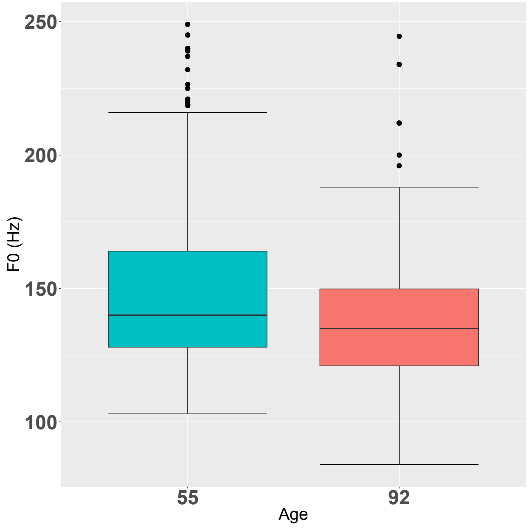

Figure 6.44: *Pitch boxplot*

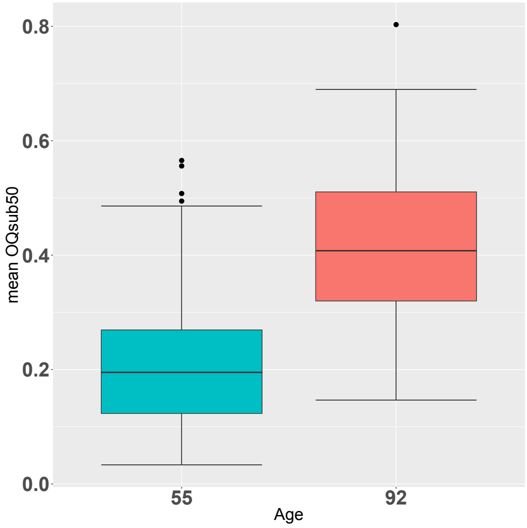

Figure 6.45: *OQsub50 boxplot*

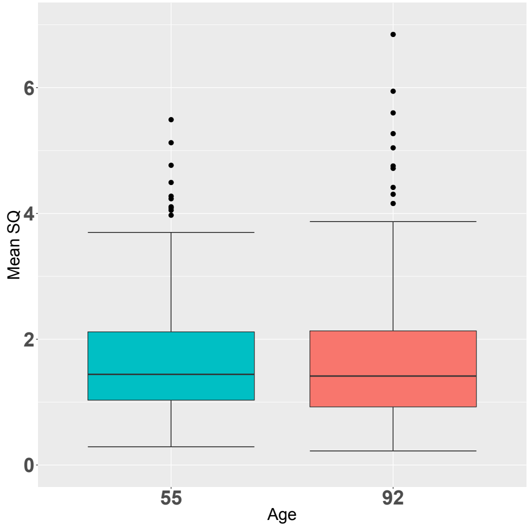

Figure 6.46: *Speed quotient boxplot*

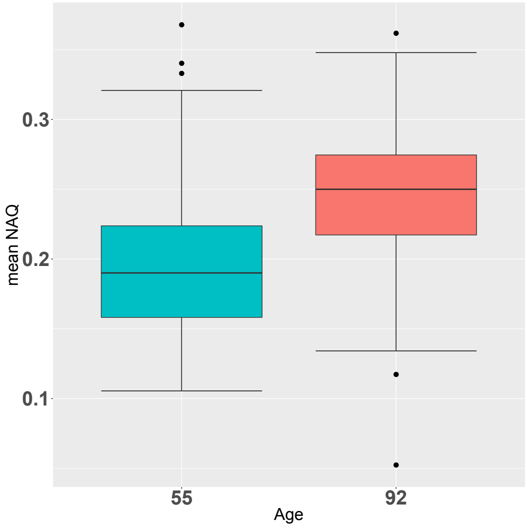

Figure 6.47: *NAQ boxplot*

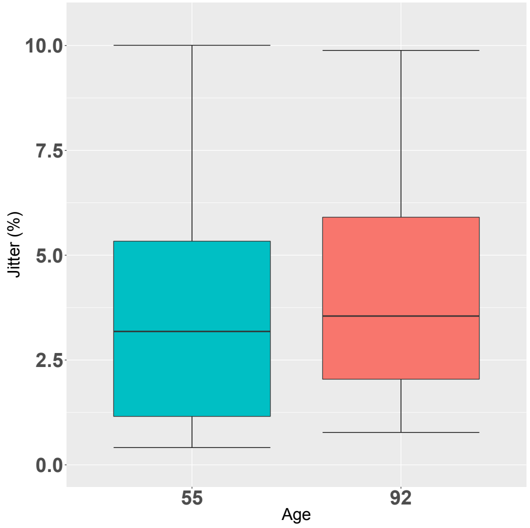

Figure 6.48: *Jitter boxplot*

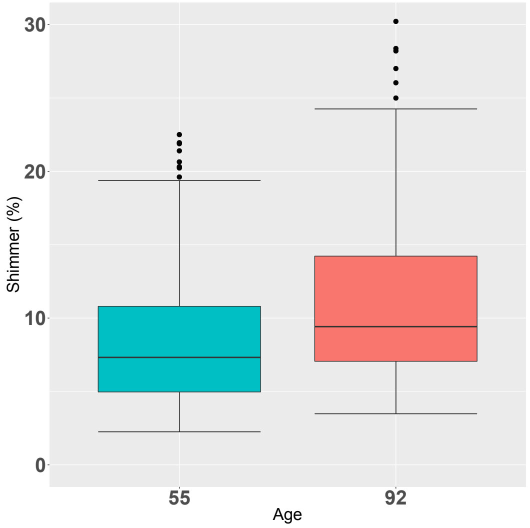

Figure 6.49: *Shimmer boxplot*

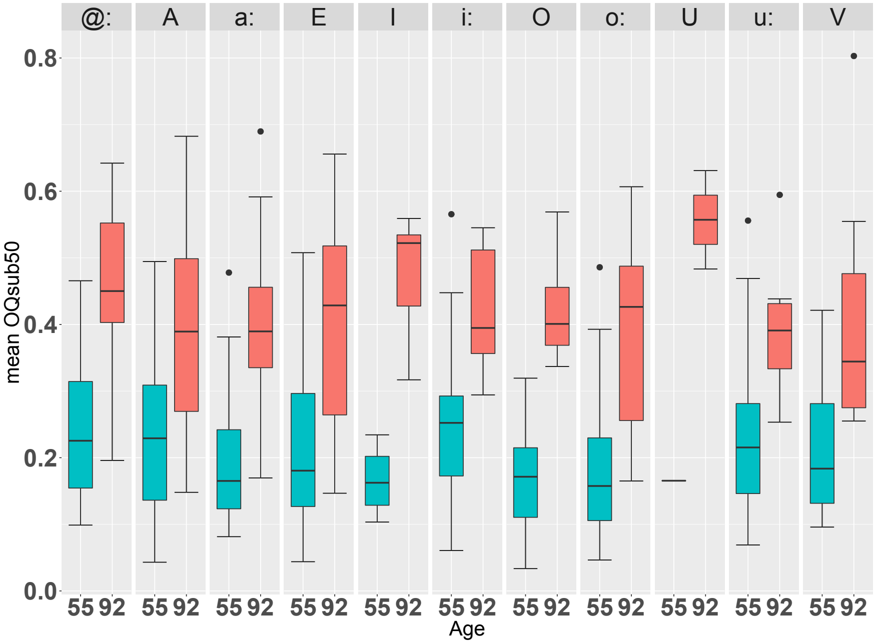

Figure 6.50: *OQsub50 boxplots for vowels interaction*

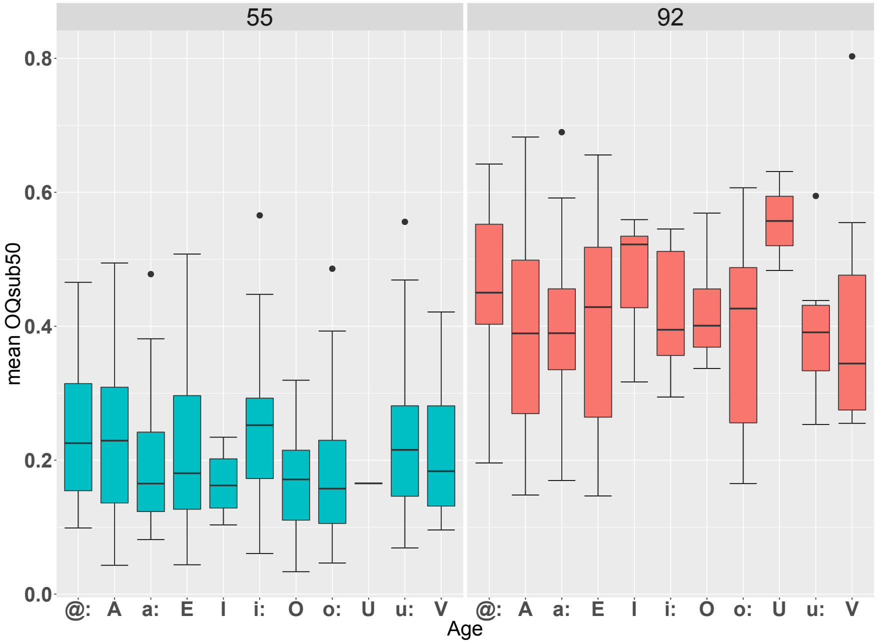

Figure 6.51: *OQsub50 boxplots for age:vowels interaction*

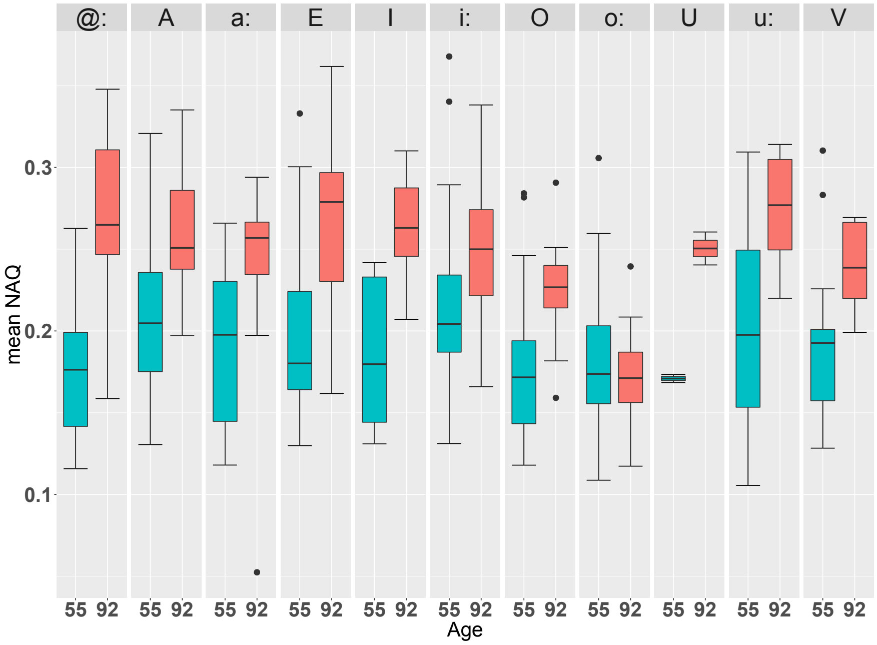

Figure 6.52: *NAQ boxplots for vowels interaction*

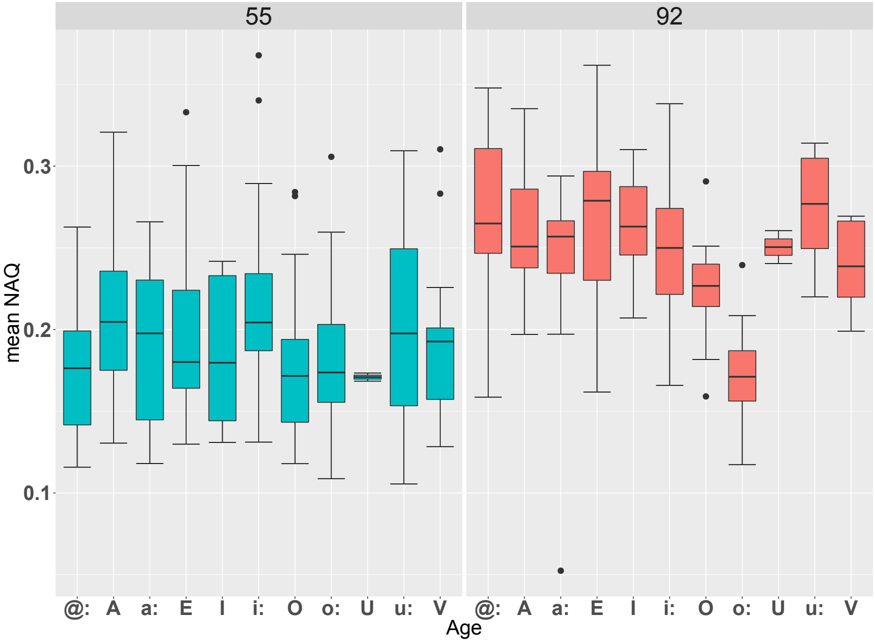

Figure 6.53: *NAQ boxplots for age:vowels interaction*

Chapter 7

Discussion

In this chapter, the findings of the previous chapter are discussed. Three key points are examined as they pertain to the glottal source: the ageing effect, the vowel effect, and the age-vowel interaction. The impact on ageing on each glottal parameter is presented and related back to the literature. Next, we elaborate on the impact of vowel interaction and how it pertains to source-filter coupling and/or mechanical coupling. Finally, age-vowel effect is discussed and its implications on vocal ageing analysis are highlighted. In addition, the limitations of the analysis are discussed, as well as the impact of formant ripples and polarity on the signal.

7.1 Age Differences

7.1.1 Ageing Effect on Pitch

The rate of vocal folds vibration during the self-oscillatory cycle of speech production accounts for the fundamental frequency of voiced speech. A comprehensive overview of the fundamental frequency trends in literature was provided in Chapter 3, Section 4.6. Although the literature offers an abundance of publications examining the changes in fundamental frequency with advanced age, the result are contradictory and inconsistent. A similar output was observed in this investigation, with multiple pitch trends observed for aged speakers.

For database A, an increase in mean fundamental frequency was observed between speakers of the young and old age groups. The older speakers produced fundamental frequency which was significantly greater ($p < 0.05$) than that of the young adult speaker. Similar age effect was found by Bier & Watson *et al* [220] in a study utilising the same database, although they used the EGG signal, not the speech signal. An increase in fundamental frequency (about 12 Hz) is in line with previous studies which examined the effect of ageing on fundamental frequency in males [139, 140]. The increase in fundamental frequency of vibration is attributed to physiological changes with senescence voice. Geriatric changes in the tissue structure of the larynx and the vocal folds had been observed in previous studies. It was shown that reduced vocal folds mass and thickness, as well as reduced elasticity of the laryngeal cartilages, accounted for increased stiffness and shortening of the vocal folds, which accounts for an increase in the rate of vocal folds vibration with advanced age [134, 133].

Database B consists of two gender subsets analysed seperately. The first subset consists solely of male speakers. The pitch findings for this subset are surprising, as no significant fundamental frequency change was observed between the age groups. Our results indicate that effectively there is no change in pitch with advanced age. This lack of pitch trend is scarce in literature, with most publications reporting an increase or decrease in pitch. However, similar findings were observed by Masaki *et al* [143]. In the study, mean fundamental frequency was obtained for Japanese speakers divided into three age groups. The results show no significant mean fundamental frequency differences between the three age groups (19 to 80 years old). Upon analysis of the results, the authors report no significant change in fundamental frequency until the 7th decade. We note that Japanese have a longer life expectancy and are therefore likely to sound "younger" for longer. Hence, the lack of difference could be attributed to physiological age rather than chronological age. With regards to the male speakers of this subset, the old group comprised of eight speakers in the age range of 55 to 80 years old. However, as only limited information was made available, it is unclear how many of the old speakers were in their 70s or older. The reference pitch interval for the old group was between 80 and 160 Hz, which is in line with the range reported by Masaki. Thus, we hypothesize that speakers in the old age group were predominantly under the age of 70. Those results contradict the hypothesis of the male-female coalescence model [119], reporting minimal-to-no change in pitch with ageing for male speakers.

A second subset of database B were the female speakers. For this subset, a decrease in fundamental frequency was observed between the young (19-35 years old) and old (55-80) age groups. This 8 Hz drop in pitch was not statistically significant. The results are indicative of the post-menopausal drop in pitch reported in literature [153, 142, 103, 144, 145]. The pitch drop is relatively small compared to the average pitch drop of 15 to 20 Hz most often cited in literature [148, 103]. This post-menopausal drop is attributed to hormonal changes due to menopause, such as edema [135]. We note that the fundamental frequency for female speakers was significantly higher than that of their male counterparts for database B. This reinforces previous findings of female speakers having higher pitch than male speakers [144]. Also note that the mean pitch reported for the female speakers in this investigation is lower than the pitch reported in previous publications [99], while it is in accordance with other studies [148, 141].

Database C consists of three age groups, for which two age groups are considered to be comprised of old speakers. Although changes were observed in the fundamental frequency for the age groups, they were not significant. The lack of significant interaction is attributed, in part, to the small sample size of this database. The results are rather surprising, as two inverse pitch trends are observed. First, consider the interaction between the young speakers and the present-day elders speakers. A pitch drop of 10 Hz was observed. A closer examination of the results reveal multiple outliers, with speakers producing pitch as high as 190 Hz for the young group, and multiple speakers with pitch lower than 80 Hz. Hence, due to the small sample size and the variability of pitch, we are reluctant to draw conclusions with regards to this interaction. Second, consider the interaction between the young speakers and the historical elders speakers. A pitch increase of 23 Hz was observed. The results show greater robustness for the historical elders group, with the majority of pitch periods between 130 and 150 Hz, and only one outlier below 100 Hz. This reinforces the pitch trend exhibited for database A. In analysing this database, it's important to consider the relation between the present-day elders (PE) and the historical elders (HE) group. Although both groups are labelled as old, there are two key differences between the speakers. First, consider the age of the speakers. All PE speakers were above the age of 65, while all HE speakers were between the ages of 62 and 77. With regards to the PE speakers, it is unknown whether those speakers were substantially older than 65 years old. This can be related back to the findings by Masaki, which suggested an increase in pitch could be observed as late as the start of the 7$^{\text{th}}$ decade. Hence, it is possible that the decrease in pitch observed for the PE group was due to the decrease in pitch from adulthood to middle-age. In turn, the HE speakers had a higher pitch as multiple speakers were older than 70 years old. This implies that an increase in pitch for male speakers could occur several decades past the middle-age point. Second, consider the socio-economics differences between the PE and HE speakers. Per King *et al* [226], HE speakers would be considered to belong to a lower socio-economic class than the PE speakers. In turn, environmental factors such as poor nutrition, or being of a large family, could had had a detrimental affect on the voice quality of the speakers. Moreover, life expectancy in the 1940s was lower than now. Hence, although the speakers are of a similar chronological age, their physiological age would have shown greater separation, which is manifested in the difference in pitch.

Database D contains speech data gathered 40 years apart, to comprise a longitudinal study of a singular male speaker. For this speaker, a decrease in fundamental frequency of 10 Hz was observed. However, this decrease was not statistically significant. We note that a potential reason for the higher pitch in the 1955 data is the digitisation procedure of the data. The 1950s recordings would have initially been stored on a record. It is possible that the record was played at a wrong turning speed while it was digitised, causing the pitch to go up. These results are consistent with those from other longitudinal studies showing decrease in pitch with advanced age [159, 158, 210]. Watson *et al* [158] reported similar findings for a longitudinal study for a New Zealand English speaker, though a larger 20 Hz drop was reported. Similarly, Decoster *et al* [159] reported a 14 Hz drop in pitch for a longitudinal study across 20 male speakers. For this, however, they note that five speakers exhibited an increase in fundamental frequency. The mean fundamental frequency per each recording session is higher than that reported by Watson, but is within the range reported by Decoster. Though this database consists of a longitudinal study, we are cautious to draw conclusions. The results could be better understand if intermediate (throughout the years) speech data was made available (i.e. examine the changes in pitch once per decade). The literature reports a gradual increase in fundamental frequency between middle-age (50s) and advanced age. However, the results are in contrast to the notion of an increase in pitch in advanced age for male speakers. Our results suggest that an increase in pitch could be observed much later life (potentially beyond 85 years of age). We note that an increase in age at such advanced age is usually attributed to female speakers [142]. Thus, we conclude that the turning-point for pitch increase with advanced age is speaker-dependent and does not necessarily correspond pre-conceived trends.

7.1.2 Ageing Effect on Perturbation Parameters

An investigation was carried out to determine the effect of ageing on the cyclical glottal perturbation measurement, jitter and shimmer. Those glottal perturbation parameters provide a quantifying tool to determine vocal instability for speakers. The results are analysed separately for male and female speakers. For male speaker across all databases, older speakers demonstrated greater cycle-to-cycle period perturbation compared to young speakers. Increased jitter for senescence voice has been observed in literature[164, 165, 166, 233]. Although an increase in jitter was observed across all databases, the effect of ageing had a significant effect only for database A ($p < 0.001$). The lack of significant jitter interaction is prominent in literature [146, 147, 107, 161, 129, 138]. The jitter values obtained in this analysis are large in view of Hollien *et al* [167]. Hollien reported jitter values in the range of 0.5 to 1 for male speakers, with no noted laryngeal pathology, using high speed photography. However, Hollien's study was centred around sustained vowels. Moreover, since a unified established jitter range has yet to emerge, those results are taken at face-value. The increase in jitter values suggests asynchronousy of vocal folds vibrtation. For older speakers, this is attributed to senescence voice qualities such as hoarseness, roughness, breathiness, vocal tremor, and hesitancy [164, 234]. Multiple publications, however, argue that ageing does not affect the synchronization of the vocal folds, thus discarding the reliability of the jitter measurement [129, 107, 165]. We are cautious to attribute vocal ageing to jitter values, due to overlap in measurements. This can be observed in database C, as shown in Figure 40 and Figure 41. Although the box-plot indicates an increasing jitter with advanced age, a scatter-plot reveals great overlap in jitter range for the three age groups. We note that this is particularly apparent in this database due to its small sample size. Similar overlap in jitter range was observed by Linville *et al* [161], which reported overlap in jitter range for young and elder women speakers.

Furthermore, studies have shown jitter to be an unreliable measurement for vocal ageing, due to its correlation with environmental factors [130]. Hence, jitter may be more greatly correlated to physiological age than chronological age. Our results support this hypothesis. Comparing the mean jitter values for database B with the mean jitter values for databases C & D, it can be readily seen that the values are approximately twice as large for the latter. The speaker in database B profession was an explorer. Although the affect of his profession on his voice quality is unknown, it can be assumed that he had been exposed to environmental factors of imbalanced nutrition and physical injuries. The HE Maori speakers in database C could experience reduced vocal quality due to low socio-economic statue and lower life expectancy compared to the PE and Y speakers. In turn, those external factors will manifest as increased jitter. The larger jitter values for database A compared to database B could be attributed to emotional speech for young speakers (excited to perform) and hesitancy for old speaker (unfamiliar with the procedure and the EGG apparatus). Finally, the large values for databases A, C, and D, compared to database B could be attributed to the procedure of speech recordings. Continuous speech is more prone to vowel coarticulation effect. This could lead to pitch variations registered as jitter.

Shimmer was observed to increase with ageing across all database. Significant age effect for shimmer values was found in database A ($p < 0.01$), database c ($p < 0.001$) for historical elders compared to young speakers and compared to the present-day elders, and database D ($p < 0.05$). Thus, shimmer exhibits age-related acoustic changes. Increased shimmer for senescence voice was reported in previous publications for both male and female speakers [99, 233] The shimmer values for databases A, B (males), C and D were higher, for both young and old speakers, compared to shimmer values reported in the aforementioned publications. Higher shimmer values were found for male speakers compared to female speakers. For female speakers in database B, the shimmer results for young speakers were higher than the values reported in previous studies [168, 99, 233], while the shimmer values for old speakers were similar to the results reported in [99, 136]. The increase in shimmer with old age might be attributed to structural age-induced changes in the larynx leading to vocal folds aperiodicity. We note that for database C, historical elder speakers exhibited extremely high shimmer values. This finding can most likely be attributed to the presence of noise in the signal, rather to vocal folds behaviour. A larger sample size will be required to effectively determine the effect of ageing on shimmer for this database. Nevertheless, shimmer was shown to reflect vocal ageing. However, a subsequent examination of signal-to-noise ratio is required prior to the establishment of shimmer reliability.

7.1.3 Ageing Effect on Open Quotient

The open quotient provides information about the width of the glottal pulse in relation to the pitch duration. The open quotient quantifies the vibration patterns of the vocal folds during speech production. The findings show an increase in open quotient with advance age across all databases. The effect of age was statistically significant for databases A ($p < 0.01$), B (males) ($p < 0.05$), C ($p < 0.001$), and D ($p < 0.001$). A slight increase with age was observed for database B (females), however, it was not statistically significant. We note that a study by Bier *et al* [220] examined the effect of ageing on vowels from database A. In the study, the contact quotient was derived from Electroglottography data. For this study, the authors report a decrease in contact quotient with age, indicating longer open phase for elder speakers. A similar effect was observed for this analysis. Increasing open phase duration (longer vocal folds abduction) with ageing was observed in other studies [235, 236, 237]. Studies have shown female speakers to exhibit higher open quotient measurements than male speakers, suggesting female voice to be naturally breathier than male voice [74, 238, 235]. However, this was not reflected in our results. In senescence voice, open quotient is often correlated to breathy register. Physiologically, breathy voice pathology is suggested to occur due to posterior opening of the glottis, allowing glottal air flow leakage during the closed phase. This air flow leakage results in turbulent flow, which then couples to the speech signal. This might be due to laryngeal muscle atrophy, which results in vocal folds bowing [239]. Moreover, increased glottal width could be attributed to increased vocal folds tension, reduced sub-glottal pressure due to reduction in lung capacity which results in reduced air flow [240]. Examination of the glottal waveforms reveal that no glottal flow exhibited a flat flow during the closed phase, which implies incomplete closure of the glottis. Alternatively, however, this could be attributed to incomplete decoupling of the glottal source and vocal tract filter, as well as the presence of noise in the signal. Alku *et al* [73] examined the glottal waveform, and showed breathy voice to result in larger open quotient for both male and female speakers compared to modal register. Breathy voice correlated to open quotient measurements in the range of 0.86 to 0.97, compared to open quotient range of 0.61 to 0.91. Childers *et al* [109] reports an open quotient value of 0.91 for breathy voice and 0.70 for modal speech via Electroglottography signals. Our open quotient results are in the range of 0.21 to 0.72. Per the aforementioned studies, this will more closely correlate to vocal fry or pressed phonation [109, 73]. Hence, we are unable to comment on the presence of breathy voice pathology. In order to determine breathiness, further examination into the results is required, in the form of Electroglottography comparison or an auditory perceptual test. However, we conclude by proclaiming age to have an increasing effect on the glottal pulse width.

7.1.4 Ageing Effect on Speed Quotient

The speed quotient is defined as the ratio between the opening phase and the closing phase. Thus, it is a glottal pulse skewness measurement, reflecting the asymmetry between the vocal folds abduction and adduction durations. A symmetrical glottal pulse will therefore have a speed quotient measurements of 1, which is highly unlikely. The speed quotient was greater than 1 across all speech corpora. This indicates vocal folds abduction, on average, to be slower than vocal folds adduction. The ageing effect on the speed quotient parameter has yet to yield a unified trend, with studies showing both increase and decrease in speed quotient [241, 237].

For this investigation, only database A showed significant age interaction for the speed quotient ($p < 0.001$). For dabase A, the speed quotient was observed to decrease with advanced age. Similar observations were made for databases B (females), although the results were not statistically significant. Physiologically, lower speed quotient corresponds to longer closing phase relative to the opening phase. This suggests older speakers exhibit reduced vocal folds efficiency [72]. This implies greater glottal symmetry, which could imply breathy voice register compared to modal voice register [73].

An increase in speed quotient values for senescence voice is an indicator voice pathologies, such as hoarseness and vocal fry. In Electroglottographic evaluation by Chen *et al* [120], increase in mean speed quotient values for vocal fry compared to modal phonation was reported. The speed quotient values were found to increase with age for databases B (males), C, and D. However, no significant age effect was observed for those databases. Physiologically, increase in mean speed quotient values implies slower vocal folds abduction. A longer opening phase in the vocal folds vibration cycle suggests reduced elasticity and increased vocal folds stiffness [120]. The mean speed quotient range observed for vocal fry for male speakers in [120] is similar to the range observed in this investigation. With advanced age, reduced vocal folds elasticity is attributed is attributed to increase in vocal folds thickness and fibrosis [134]. Moreover, it is likely that age-induced changes in the laryngeal cartilages (ossification) and muscles (atrophy) deteriorate the fine-motor control of the voice production mechanism. Such changes are more prominent in male speakers, which could account for the lack of speed quotient reduction with ageing for the female subset in database B.

Although the above findings suggest voice pathology for older speakers, we are unable to conquer those findings without further examination of the speech data (i.e. perceptually). We note that factors may influence the interpretation of the speed quotient results. When considering the speed quotient, measurements can be affected by sound pressure level. Speed quotient had been shown to increase for increasing sound pressure levels [70, 82]. Increasing sound pressure levels could be attributed to the increase in speed quotient values for database C, as the recording procedure did not take place in an isolated environment. Moreover, the speech data was extracted from continuous speech, which is more susceptible to loudness variations compared to isolated words. Though we did not examine the different sound pressure levels of the speech data, we hypothesize that the increase in speed quotient attributed to ageing effect could have resulted from variations in sound pressure levels.

7.1.5 Ageing Effect on Normalized Amplitude Quotient

The normalized amplitude quotient data with regards to vocal ageing is scarce. This is due, in part, to the fact that it was proposed relatively recently compared to the other glottal quotients used in this study. Hence, the comparison of results across multiple studies is not possible at the present time. The vocal ageing literature for the amplitude quotient is also scarce, as the parameter was devised only six years prior to its normalized iteration. Several studies had reported increase in amplitude quotient with advanced age by examining the glottal flow waveform [237]. Those findings, however, were all derived from a single speech corpus, which limits the applicability of the results. For the purposes of this investigation, the normalized amplitude quotient was examined to reinforce the findings for the open and speed quotients. Moreover, we looked for observable inter-parameter trends.

For this investigation, significant age effect on the normalized amplitude quotient parameter was observed for databases A ($p < 0.01$) and D ($p < 0.001$). An increase in mean normalized amplitude quotient values was observed with ageing. A similar age effect was observed between the young and present-day elders for database C, although the results were not statistically significant. Since the normalized amplitude quotient is an amplitude-domain representation of the closing quotient, an increase in normalized amplitude quotient values is indicative of longer closing phase for senescence voice. Longer vocal folds adduction is characterized as inefficient vocal folds behaviour, and is often attributed to breathy phonation. In a glottal analysis study by Alku *et al* [242], increased normalized amplitude quotient measurements for breathy register compared to modal register for both male and female speakers were reported. For breathy voice, NAQ might be influenced by the degree of glottis openness. Glottal air flow leakage could influence the negative amplitude of the derivative glottal flow, which in turn would increase the NAQ value [243].

Considering the results obtained for open quotient, speed quotient, and normalized amplitude quotient, observations can be made regarding the behaviour of the vocal folds, as reflected in the shape of the glottal pulse. There exists a correlation between the speed quotient and the glottal width (open quotient) [240]. The speed quotient and the open quotient can be jointly examined to deduce information regarding the duration of the closing phase. Consider two glottal signals representing young and old speakers. Assume the open quotient to increase with age and the speed quotient to reduce with age. An increase in open quotient values indicate a wider glottal pulse width. A reduction in speed quotient values indicate a longer closing duration. Hence, since the duration of the pulse increases, and the duration of the closing phase increases, the closing phase duration for old speakers is necessarily longer than that of young speakers. This will result in a larger normalized amplitude quotient for old speakers compared to young speakers. Thus, we predict the speed quotient to show an inverse trend to the normalized amplitude quotient. This inverse relationship between the speed quotient and the normalized amplitude quotient is observable for database A. For database B (males), the normalized amplitude quotient decreases with age. This indicates shorter vocal folds adduction, implying greater efficiency. However, the speed quotient increases with age, and therefore, the inverse relationship is maintained. For database B (females), an interesting observation is made. For this, the open quotient and the normalized amplitude quotient show minor changes with age, which are not statistically significant. However, the speed quotient is observed to decrease for old speakers. This indicates similar average glottal pulse width and closing phase for both age groups. Hence, it is expected that the speed quotient will not exhibit changes with ageing. Although the speed quotient changes are not significant, speed quotient decreases with age. This could imply that the speed quotient measurements were compromised due to changes in sound pressure level, or due to incorrect glottal opening and closing instances. For database C, we refrain from attributing significance to the results due to the large shimmer values. As the normalized amplitude quotient is an amplitude-based parameter, the integrity of the results is compromised. For database D, increase with age was observed for both normalized amplitude quotient and speed quotient. However, considering our previous comments with regards to speed quotient outliers, removal of those outliers will result in the hypothesized inverse relationship between speed quotient and normalized amplitude quotient.

We note that the closing quotient parameter is more widely featured in literature. A subsequent continuation of this study could examine the age effect on the closing quotient. Thus, allow comparison of results to the results obtained for the normalized amplitude quotient.

7.2 Vowel Effect

Significant vowel effect was examined in this study. The aim of the investigation was to determine the influence of the glottal flow on the production of vowels. As stated in the outset of Chapter 4, the source-filter model builds upon the foundations of complete independence between source and filter. Per this definition, the production of vowels is the result of a singular glottal source, which is then filtered through a vocal tract filter of various configurations, resulting in the production of vowels. This describes a linear system. Hence, in theory, inverse filtering of multiple speech signals for vowels will result in the same glottal source signal. Our findings, however, implicate otherwise. Please note that since different vowel symbol sets (MRPA, SAMPA) were used in this investigation, vowels will be indicated using hVd words for this section. For the mapping between the hVd words and the vowel symbols used in the various studies, refer to Table 2.1 and Table 2.2 in Chapter 2.

Significant vowel interaction was observed for the fundamental frequency of the glottal source for databases A & B (male subset only). An intrinsic pitch effect was observed for the vowel measurements. These results were related back to the glottal source signal. Considering the definition of the fundamental period for a glottal signal, the fundamental period reflects the rate of vocal folds vibration during oscillation. Here, once again, we reiterate the notion of source-filter interaction. The results suggest that different vowels require different periodicity of the excitation signal. However, per the source-filter model, the source signal is said to be independent of the vocal tract formant frequencies. Our findings suggest acoustic coupling of source and filter. In turn, such phenomenon could account for the result. Acoustic coupling might occur when the first formant is approximately similar to the fundamental frequency, resulting in an interaction with the glottal flow waveform [26]. For a relatively high first formant frequency, the vocal tract walls act to increase the interaction between the vocal tract and the vocal folds. Similarly, a low first formant will result in reduced interaction between the vocal tract and the vocal folds, yet the glottal source signals will still vary from one another.

The intrinsic pitch of vowels is attributed to the vocal tract configuration. It is the tendency of high (closed) vowels to exhibit a higher fundamental frequency than low (open) vowels [244]. There are two hypothesis in place to account for the intrinsic pitch of vowels: *source-filter coupling* and *mechanical coupling*. The hypothesis for acoustic source-filter coupling was introduced in the preceding paragraph. This hypothesis was introduced by Lieberman *et al* [245], suggesting interaction between the vocal tract and the laryngeal system. Per the hypothesis, vocal tract configurations will affect the production of vowels. It is hypothesized that high (closed) vowels will exhibit a higher fundamental frequency than low (open) vowels. This hypothesis was in part a continuation of the findings by Flanagan *et al* [25]. For this, the authors used a self-oscillating, one-mass model of the vocal folds as an excitation source for a vocal tract synthesizer, in order to determine the influence of glottal parameters on the characteristics of a speech signal. The findings of this study show a correlation between pitch and sub-glottal pressure, vocal folds tension, and vocal tract configuration. It is reported that vowels with low first-formant frequency (high, closed) will exhibit higher fundamental frequency than vowels with high first-formant frequency (low, open). This hypothesis was reinforced in a study by Inbar *et al* [246]. The authors show a significant correlation between pitch period and vocal tract configuration via mathematical modelling. For this, they show variations in vocal tract configuration to cause changes in first-formant frequency, which in turn affects the fundamental frequency. However, multiple subsequent studies had challenged the affect of acoustic coupling on the intrinsic pitch [247, 248].

As mentioned above, intrinsic pitch of vowels is also attributed to mechanical coupling. Mechanical coupling attributes intrinsic pitch to coupling between the phonatory and articulatory systems. A mechanical coupling hypothesis was introduced by Ladefoged *et al* [249], referred to as the *tongue-pull hypothesis*. Per the tongue-pull hypothesis, tongue movement results in anterior pull on the thyroid cartilage, which influences the vocal folds. The hypothesis suggests that such tongue movement will result in increased stiffness and a subsequent increase in fundamental frequency. However, this hypothesis was inconsistent with physiological data and subsequently abandoned [244]. However, a new iteration of this hypothesis was proposed by Ohala *et al* [247, 248]. For this, the authors suggest that tongue movement can indirectly influence the tension of the vocal folds. This new hypothesis relates superior tongue pull to increase in vocal folds tension, which increase the fundamental frequency in the production of high vowels. Ohala *et al* [247, 248] found high vowels to observe higher fundamental frequency than low vowels (including mid-vowels). By holding the vocal tract formants as constant, the effect of tongue-pull on intrinsic pitch was investigated. Results suggest that the tongue-pull hypothesis is responsible for this intrinsic pitch of vowels. Sapir *et al* [244] notes that neither one of the hypothesis is sufficient to account for intrinsic pitch of vowels as data examining the influence of acoustic and mechanical coupling is scarce.

In the vowel investigation carried out in this thesis work, intrinsic pitch was observed for vowels. The results are presented in Table B.1 and Table B.6 in Appendix B. The results reinforce the reported trend in literature of high vowels having greater intrinsic pitch than low vowels [250, 251, 244, 252, 220]. For database A, the high vowels WHO'D, HEAD, and HEED result in higher pitch than the low vowel HARD. For database B (male speakers), significant vowel effect was observed. The result indicate the high vowel HEAD to have greater intrinsic pitch than the low vowels HAD and HARD. Moreover, it was found that the mid-high vowels HOD, HID, and HERD also showed greater intrinsic pitch than the low vowel HARD. For database A, the vowel HOOD had the highest pitch. Moreover, its pitch was significantly higher pitch than all other vowels. Similar findings were observed by Bier *et al* [220]. While Bier obtained those findings via an EGG study, we obtained those results using the glottal flow signal. This implies that vowel effect can be observed in a multitude signals for the same speech corpus. This shows the derived signals are complementary to each other. Per the tongue-pull hypothesis, superior tongue pull increases vocal folds tension, causing increase in fundamental frequency. Alternatively, increase in pitch for high vowels could be the result of source-filter coupling. For database A, the vowel WHO'D has significantly higher pitch than HEED. We are unable to comment on the interaction between the vowels, as they are both high vowels, produced with superior tongue position. In order to determine the interaction, the formants of the vowels will need to be analysed. Although high vowels are correlated with low first-formant frequency, it is possible that the result will indicate a significant margin between the first formants of the high vowels, accounting for this interaction. We note that similar findings with regards to high vowel interaction was observed by Shadle *et al* [250], where it was found that WHO'D is generally higher than HEED. Formant analysis will also provide a looking glass into the difference in formants between mid and low vowels, which could explain why the mid vowel HOOD is significantly higher in pitch than the high vowel HEED. If formant frequency is in fact the cause for intrinsic pitch, this reinforces the acoustic coupling hypothesis. If vocal tract formants (particularly the first formant) influence the rate of vocal folds vibration, this implies source-filter coupling, which puts the fundamental assumption of the source-filter model in question. Hence, complete source-filter decoupling is not achievable due to the dependency between the two components. Thus, we suggest the vowel differences for pitch could be attributed to both the source-filter coupling hypothesis and the mechanical coupling hypothesis. Further analysis is required in order to determine the governing factor for vowel effect.

In addition to pitch-vowel interaction, vowel effect was also observed for the open quotient parameter. The open quotient quantifies the width of the glottal pulse. In literature, the dependency of open quotient on vowel type has not yielded significant interaction for male speakers [252, 253]. Significant interaction was observed for long vowels by female speakers [252]. No vowel effect was observed for the female speakers in database B. This, in part, might be attributed to the small sample size. For male speakers, vowel differences were observed in database A. The results indicate larger open quotient measurements for the high vowels HEED, WHO'D, HOARD and HEAD compared to the low vowels HUD and HARD. Moreover, larger open quotient measurements for the mid vowels HID, HERD, HOD and HOOD was observed compared to the low vowels HUD and HARD. This implies that vowels with superior tongue position result in a glottal cycle during which the vocal folds are abducted for longer. Since significant vowel differences were only observed for male speakers, this limits the generalisation of our findings.

The speed quotient is a measure of glottal pulse skewness, relating vocal folds abduction to vocal folds adduction. The effect of vowels was examined in order to identify whether vocal tract configuration indirectly affects the motion of the vocal folds during speech production. The dependency of speed quotient on vowel type is little researched with inconclusive results. Chen *et al* [120] found no vowel effect for speed quotient. The lack of vowel effect on speed quotient was reported in other studies as well [254, 253, 255]. Marasek *et al* [252], however, reported significant differences in speed quotient across vowel groups. For this, the vowel HARD showed significantly lower speed quotient than the vowel HEED, HEAD, HOARD, and WHO'D. Speed quotient was shown to increase with vowel height. This is hypothesized to be related to the tongue-pull hypothesis, where the increased tongue height results in greater vocal folds tension, causing to glottal pulse skewness. Significant vowel differences for speed quotient were observed for database B (males). Although statistically significant vowel difference was also reporting for database A and database B (females), it was reputed by the *post hoc t-test*. These findings implicate the lack of vowel effect on the skewness of the glottal pulse. However, due to the small sample size, we are cautious to accept those findings. A larger sample size is required in order to establish glottal pulse shape independence from vowel effect. Since the normalized amplitude quotient resulted in longer vocal folds closing duration during vibration, we hypothesize that a larger sample size will indicate the effect of vowels on glottal skewness.

The normalized amplitude quotient quantifies vocal folds adduction. Vowel effect was examined to determine whether vowel groups influence the duration of the closing phase. Significant difference in vowels was also observed for the normalized amplitude quotient. For database A, the high vowels HEED, WHO'D, HEAD, and HOARD had significantly higher NAQ than the mid vowel HOD. For database b (females), the high-vowels WHO'D and HEAD also had significantly higher NAQ than the mid vowel HOD. Although statistically significant vowel difference was also reporting for database B (males), it was reputed by the *post hoc t-test*. For database B, the small sample size compared to database A could be a factor for the limited number of vowel differences. The vowel effect on the normalized amplitude quotient implies vowel production to influence the shape of the glottal pulse. The findings indicate that high vowels result is longer vocal adduction duration, indicating vocal folds inefficiency. Since NAQ is an amplitude-based parameter, vowel differences might be attributed to variations in vocal intensity across vowels. This statement can be further strengthened by considering the definition for NAQ. The normalized amplitude quotient is inverse to the fundamental frequency. The results show high vowels to produce higher fundamental frequency. One would then expect NAQ to reduce for high vowels. However, the results are contrary to this. Hence, it is highly likely that high vowels also affect the vocal intensity. Furthermore, the high vowels tongue position is either front or central, while the HOD vowel is characterized by posterior tongue position. Hence, we hypothesize that tongue position also plays an integral role in the production of vowels. This tongue effect can be considered further reinforcement to the tongue-pull hypothesis.

Jitter and shimmer are quantifying measurements for ventilatory and vocal folds stability during voice production. Prior studies had examined the influence of vocal tract configuration on glottal perturbations from vowel signals. Previous studies have shown significant perturbation-vowel interaction, where lower perturbation was found for high vowels compared to low vowels [164, 162, 256]. In this study, significant vowel differences for jitter and shimmer were found across multiple database. The vowel effect was sporadic across multiple vowels, with no unified trend observed. Hence, the perturbation-vowel effect is inconclusive. As glottal perturbations are the result of vocal folds asynchronous vibration, we did not expect the vowel effect to be significant. Orlikoff *et al* [257] indicates that under controlled vocal pitch and vocal intensity levels, jitter and shimmer do not show significant vowel effect. Considering the correlation between jitter and fundamental frequencies, Titze *et al* [258] notes jitter to be inversely related to pitch, with considerable perturbation at low fundamental frequencies. For database A the vowels HUD and HOD showed significantly greater jitter that the vowel HERD. This is an accordance to the claims by Titze, as the vowel HERD showed higher pitch than vowels HUD and HOD. An examination of the results showed the vowel HID had significantly higher jitter than the vowel HERD for databases A and B (males). This might be due to lip rounding effect, resulting in greater pitch perturbation for lip rounded vowels. We suggest that natural decrease in sound propagation due to lip rounding could account for pitch perturbations. The vowel effect of lip rounding on the glottal waveform was previously investigated by Bier *et al* [220]. We conclude by acknowledging that further examination irregular pitch and amplitude is required before vowel effect can be established. Quantifying vowel effect for perturbations is an elaborate task, as other factors could influence those measurements, such as the presence and movement of mucous on the vocal folds, mucousal wave propogation, turbulent airflow, and source-filter coupling [258].

7.3 Age-Vowel Effect

Following the investigation of age effect and vowel effect, analysis was performed to identify age-vowel effect. Significant age-vowel differences was observed solely for database A. This further supports our prior comments with regards to the small sample size for all other databases. Thus far, studies have reported significant age effect and/or significant vowel effect. However, we argue that a third analysis for the interaction of the two factors must take place in order to fully realize the extent of vowel impact. Significant age-effect was observed for both the open quotient and the jitter measurements. We believe this to be the first study investigating this effect, and therefore, it was not discussed in prior publications. The results imply that while vowel effect is present for different vowel types, the impact of vowels cannot be generalized across all data. If analysis is carried out between two speaker groups for different vowels, we cannot be completely certain that any effects observed between the two sets were due to age differences or vowel difference. Hence, an age-vowel analysis is required. If no significant age-vowel effect was is identified, then the behaviour of the glottal source can be generalized across vowels for the entire database, regardless of age. A significant age-vowel effect, however, indicates that the glottal pulse shape is not only dependent on the vowel type, but is it also codependent on the age of the speaker. For example, consider the following: if for open quotient analysis for a young speakers group results in the vowel HEED being significantly different than HOARD, if no age-vowel interactions were observed, the same relationship holds true for the old speakers group as well. However, if age-vowel effects were observed, the results cannot be generalized for both age groups, unless the generalization is reflected in the age-vowel findings. We conclude this section by highlighting the importance of, and need for, this type of analysis.

7.4 Signal Processing Issues

7.4.1 Speech Corpora Limitations

The following limitations regarding the results of this study should be noted. The dynamic range of the speech recordings must be considered in the analysis. Since the speech data was obtained by separate parties throughout the years, the audio files vary in dynamic range. Database A has a dynamic range of 10 K bits, out of a possible 64 K bits. Database B has a dynamic range of 63 K bits, out of a possible 64 K bits. Database C (Y, PE) has a dynamic range of 30 K bits, out of a possible 64 K bits. Database C (HE) has a dynamic range of 10 K bits, out of a possible 64 K bits. Database D (1955) has a dynamic range of 8 K bits, out of a possible 64 K bits. Database D (1992) has a dynamic range of 55 K bits, out of a possible 64 K bits. Considering the speech data obtained in the 1940s (HE group in database C) and 1950s (1955 data in database D), the lower dynamic range is attributed to the recording equipment. The microphones used at the time were band-limited. As a result, the speech signal contains less spectral information. For database A, the low dynamic range is also attributed to low-quality recording equipment, as the microphone used was a simple lapel-microphone. The low dynamic range for the aforementioned databases means the noise component has greater impact on the glottal extraction process. The presence of large noise component will have an impact on the signal-to-noise ratio. A low signal-to-noise ratio could have a detrimental affect on the extraction of the glottal parameters.

Secondly, the nature of the phonetic environment must be taken into account. The speech data for databases C & D include vowel tokens extracted from continuous speech. Hence, the vowel tokens were obtained from a phonetically rich environment compared to the hVd words data. Since the continuous speech recordings were collected as an interview, it is highly unlikely that the majority of the vowels were extracted from hVd words. As a result, vowels might exhibit co-articulation effects, which are usually mitigated using hVd words. When vowels are extracted from continuous speech, co-articulation effects could have an impact on the glottal shape. We note that a more systematic investigation is required.

7.4.2 Glottal Source Ripples

As highlighted throughout this document, the glottal flow waveform provides invaluable information with regards to the behaviour of the vocal folds vibratory cycle. It is well established that the vocal folds adduction and abduction movement can be characterized by glottal waveform's shape. However, the glottal waveform may also provide misleading information, which in turn could lead to incorrect assumption. One of the most documented glottal waveform perturbations manifest as *ripples*. This phenomenon had been observed for multiple glottal waveforms across all speech corpora used in this investigation.

In literature, glottal waveform ripples are most often referred to as *formant ripples*. As the name suggests, those ripples are directly correlated to the formant frequencies of the vocal tract. The inclusion of formant ripples in the source signal is directly correlated to the source-filter model. Per the source-filter model, the source and the vocal tract filter and linearly separable as the system is assumed to be time-invariant for short durations. In theory, source-filter decoupling allows extraction of the glottal source via elimination of the vocal tract contributions. In practice, however, source-filter decoupling poses a challenge. In the process of glottal source extraction via inverse filtering, the vocal tract is modelled and subsequently removed from the speech signal. The accuracy of the model is attributed in part to the order of the LPC analysis. For example, a low-order LPC analysis will underestimate the vocal tract contribution [97]. Failure to sufficiently remove vocal tract contributions will result in source-filter coupling. Hence, formant ripples are the result of imprinted vocal tract formant frequencies in the source signal.

Formant ripples, also referred to as *interaction ripples*, occur due to non-linear interaction between the glottal flow and the trans-glottal pressure [243]. Effectively, formant ripples are attributed to the carry of energy from previous glottal cycles. During the vocal folds vibration cycle, the vocal tract for cycle $i+1$ will contain excitation energy from the glottal pulse in the previous cycle, i. This exponentially decaying energy, carried from one cycle to the next, is the primary source for ripples [243, 176]. The ripples are superimposed on the glottal flow to result in perturbation. It can be readily seen that this non-linear interaction contradicts the linearity proposed by the source-filter model. Indeed, Fant *et al* [259] proposed to incorporate the ripples into the source-filter model. They proposed the modelling of the vocal tract during the closed glottal phase, during which there is no source-filter interaction, and the vocal tract is assumed time-invariant. However, linear prediction analysis needs also to be carried out over the closed phase. The authors note that linear prediction analysis applied over an entire pitch period will result in errors in terms of formant bandwidth estimates. This is due to the source-filter interaction during the open phase, for which the source is not independent of the vocal tract and will exhibit formant ripples [259, 260].

The glottal ripples affect the shape of the glottal waveform. Though formant ripples were attributed to the open phase of the glottal pulse, publication have shown ripple interaction for both the open and closed phases when examining glottal volume velocity waveforms [261, 262, 263, 264]. Childers *et al* [265] modelled the effect of the source-filter interaction. For a glottal source model, the authors show source-filter interaction to manifest as ripples in the opening phase of the glottal pulse. Fant *et al* [259] notes first-formant ripples may also be observed in other glottal source phases. For this, the authors note that a first-formant ripple may manifest as an increase or decrease in glottal slope during the closed phase. We note that all of the aforementioned publications attribute glottal ripples to first-formant coupling. Fant *et al* [259] notes that ripples are primarily due to the first-formant frequency, with the second and third formants having minimal to no effect.

Although glottal ripples are widely attributed to source-filter interaction, classification of ripples due to formants may lead to overlooking voice pathologies. When considering formant ripples in the closed phase, it's important to consider the voice quality of the speaker. Veeneman *et al* [266] observed ripples in the closed phase of Electroglottographic signals. The authors attempted to determine whether the glottal flow in the closed region was the result of incomplete glottal closure or vocal tract coupling. They note that hoarse voice exhibits ripples. They hypothesize the reason is the exclusion of zeros from the vocal tract filter model, citing the presence of vocal tract zeros during nasal vowel production. However, for hoarse speech, the authors conclude that in all likeliness, the occurrence of ripples is due to incomplete glottal closure and not first-format coupling. Thus, the ripples reflected the flow of glottal air during the closed phase. Fant *et al* [243] notes that incomplete glottal closure will have the opposite effect of a ripple. For this, the authors note that incomplete glottal closure will introduce a constant leakage of air, which in turn will have a damping effect on the formants oscillation, effectively smoothing the ripples. Such phenomenon could be observed in breathy phonation, where the glottis may only exhibits partial closure due to vocal fold bowing.

Thus far, we have discussed the detrimental effect of source-filter coupling. Another potential factor which could attribute to ripples is the *mucous bridge* (See Chapter 1, Section 6.2). The mucous bridge forms a strand between the vocal folds, effectively creating a small obstruction for glottal airflow. As it occurs during the vocal folds abduction phase, it could provide false information about the behaviour of the vocal folds. Such obstruction, although minor, could result in the following glottal flow pattern: as the vocal folds abduct, the glottal flow increases, followed by a secondary spike in glottal flow at the moment of mucous bridge collapse. Moreover, it could provide false information about the moment of glottal opening. Although the mucous bridge could have an effect on the shape of the glottal pulse, it is mostly observed in Electroglottography studies [267]. For the purposes of this study, we refrain from attributing ripples to mucous bridge effect without access to corresponding EGG data.

Throughout this thesis work, glottal ripples were observed in speech recordings from all databases. We are inclined to attribute formant ripples to source-filter interaction due to inverse filtering. As previously mentioned, formant ripples are the result of partial removal of vocal tract contributions. This issue is apparent in automatic inverse filtering. As the order of LPC analysis is set relative to other estimated parameters (i.e. pitch period), the vocal tract contributions could still be present subsequent to the analysis. We hypothesize that manual setting of the order of the LPC analysis will enable better estimation of the vocal tract contributions. However, a manual process is somewhat undesirable, as it could deter inexperienced users from using our toolbox. Moreover, manually setting the order of the LPC analysis could result in increased computation time. In addition, users might be tempted to increase the order of the analysis, while unaware of the potential errors that may arise due to over-fitting the vocal tract model. Hence, we conclude that manual inverse filtering is unfavoured for this toolbox. Alternative to manual involvement, it will be more beneficial to assess the quality of the glottal flow waveforms extracted via inverse filtering. One proposed improvement will be the implementation of group delay function to assess the quality of glottal flow waveforms estimated via inverse filtering, as proposed by Alku *et al* [268]. Group delay function is used to assess the inverse filter settings and how closely they correspond the vocal tract poles. Moreover, this algorithm addresses the issue of formant ripples, by considering the superior glottal flow estimate to have a flat closed phase. The authors note that ripples manifest as group delay distortion, effectively deeming the glottal waveform as non-ideal. Aside of the effect of formant ripples on the implementation of our toolbox, it can be readily seen that the inclusion of formant ripples can have detrimental impact on our results. The inclusion of formant ripples can lead to incorrect jitter and shimmer results, as well as incorrect pitch-synchronous analysis. A formant ripple of sufficiently high energy could potentially register as a glottal cycle, which will introduce a potential error/outlier to our results. As previously mentioned, formant ripples can register as false glottal opening instances. Hence, this could impact the result for the open quotient and speech quotient parameters, which depend on the exact moment of glottal opening. A proposed solution is to conduct analysis using glottal parameters which are not dependent on the glottal opening instances, such as the closing quotient (ClQ) or the normalized amplitude quotient.

7.4.3 The Importance of Polarity

The importance of speech polarity with regards to speech processing was made clear in the earlier stages of this thesis project. As a starting off point for this project, the open quotient criteria was computed for speech corpus A. The purpose of the investigation was two-fold: test the implementation of the algorithms and compare the results to the previous Electroglottography study by Philip *et al* [269]. The results matched the open quotient measurements obtained by Philip. However, the results suggests a decreasing open quotient for speaker of advanced age compared to young speakers. Those results were in contrast to the ones reported by Watson & Bier *et al* [147], and subsequently in [220]. These afflicting results led to the re-examination of our implementation. Subsequently, the polarity of the recorded speech data was examined. Through implementation of the RESKEW polarity detection algorithm [199], multiple polarity inversions were detected. Analysis of the open quotient for speech corpus A was repeated, with the results matching those reported for the Electroglottography studies. All other speech corpora used in this study exhibited polarity inversion to some degree. other database. In speech corpus B, speech data for both male and female speakers. The inversion in polarity could be attributed to multiple recording sessions. If the speech data was gathered over multiple sessions, it is possible that an inversion of the microphone connections took place. In speech corpus C, over 20% of the speech data required polarity correction (multiplication of the speech signal by -1). Since a partial subset of the data was collected from archival recordings made for radio broadcast, and considering the insensitivity of human perception to speech polarity, it can be assumed that no preventive measurements were put in place to ensure correct speech polarity. Similarly, speech corpus D consists of archival speech recordings made for broadcast radio. Multiple speech files were observed to exhibit polarity inversion. The importance of polarity detection was concurred in multiple publications [193, 196, 201]. Therefore, it follows to reason that polarity detection should be employed in all speech analysis studies. However, reflecting back on the literature utilized in this study, it was observed that not all publication mention the use of polarity detection. This is worrisome, as it compromises the integrity of the results. Moreover, it makes the comparison of results challenging, as neglecting to perform polarity checks can skew the results, as mentioned above. Hence, we conclude by emphasizing the importance of polarity checks and the need for reliable, robust, and widely-available polarity detection algorithms.

Chapter 8

Conclusion

8.1 What Was Achieved?

Voice analysis through glottal source processing had been documented in literature for over 50 years. Although numerous studies had been published, the number of freely-available vocal analysis packages is scarce. This is in part due to the complex nature of glottal analysis, which requires computationally complex operations. However, recent advancements allow for fast glottal source estimation and parametrization. Albeit those recent advancements, the number of glottal source processing software has not increased. This thesis work strived to narrow the gap between proprietary and freely available speech processing software. In this document, the design and implementation for a glottal analysis package in R was presented. Popular speech processing algorithms were implemented in R to allow for a comprehensive analysis of the glottal source. With this package in place, we have opened the door for free voice analysis platforms in free software. In turn, the platform put in place can be used as the foundation for the implementation of a variety of glottal processing algorithms. The advancement of senescence voice analysis via glottal processing was championed in this thesis work. The influence of ageing and vowel production on the glottal source was investigated, with our initial hypothesis put into the test. Our results yielded interesting findings. Ageing was found to have an affect on the glottal source signal, implying ageing corresponds to physiological changes in the human body.

Vowel production was also found to impact the glottal source signal. The findings suggest source-filter coupling, or mechanical coupling, are the cause for the change in glottal shape across vowels. This reinforced our initial hypothesis for only partial source-filter decoupling. Finally, this work presents the first investigation into age-vowel effects. Age-vowel interactions suggest the glottal source is susceptible to simultaneous influence by age and vowel type. Those findings act as a warning against generalization in voice analysis. The findings show the importance of comprehensive data analysis for speech corpora, implying greater care needs to be put forth in voice analysis. Next, a comprehensive summary of each chapter is presented, followed by our suggestions for future work.

8.2 Revisiting The Thesis

This thesis began by outlining the speech production mechanism from an anatomic and physiologic point-of-reference in Chapter 2. This chapter aimed to provide a comprehensive literature review of the sub-glottal, glottal, and supra-glottal systems. The role of the lungs in the respiratory system was introduced in Section 2.1. The anatomy of the airways (Section 2.1.1) and respiratory movements (Section 2.1.2) were introduced, paving the way for a discussion of the respiration mechanism (Section 2.1.3). We delved deeper into the pressure principles governing respiration through discussion of Boyle's and Pascal's Laws (Section 2.1.3.1). We discussed the larynx in Section 2.2. The functional use of the larynx in respiration and phonation was highlighted in Section 2.2.1. A detailed description of the laryngeal cartilages & joints (Section 2.2.2) and laryngeal muscles (Section 2.2.3) was provided and subsequently related to the vocal folds, which is the topic of Section 2.3. In this section we stressed the importance of the vocal folds to the speech production mechanism. In Section 2.3.1 we detailed the anatomic structure of the vocal folds and introduce the *cover-body* model, upon which vocal folds mass models are based. Section 2.3.2 provided a comprehensive overview of the vocal folds vibratory cycle, where we discussed vocal folds vibration theories, parametric models for vocal folds self-oscillation, pressure principals for sustained oscillation, Bernoulli's principle, and introduce the concept of a *mucousal wave*. The vocal tract was presented in Section 2.4. The vocal tract anatomy was described in Section 2.4.1. In Section 2.4.2, we equate the vocal tract to an acoustic filter. In this section we first introduced the concept of *formants*, *area function* through the vocal tract concatenated tube model, and the relation between the two concepts. Phonemes are introduced in Section 2.4.3, where we elaborated on the different vowel types and introduced the *coarticulation effect*. We concluded this chapter with Section 2.5, where we explored the development of glottal analysis methodology, such as high-speed imaging (Section 2.5.1) and Electroglottography (Section 2.5.2).

Next, in Chapter 3, we introduced the glottal flow waveform. In Section 3.1, we presented the time-domain representation of the glottal pulse and its corresponding phases and temporal points. Subsequently, in Section 3.2 we presented the size time-domain and amplitude-domain parameters used in this study. The parameters are: speed quotient (Section 3.2.1), open quotient (Section 3.2.2), normalized amplitude quotient (Section 3.2.3), pitch (Section 3.2.4), jitter (Section 3.2.5), and shimmer (Section 3.2.6). Each parameter was described and related back to the behaviour of the vocal folds. Section 3.2.2 includes discussion of the existing open quotient criteria (Section 3.2.2.1) and an introduction of the new open quotient criteria - OQsub50 (Section 3.2.2.2). This chapter concluded with Section 3.3, where vocal quality is first introduced. Description of vocal registers was given in Section 3.3.1, with four voice registers examined in detail: modal (subsection 3.3.1.1), hoarse (subsection 3.3.1.2), creaky (subsection 3.3.1.3) and breathy (subsection 3.3.1.4). The effect of ageing on the lungs (Section 3.3.2), larynx (Section 3.3.3), and vocal folds (Section 3.3.4) was presented next. Potential non-hereditary factors, and how they pertain to vocal pathologies, were described in Section 3.3.5. Finally, three acoustic measurement were explored with regards to ageing and gender differences: pitch (Section 3.3.6), jitter and shimmer (Section 3.3.7), and speaking rate (Section 3.3.8).

In Chapter 4, we presented speech from an engineering point-of-view. We began by describing the *source-filter* model in Section 4.1. We presented the source-filter model as a system diagram composed of an excitation signal (Section 4.1.1), glottal filter (Section 4.1.2), vocal tract filter (Section 1.3), and a radiation load (Section 4.1.4). Through time-domain and spectral-domain equation manipulation in Section 4.1.5, a discrete-time speech sample is modelled using the all-pole auto-regressive model. This theory breakdown allowed us to discuss linear prediction coding in Section 4.2. This section presented the principles behind the use of a linear predictor and the derivation of the *normal equations*. The subsequent subsections provided an in-depth examination on how those equations were solved. Section 4.2.1 examined the least-squares minimization technique. Section 4.2.2 presented the auto-correlation method, which resulted in the Yule-Walker equations, which were the topic of Section 4.2.3. Section 2.4 elaborated on the rigorous mathematical theory behind the solution to the Yule-Walker equations through the Levinson-Durbin recursion. Section 4.2.4.1 described the implementation of the recursion through pseudo-code. Next, an overview of the existing inverse filtering techniques employed today was presented. Section 4.3.1 included a summary of the principles and disadvantages of closed-phase inverse filtering. Section 4.3.2 presented the iterative adaptive inverse filtering algorithm. A detailed discussion of the algorithm, its advantages, and proposed improvements was given in great detail. The chapter concludes with Section 4.4, where speech signal polarity detection methods are reviewed.

Following, in Chapter 5 we explicitly detailed the methodology put in place to create our voice analysis toolbox. The discussion started off by introducing the R environment, and a description of the various data-structures was provided (Section 5.1.1). Moreover, we introduced the emuR library and its use for speech corpora handling (Section 5.1.1). Next, the voice analysis toolbox was described in Section 5.2. An overview of the existing speech processing packages was given in Section 5.2.1, and the COVAREP project, which was monumental to the design of this toolbox, was presented in Section 5.2.2. We elaborated on the framework of our toolbox in Section 5.2.3. This section included the methodology employed for speech analysis, from speech signal to glottal source estimation and glottal parameters extraction. The implementation of the pitch detection and OQsub50 algorithms was discussed in Section 5.2.4. With the framework in place, we introduced the speech corpora used in this study in Section 5.3. A detailed description of each speech corpus followed. Database A consisted of NZE citation-form hVd words collected from two age groups of male speakers (Section 5.3.1). Database B consisted of NZE citation-form hVd words collected from two age groups of male (Section 5.3.2.1) and female (Section 5.3.2.2) speakers. Database C consisted of /a:/ vowel tokens extracted from continuous speech in Māori, collected from archival recordings and modern-day interviews, for three age groups for male speakers (Section 5.3.3). Database D was formed as a longitudinal study and contained continuous speech samples collected 40 years apart from the same NZE male speaker (Section 5.3.4).

Chapter 6 presented the outcome of our vocal ageing investigation through glottal analysis. Similarly to Section 3 in the preceding chapter, we presented our results for each speech corpus separately. Each section of this chapter presented the glottal parameter measurements derived from its respective data source, as well the results of the subsequent statistical analysis carried out through ANOVA and *post hoc t-test* based on a 95% confidence interval. In Section 6.1 we presented the results for Database A. All six parameters showed significant age interaction. All parameters showed significant vowel interaction, bare the speed quotient parameter. The OQsub30 and jitter parameters exhibited significant age-vowel interaction. In Section 6.2 we presented the results for Database B. Results were separately presented for male and female speakers. For male speakers (Section 6.3.2.1), only the OQsub50 parameter exhibited significant age interaction. The speed quotient, jitter, and pitch parameters all exhibited significant vowel interaction. No significant age-vowel interactions were observed. For female speakers (Section 5.3.2.2), no significant age effects were observed. The normalized amplitude quotient, jitter, and shimmer parameters all exhibited significant vowel interaction. No parameters showed significant age-vowel interaction. Database C was presented in Section 5.3.3. Since it includes a single vowel, no vowel interaction analysis could be carried out. Significant age-interaction was observed for the OQsub50 and shimmer parameters. Finally, Database D was presented in Section 5.3.4. Significant age interaction was observed for the OQsub50, NAQ, and shimmer parameters. No significant vowel or age-vowel interaction was observed.

In Chapter 7, we presented a comprehensive analysis of the findings obtained in the previous chapter. Section 7.1 provides a detail examination of the ageing effect on the glottal features. The ageing effect on pitch is discussed in Section 7.1.1, where multiple trends were found, from which we concluded pitch to be correlated with ageing. The ageing effect on jitter and shimmer is discussed in Section 7.1.2, where shimmer was found to be an indicator for vocal ageing, while jitter produced inconclusive results. The ageing effect on open quotient is discussed in Section 7.1.3, where it was found to be a strong discriminator for senescence voice. The ageing effect on speed quotient is discussed in Section 7.1.4, with further analysis required in order to determine how senescence voice alters the glottal pulse skewness. The ageing effect on normalized amplitude quotient is discussed in Section 7.1.5, where it was found to be complementary parameter to the open quotient and speed quotient parameters. This investigation of normalized amplitude quotient with regards to ageing is the first multi-corpus analysis for this parameter. In Section 7.2 we discussed the implications of vowel effect on the glottal source signal. Our findings suggest source-filter coupling and tongue-pull are the key contributors for the impact of vowels on glottal shape. Moreover, we found that vowel type (high, low) affects the glottal signal as well. The age-vowel effect was discussed in Section 7.3. This was the first study to investigate this interaction. Our findings suggest glottal source production is simultaneously codependent on ageing and vowel type. This chapter concluded with discussion of the signal processing issues that were present in this analysis. This includes database limitations (Section 7.4.1), glottal ripples (Section 7.4.2), and polarity inversion (Section 7.4.3).

The original content of this thesis is hereby concluded. What follows are our proposed improvements and expansions for the voice analysis toolbox. We also propose multiple directions in which this work could be continued.

8.3 What Next?

The following list contains suggestion for the continuation of this project.

- **Additional parameters.** Six time- and amplitude-domain parameters were examined in this thesis. With regards to the time-domain parameters, we propose the integration of other parameters into the toolbox, such as the closed quotient (CQ) and closing quotient (ClQ). With the glottal extraction and parametrization framework in place, the extraction of those parameters is relatively straight forward. Integration of these parameters will allow examination into the relation between parameters. We have established the inverse relationship between the speed quotient (SQ) and the normalized amplitude quotient (NAQ). Both parameters are correlated with the closing quotient. Since the NAQ quotient is an amplitude-domain representation of the closing quotient, it will be of interest to compare the robustness of the parameters for various voice pathologies, as well as for speech data recorded under adverse conditions. Incorporating the CQ quotient into the toolbox will allow comparison with the Electroglottography parameter contact quotient (Qx) (this is further discussed below).
- **Spectral parameters.** For the scope of this thesis, the glottal signal was

analysed solely in the time-domain. We propose the implementation of algorithms for spectral parameters extractions, such as spectral tilt and the formant frequencies. The literature suggests that certain voice pathologies are more pronounced in the spectral domain. Those parameters could then be compared to the time-domain parameters, thus determining which parameter best correlates to a voice quality.

- **Direct all-pole (DAP) modelling.** In Chapter 4, we explored the proposed improvements to the iterative adaptive inverse filtering algorithm, as expressed in the literature. One improvement that is of great interest is the replacement of linear predictive coding (LPC) by direct all-pole modelling. As per the literature, DAP results in a glottal signal less distorted by format coupling. This could be highly beneficial for examining vowel interactions. By minimizing the glottal-vocal tract coupling via DAP, we might be able to determine whether vowel interaction is the result of source-filter coupling, or inherent anatomical changes in the vocal system (i.e. mechanical coupling). We note that the integration of DAP into the toolbox is highly recommended.

- **Māori & New Zealand English (NZE) Corpus.** As part of the speech corpora used in this thesis work, Māori speech was examined for age-induced changes. The speech data was collected by the MAONZE project to examine changes in the pronunciation of the Māori language. For this study, only a small subset of the data was analysed. There exist, however, speech recording in both Māori and New Zealand English from bilingual speakers. We propose a subsequent study, where the glottal signal from the same speaker is analysed for both the Māori and NZE speech. Such comparison will allow us to determine whether voice pathology manifests itself regardless of the language. Such examination could imply voice pathologies to be correlated with phonetic pronunciation.

- **Electroglottography comparison.** As discussed above, the implementation of the contact quotient parameter will allow cross-examinati-on of glottal signal with Electroglottographic signals. Such examination is favourable, as it will allow comparison of information about the behaviour of the vocal folds during phonation. Moreover, comparing the glottal signal to the electroglottogram signal will allow to examine the robustness of the glottal source extraction algorithm. We propose a subsequent study, utilizing speech corpus A. As mentioned in Chapter 5, this speech corpus was originally made for EGG data analysis. Comparison of the results will allow to examine the behaviour of the vocal folds, with the EGG data taken as a reference point. In addition, it could reinforce the importance of speech data quality for glottal extraction, as the speech data was recorded using a lapel microphone (poor quality).

8.4 Final Word

We have successfully created the first voice analysis toolbox in R, and our findings highlight the robustness of the glottal source extraction process. This work presents the first all-round, multi-corpus analysis of senescence detection using the glottal source signal. We have shown ageing and vowel type to be impactful factors on the production of the glottal signal. In addition, we performed the first age-vowel glottal analysis, highlighting the imporatance of an all-round voice analysis. Please note that the toolbox is available under open-source license and can be accessed using the following link: `www.github.com/itaybendom/GVV-Toolbox.git` Overall, we have contributed to the progression of glottal processing in voice analysis.

Appendices

Appendix A

Glottal Flow Parametric Models

The glottal flow can be separated from the speech signal through *source-filter deconvolution*, which is achieved through inverse filtering technique (next chapter.) Upon extraction of the glottal source, it is of great importance to fit a parametric model to the glottal flow in order to parametrize it. The following section is concerned with the parametrization of the glottal flow. This section describes several popular parametric model which are extensively featured in literature. Various glottal models have been proposed both in time-domain and frequency-domain, with most models characterizing the glottal flow in the time-domain. The models differ by the number of parameters used, with the glottal flow typically described by 3-5 parameters. One of the advantages of using a time-domain model is its applicable comparison to electroglottogram signal. Some established time-domain models are: Fant model [270, 271], Liljencrants-Fant (LF) model [84], Rosenberg model [272], Rosenberg++ model [273], and KLGLOTT88 (Klatt) model [74]. Note that the Liljencrants-Fant model is the most widely used model and is the one utilized in this thesis work. A description of the Rosenberg-C model and the Liljencrants-Fant model ensues. Readers interested in comprehensive analysis and comparison of the models are encouraged to refer to the following publications by Duval and d'Alessandro (1999, 2006) *et al* [67, 68].

A.1 Rosenberg Model

Rosenberg *et al* [272] developed a time-domain glottal source model to be used in speech synthesis. He devised six source models (A-F), with each model consisting of four parameters. These parameters are:

- A: amplitude
- T_0: fundamental period
- T_P: opening time (correlated with positive slope)
- T_N: closing time (correlated with negative slope)

The iterations of his model differ in their function description to fit a glottal flow waveform. The proposed models were then put to evaluation, where listeners were asked to compare between synthetic voice samples produced by different source models, and note which model was perceptually superior. Rosenberg notes that the preferred models were Rosenberg-B and Rosenberg-C. The models are continuous with a first-derivative discontinuity, and are polynomial and trigonometric functions, respectively. The Rosenberg-C model function is provided in Equation A.1 [272].

$$g(t) = d \begin{cases} \frac{A}{2}\Big(1 - cos(\pi\frac{t}{T_P})\Big), & 0 \leq t \leq T_P \\ Acos\Big(\frac{\pi}{2}\frac{t - T_P}{T_N}\Big), & T_P \leq t \leq T_P + T_N \\ 0, & T_P + T_N \leq t \leq T_0 \end{cases} \tag{A.1}$$

An illustration of the Rosenberg-C glottal pulse model function with varying parameter values is provided in Figure A.1. Six potential glottal pulse shapes are presented $(a - f)$. In pulse models *a, b, c*, the glottal pulse comprises 50% of the period ($\mathrm{T_N/T_0}$ = 0.5), with opening times of 0.5, 0.667, 0.75 percent, respectively. In models *d, e, f*, the pulse comprises 75% of the period ($\mathrm{T_N/T_0}$ = 0.75), with opening times of 0.5, 0.667, 0.75 percent, respectively. Rosenberg notes that parametric model functions where the glottal pulse is symmetric ($\mathrm{T_P}$

= T_N, Speed Quotient = 1) are not preferred as excitation sources (pulse models *a* and *d* in Figure A.1) [272].

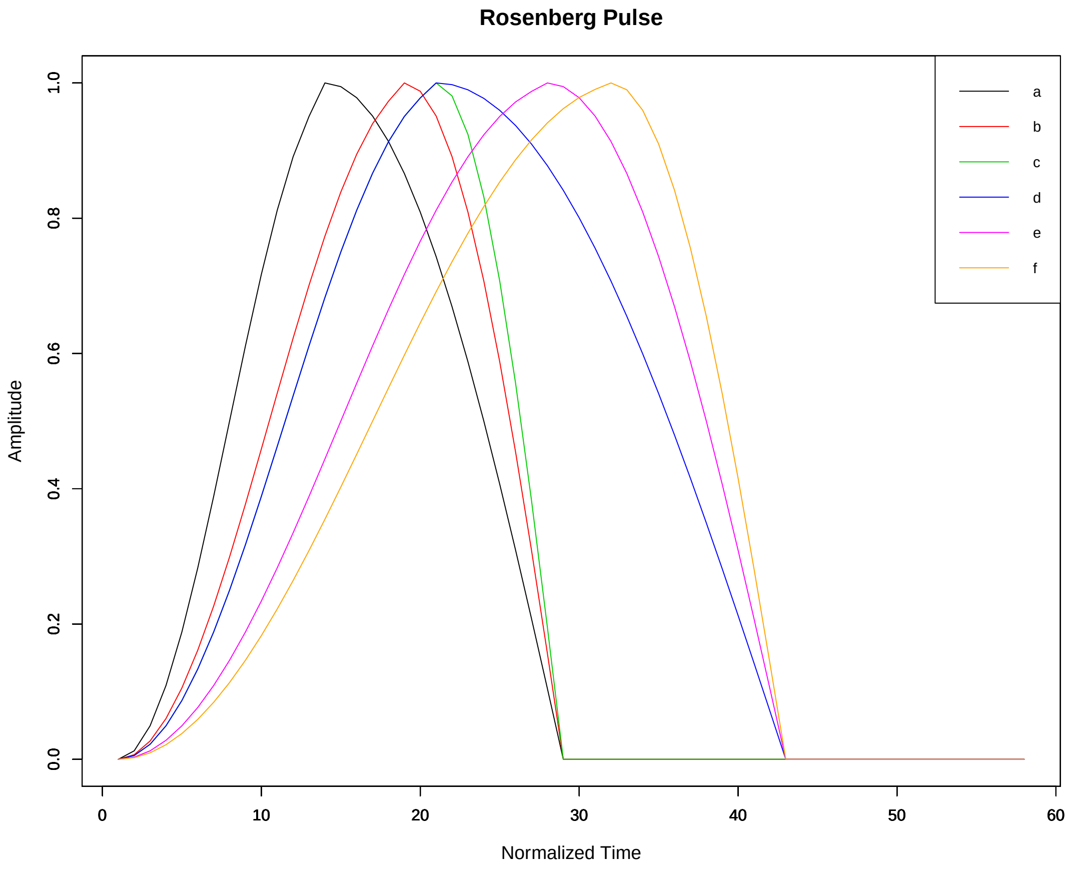

Figure A.1: *The Rosenberg-C glottal pulse model with varying glottal pulse skewness and width*

A.2 Liljencrants-Fant (LF) Model

The Liljencrants-Fant model is the most widely adopted representation for glottal source. The model builds upon the foundations of Fant's F-model and Liljencrants' L-model. Prior to delving into the LF-model, we digress to review the parametric representation of its foundation models.

The Fant model fits the glottal flow using two sinusoidal functions depicting the rising and falling branches. This representation exhibits a discontinuity at the instance of peak flow [270, 271]. The F-model parametric function is provided in Equation A.2.

$$g(t) = \begin{cases} \frac{1}{2}\Big(1 - cos(w_g t)\Big), & 0 \leq t \leq t_p \\ Kcos\Big(w_g(t - t_p)\Big) - K + 1, & t_p \leq t \leq t_c \\ 0, & t_c \leq t \leq T_0 \end{cases} \tag{A.2}$$

Where $w_g = \pi/t_p$, K is the asymmetry factor, t_p is the instant of maximum glottal flow, t_c is the instant of termination at zero, and T_0 is the fundamental period. Altering K changes the slope of the falling branch. When $K = 0.5$, the pulse is symmetrical, with the rising branch and the falling branch equal in duration. For $K > 0.5$, the falling branch reduces in duration and terminates abruptly. For $K = \infty$, the glottal pulse falling branch is in the form of a step function [270, 271]. This behaviour for altered K values is depicted in Figure A.2. Fant *et al* [84] notes that one of the major disadvantages of his F-model is the exclusion of a return phase for large K values, which implies that incomplete closure cannot be modelled for certain voice types (i.e. breathy). Subsequently, in the LF model, this problem was mitigated using an exponential function (see Equation A.4).

Liljencrants' L-model provides parametric representation for the glottal flow deriv-ative using three parameters. This continuous model is described in Equation A.3 [84].

$$\dot{g}(t) = E_0 e^{\alpha t} sin(w_g t) \tag{A.3}$$

Where E_0 is a scale factor (set to 1), α controls the rate of exponential amplitude increase, and $w_g = \pi/t_p$. This function is illustrated in Figure A.3. As evident from Figure A.3, and similar to the F-model, the L-model does not incorporate a return phase. Therefore, the L-model provides a parametric representation for the open phase only. As the return phase is excluded, the open phase is abruptly terminated and followed by the closed phase.

The L-model and F-model lead to the development of the LF-model. The LF-model better represents the behaviour of the vocal folds during speech production

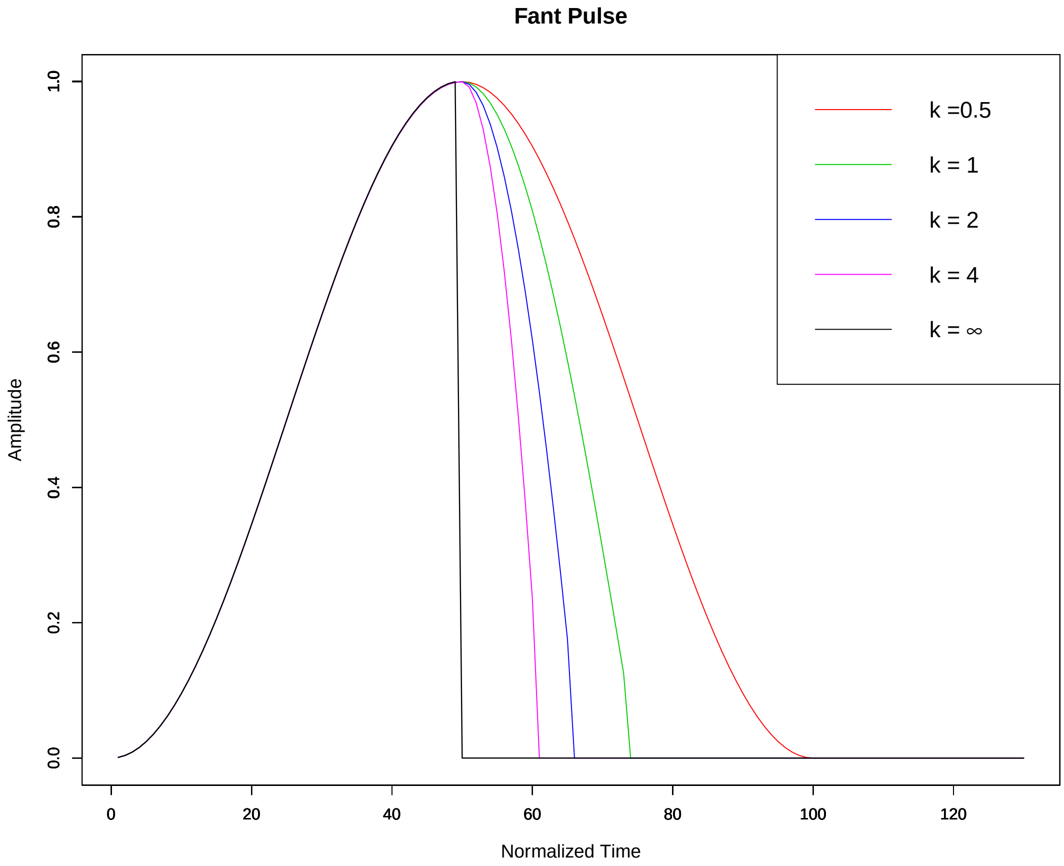

Figure A.2: *Fant glottal pulse model with varying K values*

by accounting for the return phase in the vibration cycle. Similar to the L-model, the LF-model fits a parametric function for the glottal flow derivative. The glottal flow derivative is used as it is simpler to model compared to the glottal flow [274]. Hence, the glottal flow is the integral of the LF-model glottal flow function. The LF-model function is provided in Equation A.4 [84].

$$\dot{g}(t) = \begin{cases} E_0 e^{\alpha t} sin(w_g t), & 0 \leq t \leq t_e \\ -\dfrac{E_e}{\epsilon t_a}\Big[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}\Big], & t_e \leq t \leq t_c \\ 0, & t_c \leq t \leq T_0 \end{cases} \tag{A.4}$$

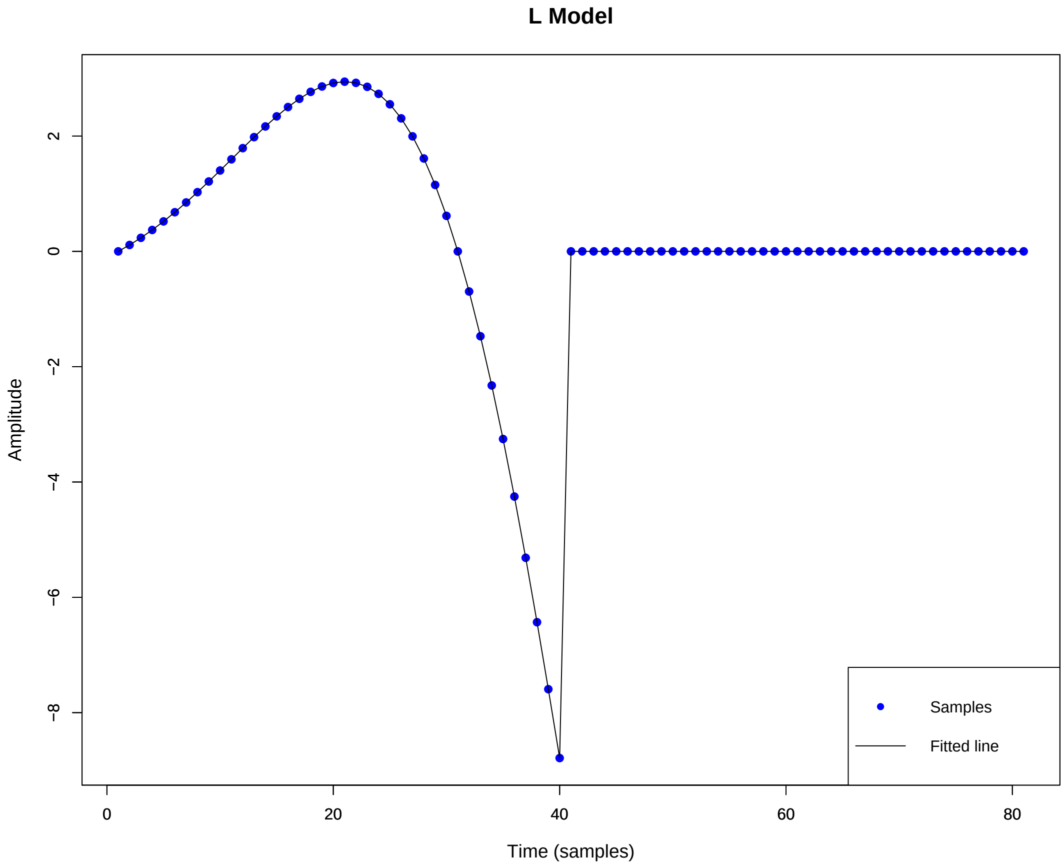

Figure A.3: *Liljencrants model*

Fant *et al* [84] describes the LF-model as a four parameter model. Fant *et al* [275], however, describes the LF-model as a five parameter model [275]. In general, researchers define the LF-model by different sets of parameters, typically between four and six parameters [84, 275, 276]. A summary of the LF-model parameters (see Equation A.4) and their definitions ensues below.

- E_0: a scale factor (set to 1)
- α: controls the rate of exponential amplitude increase
- $w_g = \pi/t_p$

- $E_e = -E_0 e^{\alpha t} sin(w_g t)$; maximum negative amplitude value of $\dot{g}(t)$
- ϵ:

$$\epsilon = \begin{cases} \dfrac{1}{t_a}, & \text{for small } t_a \\ 1 - e^{-\epsilon(t_c - t_e)}, & \text{otherwise} \end{cases}$$

- t_p: instance of maximum glottal flow
- t_e: instance of maximum glottal closure; corresponds to maximum negative amplitude in glottal derivative waveform ($\dot{g}(t_e) = -E_e$)
- t_c: instance of termination at zero; for simplicity $t_c = T_0$ [84]
- t_a: time constant; effective duration of return phase; measures the abrupt termination of the return phase

The timing parameters of the model are $t_p,\ t_e,\ t_c,\ t_a$. The synthesis parameters of the model are $E_0,\ \alpha,\ w_g,\ E_e,\ \epsilon$. These timing parameters, as well as T_0 and E_e, can be extracted from the source waveform derivative (Note: $\dot{g}(t_e) = -E_e$). The remaining synthesis parameters $\alpha,\ \epsilon,\ E_e$ can be computed by ensuring area balance (zero net gain of flow) throughout the glottal pulse cycle, per Equation A.5 [84].

$$\int_0^{T_0} E(t) = 0 \tag{A.5}$$

The LF-modelled glottal pulse is given in Figure A.4. As evident from both Equation A.4 and Figure A.4, if $t_a = 0$, the return phase is negated and the LF-model mirrors the L-model (Figure A.3) [84].

The LF-model consists of two functions:

1. *Exponentially increasing sinusoidal function*: This sinusoidal function is of frequency w_g and bandwidth α, and is scaled by a constant, E_0. This function models the glottal pulse derivative's open phase (opening + closing phase). The opening phase extends from the opening instance of the vocal

(in this case $t_open = 0$) to the instance of maximum glottal flow, t_p. The opening phase corresponds to increasing glottal flow, i.e. negative damping [84]. The closing phase extends from t_p to the instant of glottal closure, t_e. At the latter instance, the amplitude of the waveform has the value $-E_e$.

2. *Decaying exponential function*: This function models the return phase. The return phase extends from instance t_e to instance t_c. The time constant t_a is effectively the duration between those aforementioned instances. For simplicity, t_c is often equated to the fundamental period, T_0 [84].

Although the LF-model has become the reference for parametric representation of the glottal source, it has inherent problems. The LF-model requires solving non-linear equations for every new iteration of parameters, which increases its complexity and cost. Moreover, some of its synthesis parameters, as well as the t_a time parameter, do not correspond to physiological behaviour during voice production [273, 277]. As a result, advancements and improvements had been suggested. Fant *et al* [278] presented the *transformed LF* algorithm. The transformed model is a re-parametrization of the LF-model, with reduced number of parameters. Fant suggests the replacement of the time parameters with R parameters for added simplicity. As a result, in the transformed LF-model, the curve is parametrized by a single parameter, R_d. Fant argues that R_d is the most effective measure for describing voice quality using a single value [278]. A more computationally and cost effective model was presented in the form of the *Rosenberg++* model, also known as the R++ model. The Rosenberg++ model offers a replacement to the complex LF-model by a new embodiment of the Rosenberg model. The R++ model offers an extension to the Rosenberg model in the form of a return phase. In addition, it is more computationally efficient compared to the LF model, with less data dependency and shorter buffering time, thus requiring shorter processing time [273]. It is noted that the R++ model is perceptually equivalent to the LF-model when used for speech synthesis. Various other improvements and embodiments to the new models have been put forth throughout the years. Recently, a *LF-Rosenberg hybrid model* was presented [279].

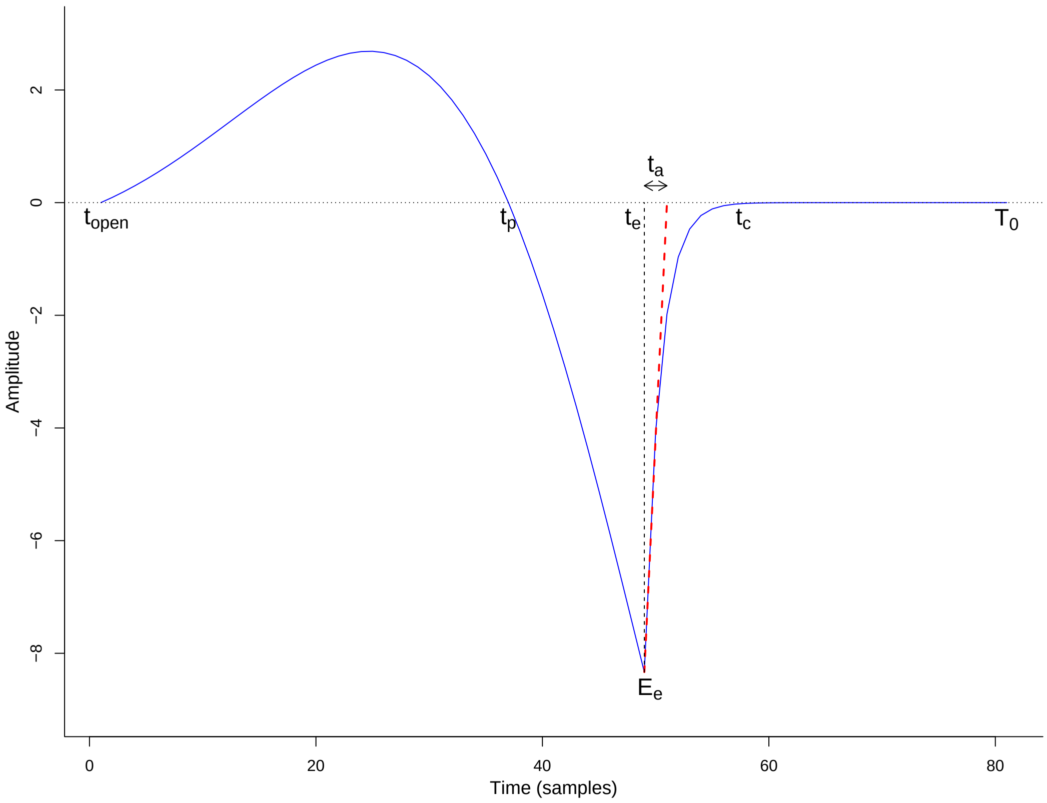

Figure A.4: *Lijencrants-Fant model*

Appendix B

Statistics

This appendix includes the *post hoc t-test* vowel interaction results for databases A, B (males), and D.

B.1 Database A

Table B.1: Database **A**: F0 vowel interaction

		post-hoc analysis	
vowels	*vowel interaction*	t-test	p-value
U	> {	$t(29) = 7.7$	***
	> }:	$t(29) = 5.0$	**
	> 3:	$t(29) = 6.2$	***
	> e	$t(29) = 6.2$	***
	> 6	$t(29) = 6.2$	***
	> 6:	$t(29) = 7.4$	***
	> I:	$t(29) = 4.2$	*
	> i:	$t(29) = 7.6$	***
	> o:	$t(29) = 7.0$	***
	> O:	$t(29) = 6.1$	***
}:	> {	$t(29) = 7.2$	***
	> 6:	$t(29) = 6.7$	***
	> e	$t(29) = 3.7$	*
	> i:	$t(29) = 5.1$	***
	> O	$t(29) = 5.7$	***
	> o:	$t(29) = 3.8$	*
e	> {	$t(29) = 6.0$	***
	> O	$t(29) = 3.7$	*
	> 6:	$t(29) = 4.9$	**
3:	> {	$t(29) = 5.9$	***
	> 6:	$t(29) = 5.9$	***
i:	> {	$t(29) = 4.1$	*

Note: *p<0.05 , **p<0.01 , ***p<0.001

Table B.2: Database **A**: OQsub30 vowel interaction

		post-hoc analysis	
vowels	*vowel interaction*	*t*-test	*p*-value
U	> 6	$t(29) = 4.9$	**
}:	> {	$t(29) = 4.8$	**
	> 6	$t(29) = 7.3$	***
	> 6:	$t(29) = 6.5$	***
	> o:	$t(29) = 4.2$	*
	> U	$t(29) = 4.0$	*
i:	> {	$t(29) = 4.5$	**
	> 6	$t(29) = 6.6$	***
	> 6:	$t(29) = 5.8$	***
	> o:	$t(29) = 3.8$	**
O	> {	$t(29) = 4.3$	**
	> 6	$t(29) = 5.5$	***
	> 6:	$t(29) = 6.5$	***
3:	> 6	$t(29) = 7.0$	***
	> 6:	$t(29) = 5.0$	**
e	> 6	$t(29) = 4.8$	**
	> 6:	$t(29) = 4.0$	**
I	> 6	$t(29) = 6.0$	***
	> 6:	$t(29) = 4.7$	**
o:	> 6	$t(29) = 5.4$	***

Note: *p<0.05 , **p<0.01 , ***p<0.001

Table B.3: Database **A**: NAQ vowel interaction

		post-hoc analysis	
vowels	*vowel interaction*	t-test	p-value
}:	> {	$t(29) = 3.7$	*
	> 3:	$t(29) = 3.9$	*
	> 6	$t(29) = 4.4$	**
	> 6:	$t(29) = 4.0$	**
	> I	$t(29) = 4.4$	**
	> O	$t(29) = 4.9$	**
	> U	$t(29) = 3.8$	*
i:	> {	$t(29) = 4.8$	**
	> 3:	$t(29) = 4.7$	**
	> 6	$t(29) = 5.1$	***
	> 6:	$t(29) = 5.1$	***
	> e	$t(29) = 4.4$	**
	> I	$t(29) = 5.2$	***
	> O	$t(29) = 5.6$	***
	> o:	$t(29) = 4.1$	*
	> U	$t(29) = 5.0$	**
e	> O	$t(29) = 4.6$	**
o:	> O	$t(29) = 4.4$	**

Note: *p<0.05 , **p<0.01 , ***p<0.001

Table B.4: Database **A**: Jitter vowel interaction

		post-hoc analysis	
vowels	*vowel interaction*	t-test	p-value
6	$>$ 3:	$t(29) = 4.2$	*
	$>$ o:	$t(29) = 3.7$	*
O	$>$ 3:	$t(29) = 4.7$	**

Note: *p<0.05 , **p<0.01 , ***p<0.001

Table B.5: Database **A**: Shimmer vowel interaction

		post-hoc analysis	
vowels	*vowel interaction*	t-test	p-value
3:	$<$ 6	$t(29) = 4.5$	**
	$<$ e	$t(29) = 3.9$	*
	$<$ I	$t(29) = 4.9$	**
	$<$ O	$t(29) = 5.8$	***
	$<$ o:	$t(29) = 5.6$	***

Note: *p<0.05 , **p<0.01 , ***p<0.001

B.2 Database B (Male)

Table B.6: Database **B** (Male): Pitch vowel interaction

		post-hoc analysis	
vowels	*vowel interaction*	t-test	p-value
A	< E	$t(14) = 7.38$	***
	< I	$t(14) = 4.47$	*
	< O	$t(14) = 7.98$	***
	< @:	$t(14) = 5.56$	**
a:	< E	$t(14) = 4.71$	*
	< O	$t(14) = 4.64$	*

Note: *p<0.05 , **p<0.01 , ***p<0.001

B.3 Database D

Table B.7: *Mean OQsub50 & NAQ per vowel*

	Age			
	1955		1992	
	OQsub50	*NAQ*	*OQsub50*	*NAQ*
I	0.234	0.207	0.301	0.218
U	0.165	0.171	0.302	0.229
E	0.287	0.267	0.247	0.211
O	0.223	0.191	0.260	0.216
V	0.258	0.211	0.248	0.203
A	0.275	0.215	0.255	0.219
i:	0.281	0.221	0.277	0.199
u:	0.271	0.211	0.262	0.204
o:	0.258	0.205	0.262	0.206
a:	0.265	0.210	0.284	0.201
@:	0.284	0.187	0.255	0.215

Appendix C

OQsub50 R Code

This appendix includes the OQsub50 function. This function is part of the glottal source voice analysis toolbox which is available under open-source license here: `www.github.com/itaybendom/GVV-Toolbox.git`.

```
OQsub50_method <- function(glottalList, threshold = 0.5){

# INPUTS

glottalList   List of glottal waveform, valleys, peaks
threshold     Amplitude threshold (used in GOI detection)

# First extract all the applicable information from the list

  winwave = glottalList$glottal  # glottal waveform
  valIdx  = glottalList$valIdx   # valley instances
  valAmp  = glottalList$valAmp   # valley amplitudes
  peakIdx = glottalList$peakIdx  # peaks instances
  peakAmp = glottalList$peakAmp  # peak amplitudes

# Next step is to set up a data-frame object containing all
    the information required for OQsub50 calculations.
```

```
# Create dataframe first 2 columns, indicating the start and
   end instances of each glottal cycle in the waveform.

  startIdx = list()
  endIdx = list()
  for (i in 1:(length(valIdx) - 1)) {
    startIdx[i] = valIdx[i]
    endIdx[i] = valIdx[i + 1]
  }

  # Create dataframe object
  output = data.frame(unlist(startIdx),
                      unlist(endIdx),
                      unlist(peakIdx)
                      )
  names(output)[1:3] = c("Start␣Idx", "End␣Idx", "Peak␣Idx")

  # The threshold value per cycle is determind by multiplying
     the threshold by the peak-to-peak amplitude value, and
     adding that value to the lowest valley.
  threVal = c()
  for (i in 1:length(peakAmp)) {
    threVal = c(threVal,  min(valAmp[i], valAmp[i + 1]) +
    (peakAmp[i] - min(valAmp[i], valAmp[i + 1])) * threshold)
  }

  # Add threshold amplitude column to data-frame
  output[, 4] = threVal
  names(output)[4] = "Threshold"

  # GOI and GCI processing

  GOI_List = list()
  GCI_List = list()
  rmIdx = c()

  for (i in 1:nrow(output)) {
    flag = c()
    # Compute GOI location
    GOItmp  = GOI_instant(output[i, ], winwave,  i)

    if (length(GOItmp > 0)) {
      GOI_List[i] = GOItmp
    } else{
      rmIdx = c(rmIdx, i)
      flag = 1
    }

    # Compute GCI location
    GCItmp = GCI_instant(output[i, ], winwave,  i)

    if (length(GCItmp > 0)) {
      GCI_List[i] = GCItmp
    } else{
      if (flag != 1)
        rmIdx = c(rmIdx, i)
    }
    rm (GOItmp, GCItmp, flag)
  }

  # Remove unused cycles from dataframe and GOI/GCI list
  if (length(rmIdx) > 0) {
    output = output[-rmIdx, ]
    GOI_List = GOI_List[-rmIdx]
    GCI_List = GCI_List[-rmIdx]
  }
  # Add GOI and GCI to dataframe object, GOI and GCI
     correspond to sample index of GVV waveform
  output[, 5] = unlist(GOI_List)
  output[, 6] = unlist(GCI_List)
  names(output)[5:6] = c("GOI", "GCI")

  ###########################
  # OPEN QUOTIENT Calculation
  ###########################

  # Compute OQ measurements via:
  # OQ = (time between GCI and GOI)/(time between consequtive
      GOIs)
  # q1 = GCI - GOI
  # cycleT = valIdx[i+1] - valIdx[i]
  # OQ = q1 / cycleT

  # Compute OQ and add measurements to dataframe:
  for (i in 1:nrow(output)) {
    OQ = (output$GCI[i] - output$GOI[i]) /
     (output$`End Idx`[i] - output$`Start Idx`[i])
     output[i, 7] = OQ
    rm(OQ)
  }
  names(output)[7] = "OQ"

  return(list(output$OQ))
}
```

References

[1] I. R. Titze, *Principles of voice production.* Prentice-Hall Inc., 1994.

[2] K. N. Stevens, *Acoustic Phonetics.* MIT Press, 2000.

[3] J. B. West, *Respiratory Physiology: The Essentials.* Lippincott Williams & Wilkins, 9th ed., 2012.

[4] H. Pulakka, "Analysis of human voice production using inverse filtering, high-speed imaging," Master's thesis, Helsinki University of Technology, Finland, 2005.

[5] M. Karjalainen, *Kommunikaatioakustiikka.* 2000.

[6] J. Kreiman and D. Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*, ch. 2, pp. 25–71. Wiley-Blackwell, 2011.

[7] E. R. Weibel, *Morphometry of the Human Lung.* Springer Berlin Heidelberg, 1963.

[8] G. J. Tortora and B. Derrickson, *Principles of Anatomy and Physiology.* USA: John Wiley & Sons, 12th ed., 2009.

[9] P. Lieberman and S. E. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics.* Cambridge: Cambridge University Press, 1988.

[10] T. J. Hixon, *Respiratory Function in Speech and Song.* Boston College-Hill, 1987.

[11] National Center for Voice and Speech, *Lung Pressure and Power: A Potpourri of Topics.* Available at `http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/lung.html`.

[12] K. Hayward, *Experimental Phonetics.* Harlow, Essex, England: Longman Linguistics Library, 2000.

[13] C. H. (StudyBlue), *Cartilages of Larynx*. Available at `https://www.studyblue.com/notes/note/n/hs-lec-30-larynx/deck/12529184`. Accessed: 10-02-2017.

[14] H. Hirose, "Investigating the Physiology of Laryngeal Structures," in *The Handbook of Phonetic Sciences* (J. L. W. J. Hardcastle and F. E. Gibbon, eds.), ch. 4, pp. 130–152, Oxford: Blackwell Publishing Ltd., 2nd ed., 2010.

[15] V. E. Negus, *The Comparative Anatomy and Physiology of the Larynx*. London: Heinemann, 1949.

[16] R. Vashishta, *Larynx Anatomy*. Available at `http://emedicine.medscape.com/article/1949369-overview`.

[17] R. A. Clark and B. Simpson, *Operative Techniques in Laryngology*. Springer-Verlag, 2008.

[18] J. M. Yoffey, "Respiratory System," in *Textbook of Human Anatomy* (W. J. Hamilton, ed.), pp. 297–328, London: Palgrave Macmillan UK, 1976.

[19] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson, "A Review of Lumped-Element Models of Voiced Speech," *Speech Commun.*, vol. 55, no. 5, pp. 667–690, 2013.

[20] R. Mittal, B. Erath, and M. W. Plesniak, "Fluid-dynamics of human phonation and speech," *Ann. Rev. Fluid Mech.*, vol. 45, p. 437467, 2013.

[21] B. H. Story, "An overview of the physiology, physics and modeling of the sound source for vowels," *Acoust. Sci. & Tech*, vol. 23, pp. 195–206, 2002.

[22] M. Hirano, "Morphological Structure of the Vocal Cord as a Vibrator and its Variations," *Folia Phoniat*, vol. 26, no. 2, pp. 89–94, 1974.

[23] B. H. Story, *Mechanisms of Voice Production*, ch. 3, pp. 34–58. John Wiley & Sons, Inc, 2015.

[24] H. J. Rubin, "The neurochronaxic theory of voice productiona refutation," *A.M.A. Archives of Otolaryngology*, vol. 71, no. 6, pp. 913–920, 1960.

[25] J. Flanagan and L. Landgraf, "Self-oscillating source for vocal-tract synthesizers," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 1, pp. 57–64, 1968.

[26] K. Ishizaka and J. L. Flanagan, "Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords," *Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[27] I. R. Titze, "The Human Vocal Cords: A Mathematical Model, Part I," *Phonetica*, vol. 28, p. 129170, 1973.

[28] I. R. Titze, "The Human Vocal Cords: A Mathematical Model, Part II," *Phonetica*, vol. 29, pp. 1–21, 1974.

[29] F. Alipour, D. A. Berry, and I. R. Titze, "A finite-element model of vocal-fold vibration," *The Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3003–3012, 2000.

[30] I. R. Titze, "Comments on the myoelastic-aerodynamic theory of phonation," *J. Speech Hear. Res.*, vol. 23, no. 3, pp. 495–510, 1980.

[31] C. H. Shadle, *The Aerodynamics of Speech*, pp. 39–80. Blackwell Publishing Ltd., 2010.

[32] B. Story and I. R. Titze, "Voice simulation with a bodycover model of the vocal folds," *The Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1249–1260, 1995.

[33] R. S. McGowan and M. S. Howe, "Comments on single-mass models of vocal fold vibration," *J. Acoust. Soc. Am.*, vol. 127, no. 5, p. 215221, 2010.

[34] I. R. Titze, "The physics of small-amplitude oscillation of the vocal folds," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1536–1552, 1988.

[35] T. Baer, *Investigation of phonation using excised larynges.* Phd dissertation, Massachusetts Institute of Technology, Boston, MA, 1975.

[36] R. Laje, T. Gardner, and G. B. Mindlin, "Continuous model for vocal fold oscillations to study the effect of feedback," *Phys. Rev. E*, vol. 64, p. 056201, Oct 2001.

[37] G. Fant, *Acoustic Theory of Speech Production.* Mouton, The Hague, Netherlands, 1960.

[38] K. Honda, *Supralaryngeal Articulators in the Oropharyngeal Region*, ch. 4, pp. 59–78. John Wiley & Sons, Inc, 2015.

[39] P. H. van Lieshout, *Jaw and Lips*, ch. 5, pp. 79–108. John Wiley & Sons, Inc, 2015.

[40] B. H. Story, "A comparison of vocal tract perturbation patterns based on statistical and acoustic considerations," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. EL107–EL114, 2007.

[41] G. Fant and S. Pauli, “Spatial characteristics of vocal tract resonance modes,” *Proc. Speech Comm. Sem*, vol. 74, pp. 1–3, 1974.

[42] L. Boé and P. Perrier, “Comments on “distinctive regions and modes: A new theory of speech production” by M. Mrayati, R. Carré and B. Guérin,” *Speech Commun.*, vol. 9, pp. 217–230, Sept. 1990.

[43] B. H. Story, I. R. Titze, and E. A. Hoffman, “Vocal tract area functions from magnetic resonance imaging,” *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[44] J. M. Heinz and K. N. Stevens, “On the derivation of area functions and acoustic spectra from cinradiographic films of speech,” *The Journal of the Acoustical Society of America*, vol. 36, no. 5, pp. 1037–1038, 1964.

[45] S. Maeda, *Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model*, pp. 131–149. Dordrecht: Springer Netherlands, 1990.

[46] S. Kiritani, Y. Tateno, and T. Iinuma, “Computer tomography of the vocal tract,” in *Dynamic Aspects of Speech Production* (M. Sawashima and F. S. Cooper, eds.), pp. 203–206, Tokyo: University of Tokyo press, 1977.

[47] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, “Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels,” *The Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 799–828, 1991.

[48] C. I. Watson, C. W. Thorpe, and X. B. Lu, “A Comparison Of Two Techniques That Measure Vocal Tract Shape,” *Australia Acoustics*, vol. 37, no. 1, pp. 7–11, 2009.

[49] C. I. Watson and C. J. Hui, “Two Short Studies in Vocal Tract Measurements,” in *Proceeding of 13th Australasian International Conference on Speech Science and Technology*, pp. 9–12, ASSTA, 2010.

[50] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing.* Upper Saddle River, NJ, USA: Prentice Hall Press, 1st ed., 2010.

[51] T. F. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice.* Upper Saddle River, NJ, USA: Prentice Hall Press, 1st ed., 2001.

[52] Wikipedia, “Ipa vowels — Wikipedia, the free encyclopedia,” 2017. [Online; accessed 10-Feb-2017].

[53] J. Harrington, *Acoustic Phonetics*, pp. 81–129. Blackwell Publishing Ltd., 2010.

[54] P. Moore, "Ultra high speed photography in laryngeal research," *Canadian journal of otolaryngology*, vol. 4, no. 5, p. 793799, 1975.

[55] S. Hertegard, H. Larsson, and T. Wittenberg, "High-speed imaging: Applications and development," *Logop. Phoniatr. Voco.*, vol. 28, pp. 133–139, 2003.

[56] S. Kiritani, K. Honda, H. Imagawa, and H. Hirose, "Simultaneous high-speed digital recording of vocal fold vibration and speech signal," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, vol. 11, pp. 1633–1636, Apr 1986.

[57] S. Kiritani, K. Honda, H. Imagawa, and H. Hirose, "Vocal cord vibration and voice source characteristics observations by a high-speed digital recording," in *Spoken Language Processing, International Conference on ICASSP '90.*, vol. 11, pp. 61–64, 1990.

[58] R. H. Colton and E. G. Conture, "Problems and pitfalls of electroglottography," *Journal of Voice*, vol. 4, no. 1, pp. 10–24, 1990.

[59] R. J. Baken, "Electroglottography," *Journal of Voice*, vol. 6, no. 2, pp. 98–110, 1991.

[60] D. G. Childers and J. N. Larar, "Electroglottography for Laryngeal Function Assessment and Speech Analysis," *IEEE Transactions on Biomedical Engineering*, vol. BME-31, pp. 807–817, Dec 1984.

[61] E. Ma and A. Love, "Electroglottographic Evaluation of Age and Gender Effects during Sustained Phonation and Connected Speech," *Journal of Voice*, vol. 24, no. 2, pp. 146–152, 1992.

[62] S. D. Bier and C. I. Watson, "Electroglottographic Analysis of Young and Old Speakers of New Zealand English," in *Proceeding of 14th Australasian International Conference on Speech Science and Technology*, (Macquarie University, Sydney, Australia), pp. 65–68, ASSTA, Dec 2012.

[63] D. G. Childers, D. M. Hicks, G. P. Moore, I. Eskenazi, and A. L. Lalwani, "Electroglottography and Vocal Fold Physiology," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 245–254, 1990.

[64] M. Rothenberg, "Some relations between glottal air flow and vocal fold contact area," in *Proceeding of the conference on the assessment of vocal pathology. ASHA Reports Vol 11*, pp. 88–96, ASHA Rockville MD, 1981.

[65] M. J. Ball and R. D. Kent, *Voice quality measurement.* San Diego : Singular Pub. Group, 1999. Available at: `http://www.laryngograph.com/papers.htm`.

[66] R. E. Orlikoff, "Assessment of the Dynamics of Vocal Fold Contact From the ElectroglottogramData From Normal Male Subjects," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 5, pp. 1066–1072, 1991.

[67] B. Doval and C. d'Alessandro, "The spectrum of glottal flow models," *Notes et documents*, vol. LIMSI 99-07, 1999.

[68] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[69] R. Timcke, H. von Leden, and P. Moore, "Laryngeal vibrations: Measurements of the glottic wave: part I. the normal vibratory cycle," *A.M.A. Archives of Otolaryngology*, vol. 68, no. 1, pp. 1–19, 1958.

[70] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.

[71] R. Timcke, H. von Leden, and P. Moore, "Laryngeal vibrations: Measurements of the glottic wave: part iiphysiologic variations," *A.M.A. Archives of Otolaryngology*, vol. 69, no. 4, pp. 438–444, 1959.

[72] J. Mahshie and A. Oster, "Electroglottograph and glottal air flow measurements for deaf and normal-hearing speakers," *STL-QPSR*, vol. 32, no. 2-3, pp. 19–27, 1991.

[73] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, no. 2, pp. 131 – 138, 1996.

[74] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[75] N. Henrich, C. dAlessandro, B. Doval, and M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1417–1430, 2005.

[76] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *The Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[77] T. Drugman and A. Alwan, "Glottal closure and opening instant detection from speech signals," in *Proc. Interspeech Conference*, pp. 2891–2894, 2009.

[78] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 34–43, Jan 2007.

[79] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Processing Letters*, vol. 14, pp. 762–765, Oct 2007.

[80] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 994–1006, March 2012.

[81] C. M. Sapienza, E. T. Stathopoulos, and C. Dromey, "Glottal airflow and electroglottographic measures of vocal function at multiple intensities," *Journal of Voice*, vol. 6, no. 1, pp. 44 – 54, 1992.

[82] C. M. Sapienza, E. T. Stathopoulos, and C. Dromey, "Approximations of open quotient and speed quotient from glottal airflow and egg waveforms: Effects of measurement criteria and sound pressure level," *Journal of Voice*, vol. 12, no. 1, pp. 31 – 43, 1998.

[83] I. Ben-Dom and C. I. Watson, "Towards vocal health assessment from the speech signal," in *The Proceedings of the 22nd Electronics New Zealand Conference (ENZCon)*, pp. 22–27, Victoria University of Wellington, Electronics New Zealand Inc., 2016.

[84] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[85] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[86] T. Backstrom, P. Alku, and E. Vilkman, "Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range,"

IEEE Transactions on Speech and Audio Processing, vol. 10, pp. 186–192, Mar 2002.

[87] X. Sun, “Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio,” in *Proc. ICASSP*, vol. 1, p. 333336, 2002.

[88] W. J. Hess, *Pitch and Voicing Determination of Speech with an Extension Toward Music Signals*, pp. 181–212. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[89] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis* (W. B. Klein and K. K. Palival, eds.), Elsevier, 1995.

[90] J. Markel, “The sift algorithm for fundamental frequency estimation,” *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367–377, 1972.

[91] A. de Cheveign and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[92] A. M. Noll, “Cepstrum pitch determination,” *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[93] A. Camacho, “Comment on ”cepstrum pitch determination” [j. acoust. soc. am. 41, 293309 (1967)],” *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 2706–2707, 2008.

[94] L. K. Saul, D. D. Lee, C. L. Isbell, and Y. Lecun, “Real time voice processing with audiovisual feedback: toward autonomous agents with perfect pitch,” in *Advances in Neural Information Processing Systems 15*, pp. 1205–1212, MIT Press, 2002.

[95] F. Sha, J. A. Burgoyne, and L. K. Saul, “Multiband statistical learning for f0 estimation in speech,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, May 2004.

[96] X. Sun, “A pitch determination algorithm based on subharmonic-to-harmonic ratio,” in *the 6th International Conference of Spoken Language Processing*, pp. 676–679, 2000.

[97] T. Drugman and A. Alwan, “Joint robust voicing detection and pitch estimation based on residual harmonics,” in *Proc. Interspeech*, p. 19731976, 2011.

[98] M. Farrus, J. Hernando, and P. Ejarque, “Jitter and Shimmer Measurements for Speaker Recognition.” Eurospeech, Antwerp, Belgium, 2007.

[99] D. Deliyski and S. A. Xue, "Effects of aging on selected acoustic voice parameters: perliminary normative data and educational implications," *Educational Gerontology*, vol. 27, no. 2, pp. 159–168, 2001.

[100] L. S. Finger, C. A. Cielo, and K. Schwarz, "Acoustic vocal measures in women without voice complaints and with normal larynxes," *Brazilian Journal of Otorhinolaryngology*, vol. 75, pp. 432 – 440, 2009.

[101] J. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis jitter, shimmer and hnr parameters," *Procedia Technology*, vol. 9, pp. 1112 – 1122, 2013.

[102] J. Teixeira and A. Goncalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Computer Science*, vol. 100, pp. 271 – 279, 2016.

[103] S. Schötz, *Acoustic Analysis of Adult Speaker Age*, pp. 88–107. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[104] H. Kasuya, H. Yoshida, H. Mori, and H. Kido, "A longitudinal study on vocal aging changes in f0, jitter, shimmer and glottal noise," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3428–3428, 2008.

[105] L. Ramig and R. Ringel, "Acoustic characteristics of speech and selected measures of body physiology," 1980. A paper presented to the convention of the American Speech-Language-Hearing Association, Detroit.

[106] L. Ramig, "Effects of physiological aging on speaking and reading rates," *Journal of Communication Disorders*, vol. 16, no. 3, pp. 217 – 226, 1983.

[107] L. Ramig and R. Ringel, "Effects of physiological aging on selected acoustic characteristics of voice," *Journal of Speech and Hearing Research*, vol. 26, pp. 22–30, 1983.

[108] J. Laver, *The Phonetic Description of Voice Quality.* 1980. pp. 115-117.

[109] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991.

[110] P. Moore and C. L. Thompson, "Comments on physiology of hoarseness," *Archives of Otolaryngology*, vol. 81, no. 1, pp. 97–102, 1965.

[111] D. S. Lundy, C. Silva, R. R. Casiano, F. L. Lu, and J. W. Xue, "Cause of hoarseness in elderly patients," *Otolaryngology - Head and Neck Surgery*, vol. 118, no. 4, pp. 481 – 485, 1998.

[112] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 344–353, 1963.

[113] J. Munoz, E. Mendoza, M. D. Fresneda, G. Carballo, and P. López, "Acoustic and perceptual indicators of normal and pathological voice," *Folia Phoniatrica et Logopaedica*, vol. 55, no. 2, pp. 102–114, 2003.

[114] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, no. 4, pp. 407 – 429, 2001.

[115] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, no. 3, pp. e111 – e116, 2012.

[116] H. Hollien and R. W. Wendahl, "Perceptual study of vocal fry," *The Journal of the Acoustical Society of America*, vol. 43, no. 3, pp. 506–509, 1968.

[117] Y. Chen, M. Blomgren, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, and physiological characteristics of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 3177–3177, 1997.

[118] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences*, 2015.

[119] H. Hollien, ""Old voices": What do we really know about them?," *Journal of Voice*, vol. 1, no. 1, pp. 2 – 17, 1987.

[120] Y. Chen, M. Robb, and H. R. Gilbert, "Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation.," *Journal of Speech, Language & Hearing Research*, vol. 45, no. 5, p. 821, 2002.

[121] I. M. V. de Leeuw and H. F. Mahieu, "Vocal aging and the impact on daily life: a longitudinal study," *Journal of Voice*, vol. 18, no. 2, pp. 193 – 202, 2004.

[122] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research*, vol. 37, pp. 769–778, 1994.

[123] "Spasmodic dysphonia." `https://www.nidcd.nih.gov/sites/default/files/Documents/health/voice/SpasmodicDysphonia.pdf`. Accessed: 13-12-2016.

[124] C. Ishi, T. Hiroshi, and N. Hagita, "Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, 2010.

[125] C. G. Henton and R. W. A. Bladon, “Breathiness in normal female speech: inefficiency versus desirability,” *Language & Communication*, vol. 5, no. 3, pp. 221–227, 1985.

[126] M. Södersten and P. Lindestad, “Glottal closure and perceived breathiness during phonation in normally speaking subjects,” *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 3, pp. 601–611, 1990.

[127] M. M. Gorham-Rowan and J. Laures-Gore, “Acoustic-perceptual correlates of voice quality in elderly men and women,” *Journal of Communication Disorders*, vol. 39, no. 3, pp. 171 – 184, 2006.

[128] J. M. Beck, *Organic Variation of the Vocal Apparatus*, pp. 153–201. Blackwell Publishing Ltd., 2010.

[129] W. B. Jr., R. J. Morris, and J. F. Michel, “Vocal jitter in young adult and aged female voices,” *Journal of Voice*, vol. 3, no. 2, pp. 113 – 119, 1989.

[130] S. E. Linville, *Vocal Aging.* San Diego, CA: Singular Thomson Learning, 2001.

[131] R. A. Dedivitis, M. A. ao M. J. Simões, O. A. Mora, and O. W. Cervantes, “Aging histological changes in the cartilages of the cricoarytenoid joint,” *Acta. Cir. Bras.*, vol. 19, pp. 136–140, 2006.

[132] A. G. Jurik, “Ossification and calcification of the laryngeal skeleton,” *Acta Radiologica. Diagnosis*, vol. 25, no. 1, pp. 17–22, 1984.

[133] F. Paulsen, M. Kimpel, U. Lockemann, and B. Tillmann, “Effects of ageing on the insertion zones of the human vocal fold,” *Journal of Anatomy*, vol. 196, no. 1, pp. 41–54, 2000.

[134] M. Hirano, S. Kurita, and S. Sakaguchi, “Ageing of the vibratory tissue of human vocal folds,” *Acta. Oto-Laryngologica*, vol. 107, pp. 428–433, 1987.

[135] I. Honjo and N. Isshiki, “Laryngoscopy and voice characteristics of aged persons,” *Archives of Otolaryngology*, vol. 106, pp. 149–150, 1980.

[136] N. Schaeffer, M. Knudsen, and A. Small, “Multidimensional voice data on participants with perceptually normal voices from ages 60 to 80: A preliminary acoustic reference for the elderly population,” *Journal of Voice*, vol. 29, no. 5, pp. 631 – 637, 2015.

[137] J. Kahane, “Connective tissue changes in the larynx and their effects on voice,” *Journal of Voice*, vol. 1, pp. 27–30, 1987.

[138] B. Das, S. Mandal, P. Mitra, and A. Basu, “Effect of aging on speech features and phoneme recognition: a study on bengali voicing vowels,” *International Journal of Speech Technology*, vol. 16, no. 1, pp. 19–31, 2013.

[139] E. D. Mysak, “Pitch and duration characteristics of older males.,” *Journal of Speech & Hearing Research*, vol. 2, pp. 46–54, 1959.

[140] H. Hollien and T. Shipp, “Speaking fundamental frequency and chronological age in males,” *Journal of Speech, Language, and Hearing Research*, vol. 15, no. 1, pp. 155–159, 1972.

[141] W. S. B. Jr., R. J. Morris, H. Hollien, and E. Howell, “Speaking fundamental frequency characteristics as a function of age and professional singing,” *Journal of Voice*, vol. 5, no. 4, pp. 310 – 315, 1991.

[142] R. Baken, “The aged voice: a new hypothesis,” *Journal of Voice*, vol. 19, no. 3, pp. 317–325, 2005.

[143] M. Nishio and S. Niimi, “Changes in speaking fundamental frequency characteristics with aging,” *Folia Phoniatrica et Logopaedica*, vol. 60, pp. 120–7, 04 2008. Copyright - Copyright (c) 2008 S. Karger AG, Basel; Last updated - 2015-05-30.

[144] P. Torre and J. A. Brlow, “Age-related changes in acoustic characteristics of adult speech,” *Journal of Communication Disorders*, vol. 42, no. 5, pp. 324–333, 2009.

[145] E. T. Stathopoulos, J. E. Huber, J. E. Sussman, and R. Schlauch, “Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4-93 years of age.,” *Journal of Speech, Language & Hearing Research*, vol. 54, no. 4, pp. 1011 – 1021, 2011.

[146] S. D. Bier and C. I. Watson, “Electroglottographic analysis of young and old speakers of new zealand english,” in *The Proceedings of the 14th Australasian International Conference on Speech Science and Technology SST*, pp. 65–68, University of Macquarie, Sydney, 2012.

[147] S. D. Bier, C. I. Watson, and C. M. McCann, “Using the perturbation of the contact quotient of the EGG waveform to analyse age differences in adult speech,” *Journal of Voice*, vol. 28, no. 3, pp. 267–273, 2014.

[148] B. J. Benjamin, “Frequency variability in the aged voice,” *Journal of Gerontology*, vol. 36, no. 6, pp. 722–726, 1981.

[149] T. Shipp, Y. Qi, R. Huntley, and H. Hollien, “Acoustic and temporal correlates of perceived age,” *Journal of Voice*, vol. 6, no. 3, pp. 211 – 216, 1992.

[150] Y. Horii and W. J. Ryan, "Fundamental frequency characteristics and perceived age of adult male speakers," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S69–S69, 1975.

[151] R. E. McGlone and H. Hollien, "Vocal pitch characteristics of aged women," *Journal of Speech, Language, and Hearing Research*, vol. 6, no. 2, pp. 164–170, 1963.

[152] M. Higgins and J. Saxman, "A comparison of selected phonatory behaviors of healthy aged and young adults," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 5, pp. 1000–1010, 1991.

[153] S. E. Linville, "Vocal aging," *Current Opinion in Otolaryngology & Head & Neck Surgery*, vol. 3, pp. 183–187, 1995.

[154] S. A. Xue, G. J. Hao, L. Xu, and T. Moranski, "peaking fundamental frequency changes in women over time," *Asia Pacific Journal of Speech, Language and Hearing*, vol. 11, pp. 189–194, 2008.

[155] A. Russell, L. Penny, and C. Pemberton, "Speaking fundamental frequency changes over time in women: A longitudinal study," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 1, pp. 101–109, 1995.

[156] O. de Pinto and H. Hollien, "Speaking fundamental frequency characteristics of australian women: Then and now," vol. 10, pp. 367–375, 1982.

[157] U. Reubold, J. Harrington, and F. Kleber, "Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers," *Speech Commun.*, vol. 52, pp. 638–651, July 2010.

[158] J. Harrington, S. Palethorpe, and C. I. Watson, "Age-related changes in fundamental frequency and formants : a longitudinal study of four speakers," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*, vol. 2, (Macquarie University, Sydney, Australia), pp. 1081–1084, August Interspeech 2007.

[159] W. Decoster and F. Debruyne, "Longitudinal voice changes: facts and interpretation," *Journal of voice*, vol. 14, p. 184193, June 2000.

[160] J. D. Harnsberger, R. Shrivastav, W. B. Jr., H. Rothman, and H. Hollien, "Speaking rate and fundamental frequency as speech cues to perceived age," *Journal of Voice*, vol. 22, no. 1, pp. 58 – 69, 2008.

[161] S. E. Linville and H. B. Fisher, “Acoustic characteristics of women’s voices with advancing age,” *Journal of Gerontology*, vol. 40, pp. 324–330, 1985.

[162] S. E. Linville, “Acoustic-perceptual studies of aging voice in women,” *Journal of Voice*, vol. 1, pp. 44–48, 1987.

[163] H. Kido and H. Kasuya, “Perceived vocal age and its acoustic correlates,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3292–3292, 2006.

[164] K. A. Wilcox and Y. Horii, “The age and changes in vocal jitter,” *Journal of Gerontology*, vol. 35, no. 2, pp. 194–198, 1980.

[165] L. O. Ramig, S. Gray, K. Baker, K. Corbin-Lewis, E. Luschei, H. Coon, and M. Smith, “The aging voice: A review, treatment data and familial and genetic perspectives,” *Folia Phoniatrica et Logopaedica*, vol. 53, no. 5, pp. 252–265, 2001.

[166] R. Vipperla, S. Renals, and J. Frankel, “Ageing voices: The effect of changes in voice parameters on asr performance,” *EURASIP J. Audio Speech Music Process.*, vol. 2010, pp. 5:1–5:10, Jan. 2010.

[167] H. Hollien, J. F. Michel, and E. T. Doherty, “A method for analyzing vocal jitter,” *The Journal of the Acoustical Society of America*, vol. 50, no. 1A, pp. 140–140, 1971.

[168] D. M. Biever and D. M. Bless, “Vibratory characteristics of the vocal folds in young adult and geriatric women,” *Journal of Voice*, vol. 3, no. 2, pp. 120 – 131, 1989.

[169] K. Leong, M. J. Hawkshaw, D. Dentchev, R. Gupta, D. Lurie, and R. T. Sataloff, “Reliability of objective voice measures of normal speaking voices,” *Journal of Voice*, vol. 27, no. 2, pp. 170 – 176, 2013.

[170] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, “Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task,” *Journal of Voice*, vol. 25, pp. 44–53, 2011.

[171] W. J. Ryan, “Acoustic aspects of the aging voice1,” *Journal of Gerontology*, vol. 27, no. 2, p. 265, 1972.

[172] D. E. Hartman and J. L. Danhauer, “Perceptual features of speech for males in four perceived age decades,” *The Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 713–715, 1976.

[173] L. Ramig, "Aging speech: Physiological and sociological aspects," *Language & Communication*, vol. 6, no. 1, pp. 25 – 34, 1986.

[174] R. J. Morris and W. B. Jr., "Age-related voice measures among adult women," *Journal of Voice*, vol. 1, no. 1, pp. 38 – 43, 1987.

[175] T. Drugman, G. Wilfart, A. Moinet, and T. Dutoit, "Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis," in *the proceedings of ICASSP*, pp. 3793–3796, 2009.

[176] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.

[177] L. Tian and C. I. Watson, "Continuous spoken emotion recognition based on time-frequency features of the glottal pulse signal within stressed vowels," in *the Proceedings of the Sixteenth Australasian International Conference on Speech Science and Technology (ASSTA)*, pp. 285–288, 2016.

[178] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classi

cation of clinical depression in speech," *IEEE Trans Biomed Eng*, vol. 55, no. 1, pp. 96–107, 2008.

[179] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1115 – 1116, 2014.

[180] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, vol. 12 of *Communication and Cybernetics*. Springer Berlin Heidelberg, 1976.

[181] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Springer Verlag, 1972.

[182] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Lfqvist, S. McCoy, D. G. Miller, H. No, R. C. Scherer, J. R. Smith, B. H. Story, J. G. vec, S. Ternstrm, and J. Wolfe, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 3005–3007, 2015.

[183] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[184] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing (2Nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1999.

[185] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[186] J. Walker and P. Murphy, “Progress in nonlinear speech processing,” ch. A Review of Glottal Waveform Analysis, pp. 1–21, Berlin, Heidelberg: Springer-Verlag, 2007.

[187] H. W. Strube, “Determination of the instant of glottal closure from the speech wave,” *The Journal of the Acoustical Society of America*, vol. 56, no. 5, pp. 1625–1629, 1974.

[188] A. Paavo, M. Carlo, Y. Santeri, B. Tom, and S. Brad, “Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering,” *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3289–3305, 2009.

[189] P. Alku, “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering,” *Speech Commun.*, vol. 11, pp. 109–118, June 1992.

[190] P. Alku, “Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production,” *Folia Phoniatr Logop*, vol. 58, pp. 102–113, 2006.

[191] P. Alku and E. Vilkman, “Estimation of the glottal pulseform based on discrete all-pole modeling,” in *Proceedings of the International Conference on Spoken Language Processing*, p. 16191622, Yokohama, 1994.

[192] A. El-Jaroudi and J. Makhoul, “Discrete all-pole modeling,” *IEEE Transactions on Signal Processing*, vol. 39, no. 2, p. 411423, 1991.

[193] W. Ding and N. Campbell, “Determining polarity of speech signals based on gradient of spurious glottal waveforms,” in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, pp. 857–860, May 1998.

[194] S. Sakaguchi, T. Arai, and Y. Murahara, “The effect of polarity inversion of speech on human perception and data hiding as application,” in *Proceeding of The International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 917–920, 2000.

[195] B. Yan, Z. Lu, J. Pan, and S. Sun, "Statistical analysis of two polarity detection schemes in speech watermarking," in *Proceedings of the 17th Conference on Computational Linguistics and Speech Processing, (ROCLING) 2005, Taiwan, ROC, 2005*, 2005.

[196] T. Drugman and T. Dutoit, "Speech polarity determination: A comparative evaluation," *Neurocomputing*, vol. 132, pp. 121 – 125, 2014. Innovations in Nature Inspired Optimization and Learning MethodsMachines learning for Non-Linear ProcessingSelected papers from the Third World Congress on Nature and Biologically Inspired Computing (NaBIC2011)Selected papers from the 2011 International Conference on Non-Linear Speech Processing (NoLISP 2011).

[197] B. Deepak and D. Govind, "Significance of implementing polarity detection circuits in audio preamplifiers," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2197–2200, Aug 2015.

[198] T. Drugman and T. Dutoit, "Oscillating statistical moments for speech polarity detection," in *Advances in Nonlinear Speech Processing: Proceedings of the 5^{th} International Conference on Nonlinear Speech Processing (NOLISP) 2011*, p. 4854, 2011.

[199] T. Drugman, "Residual excitation skewness for automatic speech polarity detection," *IEEE Signal Processing Letters*, vol. 20, pp. 387–390, April 2013.

[200] B. Abhiram, A. P. Prathosh, and A. G. Ramakrishnan, "A fast algorithm for speech polarity detection using long-term linear prediction," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, July 2014.

[201] D. Govind, P. Hisham, and D. Pravena, "A robust algorithm for speech polarity detection using epochs and hilbert phase information," *Procedia Computer Science*, vol. 58, pp. 524 – 529, 2015.

[202] D. Govind, P. Hisham, and D. Pravena, "Effectiveness of polarity detection for improved epoch extraction from speech," in *2016 Twenty Second National Conference on Communication (NCC)*, pp. 1–6, March 2016.

[203] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0. Retrieved from `http://www.R-project.org`.

[204] K. Hornik, "R FAQ," 2016. Available at: `https://CRAN.R-project.org/doc/FAQ/R-FAQ.html`.

[205] RStudio Team, *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.

[206] R. C. Team and contributors worldwide, *Base: The R Base Package*, 2017. R package version 3.4.0 – `https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html`.

[207] D. Bates and M. Maechler, *Matrix: Sparse and Dense Matrix Classes and Methods*, 2016. R package version 7.1 – `http://matrix.r-forge.r-project.org/`.

[208] U. Ligges and J. Fox, "R help desk: How can I avoid this loop or make it faster?," *R News*, vol. 8, no. 1, pp. 46–50, 2008.

[209] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB." Web page, 2005.

[210] J. Harrington, *The Phonetic Analysis of Speech Corpora*. Wiley-Blackwell, 2010.

[211] J. O. Ramsay, G. Hooker, and S. Graves, *Dunctional data analysis with R and MATLAB*. Springer, 2009.

[212] R. Winkelmann, K. Jaensch, S. Cassidy, and J. Harrington, *emuR: Main Package of the EMU Speech Database Management System*, 2016. R package version 0.2.1.

[213] R. Winkelmann, *The EMU Speech Database Management System (EMU-SDMS)*, August 2016. Available at: `https://ips-lmu.github.io/EMU.html`.

[214] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, p. 4964, 2008.

[215] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP: A collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, May 2014.

[216] C. I. Watson, W. Liu, and B. MacDonald, "The Effect of Age and Native Speaker Status on Synthetic Speech Intelligibility," in *the proceedings of the 8^{th} ISCA Speech Synthesis Workshop*, (Barcelona, Spain), pp. 195–200, August 2011.

[217] R. Tamagawa, C. I. Watson, H. Kuo, B. A. MacDonald, and E. Broadbent, "The effects of synthesized voice accents on user perceptions of robots," *International Journal of Social Robotics*, vol. 3, no. 3, pp. 253–263, 2011.

[218] L. F. Thompson, C. I. Watson, J. King, R. Harlow, M. Maclagan, H. Charters, and P. Keegan, "Phrases, Pitch, and Percieved Prominances in Māori," in *the proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*, (Florence, Italy), August 2011.

[219] S. Barreda, *phonTools: Functions for phonetics in R.*, 2015. R package version 0.2-2.1.

[220] S. D. Bier, C. I. Watson, and C. M. McCann, "Dynamic measures of voice stability in young and old adults," *Logopedics Phoniatrics Vocology*, pp. 1–11, 2016.

[221] K. Sjlander and J. Beskow, "Wavesurfer - an open source speech tool," 2000.

[222] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proc. of the ICPhS*, pp. 607–610, 1999.

[223] T. Kisler, R. U. D., F. Schiel, C. Draxler, B. Jackl, and N. Pörner, "BAS Speech Science Web Services - an update of current developments," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016. paper id 668. Available at: `https://clarin.phonetik.uni-muenchen.de/BASWebServices/#/services/WebMAUSMultiple`.

[224] C. I. Watson, "Mappings between vocal tract area functions, vocal tract resonances and speech formants for multiple speakers," in *The Proceedings of the 15th Annual Conference of the International Speech Communication Association ISCA*, pp. 1993–1997, Singapore, 2014.

[225] S. Cassidy and J. Harrington, "multi-level annotation in the Emu speech database management system," *Speech Commun.*, vol. 33, pp. 61–77, 2001.

[226] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. I. Watson, "The MAONZE project: Changing uses of an indigenous language database," *Corpus Linguistics and Linguistic Theory*, vol. 7, no. 1, pp. 37–57, 2011.

[227] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. I. Watson, "Sound change in maori and the formation of the MAONZE project," in *He Hiringa, He Pūmanawa: Studies on the Māori language* (A. Onysko, M. Degani, and J. King, eds.), pp. 33–54, Wellington: Huia, 2014.

[228] S. J. Sinclair and C. I. Watson, "The development of the Otage speech database," in *Proceedings of the Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, pp. 298–301, Nov 1995.

[229] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. I. Watson, "Changing pronunciation of the māori language: Implications for revitalization," in *Indigenous Language Revitalization Encouragement, Guidance & Lessons Learned* (J. Reyhner and L. Lockard, eds.), pp. 75–86, Northern Arizona University, 2011.

[230] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. I. Watson, "The MAONZE corpus: Transcribing and analysing Māori speech," *New Zealand Studies in Applied Linguistics*, vol. 17, no. 1, pp. 32–48, 2009.

[231] J. Harrington, S. Palethorpe, and C. I. Watson, "Capturing the vowel change in New Zealand English over a thirty year period via a diachronic study," in *Proceedings of the 10th Australian International Conference on Speech Science and Technology (SST)*, (Macquarie University, Sydney, Australia), pp. 201–206, Dec 2004.

[232] C. I. Watson, M. A. Maclagan, J. King, R. Harlow, and P. J. Keegan, "Sound change in Māori and the inuence of New Zealand English," *Journal of the International Phonetic Association*, pp. 1–34, 2016.

[233] H. Goy, D. N. Fernandes, M. K. Pichora-Fuller, and P. van Lieshout, "Normative voice data for younger and older adults," *Journal of Voice*, vol. 27, no. 5, pp. 545 – 555, 2013.

[234] B. Smith, B. Weinberg, L. Feth, and Y. Horii, "Monitoring vocal fold abduction through vocal fold contact area," *Journal of Speech and Hearing Research*, vol. 31, pp. 338–351, 1988.

[235] S. E. linville, "Source characteristics of aged voice assessed from long-term average spectra," *Journal of Voice*, vol. 16, no. 4, p. 472479, 2002.

[236] R. Winkler and W. Sendlmeier, "Egg open quotient in aging voiceschanges with increasing chronological age and its perception," *Logopedics Phoniatrics Vocology*, vol. 31, no. 2, pp. 51–56, 2006.

[237] L. A. F. Mendoza, E. Cataldo, M. Vellasco, M. A. Silva, A. D. O. Can, and J. M. de Seixas, "Classification of voice aging using ann and glottal signal parameters," in *2010 IEEE ANDESCON*, pp. 1–5, Sept 2010.

[238] E. Mendoza, N. Valencia, J. Muoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (ltas)," *Journal of Voice*, vol. 10, no. 1, pp. 59 – 66, 1996.

[239] E. Ma and A. Love, "Electroglottographic evaluation of age and gender effects during sustained phonation and connected speech," *Journal of Voice*, vol. 24, no. 2, p. 146152, 2010.

[240] D. H. Slavit, R. J. Lipton, and T. V. McCaffrey, "Phonatory vocal fold function in the excised canine larynx," *Otolaryngology-Head and Neck Surgery*, vol. 103, no. 6, pp. 947–956, 1990.

[241] G. E. Murty, P. N. Carding, and P. J. Kelly, "Combined glottographic changes in the elderly," *Clin. Ololaryngol*, vol. 16, pp. 532–534, 1991.

[242] L. Lehto, M. Airas, E. Bjrkner, J. Sundberg, and P. Alku, "Comparison of two inverse filtering methods in parameterization of the glottal closing phase characteristics in different phonation types," *Journal of Voice*, vol. 21, no. 2, pp. 138 – 150, 2007.

[243] G. Fant and M. Båvegåd, "Notes on glottal source interaction ripple," *STL-QPSR*, vol. 35, no. 4, pp. 63–78, 1994.

[244] S. Sapir, "The intrinsic pitch of vowels: Theoretical, physiological, and clinical considerations," *Journal of Voice*, vol. 3, no. 1, pp. 44 – 51, 1989.

[245] P. Lieberman, "A study of prosodic features," *Haskins Lab Status Rep Speech Res*, vol. SR-23, pp. 179–208, 1970.

[246] G. Eden and G. F. Inbar, "Physiological model analysis of involuntary human-voice tremor," *Biological Cybernetics*, vol. 30, no. 3, pp. 179–185, 1978.

[247] J. J. Ohala and B. W. Eukel, "Explaining the intrinsic pitch of vowels," *The Journal of the Acoustical Society of America*, vol. 60, no. S1, pp. S44–S44, 1976.

[248] J. J. Ohala and B. W. Eukel, "Explaining the intrinsic pitch of vowels," in *R. Channon & L. Shockey (eds.), In honor of Ilse Lehiste*, 1987.

[249] P. Ladefoged, *A phonetic study of West African languages: An auditory-instrumental survey.* Cambridge, England: Cambridge University, 1968.

[250] C. H. Shadle, J. B. Pierrehumbert, and M. Y. Liberman, "The intrinsic pitch of vowels in sentence context," *The Journal of the Acoustical Society of America*, vol. 66, no. S1, pp. S64–S64, 1979.

[251] D. OShaughnessy and J. Allen, "Linguistic modality effects on fundamental frequency in speech," *The Journal of the Acoustical Society of America*, vol. 74, no. 4, pp. 1155–1171, 1983.

[252] K. Marasek, “Glottal correlates of the word stress and the tense/lax opposition in german,” in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, pp. 1573–1576 vol.3, Oct 1996.

[253] M. Lim, E. Lin, and P. Bones, “Vowel effect on glottal parameters and the magnitude of jaw opening,” *Journal of Voice*, vol. 20, no. 1, pp. 46 – 54, 2006.

[254] A. Sluijter, *Phonetic Correlates of Stress and Accent.* Leiden: Holland Institute of Generative Linguistics, 1995.

[255] N. Paul, S. Kumar, I. Chatterjee, and B. Mukherjee, “Electroglottographic parameterization of the effects of gender, vowel and phonatory registers on vocal fold vibratory patterns: An indian perspective,” *Indian J Otolaryngol Head Neck Surg*, vol. 63, no. 1, pp. 27–31, 2011.

[256] J. E. Sussman and C. Sapienza, “Articulatory, developmental, and gender effects on measures of fundamental frequency and jitter,” *Journal of Voice*, vol. 8, no. 2, pp. 145 – 156, 1994.

[257] R. F. Orlikoff, “Vocal stability and vocal tract configuration: An acoustic and electroglottographic investigation,” *Journal of Voice*, vol. 9, no. 2, pp. 173 – 181, 1995.

[258] I. R. Titze, “A model for neurologic sources of aperiodicity in vocal fold vibration,” *Journal of Speech and Heanng Research*, vol. 34, pp. 460–472, 1991.

[259] G. Fant and T. V. Ananthapadmanabha, “Calculation of true glottal flow and its components,” *STL-QPSR*, vol. 23, no. 1, pp. 1–30, 1982.

[260] A. Krishnamurthy and D. Childers, “Two-channel speech analysis,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

[261] G. Fant and T. V. Ananthapadmanabha, “Truncation and superposition effects in voice production,” *The Journal of the Acoustical Society of America*, vol. 71, no. 1, 1982.

[262] B. Cranen and L. Boves, “On subglottal formant analysis,” *The Journal of the Acoustical Society of America*, vol. 81, no. 3, pp. 734–746, 1987.

[263] A. M. Sulter and H. P. Wit, “Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age.,” *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3360–3373, 1996.

[264] J. F. Torres, *stimation of glottal source features from the spectral envelope of the acoustic speech signals.* PhD thesis, Georgia Technical Institute, 2010.

[265] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–671, 1994.

[266] D. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 369–377, 1985.

[267] M. Rothenberg and J. Mashie, "Monitoring vocal fold abduction through vocal fold contact area," *Journal of Speech and Hearing Research*, vol. 31, pp. 338–351, 1988.

[268] P. Alku, "Using group delay function to assess glottal flows estimated by inverse filtering," *Electronics Letters*, vol. 41, pp. 562–563(1), April 2005.

[269] A. Philip, "Glottal wave extraction using inverse filtering and performing analysis of ageing effects on the glottis." unpublished, 2015.

[270] G. Fant, "Glottal source and excitation analysis," *STL-QPSR*, vol. 20, no. 1, pp. 85–107, 1979.

[271] G. Fant, "Vocal source analysis - a progress report," *STL-QPSR*, vol. 20, no. 3-4, pp. 31–53, 1979.

[272] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[273] R. Veldhuis, "A computationally efficient alternative for the liljencrantsfant model and its perceptual evaluation," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 566–571, 1998.

[274] P. K. Muthukumar, A. W. Black, and H. T. Bunnell, "Optimizations and fitting procedures for the liljencrants-fant model for statistical parametric speech synthesis," in Bimbot *et al.* [280], pp. 397–401.

[275] G. Fant and Q. Lin, "Frequency domain interpretation and derivation of glottal flow parameters," *STL-QPSR*, vol. 29, no. 2-3, pp. 1–21, 1988.

[276] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "Hmm-based speech synthesiser using the LF-model of the glottal source," in *2011 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4704–4707, May 2011.

[277] Y. Qi and N. Bi, "A simplified approximation of the four-parameter lf model of voice source," *J.Acoust. Soc. Am.*, vol. 2, pp. 1182–1185, 1994.

[278] G. Fant, "The lf-model revisited. transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[279] S. Dias and A. Ferreira, "A hybrid LF-Rosenberg frequency domain model of the glottal source," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2013.

[280] F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, eds., *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, ISCA, 2013.

[281] V. Wolfe and D. Martin, "Acoustic correlates of dysphonia: type and severity," *Journal of Communication Disorders*, vol. 30, no. 5, pp. 403 – 416, 1997.

[282] M. Airas, "TKK Aparat: an environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, pp. 49–64, 2008.

[283] M. A. Lawrence, *ez: Easy analysis and visualization of factorial experiments*, 2011. R package version 3.0 – `http://CRAN.R-project.org/package=ez`.