RECONSTRUCTION OF NATURAL SOUNDING SPEECH FROM WHISPERS

HAMID REZA SHARIFZADEH

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

2011

Abstract

This thesis explores reconstruction of natural sounding speech from whispers. As a broad research class, the generation of normally phonated speech from whispers can be useful in several types of application from different scientific fields ranging from communications to biomedical engineering. The primary focus of the thesis and current work is therefore to investigate appropriate solutions and algorithms for regenerating natural phonated speech from whispers.

Interestingly, unlike other speech processing fields, many aspects of such reconstruction, in spite of the useful applications, have not yet been resolved by researchers. In particular, the outcome of this research will find at least two immediate applications which have different forms but similar solutions: a) reconstructing natural speech for laryngectomy patients, b) restoring natural pitched speech in a cell phone/telephone communication when one party talks in a whispering mode for privacy or security reasons.

This thesis presents a solution for the conversion of whispers to fully-phonated speech through the modification of the CELP codec. We also present a novel method for spectral enhancement and formant smoothing during the reconstruction process, using a probability mass-density function to identify reliable formant trajectories in whispers, and apply spectral modifications accordingly. The method relies upon the observation that,

whilst the pitch generation mechanism of patients with larynx damage is typically unusable, the remaining components of the speech production apparatus may be largely unaffected. The approach outlined here allows patients to regain their ability to speak (simply by whispering into an external prosthesis), yielding a more natural sounding voice than alternative solutions.

Since whispered speech can be identified as the core input of the system, the acoustic features of whispers also need to be considered. Despite the everyday nature of whispering, and its undoubted usefulness in vocal communications, whispers have received relatively little research effort to date, apart from some studies analysing the main whispered vowels and some quite general estimations of whispered speech characteristics. In particular, a classic vowel space determination has been lacking for whispers. For voiced speech, this type of information has played an important role in the development and testing of recognition and processing theories over the past few decades, and can be expected to be equally useful for whisper-mode communications and recognition systems. This thesis also aims to redress the shortfall by presenting a vowel formant space for whispered speech, and comparing the results with corresponding phonated samples.

To my parents:

Sedigh & Naser

for their unconditional kindness and support

Acknowledgements

During the past 4 years, it was my great honour to work with and learn from the groups of distinguished mentors and outstanding fellows in Nanyang Technological University (NTU), Singapore.

Foremost, I would like to express my warmest gratitude to my supervisor, Associate Professor Ian McLoughlin, for his excellent guidance and kind support throughout my candidature. I would like to thank him for all his rewarding discussions and perceptive suggestions to my study and life. His extensive knowledge and sharp insight immensely influenced my research and his lasting encouragement helped me towards becoming a confident researcher. Working for him was, in fact, a delightful experience full of learning and personal development.

I also would like to express my appreciation to my colleague, Ms. Farzaneh Ahmadi, for her helpful guidance and useful discussion within my research work. Professor Martin Russell also kindly helped me during my stay at University of Birmingham while his advice and guidance efficiently improved the outcome of the research.

I also would like to thank the colleagues in Parallel and Distributed Computing Centre (PDCC) for creating a dynamic and friendly ambient. My special gratitude also goes to the technical manager, Ms. Ng-Goh Siew Lai, Irene, for her kind help and effective impression in providing all the

facilities during my research work. Furthermore, I would like to show my appreciation to all of my friends, specially my close friends in Singapore: Farnaz, Mojtaba, and Omid for their support and friendship.

Last but not least, I want to thank my parents in Iran, for their constant love and encouragement. I really appreciate them for all their support, kindness and patience during the times I was far from them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech is considered by many to be the primary means of communication between human beings. Humans are suited in several ways to using an acoustic communication system that is under voluntary control and that uses the recombination of a potentially large number of basic units. The Encyclopedia of Language and Linguistics [1] points out humans' extraordinarily good control over breathing, especially on the out breath, and also notes that humans have a vocal tract that is adapted to producing a large number of distinct sounds. In contrast with other mammals, whose vocal tracts, they say, can be considered as a single tube, the human vocal tract consists of two linked tubes including the pharyngeal cavity and the oral cavity.

To estimate the history of human speech adaptations, Kay et al. in 1998 [2] investigated the size of the hypoglossal canal (the canal through which the main nerve that innervates the tongue passes). They determined that it is larger (both absolutely and relatively) in modern humans than in modern chimpanzees and gorillas. They then compared the sizes of the hypoglossal canal in different fossils and found that those of modern and early Homo Sapiens as well as those of Neanderthals fall within the modern human range, while those of Australopithecus and Homo Babilis fall within the chimpanzee range. From this, they conclude that "human vocal abilities were present about 400000 years ago".

On the other hand, human attempts at speech processing probably began with machines that can talk in the 13th century when the German philosopher Albertus Magnus and the English scientist Roger Bacon are reputed to have constructed metal talking heads [3]. However, no primary documentation of these devices is known to exist. According to [4], the first verifiable attempts at making speaking machines came some five hundred years later when in 1769 Kratzenstein constructed resonant cavities which produced the sounds of the five vowels a, e, i, o, and u. From that date and particularly during the 20th century, speech processing has grown significantly while more and more applications have found their solutions through progressing research in this field.

Interestingly, the scope of this present thesis is somewhat similar to the first motivations of scientists to mimic human speech by machine, but in a different aspect: we aim to restore natural speech to laryngectomy (larynx surgery) patients who have lost their glottis (part of the speech production apparatus in humans).

1.1 Scope of the Research

Explaining the mechanical process of the speech production, it starts with the airflow (within the exhalation) passes the glottis to create a varying pitch signal which resonates through the vocal tract, nasal cavity and out through the mouth. To shape the speech, individual vocal organs such as oral and nasal cavities, the vellum, tongue, and lip positions have the significant roles in synthesising the sounds.

Total laryngectomy patients will have lost their glottis and also the control to pass lung exhalation through the vocal tract in many cases. Partial laryngectomy patients, by contrast, may still retain the power of controlled lung exhalation through the vocal tract. Despite loss of their glottis, both classes of patient retain the power of vocal tract modulation - in other words they maintain most of the speech production

apparatus. Despite the loss of the glottis including vocal folds, these patients retain the power of vocal tract modulation and therefore by controlling lung exhalation, they have the ability to whisper [5]. By knowing this, the focus of the thesis can, thus, be simplified to the reconstruction of natural sounding speech from whispers.

1.2 Research Motivations

Whispering is one mode of communication with reduced perceptibility and degree of understanding. The main difference between speech and whispers is the absence of significant vocal cord vibration in whispers. This situation can occur with normal physiological blocking of vocal cord vibrations or, in pathological cases, when vocal cords are blocked by disease of the vocal system or after excluding them by an operation. As a paralinguistic phenomenon, whispers can be used in different contexts. One may wish to communicate clearly, but be in a situation where the loudness of normal speech is prohibited, such as in a library. On the other hand, one may be whispering to avoid being overheard, in which case some loss of an understanding of context may be desirable.

Regarding applications, this speech modality is commonly used: general speech scientists use whispers to determine perceptual constants in the speech process [6], medical doctors want to know if whispering is safe for recovering larynx surgery patients [7], speech therapists want to learn more about this mode of speaking to help evaluate voice disorders in aphonic patients [8], and forensic scientists would like to be able to recognize speaker identities from whispered speech [9].

By and large, during several decades of work in this field, a great amount of research has been done on various aspects and properties of speech such as production mechanisms, the physiology of the auditory system, psychophysics, etc. Furthermore, the

availability of high-speed digital hardware since the 1970s has accelerated the exploitation of speech signal processing. Since then much progress has been made in speech coding for efficient transmission, speech synthesis, speech and speaker recognition, and hearing aids. Despite the progress and great achievements in natural language processing, the study of whispered speech and particularly its applications are practically absent in the majority of speech processing literature. Practically speaking, the result of this research thesis will find at least two immediate applications which have different forms but similar solutions: a) reconstructing natural speech for laryngectomy patients, b) restoring natural pitched speech in a cell phone/telephone communication when one party talks in a whispering mode for privacy or security reasons. However, the next sections and beyond concentrate on the first application: that of benefit to laryngectomy patients.

1.3 Context of The Research

Within this context, the focus of the thesis and current work is to present appropriate solutions and algorithms for regenerating natural phonated speech from whispers. This reconstruction, as a broad research class, is useful in different scientific fields from communications to biomedical engineering. We, as a demonstration for assessing the efficiency of the approach, propose a system to restore natural sounding speech for laryngectomy patients.

Furthermore, in the wider context of this research beyond the current thesis, a portable device is to be developed to ultrasonically map vocal tract shape as a patient speaks. This map plus detected lung exhalation will be applied to an artificial glottal resonance to recreate natural sounding speech within the portable electronics (in the same way that a handphone works - not using the vocal tract itself to create the sound). The glottal resonance is to be dynamically and naturally varied based on

recordings of the patients' original voice if available, leading to more natural recreated speech. This research is, in fact, part of a project funded by the Singapore National Medical Research Council, NMRC EDG (EDG07MAY002) grant.

Eventually this 'bionic voice' will return the power of natural-sounding speech to laryngectomy patients, and be useful for temporary voice rest cases. Since the project implements a modified CELP codec, it can also be used in cell phone communications.

It should also be noted that existing methods of returning speech to post-laryngectomised patients do exist and will be discussed in Chapter 2; all these current methods are medical assistive devices suitable for laryngectomy patients, mostly built upon mechanical function of glottis or rehabilitating through surgeries. By contrast, this research itself is not medical in nature: it is a speech processing approach, comprised of the communications algorithms and acoustical features of human voice (not any medical surgery or mechanical functions involved). As such an engineering background mostly based upon normal and whispered speech characteristics will also be given in detail within Chapter 3.

1.4 Thesis Organisation

The literature review, as mentioned, is presented in two separate chapters due to distinctly different topics being covered. Actually, these support chapters include background information about previous related work, as well as the development of possible algorithms that converge in the following chapters and naturally progress through the rest of the research leading to implementation and experimentation.

Since the research is based on whispered speech and is aimed at returning pitch to it, two chapters (3 and 4) are assigned to the issues related to whispers while one of the chapter reviews its characteristics and the other proposes a vowel space for it based upon a study conducted by the author at the University of Birmingham, UK.

Furthermore, the remaining three chapters discuss the system implemented based on a modified CELP codec along with experiments carried out to assess the quality of the system, as well as discussion and conclusion. The organisation of the chapters is as follows:

- Chapter 2 reviews the current rehabilitative methods used for speech regeneration in laryngectomised patients. Particularly, this chapter discusses the three main prostheses and rehabilitative techniques (primarily oesophageal speech, tracheo-oesophageal puncture (TEP) and electrolarynx) outlining the weaknesses of these methods. It justifies why a new engineering approach is necessary and later, how this approach suits the patients.

- Chapter 3 provides an overview, first on normal speech, linear prediction, pitch filters, etc and then describes whispered speech (as the non-existence of pitch) characteristics in terms of articulation and acoustic features. Furthermore, particular properties of whispers affecting digital speech are presented.

- Chapter 4 proposes a comprehensive vowel formant space for whispered speech. Whispered speech still lacks a classic published vowel diagram similar to the vowel acoustic measurements for normal speech. The purpose of this chapter is to redress this shortfall while comparing the results with corresponding phonated samples which can be helpful for the system design in terms of vowel/diphthong enhancements.

- Chapter 5 describes the system implemented as a novel method for whisper-voice conversion through a modified CELP codec. This chapter discusses all algorithms in terms of necessary modules and required modifications added/applied to a standard CELP codec to gain an acceptable output in comparison with

current medical methods, while a technical overview of the novel approaches used for the spectral enhancement of whispered speech mainly based upon line spectral pairs (LSPs) and probability mass function (PMF) are also presented.

- Chapter 6 presents the results in terms of subjective and objective assessments carried out to demonstrate the effectiveness of the methods implemented in Chapter 5. Discussing/comparing the results of these tests as well as considering the shortfalls of the system, eventually completes the chapter.

- Chapter 7 concludes the thesis and provides an insight into possible future work related to this research.

Chapter 2

Current Methods for Speech Rehabilitation in Laryngectomees

Rehabilitation of the ability to speak in a natural sounding voice, for patients who suffer larynx and voice box deficiencies, has long been a dream for both patients and researchers working in this field. Removal of, or damage to, the voice box in a surgical operation such as laryngectomy, affects the pitch generation mechanism of the human voice production system. Post-laryngectomised patients thus exhibit hoarse, whisper like, and sometimes less intelligible speech – it is obviously different to fully phonated speech, and may lack many of the distinctive characteristics of the patients' normal voice. However, these patients often retain the ability to whisper in a similar way to normal speakers.

This chapter discusses how the laryngectomy operation affects speech, before briefly reviewing the three common methods of speech rehabilitation in such patients. In the following chapters, a fourth method which is the main focus of the thesis is presented in detail as an engineering approach to providing laryngectomy patients the capacity to speak with a more natural sounding voice. As a side effect, this approach allows them to conveniently use a mobile phone for communications. The approach is non-invasive and uses only auditory information, performing analysis, formant insertion, spectral enhancements and formant smoothing within the reconstruction process. In

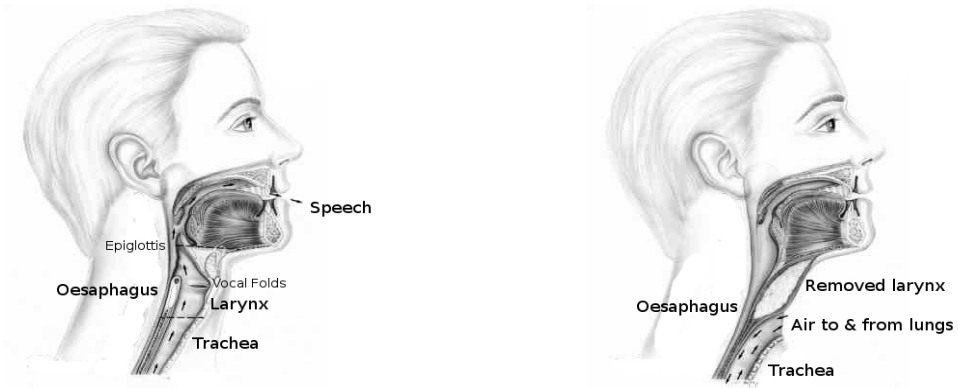
effect, natural sounding speech is obtained from spoken whispers, without recourse to surgery.

2.1 Laryngectomy

The speech voicing process relies upon modulated lung exhalation passing into the larynx where a taut glottis creates a varying pitch excitation which then resonates through the vocal tract, nasal cavity and out of the mouth. Within the vocal, oral and nasal cavities, the velum, tongue, and lip positions play crucial roles in shaping speech sounds; these are referred to collectively as vocal tract modulators [10].

Corey [11] notes that, "the larynx is the second most common site for cancer in the upper aerodigestive tract" and furthermore continues with "Laryngeal cancers account for approximately 1.2% of all new cancer diagnoses in the United States". National Cancer Institutes SEER data reveals that around 4 cases of larynx cancer appeared per 100,000 population from 1973-2000. Corey continues "Squamous cell carcinoma (SCC) is the most common histopathologic diagnosis, accounting for more than 95% of all laryngeal malignancies. Surgery, radiation, or both are the primary treatments for these cancers. Although organ preservation protocols and conservation laryngeal surgeries are in use today, patients with advanced or recurrent SCC of the larynx continue to undergo total laryngectomy in the course of their treatment." To understand the laryngectomy, figure 2.1 shows a mid-sagittal view of the vocal apparatus before and after the surgery.

Total laryngectomy patients will have lost their glottis and also the ability to pass lung exhalation through the vocal tract in many cases. Partial laryngectomy patients, by contrast, may still retain the power of controlled lung exhalation through the vocal tract. Despite loss of their glottis, both classes of patient retain the power of vocal tract modulation itself and therefore by controlling lung exhalation (or similar), they



(a) Before Laryngectomy (b) After Laryngectomy

Figure 2.1: Mid-sagittal view of human vocal apparatus showing the effects of laryngectomy operation.

have the ability to whisper [5]. In other words, they maintain control of most of the speech production apparatus. Therefore, the novel approach in this thesis is to reconstruct natural speech from the sound created by those remaining speech articulators. However since the major missing component is the pitch-generating glottis, this quest in effect is that of regenerating voiced speech from (pitch-less) whispers.

Various speech rehabilitation techniques exist such as oesophageal speech [12], tracheo-oesophageal puncture (TEP)[13], and the electrolarynx [14] (a brief review of these is presented in Section 2.2); but each suffers from weaknesses that range from learning difficulties to clumsy usage and heightened risk of infection.

Furthermore, all of these produce speech that is at best unnatural or monotonous. Concentrating on the electrolarynx as the main voice rehabilitation device adopted among laryngectomees [15], valuable efforts [16, 17] have recently been made to enhance the quality of the resulting speech by decreasing background and radiated device noise as well as to simplify its usage (i.e. producing a hands free variant). Despite these efforts, however, there has not been any effective method reported to resolve the mechanical sounding (robotised) generated voice characteristic.

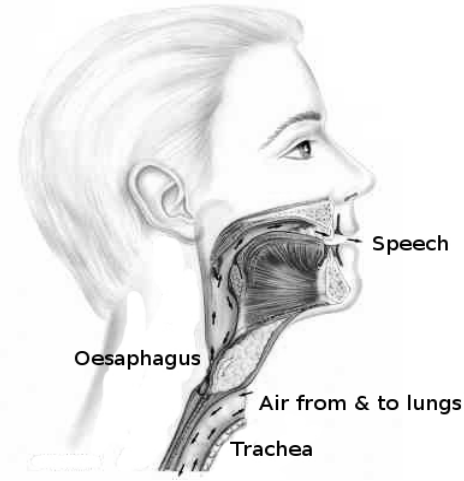


Figure 2.2: Mid-sagittal view of vocal apparatus in oesophageal speech.

2.2 Speech Rehabilitation

Existing methods of returning speech to post-laryngectomised patients are categorised under three different techniques as briefly reviewed below:

2.2.1 Oesophageal Speech

In this method, the patient is taught to use the oesophagus to expel air by means of stomach contraction rather than lung contraction [12]. The tongue must remain pressed against the roof of the mouth during this procedure to maintain an oesophageal opening. Figure 2.2 shows the status of voice production apparatus for speech generation through oesophageal speech.

Oesophageal speech can provide a harsh voice of low pitch, and loudness that is adequate for communication in small groups and quiet settings. Exceptional oesophageal speakers may have sufficient versatility and dynamic vocal range to approximate a normal voice, whereas some are unfortunately unable to master this method of communications rehabilitation [18].

Although quite difficult to learn, and often sounding unnatural, oesophageal speech is surprisingly intelligible. However a study by Hillman et al. [19] revealed that only

6% of total laryngectomy patients develop usable oesophageal speech (although five times as many do use or attempt to use it). The current status of oesophageal speech is that it has largely been eclipsed by tracheoesophageal puncture procedures and electrolarynxes [20].

2.2.2 Tracheoesophageal Puncture (TEP)

Surgical operations such as TEP [13] can produce higher quality speech and are particularly suited for those who have had a total laryngectomy and who breathe through a stoma. The TEP procedure creates a small hole to rejoin the oesophagus and trachea. This is then fitted with a one-way valve so that air from the lungs can enter the mouth through the trachea when the stoma is temporarily closed. Figure 2.3 shows the vocal apparatus using this method for speech generation.

Since the introduction of the TEP technique in 1980, numerous clinical and research studies have been published; these suggest modifications to the technique as well as studies on increasing quality and ease of speech production [20]. While TEP speech is considered by speech-language pathologists as the best method in terms of quality, only around 30% of post-laryngectomised patients use this method of alaryngeal speech [19].

The relatively good speech quality compared to the other voice rehabilitation techniques, and the high success rate of achieving usable voice requiring limited teaching are the main advantages of this method while the daily maintenance of the prosthesis by the patient, the recurrent leakage of the prosthesis after a period of time and the consequent need of replacement by the clinician (including the cost of replacement), are the disadvantages of this method. Furthermore, the prosthesis is clumsy in use and is a potential risk area for infection.

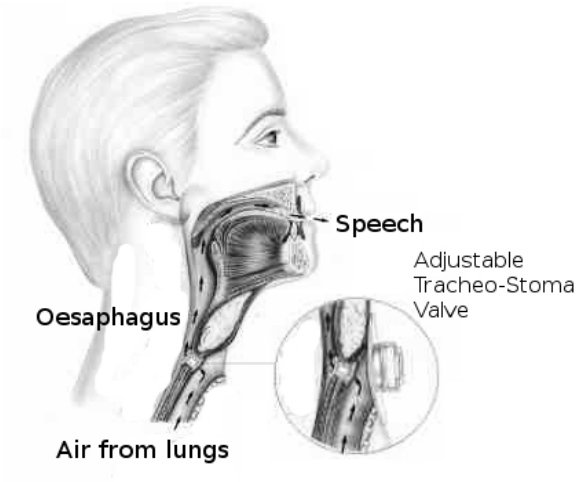


Figure 2.3: Mid-sagittal view of vocal apparatus in TEP.

2.2.3 Electrolarynx

The electrolarynx is a razor sized device that needs to be pressed against the side of the throat to resonate the vocal tract [14]. There are two types of electrolarynx: the neck and the intra-oral types, of which the former is the most widely used among the laryngectomees. During phonation, the hand-held device is held against the neck approximately at the level of the former glottis to insert a buzzing vibration into the oral and pharyngeal cavities by means of a built-in electromechanical vibrator. This sound source is transmitted through the neck tissues to resonate the vocal cavity. The user modulates this resonance to create speech by movements of articulators such as the lips, teeth, tongue, jaw and velum. Figure 2.4 demonstrates how a laryngectomee uses the neck type electrolarynx.

Speech generated by the electrolarynx is mechanical sounding and monotonous, although some modern units have a hand control to vary pitch. It has been found that the use of the electrolarynx is one of the easier methods of speech rehabilitation, and is more effective for communication in many situations [21]. Although oesophageal speech and tracheoesophageal speech are common in voice rehabilitation, electrolarynx

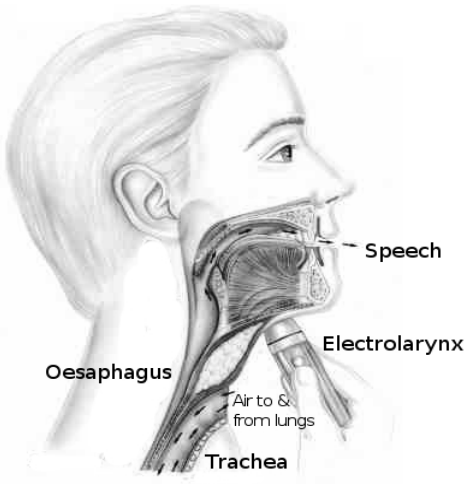


Figure 2.4: Mid-sagittal view of vocal apparatus in using electrolarynx.

phonation is the most commonly adopted method [20], with more than 55% of post-laryngectomised patients currently using it [19].

2.3 Summary

A brief medical background of laryngectomy and current speech rehabilitation techniques used to restore voice in laryngectomised patients was presented in this chapter. In particular, three methods commonly used by these patients; oesophageal speech, TEP, and electrolarynx, were discussed. Their advantages and disadvantages were outlined as well as the process/surgery required to utilise them.

By and large, all these techniques suffer from weaknesses ranging from unnatural monotonous speech to learning difficulties, clumsy usage and risk of infection. The engineering approach discussed in Chapter 5, by contrast, aims to produce higher quality speech by utilising a modified code excited linear prediction (CELP) codec to analyse, modify and reconstruct speech. For assessment purposes, the results of our system will be compared with electrolarynx output in Chapter 6.

Chapter 3

Whispered & Normal Speech Characteristics

Whispered speech as opposed to normally phonated (pitched) speech is the main focus of this research. Whispers which are simply defined as the lack of pitch in speech are considered in terms of articulation and acoustic characteristics while their particular properties affecting digital speech are presented in this chapter. Since whispered and pitched speech are mutually related (mostly and merely differing in the excitation source), first an overview of normal speech is provided and followed by whispered speech characteristics, built upon this foundation. Furthermore, phoneme classification, source-filter model, pitch filter, spectral feature extraction and so on, all play a major role in the speech regeneration methods and algorithms implemented in Chapter 5, thus, these are first described for normal speech and then their uses for whispers are distinguished accordingly.

3.1 Normal Speech

Modern digital speech communications is largely based on the knowledge of speech production. From the physiological point of view, the production of speech sounds involves the manipulation of an airstream. The air flows from the lungs along the vocal

folds in the larynx, passing through the pharynx and oral, nasal, or both cavities, and is eventually emitted from the mouth, nostrils or both. By the process of diaphragm contraction and relaxation, the lungs produce an airflow which is modulated by the larynx, processed by the vocal tract, radiated via the lips and the nostrils. The larynx provides several biological and sound production functions. In the context of speech production, its purpose is to control the stream of air that enters the vocal tract via the vocal folds [10].

In terms of sound production, the process may be described as follows: the airstream produces hardly any audible sound if the air can advance without any obstacles in its way as is the case of normal breathing. If somewhere in the airstream an obstruction occurs, it creates turbulence or blockage that is blown apart, probably repeatedly. This is a resonance which serves as the source of a sound. Before this sound reaches the listener's ears, it undergoes certain modifications by the dampening walls and resonant or shunting cavities of the vocal tract, being located behind or before the sound source [22].

Due to various mechanisms of production, speech sounds can be classified into two main categories of 'voiced' and 'unvoiced'. A sound is called voiced if the airflow is interrupted periodically by the movements of (vibration of) the vocal folds while unvoiced sounds are generated by a constriction at the open glottis or along the vocal tract causing a non-periodic turbulent airflow [10, 23]. However, there are other literature such as [24] which define voiced sounds as the interruption of airflow by the vocal folds repetitively and not necessarily periodically. This literature also states that a sound is unvoiced if the vocal folds do not interrupt the airstream repetitively, or when any other part of the vocal tract serves as a sound source.

Fortunately, this difference in definitions does not cause any confusing effect on technical concepts because both groups admit to calling 'voiced' sounds as those that

have a fundamental frequency (usually called pitch) which is the frequency of vibration of the vocal folds, either quasi-periodic or irregular as the extreme case of quasi-periodic.

According to [10], plosive sounds are caused by buildup of air pressure behind a complete constriction somewhere in the vocal tract, followed by a sudden opening. The released flow may create a voiced or unvoiced sound or even a mixture of both, depending on the actual collection of articulators operating on the airflow.

Apart from the above classification which is useful for purposes of speech signal processing, there is another categorisation in terms of linguistic characteristics in which phonemes (as the smallest vocal unit in one language) are divided into a number of groups. For example, 42 phonemes exist in English divided into 4 classes [10, 25] including vowels (such as /a/ in 'father'), diphthongs (such as /ou/ in 'boat'), approximants (such as /w/ in 'wet') and consonants (which are themselves subdivided into five subclasses including nasals such as /m/ in 'more', stops such as /t/ in 'tea', fricatives such as /f/ in 'free', aspirates such as /h/ in 'hold' and affricatives such as /t/ in 'chase'). Actually, this classification changes according to language, for example Shuzo Saito in [23] divides Japanese phonemes in two vowel and consonant classes. As an example, Figure 3.1 shows the spectral differences between a voiced (/z/) and unvoiced (/s/) phoneme in terms of pitch filter parameters (β and D) which will be explained in Section 3.1.3. The voiced phoneme has periodic variations and more energy (β) while the unvoiced sample is non-periodic and with less energy.

This phoneme feature and their corresponding classification are also important due to using of the same concept for whisper-voice conversion presented in Chapter 5 in a module called Whisper Phoneme Classification (WPC) while the focus of Chapter 4 will be analysing the vowels and diphthongs' features in whispered speech. In the

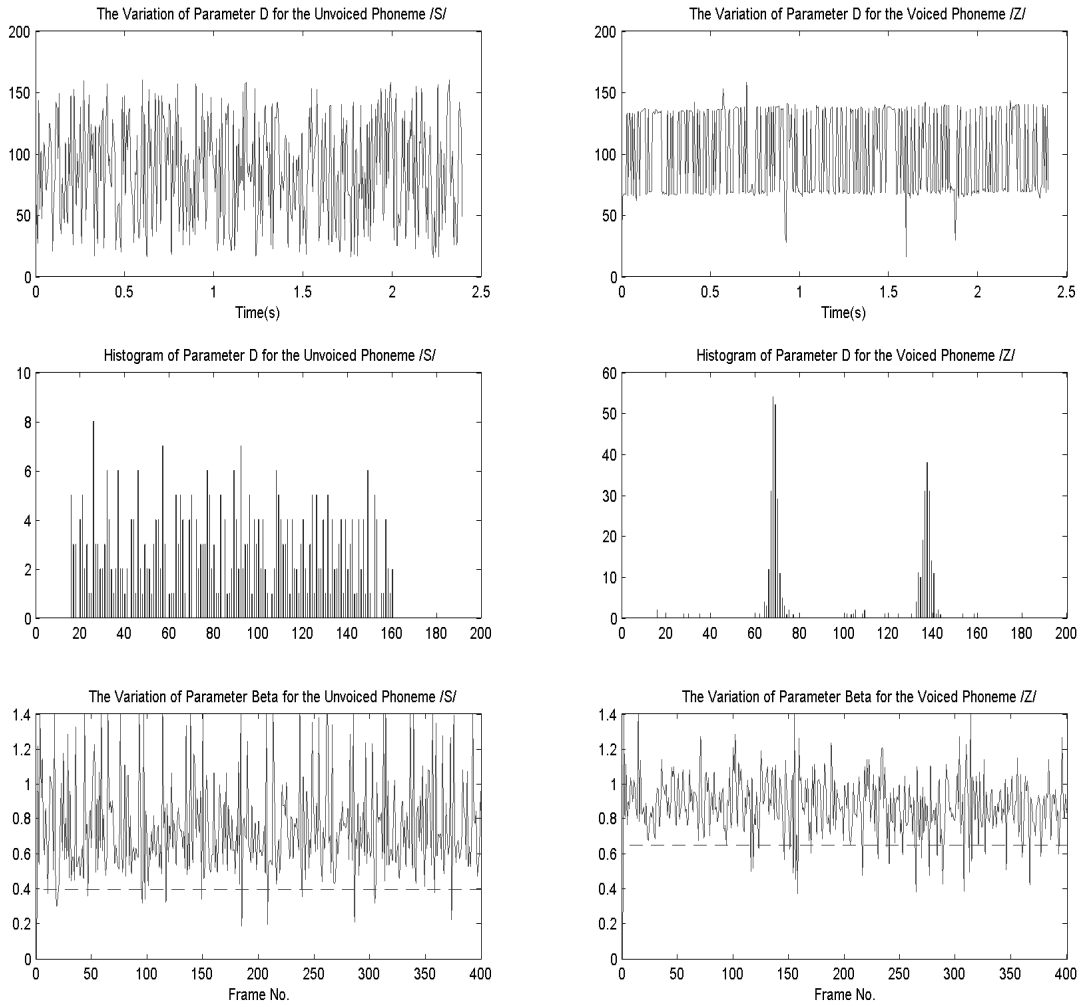


Figure 3.1: Comparison of pitch filter parameters in two phonemes: /s/ as an unvoiced sample (left) and /z/ as a voiced sample (right).

following, whispered speech in terms of acoustical and spectral features are considered mostly based on comparison with normal speech after an overview in the next subsections of the source filter model and linear prediction.

3.1.1 Source Filter Model

The source filter model proposed by Fant in 1960, explains the generation of all speech sounds in the same way. Actually the purpose of developing a model of speech production is not to obtain an accurate description of the real anatomy and physiology

of the human speech system. Hence, Fant formulated his model as described in [26] to achieve a simplifying mathematical representation for reproducing the essential characteristics of a normal speech signal.

As discussed in Section 3.1, it seems to be reasonable to design a parametric two-stage model consisting of an excitation (source) and a vocal tract filter to represent the human speech production system. Thus, the model consists of two main components: a) the excitation source featuring mainly the influence of the lungs and the vocal cords (voiced/ unvoiced/ mixed) and b) the time varying digital vocal tract filter approximating the behavior of the vocal tract (spectral envelope and dynamic transitions).

In more technical terms and by including all the processes implied in speech production, we can explain the source filter model as described in [27]: "the source-filter theory describes vocal sound production as a three step process: (1) generation of a steady flow of air from the lungs; (2) conversion of this airflow into a pseudo-periodically pulsating trans-glottal airflow, referred to as the voice source; and (3) response of the vocal tract to this excitation signal (modulation of the signal) which is characterised by the frequency curve or transfer function of the vocal tract". This theory is schematically illustrated in Figure 3.2 from [27].

In Figure 3.2, the vocal fold vibrations result in a sequence of voice pulses (bottom) corresponding to a series of harmonic overtones, the amplitudes of which decrease monotonically with frequency (second from bottom). This spectrum is filtered according to the sound transfer characteristics of the vocal tract with its peaks (i.e. the formants), and the valleys between them. In the spectrum radiated from the lip opening, the formants are depicted in terms of peaks, because the partials closest to a formant frequency reach higher amplitudes than neighbouring partials. In the simple model, the excitation source has to deliver either white noise or a periodic sequence of pitch pulses for synthesising unvoiced and voiced sounds respectively.

In this model, the source is often described in terms of a glottal flow which is modelled as a time-domain function called glottal flow model. Different models can be found in the literature proposing the parameterisation of the glottal flow such as Liljencrants-Fant (LF) model [28], Rosenberg model [29], and a more recent one by Veldhuis [30] derived from the Rosenberg model. While the well-known LF model has become a reference for glottal pulse analysis, the other two models are more common in speech synthesisers due to the less computational complexities. These models assign sequential steps for each level of opening and closing glottis usually based upon time and the amount of the passed airflow from glottis, then, a system is defined characterising input, output, and the corresponding transfer function based on these parameters.

Having a major role in prosody production as well as voice quality determination and speaker identification, glottal flow has also been considered in several studies such as [31], [32], and [33]. The glottal pulse width, the glottal pulse skewness, the abruptness of glottal closure are among the main factors found to be important for characterising the different voice types [31]. Also, the shape and periodicity of the vocal fold excitation are subject to large variations and such variations (i.e. pitch) are significant for the preservation of naturalness in speech [29]. From this point of view, a brief discussion is provided on pitch variation within the regeneration of normal voice from whispers in Chapter 5 while the glottal shape and other vocal characteristics in whispered speech are considered in Section 3.2.

As mentioned, the vocal tract is modelled as a time varying digital filter. Resonance is a key feature of the filter response. The oral, pharyngeal and nasal cavities of the vocal tract form a system of resonators. The behavior of a vocal tract resonance, or formant, is specified both in the time and the frequency domains.

Technically, the temporal and spectral shape of any speech sound is a function of both source and transfer function characteristics. As also described in [34], the source

of voiced sounds may be regenerated by submitting the speech wave to inverse filtering, which in effect cancels the poles and zeros of the transfer function [35, 36]. Since the transfer from volume velocity at the lips to the sound pressure in front of the speaker involves a differentiation, the net result of an inverse filtering without integration is to regenerate the time derivative of glottal flow. The negative-going spikes of this pulse train are the derivatives of glottal flow at instances of glottal closure. These are measures of excitation strength.

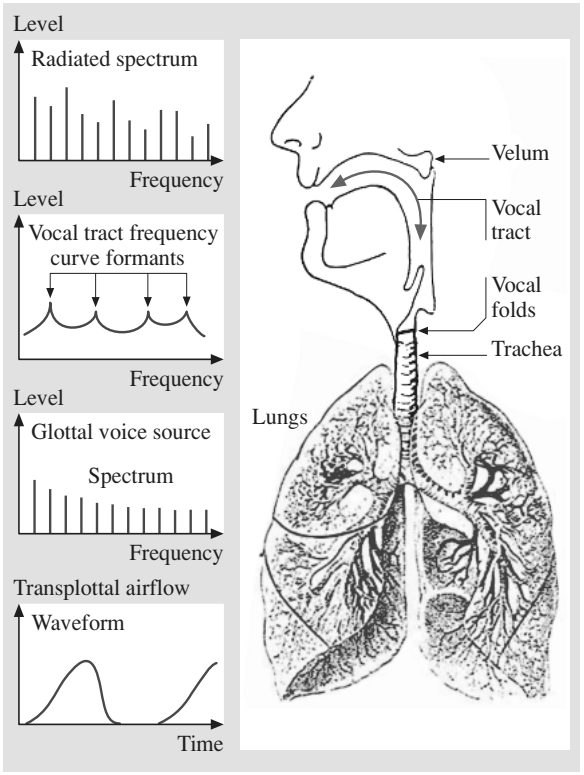


Figure 3.2: An illustration of the source-filter model corresponding to a 3-step speech production mechanism [27].

The model makes only one restriction: all systems of the model are linearly independent. However, the glottal sound source and vocal tract filter are dependent and they interact both acoustically and mechanically [37]. In short, the open glottis couples the vocal tract with the sub glottal system, while the closed glottal slit

separates both systems fairly well. Further, the changing width of the glottal opening during phonation varies the degree of coupling of the two cavities continuously. The shape and size of the filter cavities alter with the opening and closing glottal slit and therefore have varying resonance frequencies and bandwidths over one glottal period. Strictly speaking, the speech production apparatus is not a system of linear independent source and filter functions.

By and large, the source filter model provides a simple and straightforward way to look at the excitation in the larynx, independent of the modulation of the source in the vocal tract. The model has only a few restrictions and can be applied to any speech sound.

By considering speech production and source filter model sections, in short and simple words, we can define speech sounds by two main characteristics: 1) pitch, and 2) formant frequencies. From a speech compression and coding standards perspective, these two factors can be described as source and filter; i.e. pitch which simply is the frequency of the vocal fold vibrations forming a source and, the formants which are resonant frequencies of the vocal tract forming a filter.

3.1.2 Linear Prediction

Linear prediction (LP) is widely used in many speech applications including recognition, compression, modelling, etc [38, 39]. LP has also been the mainstay of speech communications technology and has been applied to speech coding since 1971. It actually relies upon several characteristics of speech derived from the fact that speech is produced by a human musculature system in which muscles shape the speech sound by their movements, hence are limited by a maximum speed similar to any other physical systems. The result is that although human speech varies considerably throughout an utterance, it actually remains pseudo-stationary for around 30 ms (because human muscle can not move infinitely fast) [40].

The speech production process is quite well modelled with LP, allowing a sampled speech signal to be represented in the following form [41, 42]:

$$x(n) = \sum_{k=1}^{P} \alpha_k x(n-k) + G.u(n) \tag{3.1}$$

where n is the time index, P represents the number of coefficients in the model, α_k, $k = 1, \ldots, P$ are defined as the linear prediction (or autoregressive) coefficients, G is the gain of the system, and $u(n)$ is the excitation signal, which can be modelled as either a quasi periodic train of impulses or a random noise source (also a combination of both for voiced fricatives such as 'v', 'z', and 'zh'). A periodic source matches voiced sounds such as vowels and nasals, while the noise source matches unvoiced or fricative sounds [43]. The parameters, α_k, determine the spectral characteristics of the particular sound for each of the two types of excitation, and are widely used directly in many speech coding schemes and automatic speech recognition systems.

Equation 3.1 can be rewritten in the frequency domain, by using the z-transform. If $H(z)$ is the transfer function of the system, we have:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} \alpha_k z^{-k}} = \frac{G}{A(z)} \tag{3.2}$$

which is an all-pole transfer function. This filter, $H(z)$, is generally considered to be a good model of the human vocal tract [38], when sampled at a sufficient rate.

As discussed earlier, LP-based speech coders, as a class, have been used successfully for many years. These coding algorithms are known as Linear Predictive Coding (LPC) based systems; which attempt to predict "as well as possible" a speech sample through a linear combination of several previous signal samples (this technique is also called forward linear prediction [43]). The number of previous signal samples used in the prediction determines the order of the LP model, which is denoted by P in equation

3.1. Examples of coders in which linear prediction is used, are: Multi Pulse Excitation coder [44] operating at around 16 kbps, Regular Pulse Excitation coder [45] operating at around 13 kbps, and Code Excited Linear Prediction [46] operating at between 4 and 16 kbps. Moreover, these LPC algorithms have led to several standards for the telecommunications industry.

LPC was originally proposed by Atal in the late 1960s and early 1970s for use in the coding and transmission of voice [47, 48]. In LPC-based coders, finding autoregressive coefficients (α_k) is the most important and sensitive part. Generally, to obtain these coefficients, the error between the speech segment and an estimate of the speech based on the prediction is minimised in the least squares sense.

As described in [49], besides the autoregressive coefficients, other parametric representations of the model can be used. Among these, the most common include the following:

- Complex poles of the prediction polynomial describe the position and bandwidth of the resonance peaks of the model.
- The reflection coefficients of the model relate to the reflections of the acoustic wave inside a hypothetical acoustic tube whose frequency characteristic is equivalent to that of a given LP model.
- Area functions describe the shape of the hypothetical tube.
- Line spectral pairs (LSPs) relate to the positions and shapes of the peaks of the LP model (LSPs are described in Section 5.3.1).
- Cepstral coefficients of the LP model form a Fourier pair with the logarithmic spectrum of the model (they can be derived recursively from the prediction coefficients).

All of these parameters carry the same information and uniquely specify the LP model by at least $P + 1$ numbers. The analytic relationships among the different sets of LP parameters are described by Viswanathan and Makhoul in [50].

LP has also played a major role in formant estimation techniques. Having a long history in the literature for the past five decades, formant frequency estimation methods are usually based on spectral analysis and peak picking/root finding techniques [51, 52, 53, 54] in which all relying upon the theory of the source filter model. These algorithms mostly exploit mathematically motivated analysis techniques such as LP, and through parametric or non-parametric methods of spectral estimations (i.e. based on Fast Fourier Transform (FFT) or autoregressive (AR) technique), the local maxima of the resulted spectrum or zeros of the LPC polynomial are identified as formants.

More recent approaches have also been proposed to improve the traditional automatic formant tracking methods in terms of accuracy, noise robustness, decreasing computational complexities, and etc. Applying Hidden Markov Model (HMM) [55], auditory models [56], dynamic programming [57], or pre-filtering/filter banks [58] are some significant examples of such efforts. In the following chapters, root finding/peak picking based upon LPC technique (as the dominant method of formant tracking) would be the core of our methods/applications within this respect while the improvements for whispered speech (Section 5.3.2) and efficiency of the current techniques for whispers are discussed as well (see Section 4.5.2 for the output of technique proposed in [58]).

To summarise, by returning to the source filter model, we can say that the glottis (source) is responsible for making pitch, creating some long term correlations in speech, while the filter part including the vocal tract, creates some short term correlations (these are primarily formants). LPC, by applying the vocal tract filter ($H(z)$ in equation 3.2), is used to predict and remove these short term redundancies while long

term correlations (basically pitch) are usually predicted by specialised pitch filters. As part of the coding process in many LPC-based compression algorithms, removing pitch from input speech leaves a much lower energy signal called the residual. This residual can be coded by LPC to parameterise the signal into a set of coefficients (usually numbering 8 to 14), determined by the order of the LP model.

3.1.3 Pitch Filters

In linear predictive coding algorithms, the influence of the pulse shape, vocal tract, and lip radiation are combined into one filter. The coding algorithm has to provide the synthesis filter with sufficient excitation. In this process, pitch filters play an important role in determining quality within medium and low bit rate speech coders.

Nowadays, a broad class of speech coders uses long term prediction (called LTP) filters to exploit the quasi periodic structure of voiced speech. In terms of implementation, there are different types of these pitch filters, but all share a basic concept which models pitch using a filter including (at least) one lag value and its appropriate amplitude (called one-tap pitch filter). In this subsection, the basic structure of a one-tap LTP filter is presented, and later, more advanced types such as multi-tap filters and pitch filters with fractional delays will be discussed.

3.1.3.1 Basic LTP Filters

Among the pitch filters used in speech coders, the basic one-tap pitch filter is widely employed in many low-bit rate coders [40, 59]. Generally, there are two kinds of pitch filter: the pitch filter at the analysis stage (coder) which is a non-recursive pitch prediction filter and the pitch filter at the synthesis stage (decoder) which is the inverse filter to the pitch prediction, i.e. it is a recursive filter.

A pitch filter typically is identified by an amplitude (β) and lag (D). The system function for a one-tap pitch predictor can be formulated as follows:

$$P(z) = \beta z^{-D} \tag{3.3}$$

where β scales the amplitude (predictor coefficient) of the pitch component and the lag D corresponds to the primary pitch period. The pitch lag D is usually updated along with other coded coefficients. The corresponding pitch synthesis filter has a system function as follows:

$$H_p(z) = \frac{1}{1 - P(z)} \tag{3.4}$$

As mentioned, in the context of speech coding, pitch predictors are most useful during voiced speech, since voiced speech is characterised as a quasi periodic signal with considerable correlation between samples that are separated by a pitch period. Pitch filters should be used in conjunction with formant predictors. Formant predictors remove the short correlations in speech to a large extent (see 3.1.2, vocal tract filter) while pitch predictors, by operating on the residual, try to remove long term correlations (hence the name 'long term'). However using the term 'pitch filter' is somewhat misleading in describing the action of this filter for unvoiced speech, hence the literature uses these terms (long term and pitch filter) interchangeably in many cases.

A simple and common method for calculating β and D is based on minimising the mean-squared error between an LPC residual (containing pitch), and the reconstructed pitch signal resulting from analysis (in an analysis-by-synthesis approach).

3.1.3.2 Advanced LTP Filters

If the pitch period spans an integral number of samples, a one tap pitch predictor can be suitable for codec purposes, but to model non-integer pitch periods, multi tap pitch predictors, sometimes with fractional delays, might be necessary. As described in [60]

and [61], a multi-tap pitch filter can be illustrated by several delays (usually bunched around the pitch lag value) along with related amplitude coefficients. For example, a third-order pitch predictor (3 tap pitch filter) can be given by:

$$P(z) = \sum_{k=-1}^{1} \beta_k z^{-D+k} \tag{3.5}$$

In this example, coefficients can have either three independent values or some related values such as being multiples of each other. This actually specifies the degrees of freedom in multi-tap pitch filters. As for this example, three non-zero coefficients at lags $D-1$, D, $D+1$, together have three degrees of freedom. Alternatively, this could be restricted to two degrees of freedom by assigning a symmetrical set of coefficients as: $\beta_{-1}= \beta_{+1}= \gamma$, $\beta_0=\beta$, both β and γ being chosen to give best performance. An accepted notation for pseudo multi tap filters is $nTmDF$, meaning n taps, m degrees of freedom [59].

For comparison purposes, Qian et al. [59] explain that the frequency response of a one-tap pitch synthesis filter shows a constant envelope while the spectrum of a conventional three-tap pitch filter often shows a diminishing envelope with increasing frequency in some voiced segments. Such a frequency response adds more pitch structure at low frequencies than at high frequencies. Consider the case of an integer lag, one-tap pitch filter. Suppose that the true pitch lag is in-between integer values. The frequency response of an integer lag filter will be up to 90 degree out of phase at the half-sampling frequency. At low frequencies such fractional lag errors do not affect the spectral fit. One effect of a shaped envelope such as that provided by a multi-tap pitch filter is that the effect of mismatches at high frequencies can be deemphasised.

Apart from multi-tap pitch filters, a fractional pitch lag is another accurate and efficient means to characterise speech periodicity in low bit-rate speech coders, especially for high pitched sounds. As a matter of fact, in a conventional pitch filter

implementation, the resolution of the delay is determined by the sampling rate. For an 8 kHz sampling rate, the resolution is not high enough for speech with short pitch periods. Thus, by using a fractional resolution for the delay, a higher prediction gain can be achieved [62, 63]. This is because it allows a better matching of the current and delayed samples, thereby reducing the prediction error. The fractional delay D can be expressed as the combination of an integer delay k and a fraction k/v, $k = 0, \ldots, v-1$, where v is the resolution of the fraction specified as a multiple of the sampling rate used for the input signal [61].

To summarise this topic, both fractional delay and higher order pitch predictors are realized with multiple taps. Consequently, there is a strong resemblance between both filters, and they produce comparable results [61].

3.2 Whispered Speech

Since the mechanism of whisper production is different from that of voiced speech, whispers have their own attributes, imposing several considerations for a pre-processing phase prior to analysis by a standard analysis-by-synthesis coder. However, the term 'whispered speech' itself encompasses two distinct classes of speech which we shall refer to as soft whispers and stage whispers [64] each differing slightly.

Soft whispers (quiet whispers) are produced by normal speakers when wishing to deliberately reduce perceptibility, such as whispering into someone's ear, and are usually used in a relaxed, low effort manner [7]. Stage whispers, on the other hand, are a combined kind of whisper one would use where the listener is some distance away from the speaker [64], but the speech is deliberately made to sound whispery. Some partial phonation, requiring vocal fold vibration [65] is involved in stage whispers. This thesis concentrates on the more common soft whispers, which are produced without vocal

fold vibration. These are often used in daily life, and furthermore closely resemble the type of whispers produced by laryngectomy patients.

In this section, characteristics of whispered speech are considered in terms of: a) acoustical features caused from the method of production (excitation, source-filter model, etc) and b) spectral features compared to normal speech.

3.2.1 Acoustical Features of Whispers

The essential physical feature of whispering is the absence of vocal cord vibration, and hence a missing fundamental frequency and harmonics [66] in speech. Using a source filter model [26], exhalation can be identified as the source of excitation in whispered speech, with the shape of the pharynx adjusted to prevent vocal cord vibration [67].

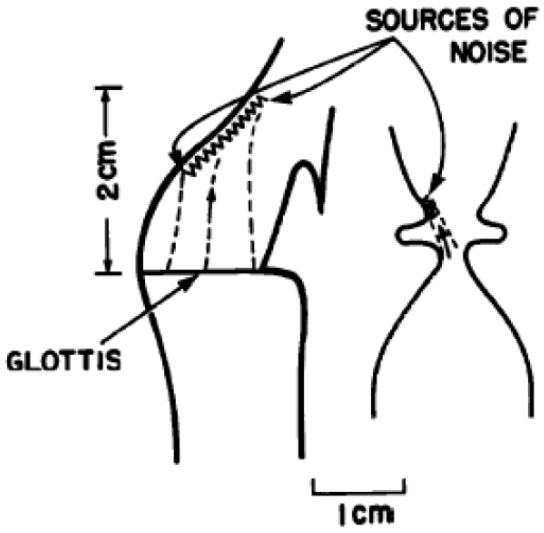


Figure 3.3: Schematised views of the laryngeal region in the sagittal plane (left) and in coronal section (right), indicating how airflow through the glottis might impinge on the surface of the epiglottis (left) or on the ventricular folds (right) to produce turbulence noise that is represented as a source of sound pressure [68].

When the glottis is abducted or partially abducted, there is a rapid flow of air through the glottal constriction. This flow forms a jet which impinges on the walls of

the vocal tract above the glottis. An open glottis in the speech production process is known to act as a distributed excitation source [68], in which turbulence noise is the primary source. Turbulent aperiodic airflow is thus the source of sound for whispers, giving them a rich 'hushing' sound [69]. In this respect, Stevens [68], has technically considered this kind of distributed source. Figure 3.3 from [68] shows a sketch of the portion of the vocal tract over a distance of a few centimetres above the glottis. A possible shape of the jet is indicated both in midsagittal section and in lateral section, and the sketches show the jet impinging on the walls of the glottal airway. Turbulence noise generated at the walls is expected to be the dominant source of noise in the vocal tract, assuming that the supraglottal airway is not constricted in the pharyngeal or oral regions.

Figure 3.3 suggests that the noise source is located either along the surface of the epiglottis about 1.0 to 2.5 cm downstream from the glottis, or at the level of the ventricular folds, which are about 0.5 cm above the vocal folds. The noise source at each of these locations can be represented as a sound pressure source. Stevens [68] makes a rough approximation of the equivalent noise source by assuming that one-third of the noise energy is located at the ventricular vocal folds (0.5 cm above the glottis), and the remaining two-thirds of the acoustic energy is generated on the epiglottis surface, distributed equally at 1.5 and 2.5 cm above the glottis. The corresponding spectra from each of these noise sources and a detailed discussion can be found in [68].

There are different descriptions at the glottal level for whispers: Catford [69] and Kallail and Emanuel [8] describe the vocal folds as narrowing, slit-like or slightly more adducted when whispering. Tartter [66] states that "whispering speech is produced with a more open glottis than in normal voices." Weitzman [64] by contrast defines the whispered vowels as "produced with a narrowing (or even closing) of the membranous glottis while the cartilaginous glottis is open."

Solomon et al.[7] studied laryngeal configuration during whisper in ten subjects from videotapes of the larynx, identifying three types of vocal fold vibration: i) the shape of an inverted V or narrow slit, ii) the shape of an inverted Y, iii) bowing of the anterior glottis. They concluded that soft whispers have the dominant pattern of a medium inverted V.

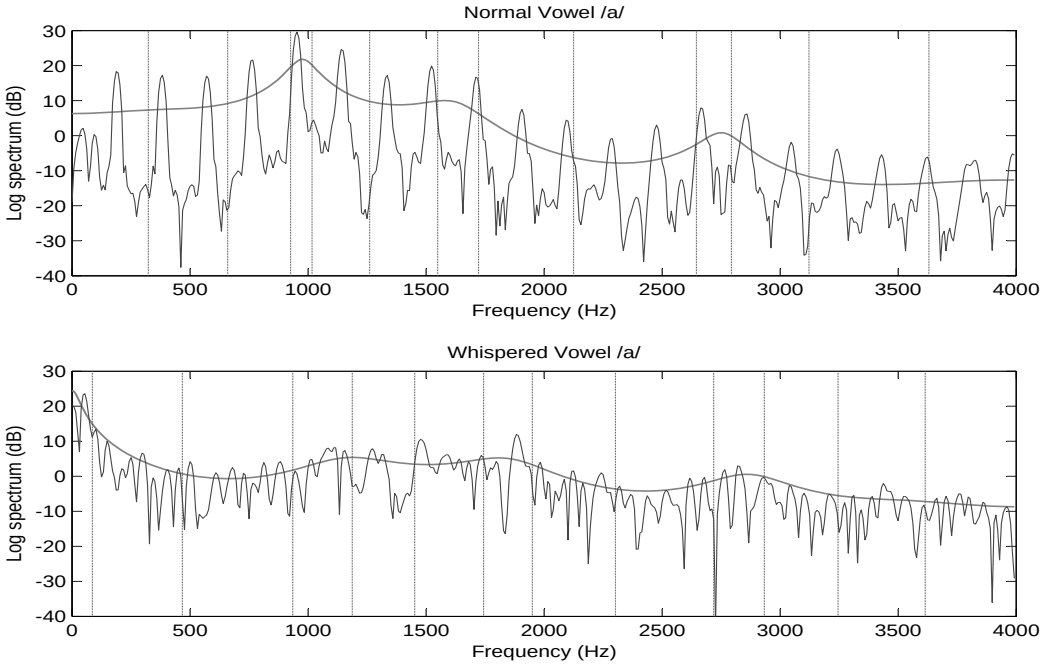


Figure 3.4: Comparison of the spectra for vowel /a/ in normally phonated speech (top) with whispered speech (bottom) for a single speaker during a single sitting. The smoothed spectrum overlay shows formant peaks existing in similar locations, but less pronounced for whispered speech. Furthermore, overlaid LSPs typically exhibit wider spacing for the whispered speech.

Morris [70] describes that the source-filter must be extended to behind the glottis to include both the glottis and the lungs in order to describe whisper speech; furthermore, he states that the source of whispered speech cannot be a single velocity source, but instead a distributed sound source must be defined to model the open glottis. Both of these suggestions are significantly different from a model of normal speech. Further glottal level analysis in whisper production, as well as physiological features of whispers, have been explained in detail in [68].

3.2.2 Spectral Features of Whispered Speech

If excitation in whisper mode speech is the turbulent flow created by the exhaled air passing through the open glottis, then the resulting signal is noise excited [67] rather than pitch excited. Another consequence of a glottal opening is an acoustic coupling of the upper vocal tract to the subglottal airways. The subglottal system has a series of resonances, defined by their natural frequencies with a closed glottis. The average values of the first three of these natural frequencies have been estimated to be about 700, 1650, and 2350 Hz for an adult female and 600, 1550, and 2200 Hz for an adult male [71], with substantial differences among the constituents of both populations.

Analysis shows that the effect of these subglottal resonances is to introduce additional pole-zero pairs into the vocal tract transfer function from the glottal source to the mouth output. The most obvious acoustic manifestation of these pole-zero pairs is the appearance of additional peaks or prominences in the output spectrum. The influence of zeros can also be seen sometimes as minima in the spectrum [68].

The spectral characteristics of whispered speech do exhibit some peaks in their spectra at roughly the same frequencies as those for normally phonated speech sounds [72]. These 'formants' occur with much flatter power frequency distribution, and there are no obvious harmonics in the spectra corresponding to a fundamental frequency [66]. Figure 3.4 shows this feature by contrasting the spectra of the vowel /a/ spoken in a whisper and in a normal voice.

Whispered vowels also differ from normally voiced vowels. All formant frequencies (including the important first three) tend to be higher [73], particularly the first formant which shows the greatest difference between two kinds of speech. Lehiste [73] reported that $F1$ is approximately 200-250 Hz higher, whereas $F2$ and $F3$ are approximately 100-150 Hz higher in whispered vowels. These approximate numbers will be

studied in Chapter 4 where a classic $F1 - F2$ vowel space is generated for whispered speech.

Furthermore, unlike phonated vowels where the amplitude of each higher formant is less than for lower formants, whispered vowels usually have second formants that are as intense as first formants. These differences mainly in first formant frequency and amplitude are thought to be due to the alteration in the shape of the posterior areas of the vocal tract, including the vocal cords which are held rigid [74].

By considering these differences between normal and whispered speech in terms of both production and spectral features, we can determine modifications which are required to be made to whispered speech to adapt this form of speech to work effectively with communication devices and applications which have been designed for normal speech as well as innovating required algorithms for whisper-voice conversion.

3.3 Summary

The human speech production mechanism, in conjunction with the well-known source filter model, was explained in the first part of this chapter. Following this, an overview of pitch prediction filters used in speech codecs was presented while eventually in the second part of this chapter, whispered speech characteristics in terms of acoustic and spectral properties were described.

These concepts of pitch and vocal tract filters along with the whispers' features would be our main base within the implemented framework discussed in Chapter 5 based upon a modified CELP codec. By applying spectral enhancement along with some other modules corresponding to whispered speech features described in this chapter, a whisper to voice conversion method will be explained in the following chapters.

Chapter 4

Vowel Space for Whispered Speech

As described in the preceding chapter, whispered speech is known as a relatively common form of communication, used primarily to selectively exclude or include potential listeners from hearing a spoken message. Despite the everyday nature of whispering, and its undoubted usefulness in vocal communications, whispers have received relatively little research effort to date, apart from some studies analysing the main whispered vowels and some quite general estimations of whispered speech characteristics as discussed in Chapter 3. In particular, a classic vowel space determination has been lacking for whispers. For voiced speech, this type of information has played an important role in the development and testing of recognition and processing theories over the past few decades, and can be expected to be equally useful for whisper-mode communications and recognition systems.

The purpose of this chapter is to redress the shortfall through a vowel formant space for whispered speech, while comparing the results with corresponding phonated samples. As part of this work, the author travelled to the UK to be a short term visiting researcher at the University of Birmingham (School of Electrical, Electronics, and Computer Engineering). There, the well-known speech expert Professor Martin Russell, advised him on constructing and running the experiments presented in this chapter. This work also led to our publication of the first comprehensive vowel space

for whispered speech.

Since the study was conducted using speakers from Birmingham, the analysis here also briefly considers the possible effect of their British West Midlands (WM) accent in comparison with a Standard English (or received pronunciation, RP) accent. Thus, the chapter presents the analysis of formant data showing differences between normal and whispered speech while also considering accentual effect on whispered speech. Furthermore, a brief discussion on automatic method of vowel segmentation/extraction based upon Hidden Markov Models (HMMs) through HTK and ESPS toolkits for formant frequency measurements is also presented, since this was fundamental to the analytical method.

The acoustic analysis including details of the recording, speakers, equipment and measurement methods, are described in the following sections, while the results are outlined separately for men and women as well as a discussion provided on findings (including the consideration of possible accentual effects British West Midlands accent in whispers and normal speech).

4.1 Importance of Formant Study

Acoustic measurements of phonated vowels and diphthongs form foundational material for the speech processing and recognition fields. Significant research effort [75, 76, 77, 78, 79, 80, 81], mainly based upon acoustic characteristics of normal vowels, shows the importance of these measurements, while numerous studies [82, 83, 84], in turn, have considered formant patterns in terms of vowel diagrams and the corresponding characteristics of normal vowels.

Despite this extremely strong literature supporting normal vowels, very little published material can be found on whispered speech, relating to vowel space. Apart from some studies describing the vocal mechanism of whispers' production mostly on

a glottal level [7, 69, 85, 86], as well as a recent study on whispered consonants [87], the few notable studies on whispered vowels [8, 88, 89] are mainly concentrated with just a few main vowels /ɪ,ɛ,æ,ʌ,ʊ/, and concluding with general comments on vowel placement such as "higher formants in comparison with normal vowels". However, accurate acoustic measurements of the precise amount of shift for each vowel/diphthong are lacking. Thus, whispered speech still lacks an acoustic vowel space determination (a classic $F2$x$F1$ plane) for researchers to refer to. Whisper vowel diagrams would be useful not only for common speech processing/recognition applications that may need to work with portions of whispered speech, but also can help those working in the biomedical engineering field of whisper-to-voice reconstruction, particularly rehabilitation of post-laryngectomised patients through restoring their normal sounding speech [90, 91].

Therefore, by considering the importance of having a comprehensive formant study on whispers and based upon the literature mentioned in Chapter 3 outlining acoustic and spectral features of whispered speech, the aim of this chapter is, as mentioned, to establish a classic formant plane for all 11 whispered vowels in common use for English speech, through analysing the formant contours of whispered samples in a /hVd/[1] structure.

4.2 Birmingham and RP Accents

Accent is known as a distinctively characteristic manner of pronunciation, usually associated with a community of people with a common regional or social/cultural background [92]. Due to its subjective nature, accent is a difficult entity to quantify. Geographical classifications for describing/identifying accents might suggest a partial

[1] A group of words starting with /h/ and ending with /d/ with a vowel inserted in between. The reason for using a /hVd/ carrier is discussed in Section 4.3.

solution to this, however the neat borders which separate areas on a map are not as clearly applied to accents. A listener's accent is also another important factor in distinguishing the existence of an accent in another person [93]: if people are not from the same accent group, they will most likely identify the existence of an accent, yet they may not notice the subtle differences between accents in that group.

Much of the existing literature on accents in relation to formant study has focused on the comparing main accents (which are easier to identify) such as accents of British RP (received pronunciation) with general American, Australian or non-native speakers of English [92, 94, 95, 96]. However, different British accents across the British Isles have also been studied [97, 98, 99].

Since British RP and Birmingham accents were considered for the purpose of this study, a brief vowel characteristics/introduction of these accents are presented in the following; mostly extracted/summarised from the works of Wells [97] and Clark [100].

RP might be what anyone living in the United Kingdom hears constantly from radio and television announcers and news readers and from many other public figures. Everyone in Britain has a mental image of RP, even though they may not refer to it by that name, and even though the image may not be accurate. It has been estimated that only about three percent of the English population speak RP [98]; yet it is the accent mostly taught to foreign learners due to it having the best chance of being understood among all other British accents.

There are various kinds of RP in the literature which are commonly classified into: upper-class RP (U-RP), mainstream RP, adoptive RP, near RP, conservative RP, and advanced RP. All of these sets of distinctions within RP have a central tendency to mainstream RP, so the results and discussion presented in Section 4.6 will be focused on this type of RP.

Today, the term West Midlands (WM) is generally used to refer to the conurbation that includes the major cities of Birmingham, Wolverhampton, Walsall, West

Bromwich, Coventry and many surrounding towns, and can also be used to refer to speech associated with the modern urban area. According to Clark [100], since it is unclear whether and if so to what degree the dialect of the large but geographically distinct city of Coventry may differ from other West Midlands varieties, the term West Midlands will be taken to refer to Birmingham and the wider Black Country, unless explicitly stated otherwise.

Trudgill [101] provides these main diagnostic features for the WM accent:

- lacking a FOOT-STRUT distinction
- lacking a Trap-Bath distinction
- having happy-tensing (this refers to the process in which final lax vowel /ɪ/ becomes tense and closer to /i/ in words like 'happy', 'duty', and 'city' [102].)
- being non-rhotic
- distinguishing FOOT from GOOSE and LOT from THOUGHT
- having /h/-dropping as a normal feature
- having broad diphthongs for FACE and GOAT

The features corresponding to vowels/diphthongs from RP and WM accents are considered in Section 4.6 where the results of the experiments are discussed. However, our main focus throughout this thesis will be the whisper-speech characteristics rather than these regional accents of the test participants.

4.3 Subjects and Recordings

Speakers of this study consisted of ten middle-aged volunteers (5 men and 5 women, 35 to 45 years old) raised and living in Birmingham all of their lives. An additional

criterion of one's parents having lived in the area most of their lives was also used for the selection of volunteers.

Audio recordings were made of subjects reading lists containing 11 framed vowels (/ɪ,i,ɛ,æ,ɑ,ʌ,ɒ,ə,ɔ,ʊ,u/) in an anechoic chamber, five times with normal phonation and five times in whispered mode (total 10 times).

Subjects read from five different randomisations of a list containing the words 'heed', 'hid', 'head', 'had', 'hard', 'hudd', 'hod', 'heard', 'hoard', 'hood', and 'who'd'. Since the objective is to find out how *ordinary* people from Birmingham speak the vowels in the specific words, /hVd/ carrier gives actual/meaningful words in most cases, except of 'hudd' (although this does occur as a reasonably common family name) while this also keeps the current study aligned with the previous acoustic studies on vowels [82, 83, 84] through following the same pattern.

Furthermore, having a plosive phoneme such as 'd' at the final syllable, makes it simple to detect vowels in between from carriers within both automatic or manual methods of extraction; particularly, due to showing a peak of energy in both whispered and spoken modes after a very short silence, 'd' can be a good choice for the final syllable.

Recordings were made of 5 readings of the list in each whisper and normal modes (total $5 * 11 * 2 * 10 = 1100$ samples). The details of the interface is described in Section 4.4. If the subjects stumbled over the samples, re-recording of the samples was allowed. Speakers could repeat the sample until an accurate pronunciation was achieved.

4.4 Equipments and Interface

Speech was read, and recorded directly onto a laptop computer in an anechoic chamber. The microphones used were an Emkay head mounted microphone and a Telex desk

microphone (for near and far field recording, respectively). An Edirol UA-5 USB sound card interface bypassed the sound card of the laptop, removing any variation in the recordings due to different hardware. An Emkay VR3294 Battery Box provided a stable bias voltage for the microphones.

A special prompt-based recording software, developed by the University of Birmingham, was used as the recording application. Any set of prompts specified in a separate xml-formatted file with different login options, can be loaded into the prompt-recorder at run time; so the randomised lists of vowels in /hVd/ carriers as mentioned in Section 4.3 were customised using this application. Appendix A shows the corresponding prompt.

The recorded speech was sampled at a rate of 22050 Hz with 16 bit resolution. The ABI application included level meters for both microphone inputs. At the beginning of each recording session, the subject spoke for a few seconds while the input levels were adjusted on the USB sampler to achieve a peak SNR of - 12 dB. The record, stop, play and accept buttons were all controlled by the person recording the speech, so the subjects need only concentrate on reading the text in front of them. Subjects were seated in front of the laptop and the headset microphone placed on their head with the microphone angled about 5 cm away from the right corner of their mouth. The desk-mounted microphone was placed to the left of the laptop.

4.5 Formant Contours

Segmentation/extraction of normal speech at a phonemic or sub phonemic level has been generally an attractive research field in speech recognition. The different approaches and solutions proposed by researchers can be mainly categorised into two major recognition classes: manual and automatic [93, 103, 104, 105, 106]. Based upon specific acoustic cues, these approaches try to identify/classify different groups

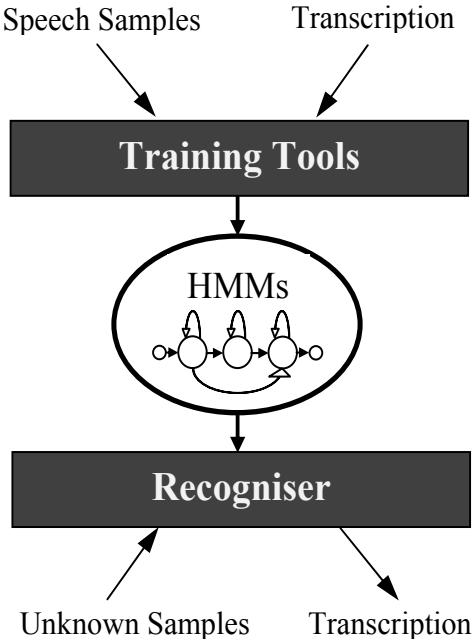


Figure 4.1: Block diagram of HTK toolkit, as used for speech analysis.

of phonemes. The process of segmentation in whispered speech is more complicated [87]. Although for the purpose of this research, a combined approach based on manual observation has been finally taken, an automatic method was firstly considered/implemented. This automatic approach based on forced alignment through Hidden Markov Model (HMM) is described in the following section.

4.5.1 Automatic Approach Using HMMs

An automatic approach to formant analysis based on forced alignment using single emitting state phone-level HMMs to detect the vowel centres and ESPS for formant frequency measurement was implemented. For this purpose, a toolkit called HTK for building HMMs was used. HMMs can be used to model any time series and the core of HTK is similarly general-purpose. However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognisers. Figure 4.1 shows the block diagram of HTK toolkit modules, while the core of the system is built upon

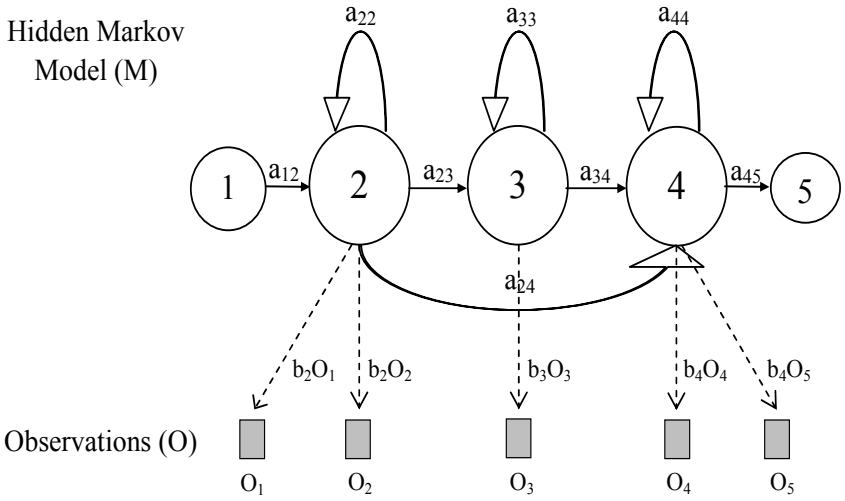


Figure 4.2: Typical Hidden Markov Model used for speech recognition.

HMMs. Since our experiment deals with isolated word recognition, in the following, HMMs are briefly described from this aspect.

HMMs are statistical models that are commonly used in speech recognition [107]. The statistical nature of HMMs makes them appropriate for modelling a continuously varying sequence of observations of the type which represents a speech signal [108].

Briefly speaking, in HMM based speech recognition, it is assumed that the sequence of observed speech vectors corresponding to each word is generated by a Markov model, M. A Markov model is a finite state machine which changes state once every time unit and at each time, t, that a state j is entered, a speech vector o_t is generated from the probability density $b_j(o_t)$. Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}.

The generative nature of HMMs means that Bayes' rule, equation 4.1, can be applied for recognition:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{4.1}$$

Where the posterior probability $P(W|O)$ of the word sequence, W, being the true transcription of the given utterance O (the set of feature vectors observed), equals

the likelihood $P(O|W)$ of O being observed if w was spoken, multiplied by the prior probability $P(W)$ of W and divided by a normalising term. The posterior probability $P(W|O)$ gives information about which word sequences are being said. The W for which $P(W|O)$ is largest is the one most likely to have been said, thus, if the speech recogniser outputs this W, it will have the greatest chance of being correct (O is the vector sequence $O = \{o_1, o_2, \ldots, o_T\}$). Having a model, M, probability of $P(W)$ can be equated with $P(M)$. The acoustic model $P(O|M)$ is generally based on HMM.

A language model (LM) uses the linguistic constraints of language to calculate the probability of a word sequence occurring. N-gram LMs assume that the probability of a word occurring depends only on the previous $N-1$ words. For $N = 1, 2, 3$ the LM is referred to as Uni-gram, bi-gram and trigram LMs respectively.

The diagram in figure 4.2 illustrates the concept of a typical HMM used in speech processing. This model shows 5 states (3 emitting states), S_i, and the permitted transitions between these states, a_{ij} in order to generate the sequence o_1 to o_5. When dealing with speech, it is reasonable to only consider left to right models [107], so backward transitions are not permissible but state repetitions are (in HTK, the entry and exit states of a HMM are non-emitting).

The joint probability that O is generated by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities and the output probabilities. For example for the state sequence X in figure 4.2, we would have:

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)a_{34}b_4(o_4)a_{44}b_4(o_5)a_{45} \tag{4.2}$$

However, in practice, only the observation sequence O is known and the underlying state sequence X is hidden. This is why it is called a Hidden Markov Model.

Given that X is unknown, the required likelihood is computed by summing over all possible state sequences $X = x(1), x(2), x(3), \ldots, x(T)$, that is:

$$P(O|M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \tag{4.3}$$

where $x(0)$ is constrained to be the model entry state and $x(T+1)$ is constrained to be the model exit state. If $\sum_X$ is replaced by max_X, the expression is called 'Viterbi' likelihood [109], which is the likelihood that considers only the best path. This is usually close to the likelihood obtained by summing over all possible sequences, because the sum tends to be dominated by the largest term:

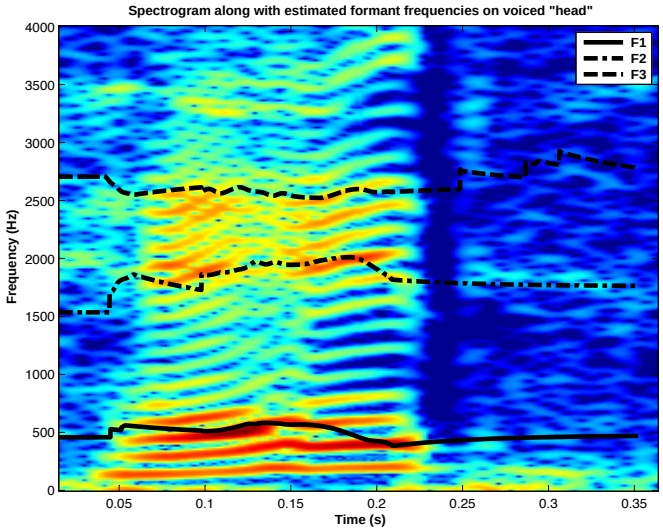
$$\hat{P}(O|M) = \max_{X} \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right\} \tag{4.4}$$

Then within the training phase, model parameters are usually being identified through a so-called Baum-Welch algorithm [110] based on Estimation-Maximisation (E-M) recursive process.

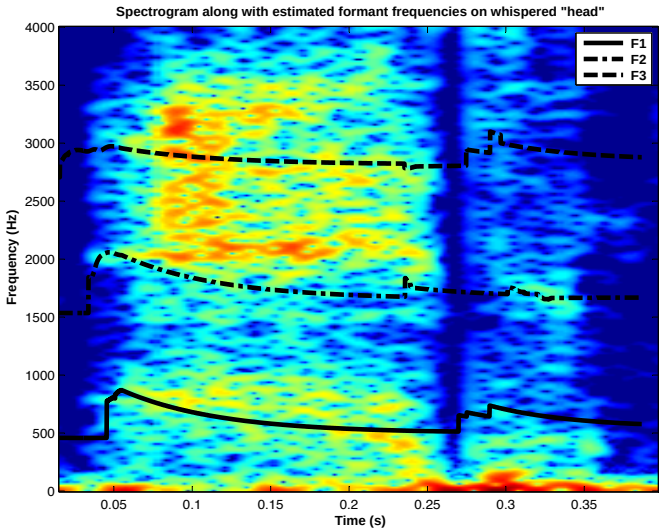
For the purpose of this study, an automatic method of estimating the first 3 formant values for each vowel was implemented. The formant values for vowels are generally estimated from the steady state portion of their spectrogram. In ASR terms, this might be expected to correspond to the centre state of a 3 state HMM of the vowel in question. Thus, it is sensible to build a set of 3 emitting state, triphone[1], HMMs to perform forced alignment of the data, but according to the previous experiments [93], the emitting state numbers were reduced to 1.

Forced alignment produces state-level, time-aligned transcriptions of the data, which provided the mid-point timing of the centre states of the vowel models. D'Arcy

[1]A triphone is a model of the acoustic realisation of a phone given the immediately preceding and proceeding phones. In fact, phones are small units of sounds in language, and are defined by the International Phonetic Association's (IPA) alphabet whereas the single phones from this alphabet are simply called monophones.



(a) Formant trajectories for voiced "head"



(b) Formant trajectories for whispered "head"

Figure 4.3: First three formants detected by the algorithm proposed in [58] overlaid on the corresponding spectrograms for a) voiced "head", b)whispered "head". The formant trajectories are accurate for voiced sample while formants of whispered samples (particularly first and second formants) are not detected accurately.

[93] explains that many misalignments (even in the restricted vocabulary) were found in the transcriptions after inspection of the output files: "The start and end times of vowels were identified by sight from their spectrograms and it was found that for many examples the automatic estimations had been inaccurate"[93]. Then, she reduced the

number of emitting states to 1, and through testing with a large number of Gaussian mixture components (using the ABI corpus training set [111]), she found that "the most accurate phone level, time aligned transcriptions of the words" are produced [93]. Therefore, the same 1 emitting state was created and trained with the collected data from Section 4.3. Using this data the start and end time of the portion of the vowel from which the formants were to be measured was identified.

The ESPS toolkit uses Linear Predictive Coding (LPC), as described in Chapter 3 to measure the formant frequencies of speech. This produced a list of the first 3 formant values for each successive time frame over the duration of the vowel. ESPS calculates the value of the features at regular (specified) intervals over given windows of time, in this case 10 ms. The formant values were taken at the centre of each vowel, according to the forced alignment. ESPS outputs a parameter measurement for successive speech segments over the whole utterance, and the average value was calculated for each vowel.

However, due to many outliers resulting from the use of whispered speech, the more time consuming manual methods explained in the following subsection were preferred. In fact, by looking at the output generated by ESPS, most of the measured formant frequencies were not accurate; so all results have been verified manually one-by-one.

4.5.2 Manual Method

Different methods were combined for accurate extraction of the first three formant frequencies for each sample in the normal and whisper modes. After manually clipping the steady state of vowel duration by removing the /h/ and /d/ carriers, the analysis methods, which are mainly based upon manual observation of the results, outlined as follows:

- peak-findings through direct observation of 12-pole, 128-point linear predictive coding (LPC) spectra on every 6 ms over 12 ms Hamming windowed segments (256-point)
- looking at the results of the robust formant tracker implemented in [58]
- observation of the gray scale spectrograms (both wide and narrow band)

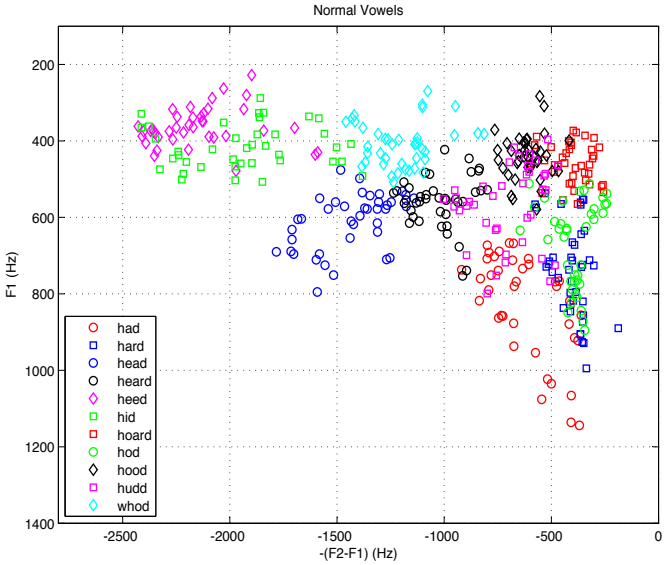
For example, figure 4.3 demonstrates the formant trajectories resulting from the formant tracker introduced by Mustafa and Bruce [58] for a voiced and whispered vowel. As can be seen, the tracker works well on voiced samples but does not show accurate results for the whispered sample on detecting formants (the detected first and second formant trajectories have around 200 Hz offset). This will be one of the main aims of Chapter 5 in which designing an efficient formant tracker for whispered speech is discussed.

Finally, decisions about formant frequencies were determined by the outcome of the methods mentioned, as well as by comparing the results to select the most accurate representation (a general knowledge of acoustic phonetics, such as the close proximity of $F2$ and $F3$ in /ɪ/ and /ɛ/, also played a role in this process). Figure 4.4 shows the individual data points of the measured first and second formants through this approach for a)normal samples and b)whispered data while a few redundant points have been omitted to be clearer.

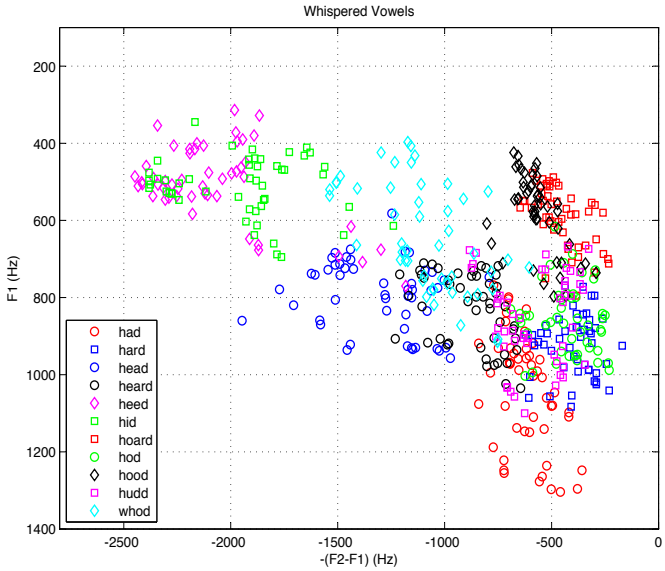
4.6 Results and Discussion

4.6.1 Results

Acoustic measurements on formant values of the /hVd/ samples for both normal and whisper modes separately for men and women are presented in this section. Since



(a) Normal data points



(b) Whisper data points

Figure 4.4: Values of $F1$ and $F2 - F1$ for 11 vowels from five men and five women recorded 5 times voiced and 5 times whispered. A few redundant data points have been omitted for better clarity. The words heed, hid, head, had, hard, hudd, hod, heard, hoard, hood, and who'd contain vowels /ɪ,i,ɛ,æ,ɑ,ʌ,ɒ,ə,ɔ,ʊ,u/ respectively.

the data were collected in the West Midlands, with particular accent characteristics throughout the conurbation of Birmingham (see Section 4.2), the amount of vowel variation in the Birmingham accent, compared with Standard English (Received Pro-

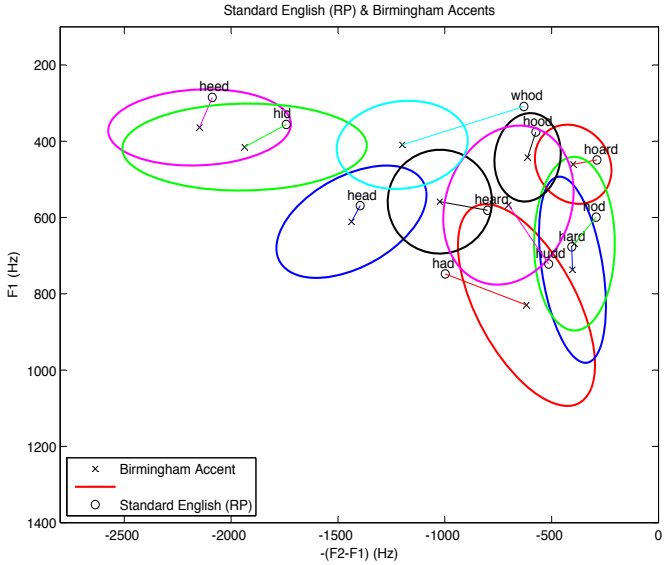


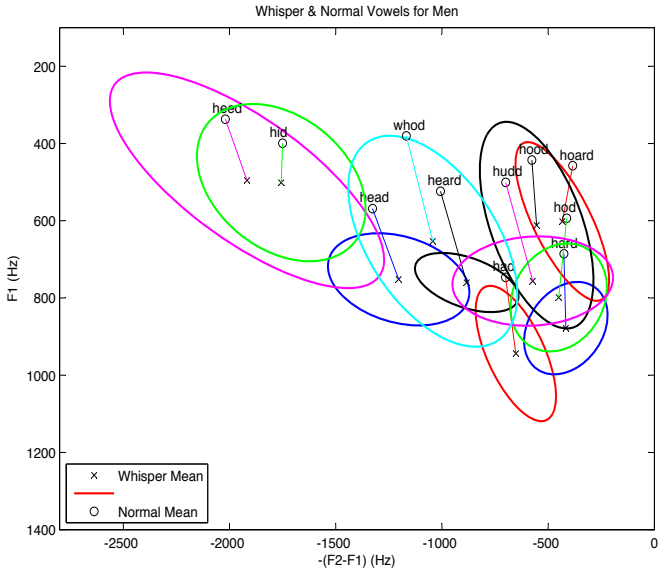
Figure 4.5: Average values of $F1$ and $F2 - F1$ for standard English and Birmingham accents. Ellipses fit to each vowel category in Birmingham accent. The average shift amounts also have been joined by a line. The words heed, hid, head, had, hard, hudd, hod, heard, hoard, hood, and who'd contain vowels /ɪ,i,ɛ,æ,ɑ,ʌ,ɒ,ə,ɔ,ʊ,u/ respectively.

nunciation, RP) is also provided for referencing purposes, in addition to normal and whisper variations which are the primary aim of the chapter. As mentioned before, RP formant values were obtained from Wells' work [97].

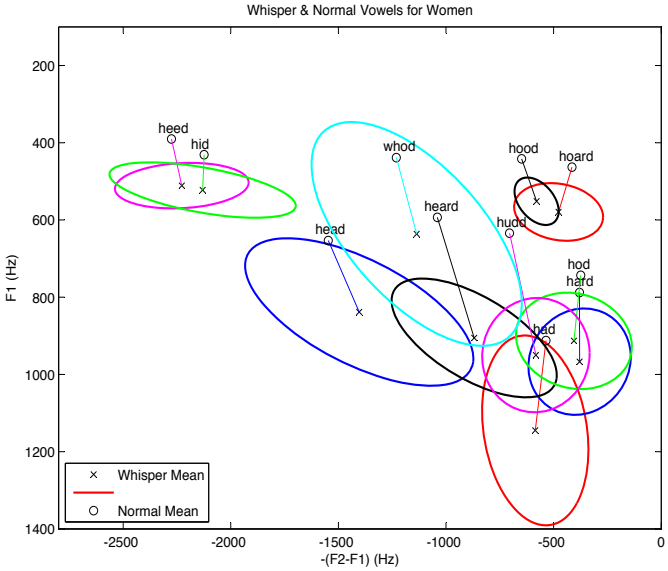
Figure 4.5 shows the average frequencies for $F1$ and $F2 - F1$ along with ellipses indicating the standard deviation within each vowel category. The variations between Birmingham accent and average formant frequencies in RP accent have been shown. As can be seen, /ɪ/ and /i/ as well as /ɑ/ and /ɒ/ show a high degree of overlap in this accent while shift of vowels have been identified. The detailed discussion and the verification of these data in comparison with RP is presented in 4.6.2.1.

In figure 4.6, the variations between normal and whispered vowels are separately illustrated for male and female speakers, while the corresponding acoustic vowel diagrams on a $F1$x$F2$ space are presented in figure 4.7 based on average formant frequency. Again, this shows normal and whisper samples for a)men, and b)women.

Figure 4.8 compares average values of third formant in normal samples with whis-



(a) Whisper vowels versus normal vowels for men



(b) Whisper vowels versus normal vowels for women

Figure 4.6: Average values of $F1$ and $F2 - F1$ for normal and whispered vowels in: a)men, b)women. Ellipses fit to each vowel category in Birmingham accent for a)men, b)women. The average shift amounts also have been joined by a line. The words heed, hid, head, had, hard, hudd, hod, heard, hoard, hood, and who'd contain vowels /ɪ,i,ɛ,æ,ɑ,ʌ,ɒ,ə,ɔ,ʊ,u/ respectively.

pered ones. Overall, the $F3$ values from both normal and whisper samples are quite similar, with all vowels except /u/, averaging less than 100 Hz shift from normal to whispered data. These slight shifts of $F3$ average 2.0% (σ : 2.4%) for men and 2.2%

(σ : 1.6%) for women. Thus, third formants have not been much affected by whispering as compared to first and second formants.

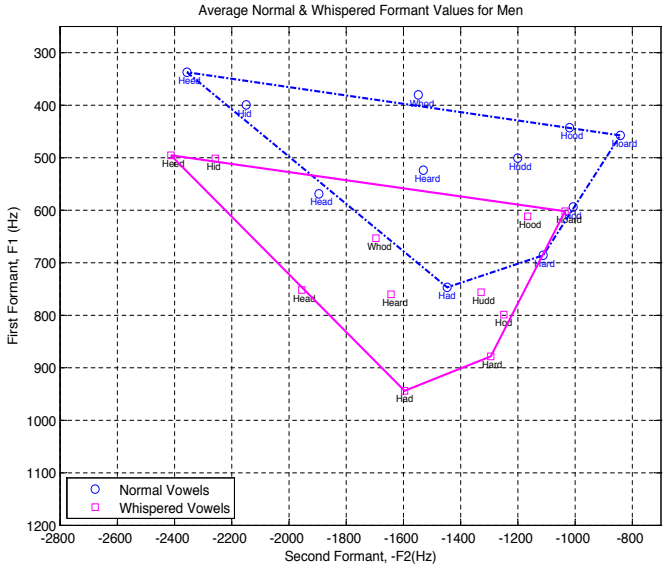
4.6.2 Discussion

4.6.2.1 RP and Birmingham Accents

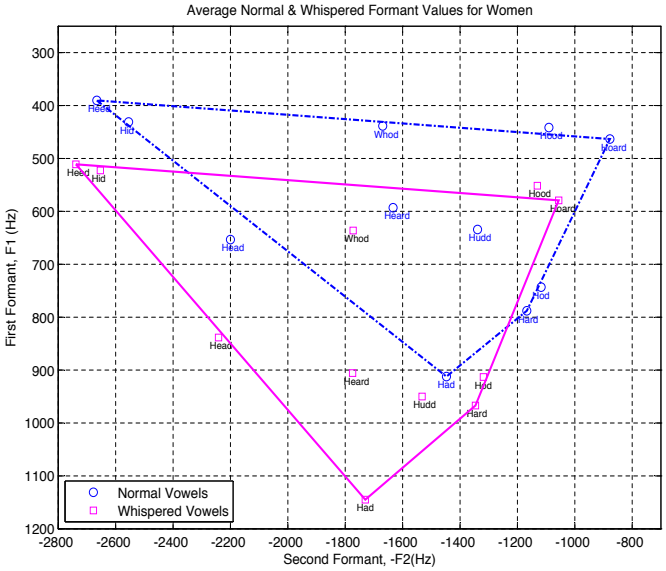
As mentioned in Section 4.2, Hughes and Trudghill [98] lists the lack of distinction made between /ʌ/ and /ʊ/, and /i/ becoming very close (shifted up) toward the vowel space occupied by the /ɪ/ as the main vowel characteristics of the Birmingham accent. Both these characteristics can be observed from figure 4.4(a) and figure 4.5. The $F1$ values of /ʌ/ decrease from 722 Hz (RP) to 567 Hz while the $F2-F1$ formant increase to 702 Hz from 514 Hz (RP). These values are now more similar to the $F1$ and $F2-F1$ values for /ʊ/ in Birmingham normal samples, 442 Hz and 612 Hz respectively compared to the RP $F1$ and $F2-F1$ values of 376 Hz and 574 Hz respectively.

The distribution for /ʊ/ appears to be a subset of the distribution for /ʌ/. This might be because although the pronunciation of /ʊ/ is relatively unambiguous for a Birmingham speaker, the pronunciation of /ʌ/ is perhaps more ambiguous - the Birmingham speakers may be aware that there is a 'proper' pronunciation of /ʌ/ and they may be trying to approximate it.

The West Midlands pronunciation of the vowels in 'bath' and 'trap', as mentioned, are identical (following the short vowel, /æ/ from 'trap'). The closest (not the same) we have to this is 'hard'(/ɑ/) versus 'had'(/æ/). The overlap between these two distributions increases in normal speech, as in figure 4.4(a) and figure 4.5. The average $F1$ value of /æ/ increases from 748 Hz (RP) to 829 Hz while the average $F2-F1$ falls significantly down from 998 Hz (RP) to 618 Hz which are now closer to the corresponding amounts of $F1$ and $F2-F1$ in /ɑ/ with 736 Hz and 403 Hz in Birmingham normal samples.



(a) Vowel diagrams from normal and whispered vowels for men



(b) Vowel diagrams from normal and whispered vowels for women

Figure 4.7: Acoustic vowel diagrams showing average formant frequencies from normal and whispered vowels for a)men, b)women. The words heed, hid, head, had, hard, hudd, hod, heard, hoard, hood, and who'd contain vowels /ɪ,i,ɛ,æ,ɑ,ʌ,ɒ,ə,ɔ,ʊ,u/ respectively.

The overlap between the $F1$ and $F2-F1$ values for the vowels in 'hod' (/ɒ/) and 'hard' (/ɑ/) is very striking in figure 4.5. The average value for /ɒ/ moves to the RP value for /ɑ/, while the value for Birmingham shows a slightly higher F1 than the RP version.

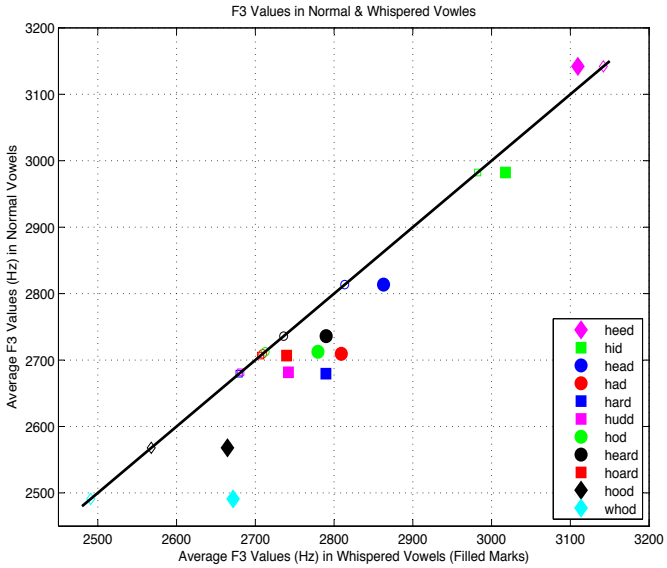


Figure 4.8: Average values of $F3$ for normal and whispered vowels. The solid line indicates the $F3$ values for normal vowels while filled marks shows the corresponding average $F3$ values from whispered samples across the horizontal axis. The words heed, hid, head, had, hard, hudd, hod, heard, hoard, hood, and who'd contain vowels /ɪ,i,ɛ,æ,ɑ,ʌ,ɒ,ə,ɔ,ʊ,u/ respectively.

A closer version of 'heard' is mentioned in Wells' discussion [97] for certain urban accents in the West Midlands. The space occupied by 'heard' (/ə/) in figure 4.5 is indeed more closed than that expected for RP, as the $F1$ values slightly decreased from 581 Hz (RP) to 558 Hz and $F2-F1$ values increase from 800 Hz (RP) to 1023 Hz. Listed by Clark [100] as another characteristic in WM dialect, 'lot' is distinguished from 'thought'; the closest samples we have are 'hod' (/ɒ/) and 'hoard' (/ɔ/) which can also be identified in figure 4.5.

By and large, the main characteristics of WM accent noted by the literature can be seen in the figures 4.4(a) and 4.5 as compared to RP accentual effects; the consistence of this study with previous ones also improves confidence in the reliability of the study.

4.6.2.2 Normal and Whisper Vowels

More convergence of adjacent vowels is evident in the whispered samples both for men and women. As shown in figure 4.6, vowel groups such as {/ɔ/ and /ʊ/} or {/ʌ/ and

/ɑ/ and /ɒ/} or {/ɪ/ and /i/} become similar in terms of formant characteristics while the back vowels show greater amounts of shift compared to front-closed vowels. This means that vowel durations as well as small changes in the shape of the vocal tract for differentiating these vowels in normal speech, are less significant in their whispered counterparts.

Another main difference for whispered speech appears to be that /ʌ/ has moved away from /ʊ/ and now overlaps instead with /ɑ/ and /ɒ/ for both men and women. This is interesting, since one of the Birmingham accent clues (as mentioned in Section 4.2) seems to have disappeared, while the clue for /ɑ/ and /æ/ maintains approximately the same relationship as seen in normal speech.

The acoustic vowel diagrams presented in figure 4.7 show that for both men and women, the effect of whispering is greater for the first formants than the second formants, however $F2$ values have also been shifted proportionally. Interestingly, the extreme front-back and open-close vowels show almost consistent shift for both men and women but more significant shifts appear in central open-mid and close-mid vowels in which women show more tendency toward lower tongue positions in whisper mode. For example, first formants of mid-central vowels such as /ə/ and /ʌ/ in women shift up from 593 Hz and 634 Hz to 905 Hz and 949 Hz, respectively while the amount of shifts for the vowels at the extremes are significantly less; for example, from 390 Hz and 463 Hz to 511 Hz and 579 Hz in /ɪ/ and /ɔ/, respectively for women.

Apart from the shifting, the size of quadrilaterals in figure 4.7 show different changes for men and women within moving from spoken vowels to whispers. While the area remains almost the same for men, the significant change appears in women's diagrams particularly on the height of the quadrilateral corresponding to whispers.

The average change in size of quadrilateral for men is 9% in both height and width at the extremes while the height shows increase but width decreases by this amount

Table 4.1: *Average formant values in normal and whispered vowels for men*[a]

		/ɪ/	/i/	/ɛ/	/æ/	/ɑ/	/ʌ/	/ɒ/	/ə/	/ɔ/	/ʊ/	/u/
	N	337	399	568	746	685	500	593	523	457	442	380
F1	W	495	501	751	943	878	756	798	760	601	611	653
	S.A	0.468	0.255	0.321	0.263	0.289	0.510	0.345	0.451	0.315	0.381	0.717
	N	2356	2149	1895	1447	1111	1200	1006	1530	841	1019	1548
F2	W	2412	2257	1955	1595	1294	1328	1249	1643	1035	1165	1696
	S.A	0.023	0.050	0.031	0.102	0.165	0.106	0.240	0.073	0.229	0.143	0.095
	N	3035	2876	2643	2586	2505	2537	2536	2606	2605	2434	2348
F3	W	2979	2894	2717	2688	2628	2639	2587	2682	2550	2502	2583
	S.A	-0.018	0.006	0.028	0.039	0.049	0.040	0.020	0.029	-0.021	0.028	0.100

[a]N: Normal, W: Whisper, S.A: Shift amount in %

when moving from voiced to whispered mode. These amounts are 6% decrease in width and 21% increase in height for quadrilaterals of women's vowels. In fact, the significant change occures in height of the female diagram by increasing 21% in whispered speech.

Figure 4.7 also demonstrates that the vowels for men and women both, by and large, occupy similar relative positions in whispered and normal modes, however, it is

Table 4.2: *Average formant values in normal and whispered vowels for women*[a]

		/ɪ/	/i/	/ɛ/	/æ/	/ɑ/	/ʌ/	/ɒ/	/ə/	/ɔ/	/ʊ/	/u/
	N	390	431	653	912	787	634	743	593	463	441	438
F1	W	511	522	838	1144	966	949	913	905	579	551	636
	S.A	0.309	0.211	0.283	0.255	0.227	0.497	0.228	0.526	0.250	0.248	0.450
	N	2665	2554	2200	1447	1167	1338	1117	1633	877	1090	1669
F2	W	2738	2653	2241	1730	1346	1532	1318	1774	1056	1130	1772
	S.A	0.027	0.038	0.018	0.195	0.153	0.144	0.180	0.086	0.203	0.036	0.061
	N	3248	3087	2984	2832	2853	2825	2888	2865	2807	2701	2633
F3	W	3239	3140	3007	2929	2950	2844	2971	2897	2928	2826	2759
	S.A	-0.002	0.017	0.007	0.034	0.034	0.006	0.028	0.011	0.043	0.046	0.048

[a]N: Normal, W: Whisper, S.A: Shift amount in %

difficult to arrive at the simple summary of the differences that are seen in these figures while a few general observations have been mentioned to the extent that conventional articulatory interpretations of formant data are valid.

The diverse amount of shifts in figure 4.6 shows each vowel has its own variation when converting to whispered speech and this amount also varies in terms of formant number. Tables 4.1 and 4.2 summarise these variations for the first three formants in whisper and normal speech for men and women, respectively.

From Tables 4.1 and 4.2, it can be seen that all first and second formants are shifted up, by amounts ranging from 25% in /i/ to 71% in /u/ for men and from 21% in /i/ to 52% in /ə/ within the first formants and from 2.3% in /ɪ/ to 24% in /ʌ/ for men and from 2.7% in /ɪ/ to 20% in /ɔ/ for women within the second formants. Furthermore, significant shifts occur in the first formants with averages of 39% (σ : 13%) and 31% (σ : 11%) while these numbers are 11% (σ : 7%) and 10% (σ : 7%) for the second formants for men and women, respectively. As mentioned before, third formants are almost consistent between normal and whiper modes.

4.7 Summary

A study to establish a vowel formant space for whispered speech has been carried out. By comparing whispered vowels with the corresponding phonated samples separately for men and women, a table outlining the amount of shift for each vowel and formant was presented, while distribution of formant values for normal and whispered samples was illustrated.

Acoustic vowel diagrams were presented showing that more shift occurs within central open-mid and close-mid vowels rather than the extreme front-back and open-close vowels both in men and women in whisper mode. In fact, shift amounts in whispered vowels depend on the method of articulation in normal mode.

Since the study was conducted on speakers from Birmingham, the analysis also briefly considered the effect of British West Midlands (WM) accent in comparison with Standard English (RP). Some of these accentual effects in the whispered speech, as observed in the data, were also discussed.

An automatic method of segmentation/extraction based upon single emitting state HMMs was also discussed while HTK and ESPS toolkits were considered for this purpose.

Chapter 5

Speech Regeneration from Whispers Through a Modified CELP Codec

Reconstruction of natural sounding speech from whispers is useful in several applications in different scientific fields ranging from communications to biomedical engineering. Many aspects of such reconstruction, in spite of their potential, have not yet been resolved by researchers, and this type of speech regeneration has received relatively little research effort to date.

Patients who have undergone partial or full laryngectomies are typically unable to speak anything more than hoarse whispers without the aid of prostheses or by learning specialised speaking techniques. As described in Chapter 2, each of the current prostheses and rehabilitative methods for post-laryngectomised patients, (primarily oesophageal speech, tracheo-oesophageal puncture (TEP) and electrolarynx) have particular disadvantages, prompting new work on non-surgical, non-invasive alternative solutions. One such solution, described in this chapter, combines whisper signal analysis with direct formant insertion and speech modification located outside the vocal tract. This approach aims to allow laryngectomy patients to regain their ability to speak with a more natural voice than alternative methods, by whispering into an external prosthesis which then recreates and outputs natural sounding speech. It relies

on the observation that whilst the pitch generation mechanism of laryngectomy patients is damaged or unusable, the remaining components of the speech production apparatus may be largely unaffected.

This chapter aims to discuss an innovative approach to the near real-time conversion of whispers to normal sounding phonated speech using a modified CELP codec (see 3.1.2), presenting a novel method for spectral enhancement and formant smoothing during the reconstruction process. The proposed approach uses a probability mass-density function to identify reliable formant trajectories in whispers, and apply spectral modifications accordingly.

In addition to a brief review on the analysis-by-synthesis approach utilised in many codecs such as CELP, the modifications required for a typical CELP framework for whisper-speech reconstruction are described in the following sections, with the three most important structural additions being the whisper activity detector (WAD), the whisper phoneme classifier (WPC), and spectral enhancement. A brief description on modifications based on line spectral pairs (LSPs, see 3.1.2) are also provided as a preliminary method for spectral enhancement of whispered speech while the method, later, is replaced with a more efficient technique based on the probability function. Furthermore, methods of pitch generation and variation for the voice regeneration are pointed out, one based on the basic LTP filter, which was previously described in Chapter 3, and the other for pitch variation based upon formant locations and amplitudes.

Finally, the evaluation of the system using both subjective and objective tests are presented in Chapter 6 where it compares the resulting speech quality against electrolarynx speech as well as describing the current limitations of the system. Before going through these sections in detail, first an overview of the system implemented outlining the entire framework as well as a brief review on the corresponding methods in terms of signal processing techniques are presented in the following section.

5.1 System: Big Picture

During normal phonation, modulated lung exhalation passes a taut glottis to create a varying pitch signal that resonates through the vocal tract, nasal cavity and out through the mouth. Within these cavities, vocal tract modulators such as the velum, tongue, and lips each play a part in shaping speech sounds. Unphonated phonemes, by contrast, dispense with a glottal pitch source, instead relying on the broadband excitation due to exhaled turbulent airflow.

When whispering, no phonation takes place, even for phonemes that are normally strongly phonated: whispers involve almost no vocal cord vibration as described in detail in Chapter 3. Non-phonation can occur with voluntary physiological blocking of vocal cord vibration during whispers, or in pathological cases, when vocal cords are damaged by disease or even removed due to surgical treatment of disease. Whispering leads to a situation of reduced speech perceptibility - something which normal speakers may be aiming for when they whisper, but which is an unwanted side effect for those with pathological vocal cord damage.

As described in Chapter 2, total laryngectomy patients, lacking a functioning glottis, may also lack the ability to pass lung exhalation through the vocal tract. Partial laryngectomy patients, by contrast, can retain the power of controlled lung exhalation through the vocal tract (VT). Despite the effective removal of their glottis, both classes of patient retain most of the remaining vocal tract modulators themselves. These, plus controlled lung exhalation (or by similar means such as oesophageal/stomach contraction), provide the ability to produce unphonated speech, effectively the same as normal whispers [5].

In other words, post-laryngectomised patients retain control of most parts of their vocal production apparatus, but have lost one vital element which forces them into whisper-mode speech. An approach of artificially reconstructing the role of this missing

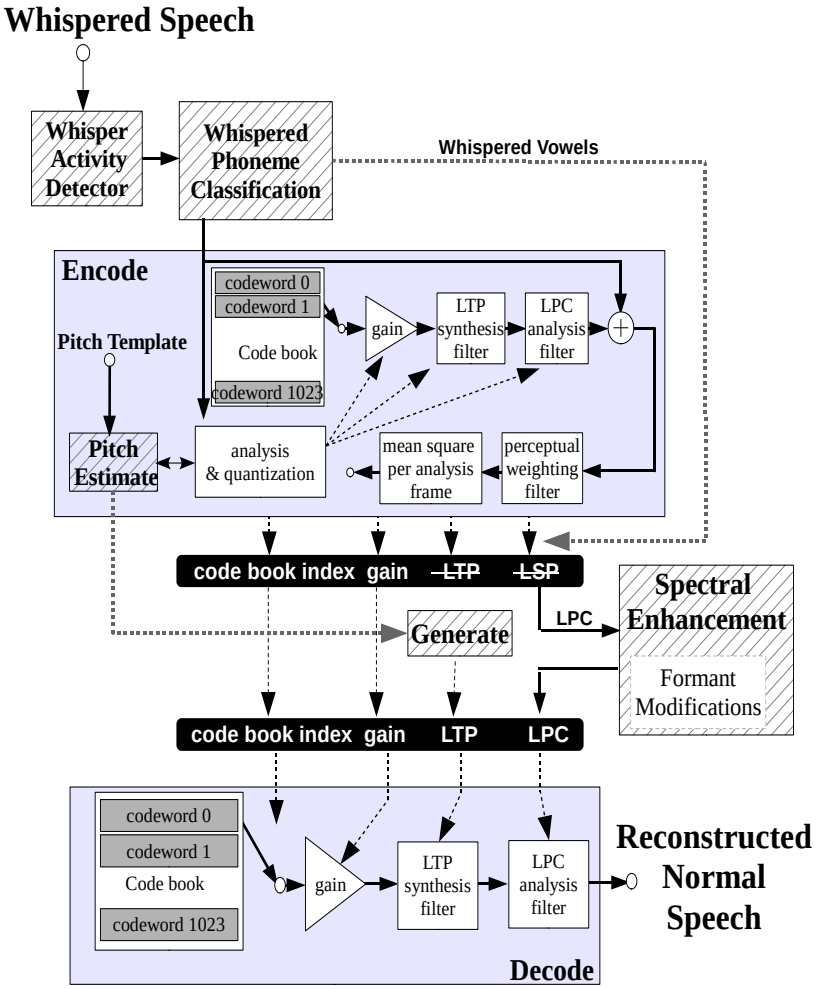


Figure 5.1: Block diagram of the proposed vocal reconstruction process, showing a modified CELP codec augmented with additional processing units (the units drawn with hatched backgrounds).

element from the analysis of the sounds created by the remaining speech articulators, plus associated information, is the main aim of the proposed system. This quest, thus, is that of regenerating speech from whispers: the analysis of Kazi et al. [112] on post-laryngectomised patients' voiced vowels (/i/), shows that the predominant spectral features have similar characteristics to those of whispered speech, whereas non-voiced phonemes are naturally similar.

Currently, various techniques such as oesophageal speech [12], tracheo-oesophageal puncture (TEP)[13], and the electrolarynx [14] are employed as speech aids by post-

laryngectomised patients; but each, as described in Chapter 2, suffers from weaknesses that range from learning difficulties to clumsy usage and heightened risk of infection. Furthermore, none of those techniques sounds particularly natural and all tend toward monotonous speech. The electrolarynx, as the most common voice rehabilitation aid [15], has recently attracted research attention to improve its quality - by decreasing background and radiated noise as well as to simplify its usage [16, 17]. Despite these efforts, there has not been any effective method reported to resolve the issue of mechanical sounding (robotised) output speech.

By comparing with these methods, the speech processing approach discussed in this research is rather different: it aims to produce more natural characteristics, by an analysis-by-synthesis framework based on the CELP codec to analyse whispers, and reconstruct the missing pitch elements from the whispered speech (as well as perform some of the other adjustments necessary to account for the subglottal coupling impedance, and changed glottal shape). A notable example synthesising normal speech from whispers within a MELP[1] codec has been suggested by Morris [70, 91]. Although his proposed approach performs a fine spectral enhancement, the mechanisms of reconstruction and pitch insertion are not well suited for real time applications; since for pitch prediction and spectral enhancement, the method relies upon comparison of normal speech samples with whispered samples to train a jump Markov linear system (JMLS) for estimating pitch and voicing parameters accordingly. In the applications which we target, i.e. laryngectomy patients and possibly in private mobile phone communications, the corresponding normal speech samples would not be available for comparison and regeneration purposes. Thus, our approach works without deploying normal speech samples, utilising a CELP codec to adjust whispered speech to sound

[1]Mixed Excitation Linear Prediction (MELP) is used for low bit-rate coding based upon a source-filter decomposition of the speech signal with a parametric excitation model. In this codec, the excitation is created by mixing periodic and random signals in different frequency bands [113].

more like fully phonated speech. Some other authors have proposed signal processing based speech prosthesis, such as Sugie and Tsunoda [114] who attempted to synthesise Japanese vowels by detecting mouth movements from electromyograms (EMG). Whilst this is welcome, it does not address the issue of continuous whispered speech, and suffers from the need for handling very sensitive EMG input signals.

Our proposed system operates as a non-invasive hands-free wearable device able to reconstruct normal speech from whispers in near real time. In particular, we now present the analytical structure of the system, whisper activity detection, whisper phoneme classification and a novel method for the spectral enhancement of reconstructed speech.

The basic analysis-by-synthesis structure employed in this research is presented in figure 5.1, where a CELP-based system is shown along with the primary modifications required for whisper-speech reconstruction. Basically, the standard CELP codec utilises a source-filter model [26] to parameterise speech with gain, vocal tract, pitch and lung excitation information based on a linear predictive model of speech production. The original speech is first segmented into analysis frames and then LPC analysis is performed on each frame to give a set of coefficients which are used in a short term predictor to model the spectral envelope of the speech. Long term prediction (LTP) then yields pitch prediction filter parameters. Finally, the excitation is determined from a codebook of random white Gaussian sequences using an analysis-by-synthesis approach to minimise the squared objective error. Depending on the inclusion of LTP filter within the analysis-by-synthesis process which leads to the addition of an adaptive codebook, two structures called open-loop and closed-loop are identified. In the decoder, synthetic speech is generated by filtering the scaled optimum codebook sequence through the LTP and LPC filters without any perceptual weighting. Each of these modules is described in detail in the following sections while an open-loop structure (only one codebook) for the CELP codec is used.

5.2 Pre-Processing Modules

Prior to entering the modified CELP codec which provides the analysis/synthesis framework, various pre-processing modules are required for the enhancement of whispers. These cater to the special characteristics and spectral features of whispers that were summarised in Section 3.2, primarily related to overall gain and spectral tilt.

By a simple classification of sound frame average energy, if plosives or unvoiced fricatives are detected, the CELP algorithm can operate largely unmodified since whispered plosives and unvoiced fricatives are largely identical to those in normal speech. Only gain adjustments and direct segment re-synthesis are necessary; otherwise, the segment of speech is considered to be potentially voiced but missing pitch, hence the algorithm performs gain modification and applies spectral enhancement. The main pre-processing modules are as follows:

5.2.1 Whisper Activity Detector (WAD)

The standard G.729 voice activity detector [115] has here been employed for the detection of whispered speech: two detection mechanisms are applied in parallel to yield a joint decision of activity. This is also used to modify the statistics of the noise thresholds, in the absence of speech.

The two classifiers are based on signal power, E, and zero crossing rate, zcr, functions applied on each speech segment, S_n, as follows:

$$WAD(S_n) = \begin{cases} +1 & \text{if } E(S_n) \geq \xi_2 \\ -1 & \text{if } E(S_n) < \xi_1 \\ \text{sign}\{\text{zcr}(\text{S}_\text{n}) - \xi_{zcr}\} & \text{if } \xi_1 \leq E(S_n) < \xi_2 \end{cases} \tag{5.1}$$

- The power classifier considers the smoothed differential power [116] of the whispered signal. It compares time domain energy with two adjustable thresholds (ξ_1, ξ_2) to identify whispered signal, noise and silence regions.

- The zero crossing detector [115] uses an adjustable threshold crossing rate (ξ_{zcr}) to improve upon the decision of the power classifier for whispered signal detection.

While numerous studies during recent years have developed different approaches for detecting speech on a noisy signal [117, 118], we merely employ the common standard accepted by ITU-T for this purpose which has already been tested/shown to be quite effective in the environments without background noise [119].

5.2.2 Whispered Phoneme Classification (WPC)

Whispered phoneme classification replaces the standard voiced/unvoiced detection unit found in typical codecs. Since there is actually no voiced segment in the whispered speech, the module is better described as a phoneme classifier which applies a voicing target class weighting accordingly. In other words, the weight of unvoicing is high when the algorithm detects a plosive or an unvoiced fricative and becomes low as the algorithm detects vowels. Candidate pitch insertion, performed at a later stage, is then sensitive to this weighting.

The WPC unit first tries to detect fricatives by comparing the power of whispered frames in bandwidths above and below 3 kHz. Then a set of bandpass filters compares signal energy ratios in small bands of high and low frequency to identify plosives and vowels, i.e. energy concentration in the 1-3 kHz range in comparison with 6-7.5 kHz is a possible indicator of a vowel sound. Furthermore, other information such as detecting the burst of energy after a small silence, provides evidence of a plosive. This, and several similar heuristics are used to yield more accurate results. In fact, plosives are detected/confirmed by comparing signal energy ratios in small bands of low and high frequency as well as considering the small silence (low energy) in the previous segment to confirm the decision. This is a mixed time/frequency domain approach. Pseudo

Procedure 5.1 WPC Outline

Initialise filters {highpass, lowpass, and bandpass filters}
Filter signal
$X_{high} \Leftarrow E(X_{high})$ {frame power in bandwidth above 3 kHz}
$X_{low} \Leftarrow E(X_{low})$ {frame power in bandwidth below 3 kHz}
$X_{bandL} \Leftarrow E(X_{bandL})$ {frame power in bandwidth between 1-3 kHz}
$X_{bandH} \Leftarrow E(X_{bandH})$ {frame power in bandwidth between 6-7.5 kHz}
$Threshold1 \Leftarrow X_{high}/X_{low}$
$Threshold2 \Leftarrow X_{bandL}/X_{bandH}$
if $Threshold1 > \Upsilon$ **then**
  $Flag = 0.5$ {fricative identified}
else if $Threshold1 \leq \Upsilon$ and $\Gamma_1 \leq Threshold2 \leq \Gamma_2$ **then**
  $X_{previous} \Leftarrow$ average energy of few previous frames
  if Energy of current frame $> \Phi \times X_{previous}$ **then**
    $Flag = 1$ {plosive identified}
  end if
else
  $Flag = 0$ {vowel identified}
end if
return $Flag$

code presented in Procedure 5.1 briefly outlines the routine (merely the logic with symbolic thresholds) for each speech frame.

Figure 5.2 shows the output of the algorithm for a sentence from the TIMIT database ("she had your dark suit in greasy wash water all year") whispered word by word in an anechoic chamber. The top plot overlays the WAD output while the bottom one demonstrates the result of the phonetic classification (1 for plosives, 0.5 for fricatives and 0 for vowels).

Classification thresholds are presently speaker dependent, determined manually (which would translate to the prosthesis requiring a set-up phase conducted by a speech therapist). This could be improved upon for a real medical product, however this research focuses primarily upon speech reconstruction rather than addressing existing limitations with the phonetic classification. The effect of WPC unit accuracy will be discussed in Chapter 6.

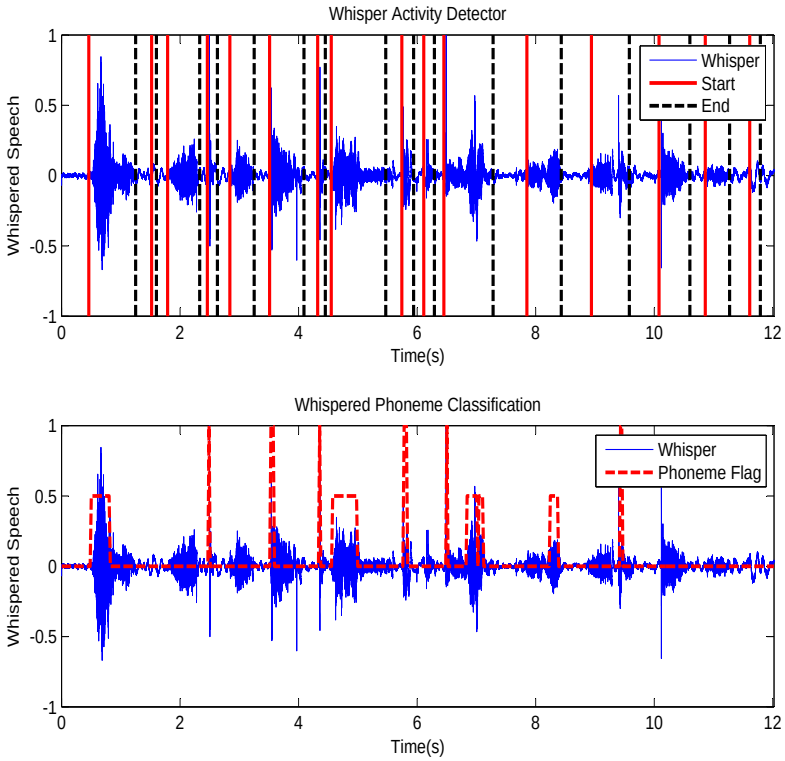


Figure 5.2: Outcome of whisper activity detector (top) and whispered phoneme classification (bottom) modules for a sentence from TIMIT database ("she had your dark suit in greasy wash water all year") whispered slowly in an anechoic chamber. The lower plot dotted line overlay indicates detected plosives specified by value 1, fricatives by 0.5 and vowels by zero.

5.3 Spectral Enhancement of Whispers

Reconstruction of phonated speech from whispers evidently involves spectral modification. In part due to the significantly lower SNR of recorded whispers compared with normally phonated speech, estimates of vocal tract parameters for such speech have a much higher variance than those of normal speech. As mentioned in 3.2, VT response for whispered speech is noise excited and this differs from the normal response when excited with periodic pitch pulses. In addition to the reported difficulties for formant estimation in low SNR and noisy environments [120, 121], the essence of whispered speech, as described, also causes inaccurate formant calculation due to tracheal coupling [88, 122]. Increased coupling between the trachea and the VT created by the

open glottis (similar to the aspiration process) leads to the formation of additional poles and zeros [71] in the VT transfer function.

Such issues become more significant in vowel reconstruction, where any instability of the resonances in the VT (i.e. formants) tends to be more obvious to a listener. To prepare whispered speech for pitch insertion, consideration is therefore required for the enhancement of the spectral characteristics regarding disordered and unclear formants caused from the noisy substance, background and excitation evident in whispers. Two approaches for this kind of enhancement are described in this section; one is based on LSP narrowing-shifting and the other is based on probability mass function (PMF). LSP modification has been introduced/employed for various purposes in speech processing, but its application for spectral enhancement of whispered speech is new. Although this method produces sensible results, the output is not reliable enough to work as an independent module in the system; thus, it was replaced by an entirely novel and efficient method based on PMF. Both of these methods are described in the following subsections.

5.3.1 LSP-Based Enhancement

In the CELP codec, as well as many other low bit-rate coders, linear prediction coefficients are often transformed into LSPs [123]. LSPs collectively describe the two resonance conditions of an interconnected tube model of the human vocal tract. The two conditions are those that describe the modelled vocal tract being either fully open or fully closed at the glottis respectively. In reality, the human glottis is opened and closed rapidly during speech and thus the actual resonances occur somewhere between the two extreme conditions. The LSP representation, thus, has a significant physical basis [124]. However, this rationale is not necessarily true for whispered speech (since the vocal folds do not vibrate), and thus adjustments to the model are required. In

fact, LSPs have been in use for a long time but applying LSPs to whispered speech for the purpose of the spectral enhancement has not been found in literature.

We define LPC polynomial, $A_p(z)$, with analysis order p and LPC coefficients α (see 3.1.2) as follows:

$$A_p(z) = 1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \ldots + \alpha_p z^{-p} \tag{5.2}$$

The two palindromic and anti palindromic polynomials, P and Q respectively, represent LSPs as follows:

$$\begin{cases} P(Z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \\ Q(Z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \end{cases} \tag{5.3}$$

Within a typical CELP speech codec, LSP data is derived from an LPC analysis of speech. The LSPs are quantised and transmitted from encoder to decoder where they are reconstructed and used within a synthesis filter.

There are technically two methods commonly used for computing LSP parameters: one is based upon Chebyshev polynomials proposed by Kabal [125] and the other is based upon a modified Levinson Durbin algorithm proposed by Saoudi [126]. Kabal method is known to be the most widely used method of computing LSP parameters [123].

LSPs are used in speech coding due to their excellent quantisation characteristics, their residence in the frequency domain and the ready interpretation of LSP positioning. A third ability, that of being able to safely adjust their locations without forming unintentional resonances (including shifting, clustering and narrowing) is the main reason for their use in this application for recognising formant locations and performing spectral enhancement on whispers.

Spectral peaks are generally bracketed by LSP line pairs, with degree of closeness being dependent upon the sharpness of the spectral peak between them, and its amplitude [127]. However, as previously discussed in section 3.2, whispered speech has few

significant peaks in the spectrum which implies wider distances between LSP lines. Hence to emphasise formants, it is necessary to narrow the LSP lines corresponding to likely formants, i.e. operate on the 2 or 3 narrowest pairs.

As illustrated in figure 5.3 (top), where the linear prediction spectrum obtained from analysing a typical segment of whispered speech is plotted, and dashed lines drawn at the LSP frequencies derived from the linear prediction parameters are overlaid on it, there are very small formants (very mild peaks) between the 3 narrowest pairs of LSPs (the pairs {1:2}, {7:8} and {11:12} respectively).

Now if ω_i represent the p LSP frequencies (where order p is 12 in this experiment) and ω'_i the altered frequencies, then narrowing line pair, e.g. {1:2}, by degree α would be achieved by:

$$\begin{aligned} \omega'_1 &= \omega_1 + \kappa(\omega_2 - \omega_1) \\ \omega'_2 &= \omega_2 - \kappa(\omega_2 - \omega_1) \end{aligned} \tag{5.4}$$

Since altering the frequency of lines may lead to formation of unintentional peaks by narrowing the gap between two irrelevant pairs, it is important to choose the paired lines corresponding to likely formants. As figure 5.3 (bottom) shows, the result for κ=0.3 have peaks provided between desired line pairs which rather satisfy the required formant emphasis for the proposed method.

Since adjusting the placement of individual LSPs may lead to the formation of spurious unintentional peaks by narrowing the gap between two irrelevant pairs, it is important to carefully choose the pair of lines corresponding to likely formants. As mentioned, this might be done by choosing the three narrowest LSP pairs which works well when the signal has fine peaks, but in case of the expansion of formant bandwidths (common in whispered speech), leading to an increased distance between the corresponding LSPs, the choice of the 3 narrowest LSPs will often not identify the three formant locations correctly, particularly for vowels (for example, {5,6} instead

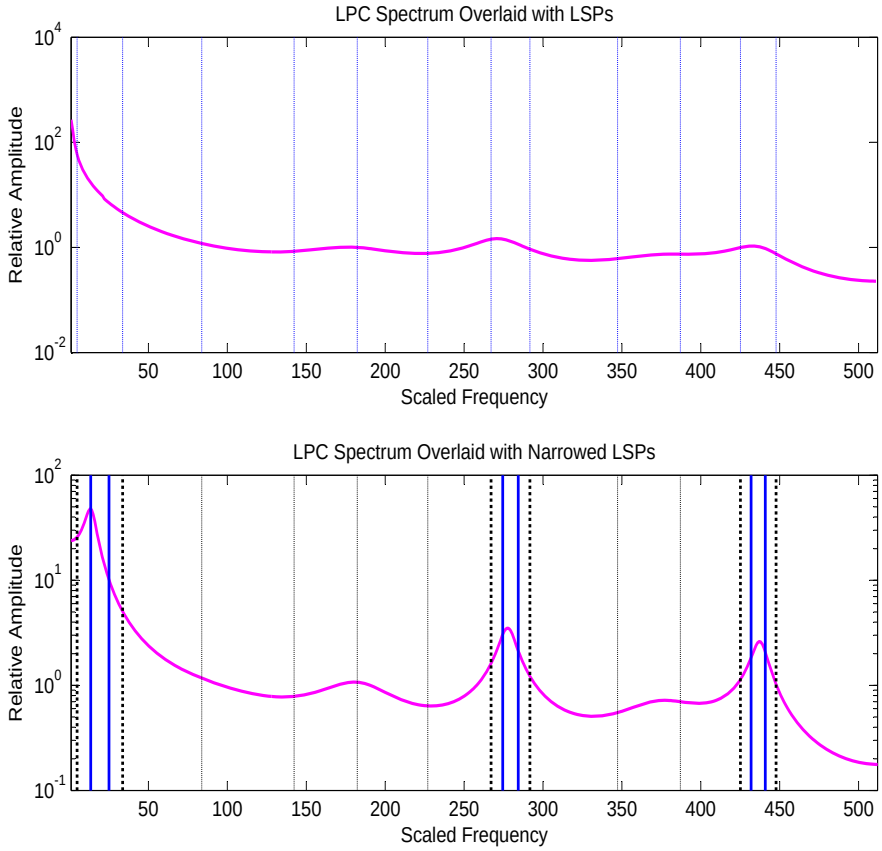


Figure 5.3: An LPC spectrum plot showing LSP positions overlaid for the whispered vowel /a/ (top) and reconstructed LPC spectrum after applying LSP narrowing (bottom), showing that the relatively flat spectrum has been enhanced into three formant peaks. The dotted lines in the bottom figure correspond to the locations of the original LSPs whereas bold lines are modified LSPs.

of {1,2} in figure 5.3 can be identified as the likely formants). Therefore, the enhancement procedure is designed carefully to perform effectively on all whispered vowels and diphthongs. This new method, which is more complex and reliable according to subjective/objective tests (Chapter 6), is described in the following subsection. (The author's paper [128] discusses this LSP-based method more in details while McLoughlin [124] reviews LSPs in the big picture, so the thesis continues with focus on the more efficient PMF-based method.)

5.3.2 PMF-Based Enhancement

Since it has been established that formant spectral locus is of greater perceptual importance than formant bandwidth [129], our improved computational strategy in this method applies a formant track smoother to ensure smooth formant trajectories without extreme frame-to-frame stepwise variations. The module tracks the formants of a whispered voiced segment and low-pass filters the trajectory of formants in subsequent blocks of speech, using oversampled and overlapped formant detection.

The formant tracking is based on the common method of linear prediction (LP, see 3.1.2) coefficient root solving. The auto-regressive LP system assumes that spectral poles correspond to formants in the speech spectrum. When the LP coefficients, α_i, are derived by analysis, the roots of the following complex equation determine these poles:

$$1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots + \alpha_p z^{-p} \Longrightarrow$$
$$z^p + \alpha_1 z^{p-1} + \alpha_2 z^{p-2} + \cdots + \alpha_{p-1} z + \alpha_p = 0 \tag{5.5}$$

Equation (5.5) is the p^{th}-order polynomial with real coefficients and generally has $p/2$ complex conjugate roots. If a pole is written as $z_i = r_i e^{j\theta_i}$, then the formant frequency, F and bandwidth B corresponding to the i^{th} root is obtained from (5.6) and (5.7) as follows:

$$F_i = \frac{\theta_i}{2\pi} f_s \tag{5.6}$$

$$B_i = \arccos\left(\frac{4r_i - 1 - r_i^2}{2r_i}\right)\frac{f_s}{\pi} \tag{5.7}$$

where θ and r denote respectively the angle and radius of a root in the z-domain and f_s is the sampling frequency. By substituting $\cos^{-1}(z) = -jln(z + \sqrt{z^2 - 1})$, the expression of (5.7) is simplified to:

$$B_i = -ln(r_i)\frac{f_s}{\pi} \tag{5.8}$$

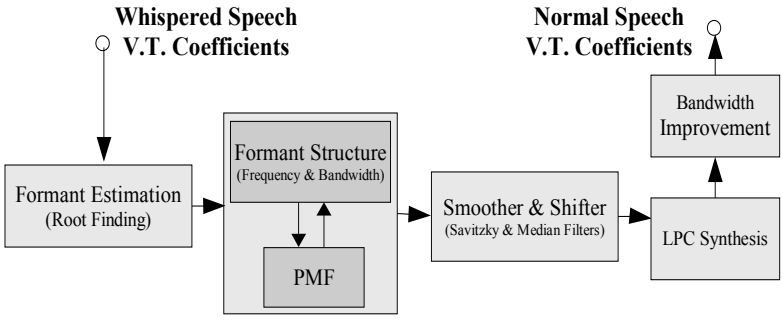


Figure 5.4: Block diagram of the whisper formant modification method (PMF is an abbreviation for probability mass function).

Evidently, when p is larger than the number of formants, roots include not only formants but also some spurious poles; a formant, hence, is approximated by the phase of the complex pole that has the smallest bandwidth (calculated at the - 3 dB point) in a cluster of poles. The bandwidth to peak ratio is then determined, and the roots with a large ratio (common in whisper and post-laryngectomised speech) or those located on the real axis, are classified as being spurious. The remaining roots may represent formant locations with high variance, as a result of noisy excitation in whispers. It is therefore necessary to eliminate the effects of this variance by applying modifications such that the de-noised formant track is more accurate concerning the formant frequency rather than the corresponding bandwidth.

To summarise the formant modification process in this method, first the formant frequencies are adjusted based on pole densities, then corresponding bandwidths are altered based on *a priori* power spectral difference analysis between whispered and phonated speech. The block diagram of the process steps are outlined in figure 5.4, and are described further as follows.

5.3.2.1 Formant Frequency Modification Based On Probability Mass-Density Function

A novel approach has been found useful in fulfilling the goal of whisper formant smoothing. This is to generate a formant trajectory based upon the probability func-

tion computed from an array of noisy formant candidate locations. Before describing the proposed method, since the words 'frame' and 'segment' are used frequently for explaining the algorithm, these are first defined as follows:

- A frame is a block of s ms Hamming windowed speech.
- A segment is a sequence of M overlapping frames (up to 95% overlap).

Performing the standard method of root finding on each frame by using equations (5.6) and (5.8), results in N formant frequencies and their corresponding bandwidths:

$$[F_1, \ldots, F_N] \quad , \quad [B_1, \ldots, B_N] \tag{5.9}$$

the formant structure is defined as $S = [F, B]^T$ for each frame, and for each segment, the resulting formant structure is denoted by F and B matrices:

$$F = [F_{n,m}]_{N \times M} \quad , \quad B = [B_{n,m}]_{N \times M} \tag{5.10}$$

The rows of formant track matrix F in (5.10) are taken as tracks of N formants across a segment of phonated speech corrupted by noise.

Matrix F is subsequently processed by the smoother which evaluates the probability mass function (PMF) of formant occurrences, $p(f)$, in frequency ranges below 4 kHz, then applies a heuristic which extracts the highest constraints of location for the first three formants, and marks extra margins as inappropriate regions. The resulting PMF is formulated as:

$$p(f) = \frac{1}{MN} \sum_i \sum_j Pr(F_{(i,j)} = f) \tag{5.11}$$

and then a density function, $D(f)$, in a specific narrow frequency band is defined as:

$$\sum_x^y p(f) = D(f), \quad y - x = \delta \tag{5.12}$$

where δ is a small frequency distance. Now the first three formants can be described as the highest densities in each band of the standard formant frequencies while less dense margins arising from whispery noise are considered inappropriate as formants and are thus ignored:

$$\begin{cases} F1 = \arg\max\{D(f)\} & f \in [\lambda_1, \lambda_2] \\ F2 = \arg\max\{D(f)\} & f \in (\lambda_2, \lambda_3] \\ F3 = \arg\max\{D(f)\} & f \in (\lambda_3, \lambda_4] \end{cases} \tag{5.13}$$

where λ_1,λ_2,λ_3,λ_4 specify the standard ranges for the first three formants (for example, 200 and 1500 for λ_1 and λ_2, respectively), and are specific to the vocal characteristics of each speaker, determined a prior or through training.

To avoid hard thresholding limitations, some points should be noted: multiple assignments, merging and splitting of $D(f)$ peaks are considered options for producing the three most significant formant regions; for example, multiple assignments to a region are allowed if the adjacent margin has no significant peak. In case of close formant adjacency, the margins would overlap and are separated through proper decisions on the boundary of overlapping margins. Another issue is the over-edge formant densities which are resolved by conditional merging and splitting of formant groups.

In fact, λ values in (5.13), per se, are not important but can initialise the commencing points of the heuristic process of finding the dense formant regions. In other words, these values can be any numbers as long as they refer to the three sequential ranges spanning 100 Hz to 4 kHz. Practically speaking, by applying bundling/unbundling procedures in the sequential regions, the most three denser regions will be selected regardless of these specified borders (for example, see Figure 5.7 or Figure 6.1 clearly showing the transition between initialised λ values). Pseudo code presented in Procedure 5.2 briefly outlines the process of PMF approach.

Procedure 5.2 PMF Outline

Find formants through LPC-root finding method
Create formant structure in highly overlapped frames
Finding the dense regions of $D(f)$ {peaks of PMF in narrow defined bands δ}
Require: $f \in [100 \quad 4000]$
Filter peaks through Savitzky-Golay
Ensure: $max\{D(f)\}$ in sequential regions
bundle/debundle to find the best peaks in ascending ranges
$F1, F2, F3 \Leftarrow$ the three sequential peaks
Median filtering the resulted formants

Once the margins of the first three formants are determined by (5.13), the columns of F are rearranged to be placed in the correct regions, and modified formants become:

$$F_i^M = \frac{\theta_i^M}{2\pi} f_s \quad i = 1, 2, 3 \tag{5.14}$$

Figure 5.5 shows the relation of $D(f)$ with formant location pattern for a whispered speech frame.

To put it simply, the implemented idea tries the elimination of the frequencies arising from spurious formants marked and observed with high variance within each highly overlapped heuristic while potentially proper formants creating denser frequency ranges point the non-spurious formants.

Finally, a smoothing algorithm encompassing Savitzky-Golay filtering [130] is applied to each margin to reduce the effect of noise in peak finding process, then a median filtering stage is used to derive the final formants. These enhanced formant frequencies can also be shifted down based upon the study presented in Chapter 4 showing the differences between long-term average phonated and whisper formant locations, i.e. first and second formants of different vowels can be shifted down directly by 100 to 300 Hz accordingly. Obtaining all the modified formants through equation (5.14), the LPC coefficients of the transfer function of the vocal tract are synthesised from six complex conjugate poles representing the first three smoothed formants, plus six other non-formant poles spaced apart across the frequency band.

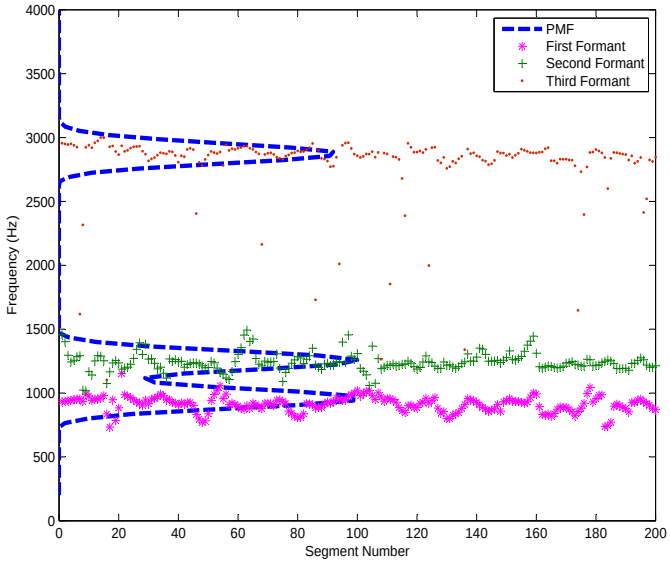


Figure 5.5: The relation of formant trajectory with probability mass function (PMF) for whispered vowel (/a/). Note the relationship between the peaks of the PMF and the corresponding formant trajectory variance.

Although this kind of formant modification is a direct modification approach, bundling the formant frequencies and weighting them based on their probabilities allows the pole interaction problem [131, 132] to be avoided. The next subsection describes bandwidth improvement while preserving the new de-noised phases (and not ignoring potential pole interaction problems).

5.3.2.2 Bandwidth Improvement Based On Spectral Energy

By obtaining new modified formant frequencies (i.e. the modified phases of the poles in the root-finding method, θ_i^M) as described in the previous section, it is necessary to apply a proportionate improvement regarding the corresponding bandwidths (i.e. the radii of the poles, r_i). This improvement should be done in such a way that not only are formant frequencies retained, but also their energy should be improved to prevail over attenuated whispers. For this purpose, a suggestion made by Hsiao and Childers [131] was considered, and then led to the use of different spectral energy criteria for whispered and normal speech. A pole specified with characteristics in

equations (5.6-5.8), has a transfer function and power spectrum of:

$$H(z) = \frac{1}{1 - re^{j\theta}z^{-1}} \tag{5.15}$$

$$|H(e^{j\phi})|^2 = \frac{1}{1 - 2r\cos(\phi - \theta) + r^2} \tag{5.16}$$

and when there are N poles:

$$|H(e^{j\phi})|^2 = \prod_{i=1}^{N} \frac{1}{1 - 2r_i\cos(\phi - \theta_i) + r_i^2} \tag{5.17}$$

The radii of the poles are then modified such that the spectral energy of the resulting formant polynomial is equal to a specified spectral target value based upon the spectral energy difference between normal and whispered speech (according to [88], whispered speech has 20 dB lower power than its equivalent phonated speech).

Suppose there is a formant pole with given radius r_i and angle θ_i, using equation (5.17) the spectral energy of the formant polynomial, $H(z)$, at the modified angle θ_i^M is given by:

$$|H(e^{j\theta_i^M})|^2 = \frac{1}{1 - r_i^2} \prod_{j \neq i}^{N} \frac{1}{1 - 2r_j\cos(\theta_i^M - \theta_j^M) + r_j^2} \tag{5.18}$$

where $|H(e^{j\theta_i^M})|^2$ is the specific power spectrum value at angle θ_i^M and N is the total number of modified formant poles. There are two spectral components on the right side of the equation, one produced by the pole itself, and the other being the effect of the remaining poles with modified angles. By solving equation (5.18), we can find a new radius for the i^{th} pole while retaining its modified corresponding angle, θ_i^M, obtained from the previous section. For maintaining stability, if r_i exceeds unity, we use its reciprocal. The modified radius, r_i^M, for each pole is then obtained:

$$r_i^M = 1 - \left\{ \frac{1}{H_i^M} \prod_{j \neq i}^{N} \frac{1}{1 - 2r_j\cos(\theta_i^M - \theta_j^M) + r_j^2} \right\}^{1/2} \tag{5.19}$$

where H_i^M represents $|H(e^{j\theta_i^M})|^2$, the target spectral energy for each pole.

Since the formant roots are complex-conjugate pairs, only those that have positive angles are applied in the algorithm, and their conjugate parts are obtained readily at the final stage. The radii modification process using equation (5.19) starts with the pole whose angle is the smallest and continues until all radii are modified.

By inserting the modified values of angles and radii in (5.9), the improved and smoothed formant structure, S^M, for whispered speech is obtained; S^M is similar to formant structures of normally phonated speech utterances as used in several speech codecs, speech recognition engines and other applications which have been developed for normal speech.

Figures 5.6 and 5.7 demonstrate the formant trajectory for a sustained whispered vowel (/i/) and a sustained whispered diphthong (/ie/) before applying the spectral enhancement and the resulting smoothed formant trajectory after the implementation of the technique (without shifting the formants). These show the effectiveness of the method even for the transition modes of formants spoken across diphthongs.

5.4 Modified CELP Codec

Code (Codebook) Excited Linear Predictive (CELP) coding is based on the linear predictive model of speech production, in which speech is generated by filtering an excitation signal as described in Section 5.1. Amongst the variants of LPC schemes, the most widely reported standard for operation at low bit rates is the CELP codec [133]. The CELP scheme was cited by Atal as far back as 1982 [134], but it took several years for realistic working implementations of the basic CELP algorithm to be reported, the most prominent being the U.S. Department of Defense proposed Federal standard 4.8 kb/s CELP coder [135]. In figure 5.8, a block diagram of a basic CELP codec is shown.

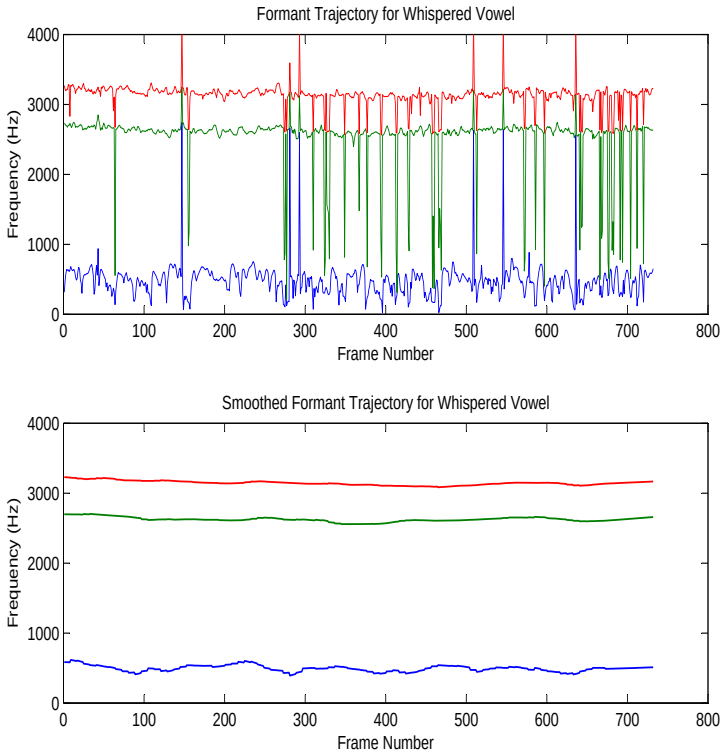


Figure 5.6: The derived F_i frequency tracks over time for sustained whispered vowel /i/, showing the frequency obtained from the initial root-finding analysis (top) compared to the smoothed vector (bottom).

LPC analysis, LTP analysis and codebook search are three main computational blocks in a basic standard CELP codec. We can briefly describe the operation of a basic CELP codec including these main blocks as follows:

- The original speech is first segmented into analysis frames of around 20-30 ms and then LPC analysis is performed on each frame to give a set of LPC coefficients which are used in short term predictor to model the spectral envelope of the speech.

- After finding the LPC coefficients, long term prediction (LTP) can proceed to find pitch prediction filter parameters (see Section 3.1.3). The LTP analysis can be performed on the residual generated by an inverse filter with the derived

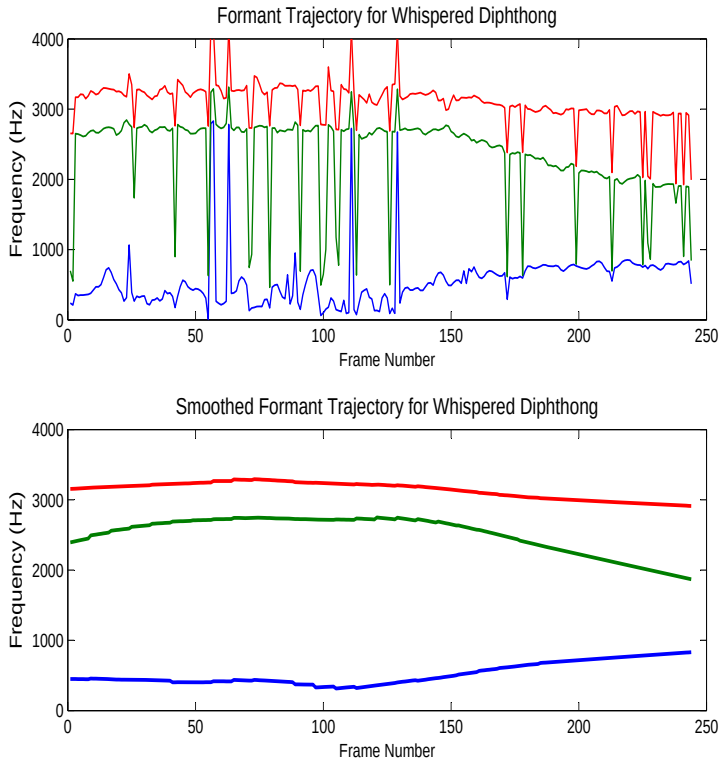


Figure 5.7: The derived formant trajectory over time for sustained whispered diphthong /ei/, showing the original root-finding frequency (top) and the smoothed vector (bottom). Note the diphthong transition beginning around frame 150.

LPC coefficients or on the original speech. Both analyses lead to a delay, D, and associated coefficient(s), β_k, k representing the number of filter taps.

- Once the parameters of the two filters are found, the excitation is determined. In the standard CELP codec, the excitation is selected from a codebook of random white Gaussian sequences. The search procedure to find the best excitation involves three basic steps: a) generation of an ensemble of filtered Gaussian sequences (synthetic speech), b) computation of the objective error for each sequence, and c) selection of the sequence with minimum error. Thus the codebook vector which produces the minimum squared objective error and the corresponding scaling factor is selected.

- In the decoder, the synthetic speech is generated by filtering the scaled optimum

codebook sequence through the filters without any perceptual weighting.

For the purpose of whisper speech reconstruction, it is necessary to modify the CELP codec as described in this section (refer to figure 5.1 for a block diagram of this modified codec). In comparison with the standard CELP codec, a Pitch Template, corresponding to the Pitch Estimate unit, exists to generate pitch factors (see Section 5.5) while Spectral Enhancement in this model is used to apply necessary spectral modifications. Also to attain more natural sounding speech, a probability mass function (PMF)-based formant trajectory modification is used, as described in Section 5.3.2.

5.5 Pitch Insertion & Variation

Two methods of pitch generation/insertion for the whispered speech conversion within the framework described in the previous sections are pointed out in this section. The simple pitch insertion technique described first was used for the listening test evaluations presented in Chapter 6. This enables the outcome of the whisper spectral enhancement to be demonstrated in like-for-like fashion against the fixed-pitch electrolarynx: i.e. in the absence of possible advanced pitch modification effects, such as natural pitch templates, in the system described. However, a more flexible technique for pitch variation based on formant locations to avoid monotonous pitch has been implemented on whispered vowels and is described in the next subsection.

5.5.1 Pitch Insertion

As described in Section 3.1.3.1 for the basic LTP filter, the formulation used for the LTP in CELP, which generates long-term correlation, whether due to a real pitch excitation or an artificial pitch, is denoted by:

$$P(z) = 1 - \sum_{i=0}^{I} \beta_i z^{(-D-i)} \tag{5.20}$$

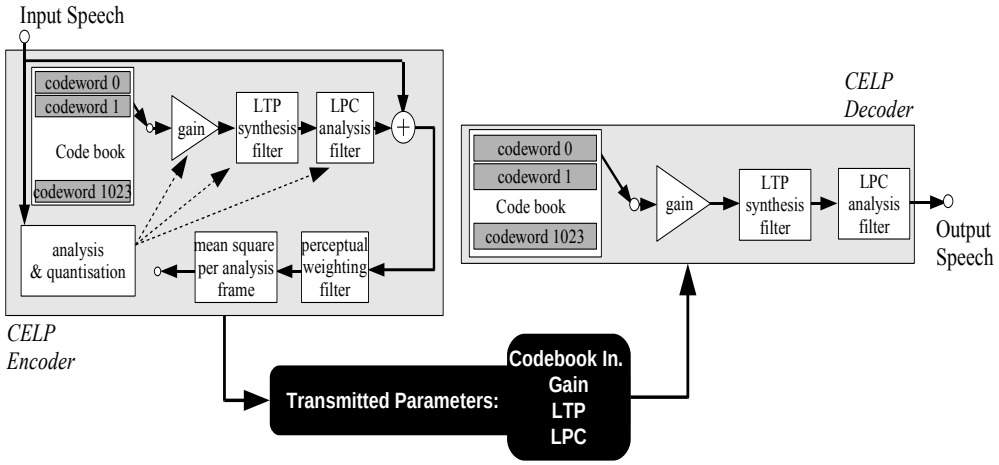


Figure 5.8: Block diagram of a basic CELP codec.

where β is the pitch scaling factor (the strength of the pitch component), D corresponds to the pitch period and I represents the number of taps (see Section 3.1.3.2 for the explanation).

The investigation of β and D in normally phonated speech shows that in an unvoiced sample of speech, D is almost random, while β remains small, but in a voiced sample, D is set by the value of the pitch delay or its harmonics while β maintains larger values (see Section 3.1 for the corresponding graphs). To derive the mechanism of pitch insertion, the results of the whispered phonemes classification (as described in Section 5.2) is used for decision making of voiced/unvoiced speech. (A formant count procedure can also be used to help on the accuracy of the decision, but having more heuristic processes leads to more computing cycles and thus, further from the intended real time system; so it was avoided). Later, a randomly biased D around the long-term average is applied to equation (5.20) for the unvoiced segments. As mentioned before, this method was used for the experiments done in Chapter 6.

5.5.2 Pitch Variation

For each individual speaker, it has been known that there is correlation between fundamental frequency (i. e. vocal pitch) and formant frequencies [26, 136, 137]. More specifically, average of fundamental frequency and formants are mutually correlated. Furthermore, conducted experiments show that naturalness of speech depends upon the proportionate amounts of pitch and formants [138]. Thus, in this section, an approach for pitch contour variation in reconstructed speech is presented accordingly.

This method extracts voice factors from the whispered signal and applies these to the reconstructed speech. The method is based upon the structure described in the preceding sections which implements an analysis-by-synthesis approach to voice reconstruction using a modified CELP codec.

By extraction of voice factors including formant location and amplitude, this method tries to regenerate natural pitch contours particularly in vowels. By consideration of formant locations as well as amplitudes, a formula for generating pitch parameters in a CELP-type codec is proposed.

As described in Section 3.1.3, long term predictor delay, D, has the main role in pitch determination in the CELP codec. Pitch contour variation, hence, can be achieved by variation of this parameter which, in our proposed method, is based on formant frequencies and amplitude according to equation 5.21 as follows:

$$D(n) = \begin{cases} D_{n-1} + \bar{D} + \alpha(\theta_n - \bar{\theta}) + \lambda h + S & h > 0 \\ D_{n-1} + \bar{D} + \alpha(\theta_n - \bar{\theta}) & h < 0 \end{cases} \tag{5.21}$$

in which n represents the index of the current speech frame, h shows the instant amount of $\Delta F1_n \Delta F2_n$ which calculates the covariance of the first and second formants, S is the average value of formants over L previous frames (calculated based on equation 5.22), α and λ are factors to be set for generating the most natural voice and θ is the

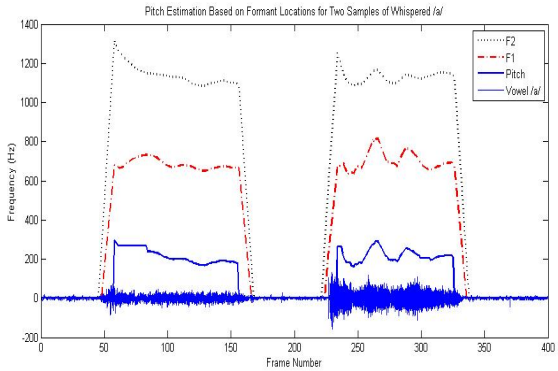


Figure 5.9: Pitch estimation based on the first 2 formants of whispered vowel /a/, showing the regenerated pitch for two different samples.

gain value in the CELP codec.

$$Q = \frac{\sum_{i=n-L}^{n-1} F1_i + F2_i}{2} \tag{5.22}$$

Figure 5.9 illustrates the regenerated pitch contours of two samples of whispered vowel /a/ based on the technique. Figure 5.10 demonstrates $F1$, $F2$, and pitch contour of a whispered and the corresponding reconstructed vowel.

5.6 Summary

This chapter has described the reconstruction of natural sounding speech from whispers using a novel speech prosthesis based upon a modified CELP codec. This uses

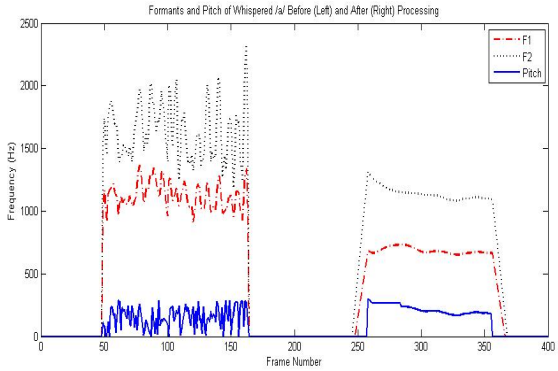


Figure 5.10: Formants and pitch values for whispered vowel /a/ before (left) and after (right) applying the pitch variation, showing regenerated pitch which varies based on the locations of the formants.

formant and pitch analysis allied with synthesis and reconstruction methods for missing pitch fundamental and vocal tract resonances (formants). Implementation of this engineering approach for whisper-to-normal speech reconstruction can achieve near real-time synthesis of normal speech from whispers and then, as a wearable, non-invasive and non-surgical aid to post-laryngectomised patients, such a prosthesis is potentially of great benefit.

As the decision points to switch to smoothed trajectories have great impact on the efficiency and accuracy of the method, two supporting modules, namely whisper activity detector and whispered phoneme classifier, were described as part of the pre-processing necessary for whispered speech analysis. Due to the approach used for spectral enhancement, the system is very sensitive to the performance of the WPC module particularly on reconstruction of the sentences. The limitations of the current WPC module will be discussed in detail in Chapter 6.

To prepare whispered speech for pitch insertion, consideration is also required for the enhancement of the spectral characteristics regarding disordered and unclear formants caused from background noise and excitation noise evident in whispers. Two approaches for this kind of enhancement were described while one was based on LSPs narrowing-shifting and the other one based on probability mass function (PMF); of which the latter produces the most reliable and accurate results. In fact, LSP based methods cannot work accurately enough, due to the wide bandwidth of formants in whispered speech. Finding the proper formants in whispers also needs more consideration; both of these problems have been addressed by the novel PMF based method presented here.

Furthermore, two methods of pitch generation/insertion for the voice regeneration were pointed out, one based on the basic LTP filter and the other for pitch variation based on formant locations and amplitudes. The next chapter demonstrates the

effectiveness of the approach by conducting experiments as well as designing objective/subjective tests on the combined system modules.

Chapter 6

Evaluations, Results, and Discussion

The evaluation of speech quality is highly important in today's telecommunication systems. To measure the perception of quality of a speech signal, traditionally, subjective tests in which humans listen to a live or recorded speech signal and assign a score to it, are known as the most reliable method of assessment; but these subjective tests are time consuming and expensive. Thus, attempts have been made to replace subjective evaluation with objective methods. Objective evaluation of speech quality is mostly based on comparison of original speech signal with the speech being evaluated (i.e. the distorted/decoded/reconstructed speech derived from the system being tested). Note that this often requires the two signals to be synchronised. Objective methods for estimating speech quality were primarily developed with the aim of predicting the result of subjective quality tests, but be more convenient, cheaper to run and more repeatable.

In this chapter, after a brief review of quality assessment techniques, the system implemented in Chapter 5 is evaluated with both subjective and objective tests and the results discussed accordingly. By considering the output, a detailed discussion is also presented to point out the deficiencies of the current system.

6.1 Speech Quality Assessment

6.1.1 Subjective Tests

The quality assessment of speech has been traditionally made through subjective tests, where a panel of several listeners, usually placed in a soundproof room, rate the quality of speech signals under evaluation. In a typical listening test, subjects hear recordings and vote on a simple opinion scale such as the common 5-point listening quality (LQ) scale. This test (including all rating types) has been standardised by the International Telecommunication Union (ITU) in recommendation P.800 [139].

Within this test, a single speech sample is played to the subjects who are asked to answer the question "what is your opinion on the quality of speech?", and they then score it using a 1 to 5 scale. In terms of rating, there are different procedures for scoring subjective tests which each have different meanings:

- Absolute Category Rating (ACR): listeners are required to make a single rating for each speech sample having been played to them. The scores are 5 for excellent, 4 for good, 3 for fair, 2 for poor, and 1 for bad. Then, an average of all votes is the overall score. This method of rating, which is called mean opinion score (MOS) is known as the most popular type of subjective test [140, 141]. For the purpose of this research as described in Section 6.2, this method of rating has been used.

- Degradation Category Rating (DCR): listeners are presented with the original signal as a reference, before they listen to the synthetic signal, and are asked to compare the two and give a score according to the amount of degradation perceived. The scores are 5 for not perceived, 4 for perceived but not annoying, 3 for slightly annoying, 2 for annoying, and 1 for very annoying. The average

of all votes which is called the degradation mean opinion score (DMOS) is then obtained.

- Comparison Category Rating (CCR): listeners are presented with a pair of speech samples on each trial. In the DCR procedure, as mentioned above, a reference (unprocessed) sample is presented first, followed by the same speech sample, which has been processed by some technique. In the DCR method, listeners always rate the amount by which the processed (second) sample is degraded relative to the unprocessed (first) sample. However, in the CCR procedure, the order of the processed and unprocessed samples is chosen at random for each trial. On half of the trials, the unprocessed sample is followed by the processed sample. On the remaining trials, the order is reversed. The scores are 3 for much better, 2 for better, 1 for slightly better, 0 for about the same, -1 for slightly worse, -2 for worse, and -3 for much worse.

Table 6.1 summarises these rating systems with their corresponding meaning.

Table 6.1: *Different rating procedures in subjective tests.*

Rating Score	ACR (MOS)	DCR (DMOS)	CCR
5	excellent	inaudible	N/A
4	good	audible, not annoying	N/A
3	fair	slightly annoying	much better
2	poor	annoying	better
1	bad	very annoying	slightly better
0	N/A	N/A	about the same
-1	N/A	N/A	slightly worse
-2	N/A	N/A	worse
-3	N/A	N/A	much worse

To obtain accurate and repeatable results, the number of test subjects needs to be large enough, and environmental conditions and control characteristics need to

be the same for all subjects. Furthermore, another important variable that is only partially controlled is the subjective scale itself: depending on the range of conditions, and the subjects' cultural interpretation of terms such as excellent or worse, there can be systematic offsets particularly in ACR as large as 1.0 MOS between tests with different listeners [141]. Therefore, subjective tests are identified not only as expensive, highly time consuming, and labour-intensive assessment techniques [142] but also the presence of some variables such as cultural interpretations which cannot be fully controlled, have led to consideration of alternative methods of assessment. So in recent years, several objective quality algorithms have been developed which try to replace/simulate the results of subjective methods. In the next subsection, these objective methods are briefly considered.

6.1.2 Objective Tests

The major aim of objective tests are to estimate subjective scores (mostly MOS) automatically based on measurements of a speech system. Objective methods are classified into two main categories of intrusive and non-intrusive methods. Intrusive models compare an original signal with its degraded version that has been processed by a system (these models have also been called comparison-based, or full-reference [143]). On the contrary, non-intrusive models try to predict speech quality through the computational model of either output speech or system properties without having the original signal and are divided into two broad classes of signal-based and parameter-based methods [141].

6.1.2.1 Intrusive Models

Intrusive test methods pass a known (reference) signal through the system under test, capture the processed (degraded) signal, and compare the two to derive a quality score that should correlate well with MOS. In fact, these models mostly work by

transforming both signals using a perceptual model to reproduce some of the key properties of hearing, then computing distance measures in the transformed space and using these to estimate MOS. So, both the original and processed signals are required to be available to the models.

Within the last two decades, many intrusive speech quality estimation models such as those mentioned in [142, 144, 145, 146, 147] have been proposed. Finally, a simple approach known as the perceptual speech quality measure (PSQM) was found to be the most accurate model with both known and unknown subjective MOS data. It was standardised by ITU-T in recommendation P.861 [148] as the first standard perceptual model. Then, in 2001, due to some limitations, ITU replaced it with the standard P.862 [149], in which a new model for speech quality evaluation called perceptual evaluation of speech quality (PESQ) was introduced. In fact, PESQ is based on integration of PSQM and a perceptual analysis measurement system (PAMS) based on Bark spectral distortion (BSD) as introduced by Rix and Holler [150]. PESQ is intended to be used for measuring one way quality on narrow band telephone signals although it is also applicable to other speech processing systems such as speech codecs.

While the above algorithms are known as the ITU-T standards, there are some other simple common algorithms which are used in speech enhancement. These algorithms try to find the distances/errors between two speech signals based on the same logic of intrusive models: comparing reference with degraded/coded signal in time or frequency domain frame-by-frame and assign an amount showing the distortion/difference between two signals; the greater the difference, the worse the quality.

These are mainly applicable to quality measures for speech enhancement algorithms such as the system we have implemented in this thesis. The types of distortion introduced by speech enhancement algorithms can be categorised into two major classes: the distortions affecting the speech signal itself (speech distortion) and the distortions

affecting the background noise (noise distortion) [151]. Among many techniques introduced in this approach, objective measures such as spectral distance measures, LPC measures (e.g., Itakura-Saito) and time-domain measures [e.g., segmental signal-to-noise ratio (SNR)] are more common. Some of these measures which are based on linear prediction, LP, and hence, helpful for our experiments, extracted/summarised from Quackenbush's book [152], are outlined as follows:

- Itakura-Saito Distortion Measure: For an original frame of speech with LP coefficient vector, $\vec{a_c}$, and coded/processed speech coefficient vector, $\vec{a_p}$, the Itakura-Saito distortion measure is defined as:

$$d_{IS}(\vec{a_p}, \vec{a_c}) = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a_p}\boldsymbol{R}_c\vec{a_p}^T}{\vec{a_c}\boldsymbol{R}_c\vec{a_c}^T}\right) + log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \tag{6.1}$$

  where $\boldsymbol{R}_c$ is the autocorrelation matrix of the original speech signal[1], and σ_c^2 and σ_p^2 represent the LPC gains of the original and processed speech frame respectively.

- Log-Likelihood Ratio Measure: The LLR measure is also referred to as the Itakura distance and is defined as follows:

$$d_{LLR}(\vec{a_p}, \vec{a_c}) = log\left(\frac{\vec{a_p}\boldsymbol{R}_c\vec{a_p}^T}{\vec{a_c}\boldsymbol{R}_c\vec{a_c}^T}\right) \tag{6.2}$$

  note that the IS measure incorporates the gain estimate using variance terms, while the LLR does not; this influences how each measure emphasises differences in general spectral shape versus an overall gain offset.

[1]Generally, the autocorrelation as a measure of periodicity for a section of signal shows how well the waveform correlates with itself at a range of different delays, the corresponding calculation for finding the autocorrelation matrix on a windowed speech frame can be found in [38].

- Log-Area-Ratio Measure: The LAR measure is also based on dissimilarity of LP coefficients between original and processed speech signals. The LAR parameters are obtained from the P^{th} order LP reflection coefficients of the original, $r_c(j)$ and processed, $r_p(j)$ signals for frame j. The objective measure is defined as follows:

$$d_{LAR} = \left\{ \frac{1}{M} \sum_{i=1}^{M} \left(log \frac{1 + r_c(j)}{1 - r_c(j)} - log \frac{1 + r_p(j)}{1 - r_p(j)} \right)^2 \right\}^{1/2} \tag{6.3}$$

  where M denotes the number of frames in the speech signal.

- Cepstrum Distance Measure: cepstrum distance provides an estimate of the log spectral distance between two spectra. The cepstrum coefficients can be obtained recursively from the LPC coefficients, a_m, using the following expression:

$$c(m) = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c(k) a_{m-k} \quad 1 \leq m \leq p \tag{6.4}$$

  where p is the order of the LPC analysis. Now, an objective measure based on cepstrum coefficients can be computed as follows:

$$d_{CEP}(\vec{c_c}, \vec{c_p}) = \frac{10}{log10} \left\{ \sum_{k=1}^{p} \left(c_c(k) - c_p(k) \right)^2 \right\}^{1/2} \tag{6.5}$$

  where $\vec{c_c}$ and $\vec{c_p}$ are the cepstrum coefficient vector of the original and enhanced signals, respectively.

For the purpose of this research, the LPC-based objective LLR was chosen for assessment purposes since the proposed enhancement approach in this system is mainly based upon LP analysis, and the LLR is widely adopted in comparable research literature [152].

6.1.2.2 Signal-Based Models

Non-intrusive estimation of speech quality from speech waveforms is a challenging problem in that the estimation of speech quality has to be performed with the speech signal under test only, without using a reference [143]. Non-intrusive signal-based models (also known as no-reference or single-ended models), which currently are not as mature as intrusive models, estimate MOS by processing the degraded output speech signal of a live network (for example, in [153] and [154]).

Several signal-based methods focus on models of speech production or speech signal likelihood, although many exploit some aspects of perception such as noise loudness. Finally, ITU-T standardised such a model in 2004 in recommendation number P.563 [155]. In P.563, a set of key parameters are extracted for the analysis of a) vocal tract and unnaturalness of speech, b) strong additive noise, and c) interruptions, mutes, and time clipping. Based on these parameters, the intermediate speech quality is estimated for each distortion class, and the overall quality is obtained by a linear combination of intermediate speech quality with 11 additional signal features.

6.1.2.3 Parameter-Based Models

Simple computational models have been proposed to estimate the quality of a network without the need to run subjective or intrusive tests, for network planning or non-intrusive measurement [141]. Non-intrusive parametric models generally have no sound signal to process (and so make limited use of perceptual techniques), but instead estimate MOS from measured properties of the underlying transport and/or terminal, such as echo, delay, speech levels and noise. The approach has also more recently been applied to real-time assessment of VoIP systems where the dominant distortions, packet loss, jitter, and the codec, can be accurately modelled by a small number of statistical measures.

This last approach requires a full characterisation of the system under test and is, therefore, sometimes referred to as a glass box approach while methods for which no knowledge of the system under test is required are referred to as black box approaches [143]. This parametric approach has been standardised by ITU-T in recommendation number P.562 [156] in the year 2000.

6.2 Experiments and Results

As mentioned in Section 6.1, the most reliable method for assessing perceptual quality is to employ subjective assessment, and in this case to directly evaluate quality. Two separate subjective testings were employed based on the ACR method described in Section 6.1.1. These tests were conducted for the purpose of assessment of (i) the efficiency of PMF-based spectral enhancement as described in Section 5.3.2 and, (ii) the performance of the WPC and WPA modules along with spectral enhancement. While subjective assessments are discussed first, objective measurements with the same assessment patterns are also provided in the following subsections to technically show the amount of improvement mostly based on spectral distances.

In the testing of our system whisper samples from normal speakers were recorded in an anechoic chamber using a sound recorder (ZOOM H4N) at 10 cm distance from the mouth of each speaker. An electrolarynx (SERVOX digital) device was then employed by the same speakers to obtain samples for comparison. Speakers were allowed several practice attempts prior to recording for both experiments to pronounce samples accurately and all subjects were trained in the use of the electrolarynx (EL) before the tests. The corresponding ethics approval for conducting the experiments which was obtained from NTU, has been attached in Appendix B. During the experiments, the main technical parameters of the system were adjusted as shown in Table 6.2.

Table 6.2: *Technical parameters of the system.*

Parameter	Value
Sample rate	$Fs = 16\,\text{kHz}$
LPC Order	$p = 12$
Frame Duration	$s = 20\,\text{ms}$
Overlap	$M = 95\%$
Density Function $D(f)$	$\delta = 40$
Pitch	β , $D = 180\,\text{Hz}$ EL = 180 Hz

The system was examined by the adjusted parameters and figures 6.1, 6.2, 6.3 demonstrate the system output for a diphthong and a sentence.

Figure 6.1 plots segments from the spectrograms of whispered and regenerated diphthong /ei/ enhanced in the previous chapter (see figure 5.7). The evolution of the formant peaks over time can be clearly seen in the regenerated plot (bottom), and also observed in the plot of original whispered speech. The strength of the $F1$ contribution is clear in the regenerated plot, as is the sharper frequency definition and stronger pitch component. This has resulted in darker, more pronounced formant lines in the regenerated system rather than the blurred frame-to-frame variations in the original plot.

Figure 6.2 demonstrates the formant trajectory for a TIMIT sentence repeated as a whisper in an anechoic chamber under the experiment conditions. In this figure, the top graph plots the original whisper formant trajectory before spectral enhancement while the bottom plot demonstrates the smoothed trajectory after applying spectral enhancement and formant modification. Figure 6.3 shows spectrograms of the same whispered sentence before and after the reconstruction process. As shown in figure 6.3 (bottom), the vowels and diphthongs are effectively reconstructed with formant modifications and necessary shifting considerations within whisper-voice conversion.

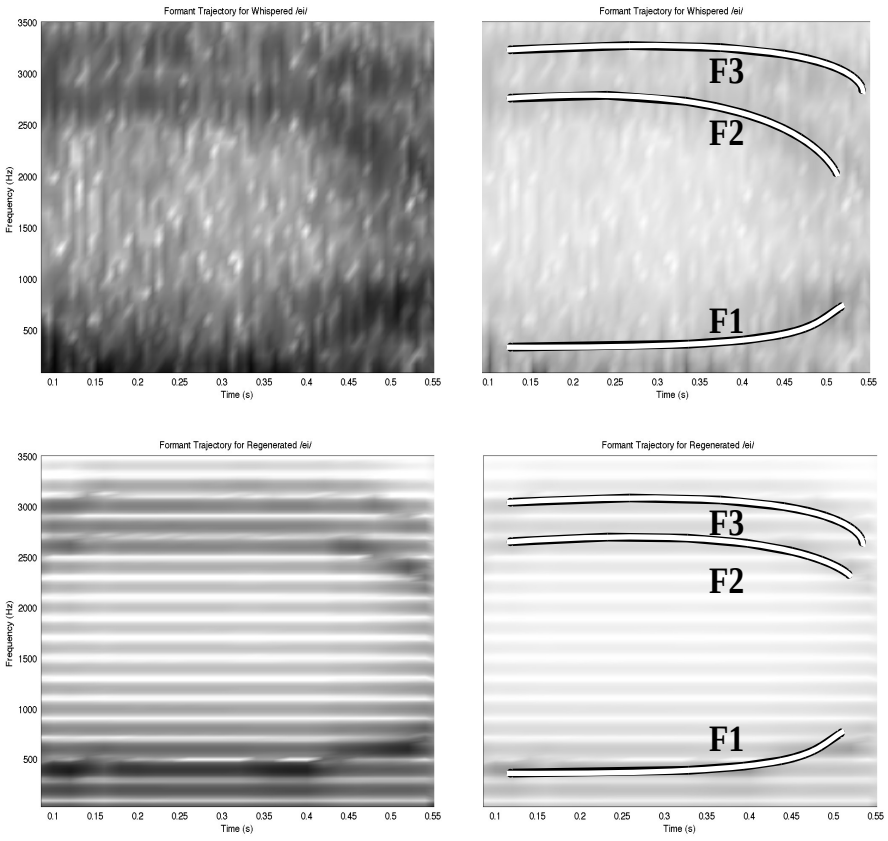


Figure 6.1: Spectrogram plots of the whispered (top) and reconstructed (bottom) /ei/ diphthong, clearly showing the position of the formant tracks (right), and incidentally also the presence of the strong artificial pitch excitation harmonics, as horizontal stripes.

6.2.1 Evaluation of PMF-Based Spectral Enhancement

For evaluation of the reconstructed vowels and diphthongs, similar to the study of Hillenbrand et al. [83], recordings were made of 7 normal speakers (3 men and 4 women, with ages between 25 and 30) "whispering" the 12 English vowels and diphthongs twice in a /hVd/ carrier structure in a sustained manner (unlike the experiments carried out in Chapter 4). The corresponding steady state whispered vowels were then manually extracted and processed. The same speakers generated the same words using the electrolarynx, and this audio was recorded and steady state vowels extracted in

the same way.

To evaluate the performance of the PMF-based reconstruction algorithm, a subjective listening test was conducted for 7 normal-hearing listeners who were proficient in English with ages ranging from 28 to 36. None of the listeners had reported a history of hearing problems and were unfamiliar with EL speech so that they could be safely regarded as naive listeners. The subjects were randomly presented with the electrolarynx generated samples and PMF-based reconstructed vowels and diphthongs, both normalised with monotonous 180 Hz pitch frequency (as mentioned in Section 5.5 in Chapter 5, the more natural variable pitch templates available to our system were overridden with fixed pitch similar to that of the EL). Each subject scored the corresponding regenerated and EL speech samples for quality over a five point scale described previously in Section 6.1.1 (5: excellent, 4: good, 3: fair, 2: poor, and 1: bad). The results summarised for vowels and diphthongs are shown in figure 6.4 (right).

It can be seen that, with a mean rating somewhere between 'fair' and 'good', the results show that the quality of speech obtained from the PMF method is significantly better than the EL samples (averaging 3.6 and 2.5 respectively). The PMF-based reconstruction system improves upon the electrolarynx for all tested vowels and diphthongs with a mean improvement across all tested material of 1.1 points. Interestingly, the results show that diphthongs in general score better than single vowels (3.7 versus 3.4 respectively), and that both averages are improved by the PMF-based method.

6.2.2 Evaluation of the WPC Module

To evaluate the performance of the pre-processing modules, full sentence regeneration was tested. For this purpose, another subjective test consisting of four speakers and four listeners (two male and two female, with the same particulars as in the previous subsection) was conducted with one ITU-proposed sentence [139] selected for the

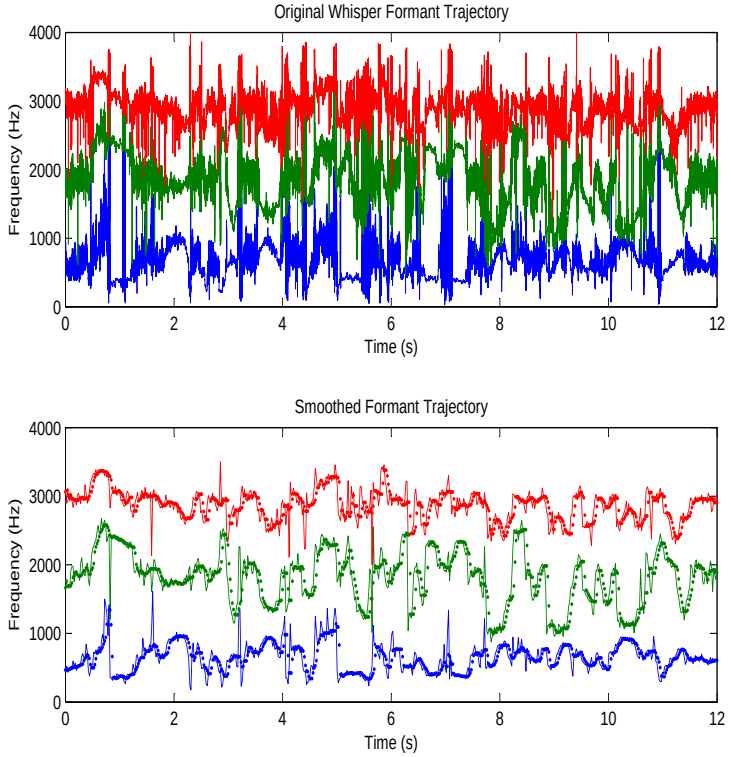


Figure 6.2: Formant trajectory for the whispered sentence, showing the original frequency against time (top) and the smoothed vector (bottom). The sentence is "she had your dark suit in greasy wash water all year" from the TIMIT database, uttered slowly and clearly in an anechoic chamber.

recordings. The speakers first slowly whispered the following sentence: "You will have to be very quiet" and then generated the same sentence using the EL.

To obtain the mean opinion score (MOS), the subjects were randomly presented with the EL and PMF reconstructed sentences, both using monotonous 180 Hz pitch excitation. Each subject scored the regenerated and corresponding EL speech samples based upon quality over a five point scale. The average MOS of the regenerated speech was found to be 2.1 using PMF and 1.6 using the EL. This is a meaningful but proportionally smaller improvement than in the vowel test, as discussed in the following section (6.3).

Furthermore, to check the performance of the system, avoiding the errors arisen

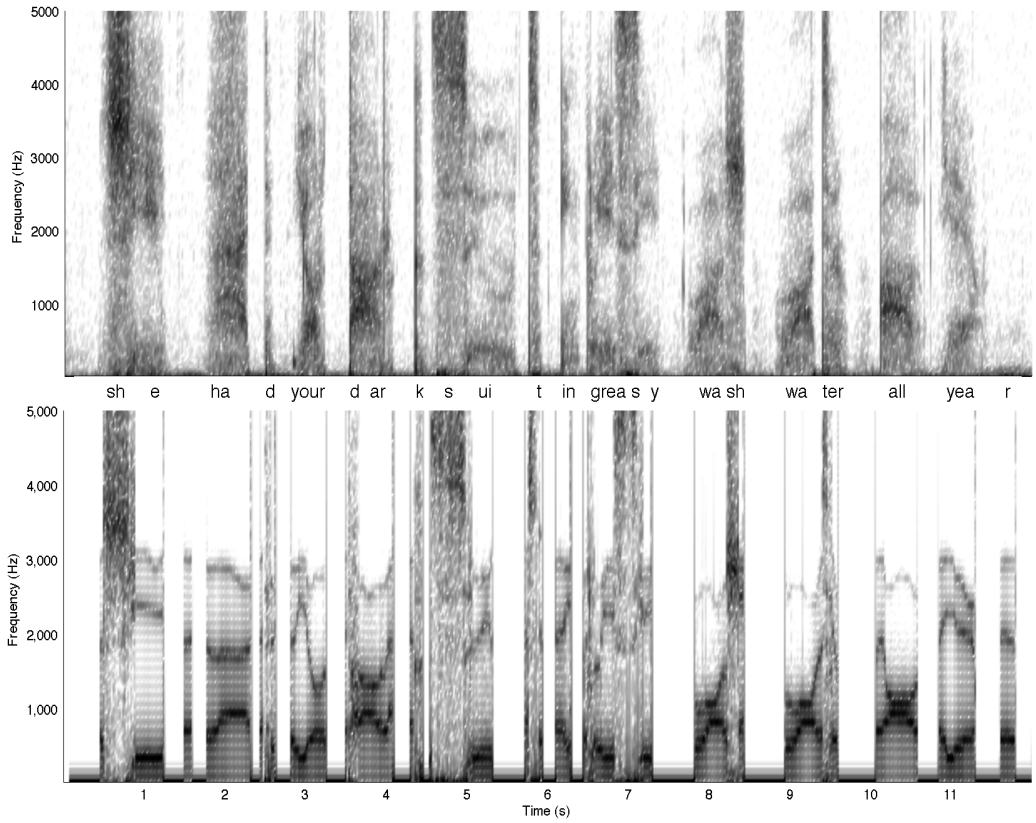


Figure 6.3: Spectrograms for whispered (top) and reconstructed (bottom) sentence from TIMIT database. The sentence is 'she had your dark suit in greasy wash water all year' uttered word by word.

from WPC and whispered speech, two other subjective tests along with the corresponding objective tests were also conducted with the same sentence recorded this time in the normal phonation. The two assessments are as follows: one included the WPC module operating automatically while the second was carried out with manual mode of the WPC (i.e. it was adjusted manually allowing the vowels to go through the smoother without any shifting/improvement while the rest are just encoded/decoded). In fact, the latter shows the maximum achievable performance of the system in absence of the errors from classification and whispered speech.

Each subject again scored the output of the system in these modes based upon quality. The average MOS of the normal speech was found to be 4.4 using manually

adjusted WPC and 2.3 using automatic WPC. On the other hand, the average LLR results were found to be 0.14 ($\sigma : 0.05$) and 1.45 ($\sigma : 0.05$) for manual WPC and automatic WPC, respectively; which also aligns with subjective scores: automatic WPC has higher LLR values reflecting more differences between input and the processed sentences (see the following section for a detailed explanation of the LLR method applied; contrary to the evaluations in Section 6.2.3, there is no timing misalignment, pronunciation mismatches, etc. for the intrusive tests reported above due to the identical sentences in terms of timing and speaker).

In other words, results of the both subjective and objective tests show that accuracy of WPC has significant effects even on normal speech in which it allows almost all the phonemes to go through unnecessary smoothing procedure. As mentioned, the discussion presented in Section 6.3 tries to cover the sensitive role of WPC.

6.2.3 Objective Measurements

As described in Section 6.1.2.1, objective assessments are widely used measures of speech distortion/enhancement in codecs [151, 157]. While our system is not typical of codecs which can directly compare input and output speech to determine distortion performance, it is nevertheless useful to apply similar objective measurements to assess the system. In fact, both known (reference) and processed (degraded) signals in typical intrusive tests are identical in terms of time-alignment, pronunciation, speaker, etc. while the latter has just been processed by a system/codec in which the objective measurement tries to evaluate the effect of that process on the same signal. In our experiments, on the other hand, the reference signal (input of the system) is whispered speech which is itself noisy and the output is a reconstructed voiced signal that for evaluation needs to be compared with normal/phonated speech and not the same input of the system.

In this case, we attempt to determine the similarity between vowels/diphthongs sets across four recordings covering original speech, whispered speech, electrolarynx speech and the reconstructed speech from our system. Before doing so, a similarity measure needs to be selected. We cannot use mean-square or similar measures because of lack of time-domain alignment, so we turn to frequency domain approaches presented in Section 6.1.2.1. Since the proposed enhancement approach in this system is mainly based upon linear prediction analysis, the LPC-based objective Log-Likelihood Ratio (LLR) was chosen for assessment purposes.

Since it is necessary to compare recordings done by the same speaker at different times (normal speech, electrolarynx speech, and whispered speech which is reconstructed through the system), it is important to determine a baseline for the repeatability of the test. Thus, first the score from several similarity measures on same-speaker in different-occasion recordings of normal speech (5 times) was computed to find an estimate of repeatability by comparing these samples with each other (i.e. 10 states per person from comparison of 5 normal samples with each other). This was repeated for the same 7 participants in the subjective tests, from which we selected the four most consistent speakers from the group for the evaluation. The mean and standard deviation of the LLR scores on normal phonation of vowels/diphthongs for these persons are 0.49 and 0.21, respectively.

We then computed the score differences between the normal samples and the corresponding a)whispered, b)electrolarynx, and c)reconstructed speech. Results are shown in figure 6.4 (left), indicating that the reconstructed samples are more similar to their corresponding normal samples in comparison with those whispered and electrolarynx samples. The average LLR of normal-reconstructed speech within all vowels and diphthongs is 2.57 ($\sigma : 0.42$) while these numbers are 3.33 ($\sigma : 0.48$) and 3.74 ($\sigma : 1.12$) for normal-whisper and normal-electrolarynx respectively.

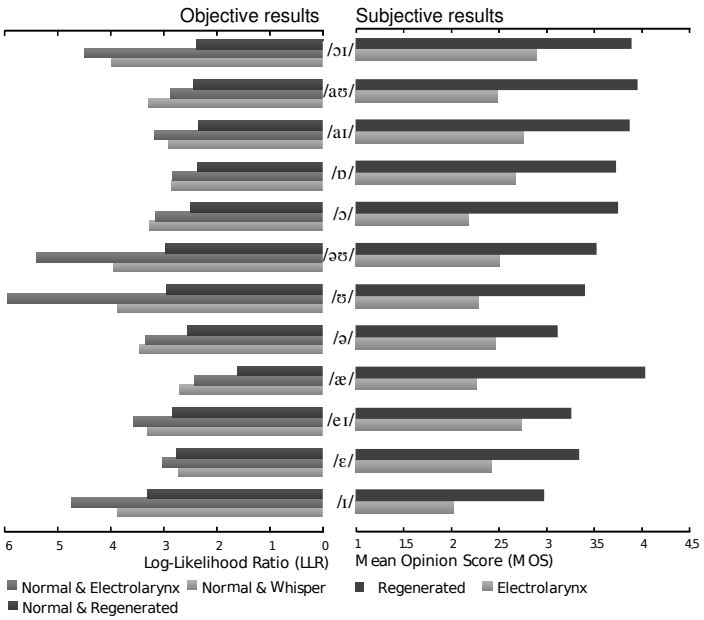


Figure 6.4: Results of the subjective (right) and objective (left) evaluations comparing samples generated by electrolarynx and the PMF-based method.

6.3 Discussion

In the PMF method, natural sounding speech is ideally obtained from spoken whisper-like speech within an analysis-by-synthesis framework. Apart from effective spectral enhancement, this class of CELP-based reconstructive method relies upon effective phoneme identification, and therefore an accurate whispered phoneme classification unit.

Since the WPC module plays a sensitive role in the proposed system, providing a robust method of phoneme classification is necessary to gain results at least as good as those results we had for vowels and diphthongs. In fact, the technique shows significant improvement over the EL in regenerating vowels and diphthongs, but less so in regenerating complete sentences; due to the importance of the classification module. The author's conjecture is that by developing a WPC unit to cater for whispered glide and nasal identification (see Subsection 6.4.1), and furthermore, for classification

thresholds to be more accurately determined, the quality of the regenerated speech could be further improved. On the other hand, EL scores may also be improved if the tests involve more experienced total laryngectomy patients due to reduced radiated noise from their more supple neck tissue (caused by cartilage and muscle removal [16]), and possibly greater EL operation skills.

Whispered speech *per. se.* is often of substantial intelligibility [87, 158], therefore the overall intelligibility of reconstructed output largely depends upon the processing methodology, rather than limitations with the input source. The observation of the author is that the largest system factor to degrade intelligibility is when the phoneme classifier mis-identifies a plosive or fricative as a vowel/diphthong, causing incorrect enhancement; however, this issue was not observed during the subjective testing reported here.

To summarise, this chapter along with the previous chapter have described the *basis* of a novel speech prosthesis that recreates a voiced version of whispered speech input. It is judged by subjective testing to be more natural sounding than speech from an EL. The formant and spectrogram plots indicate relatively clear speech, however the regeneration method could be improved further by moving towards more naturalness in pitch variation, and better support of fast continuous speaker-independent speech by the phoneme classifier. Furthermore, smoother transitions between voiced and unvoiced phonemes would be preferable for greater quality improvement.

6.4 Further Considerations

The last two chapters have described the basic framework and techniques able to perform whisper-speech conversion with a certain degree of quality. However, there are several barriers to further naturalness quality improvement. In this section, the two main system limitations corresponding to nasal detection/enhancement and speaker

individuality are discussed: although they are outside the main scope of research, they would be important for developing a commercial prosthesis based upon the techniques discovered during the course of this research.

6.4.1 Problem Phonemes

Not all speech phonemes have an exact counterpart in whispers, and some whispered phonemes may 'map' to two different spoken phonemes. One particular issue is with nasals. Voiced nasals add zeros to the frequency spectrum, whereas the proposed method is mainly based on peak finding and peak smoothing: we currently handle nasals in the same way as almost all LPC-based systems – approximate them using additional poles. However, our conjecture is that a similar smoothing method which is now applied to the poles of the system could also be applied directly to the valleys of the spectrum in the case of nasals. The main issue would be: how to detect whispered nasals and effectively identify /m/ from /n/.

In testing the current system with many words including nasals (such as hand, hind, hend, etc.), two problems were detected: a) the very low energy of whispered nasals being interpreted as silence by the WAD unit, and b) high energy nasals (emphasised/sustained /m/ and /n/) are typically quite poorly enhanced by the spectral peak picker and smoother. Therefore, the addition of a dedicated nasal detector, and enhancer to the system may be advantageous.

6.4.2 Speaker Individuality and Other Differences

Voice individuality factors are the other main issue to be considered in any voice replacement system. Being categorised into two main classes of physiological versus psychological (sociological) factors [159], speech individuality drivers can be summarised as follows: social status, dialect, age, sex, person community, and acoustic features

such as average pitch frequency, pitch contour, glottal wave shape, spectral tilt, absolute values of formant frequencies, formant trajectories and formant bandwidths.

The system implemented keeps some of the acoustical features of voice-personality while some parts are lost for post-laryngectomised patients. Technically speaking, factors related to the lung excitation and vocal tract characteristics are maintained while those from glottal sources lack in the source whispers and therefore require explicit generation.

Since CELP utilises a separate LTP filter for pitch parameterisation (as described in Section 5.5), generating complex individual pitch patterns is feasible by extending the pitch template unit and exploiting the capabilities of multi-tap LTP filters. In practical terms, having voiced samples of the patients' voice (before surgery) could allow the exact pitch characteristics of their speech to be extracted and preserved. This is beyond the scope of the current work, yet would be a significant improvement to the usability and features of a system for patients' use.

Also, implementing a technique which efficiently considers slight differences between whispered and normal speech duration [87] might further improve the system: applying duration adjustment particularly on consonants (which are 10% longer in whispers [87]) could help to achieve this.

6.5 Summary

Due to the importance of speech quality assessment in telecommunication systems, subjective and objective tests were conducted to evaluate the output of the system designed and implemented in Chapter 5. In this chapter, the results of these tests were demonstrated while a brief review on current standards and methods of speech quality testing was provided. Categorising into two main broad classes of intrusive and

non-intrusive methods, objective tests were reviewed as well as evaluative techniques for speech enhancement outlined.

To demonstrate the effectiveness of the PMF-based enhancement method, the formant trajectory for a whispered sentence before and after spectral enhancement as well as a reconstructed spectrogram for the same sentence was illustrated. Furthermore, two subjective and objective tests were described to evaluate the output of the system. Results show a significant preference for PMF reconstructed vowels and diphthongs over electrolarynx samples, but also highlight the importance of the WPC module classification result to overall output quality.

By considering the output, a discussion was also presented highlighting current system limitations. There are several barriers to further naturalness quality improvement, but the two main system limitations correspond to nasal detection/enhancement and speaker individuality, particularly on sentence regeneration (the vowel/diphthong regeneration phase along with regenerating other phonemes works well). To identify nasals properly, designing a dedicated nasal processing module would be helpful, while considering voice individuality factors could improve the naturalness of the generated speech.

Chapter 7

Conclusions & Future Work

By focusing on reconstruction of phonated speech from whispers, the thesis reviewed the current methods of speech rehabilitation for laryngectomy patients and then in the context of signal processing, whispered speech and its characteristics were presented. In particular, three methods commonly used by these patients including oesophageal speech, TEP, and electrolarynx, were discussed and outlined that these techniques suffer from weaknesses ranging from unnatural monotonous speech to learning difficulties, clumsy usage and risk of infection. Therefore, a novel engineering approach aiming to produce higher quality speech for these patients within a modified framework of code excited linear prediction (CELP) codec was pursued to convert whispers to voice for reconstruction of normal sounding speech.

To cover acoustic features of whispered speech for this purpose, a vowel space for whispered speech was established according to the data collected by the author during a six week research visit to the University of Birmingham, UK. Although a classic vowel space for voiced speech, has played an important role in the development and testing of recognition and processing theories, this type of information has been lacking for whispers. Currently, by establishing the results of this study, whispers-mode communications and recognition systems can be expected to be equally benefited from it.

Regeneration of phonated speech was discussed in detail, outlining the required modules including whisper activity detector (WAD), whispered phoneme classification (WPC), spectral enhancement, pitch insertion based upon a CELP codec. Two methods for spectral enhancement were developed one based on line spectral pair (LSP) narrowing technique and the other based on a novel approach of applying probability mass function (PMF) to find the formant trajectories which the latter showed the more accurate and reliable results to be used in the implemented system. Regarding pitch module, two methods of pitch generation/insertion for the voice regeneration were pointed out, one based on the basic LTP filter and the other for pitch variation based on formant locations and amplitudes while the final system was tested by using the basic LTP filter.

To evaluate the quality of regenerated speech, subjective and objective tests were also conducted and the results were discussed in detail showing a significant improvement over electrolarynx (EL) voice in vowels/diphthongs regeneration while the reconstructed sentences need to be improved (however, it is still better than EL). In fact, both subjective mean opinion score (MOS) test and LPC-based objective Log-Likelihood Ratio (LLR) assessment technique show a significant preference for PMF reconstructed vowels and diphthongs over electrolarynx samples, but also highlight the importance of the WPC module classification result to overall output quality.

7.1 Author's Publications

Journals

J-1 H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Reconstruction of Normal Sounding Speech for Laryngectomy Patients through a Modified CELP Codec," IEEE Transactions on BioMedical Engineering, vol. 57, issue 10, 2010, pp. 2448-2458 (2009 Impact Factor: 2.154)

J-2 H. R. Sharifzadeh, I. V. McLoughlin, M. J. Russell, "A Comprehensive Vowel Space for Whispered Speech," Journal of Voice, Accepted December 2010 (2009 Impact Factor: 1.587)

J-3 H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Voiced Speech from Whispers for Post-Laryngectomised Patients," IAENG International Journal of Computer Science, vol. 36, issue 4, 2009, pp. 367-377

J-4 F. Ahmadi, I. V. McLoughlin, H. R. Sharifzadeh, "Linear Predictive Analysis for Ultrasonic Speech," IET Electronic Letters, vol. 46, issue 6, 2010, pp. 387-388 (2009 Impact Factor: 1.007)

Book Chapters

B-1 H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Speech Rehabilitation Methods for Laryngectomised Patients," Electronic Engineering and Computing Technology, Springer, Netherlands, April 2010, pp. 597-607

B-2 I. V. McLoughlin, H. R. Sharifzadeh, "Speech Recognition for Smart Homes," Speech Recognition Technologies and Applications, InTech, Austria, 2008, pp. 477-494

Refereed Conferences

C-1 H. R. Sharifzadeh, I. V. McLoughlin, M. J. Russell, "Toward a Comprehensive Vowel Space for Whispered Speech," International Symposium on Chinese Spoken Language Processing (ISCSLP), Taiwan, December 2010

C-2 F. Ahmadi, I. V. McLoughlin, H. R. Sharifzadeh, "Autoregressive Modelling for Linear Prediction of Ultrasonic Speech," InterSpeech 2010, Japan, September 2010

C-3 H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Spectral Enhancement of Whispered Speech Based on Probability Mass Function," Advanced International Conference on Telecommunications (AICT), Spain, May 2010

C-4 H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Speech Reconstruction in Post-Laryngectomised Patients by Formant Manipulation and Pitch Profile Generation," International Conference of Systems Biology and Bioengineering (ICSBB), United Kingdom, July 2009*

**Best Paper Award*

C-5 H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Regeneration of Speech in Voice-Loss Patients," International Conference on Biomedical Engineering (ICBME), Singapore, December 2008

C-6 F. Ahmadi, I. V. McLoughlin, H. R. Sharifzadeh, "Analysis-by-Synthesis Method for Whisper-Speech Reconstruction," IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Macao, November 2008

C-7 H. R. Sharifzadeh, I. V. McLoughlin, "Speech Recognition Engine Adaptations for Smart Home Dialogues," IEEE International Conference on Information, Communications and Signal Processing (ICICS), Singapore, December 2007

7.2 Future Works

A discussion on the current system limitations were provided in Chapter 6 which mainly shapes the direction/aim of the future work that would need to be accomplished to improve the commercial or medical usefulness of this research. Apart from effective spectral enhancement, this class of CELP-based reconstructive method implemented in this study relies heavily upon effective phoneme identification, and therefore requires an accurate whispered phoneme classification unit. To have good results

for reconstructing complete sentences (at least, similar to those we have for vowels/diphthongs), an upgraded robust module for classification of whispered phonemes need to be designed.

One other important issue, as was mentioned in Chapter 6, is with nasals; since some whispered phonemes may map to two different spoken phonemes, detecting/identifying the correct phonemes in this case is difficult. In testing the current system, the main issue in this aspect was in how to detect whispered nasals and effectively identify /m/ from /n/.

Simplifying the MATLAB algorithms, or implementing on a DSP, would be necessary to make the system operate in real time. The current system can be known as a near real time whisper-voice conversion system which can be improved through decreasing the code complexity as well as through further simplification and optimisation of the modules.

The naturalness of the output can still be increased through using the advanced pitch filters (i.e. multi-tap filters). Furthermore, by introducing new commercial text to speech kits such as JayBee in [160] which can produce each individual's voice by having the speech samples of that person, naturalness should not be a concern as long as we can have efficient whisper-voice conversion in which the output can go through these off-the-shelf voice synthesis technologies (which also work phoneme based) to produce the most natural speech.

References

[1] B. de Boer. **The evolution of speech**. *Encyclopedia of Language and Linguistics*, **4**:335–338, 2006. 1

[2] R. F. Kay, M. Cartmill, and M. Balow. **The hypoglossal canal and the origin of human vocal behavior**. *Proceedings of the National Academy of Sciences*, **95**:5417–5419, 1998. 1

[3] J. Benesty, M. M. Sondhi, and Y. Huang. *Springer Handbook of Speech Processing*, chapter Introduction to speech processing, pages 1–4. Springer, New York, 2008. 2

[4] H. Dudley and T. H. Tarnocz. **The speaking machine of Wolfgang von Kempelen**. *Journal of the Acoustical Society of America*, **22**:151–166, 1950. 2

[5] R. Pietruch, M. Michalska, W. Konopka, and A. Grzanka. **Methods for formant extraction in speech of patients after total laryngectomy**. *Biomedical Signal Processing and Control*, **1**:107–112, 2006. 3, 10, 61

[6] V. C. Tartter. **Identifiability of vowels and speakers from whispered syllables**. *Perception and Psychophysics*, **49**:365–372, 1991. 3

[7] N. P. Solomon, G. N. McCall, M. W. Trosset, and W. C. Gray. **Laryngeal configuration and constriction during two types of whispering**. *Journal of Speech and Hearing Research*, **32**:161–174, 1989. 3, 29, 32, 37

[8] K. J. Kallail and F. W. Emanuel. **Formant-frequency difference between isolated whispered and phonated vowel samples produced by adult female subject**. *Journal of Speech and Hearing Research*, **27**:245–251, 1984. 3, 31, 37

[9] S. T. Jovicic and M. M. Dordevic. **Acoustic features of whispered speech**. *Acustica- acta acustica*, **82**:228–236, 1996. 3

[10] P. Vary and R. Martin. *Digital Speech Transmission*. John Wiley & Sons Ltd, West Sussex, 2006. 9, 16, 17

[11] C. L. Corey. **Voice rehabilitation after total laryngectomy**. Technical report, Baylor College of Medicine, 2005. 9

[12] M. Azzarello, B. A. Breteque, R. Garrel, and A. Giovanni. **Determination of oesophageal speech intelligibility using an articulation assessment**. *Revue de laryngologie, otologie, rhinologie*, **126**:327–334, 2005. 10, 11, 62

[13] V. Callanan, P. Gurr, D. Baldwin, M. White-Thompson, J. Becksinsale, and J. Bennet. **Provox valve use for post-laryngectomy voice rehabilitation**. *Journal of Laryngology and Otology*, **109**:1068–1071, 1995. 10, 12, 62

[14] J. H. Brandenburg. **Vocal rehabilitation after laryngectomy**. *Archives of Otolaryngology*, **106**:688–691, 1980. 10, 13, 62

[15] H. L. Morris, A. E. Smith, D. R. Van Demark, and M. D. Maves. **Communication status following laryngectomy: The Iowa experience 1984-1987**. *Annals of Otology, Rhinology and Laryngology*, **101**:503–510, 1992. 10, 63

[16] H. Liu, Q. Zhao, M. Wan, and S. Wang. **Enhancement of electrolarynx speech based on auditory masking**. *IEEE Transactions on Biomedical Engineering*, **53**:865–874, 2006. 10, 63, 106

[17] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley, and R. E. Hillman. **Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity**. *IEEE Transactions on Biomedical Engineering*, **51**:325–332, 2004. 10, 63

[18] G. A. Gates, W. Ryan, J. C. Cooper, G. F. Lawlis, E. Cantu, T. Hayashi, E. Lauder, R. W. Welch, and E. Hearne. **Current status of laryngectomee rehabilitation: I. Results of therapy**. *American Journal of Otolaryngology*, **3**:1–7, 1982. 11

[19] R. Hillman, M. Walsh, G. Wolf, and S. Fisher. **Functional outcomes following treatment for advanced laryngeal cancer. Part 1. Voice preservation in advanced laryngeal cancer. Part II. Laryngectomy rehabilitation: the state-of-the-art in the VA system**. *Annals of Otology, Rhinology and Laryngology*, **107**:1–27, 1998. 11, 12, 14

[20] G. Culton and J. Gerwin. **Current trends in laryngectomy rehabilitation: A survey of speech language pathologists**. *Otolaryngology - Head and Neck Surgery*, **115**:458–463, 1998. 12, 14

[21] H. Liu and M. Ng. **Electrolarynx in voice rehabilitation**. *Auris Nasus Larynx*, **34**:327–332, 2006. 13

[22] I. H. Witten. *Principles of Computer Speech*. Academic Press, New York, 1982. 16

[23] S. Saito and K. Nakata. *Fundamentals of Speech Signal Processing*. Academic Press Inc, Tokyo, 1985. 16, 17

[24] W. J. Hess. *Pitch Determination of Speech Signals, Algorithms and Devices*. Springer, 1983. 16

[25] A. C. Gimson and A. Cruttenden. *Gimsons Pronunciation of English*. Arnold, London, 1994. 17

[26] G. Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, second printing, 1970 edition, 1960. 19, 30, 64, 85

[27] B. Lindblom and J. Sundberg. *Springer Handbook of Acoustics*, chapter The human voice in speech and singing. Springer, New York, 2007. 19, 21

[28] G. Fant, J. Liljencrants, and Q. Lin. **A four-parameter model of glottal flow**. Technical report, Speech Transmission Laboratory, Royal Institute of Technology, 1985. 20

[29] A. E. Rosenberg. **Effect of glottal pulse shape on the quality of natural vowels**. *Journal of the Acoustical Society of America*, **49**(2):583–590, 1971. 20

[30] R. Veldhuis. **A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation**. *Journal of the Acoustical Society of America*, **103**(1):566–571, 1998. 20

[31] D. G. Childers and C. K. Lee. **Vocal quality factors: analysis, synthesis, and perception**. *Journal of the Acoustical Society of America*, **90**(5):2394–2410, 1991. 20

[32] K. E. Cummings and M. A. Clements. **Application of the analysis of glottal excitation of stressed speech to speaking style modification**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 207–210, 1993. 20

[33] 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*, **7**(5):569–586, 1999. 20

[34] G. Fant. *Speech Acoustics and Phonetics*. Kluwer Academic Publishers, Netherlands, 2004. 20

[35] G. Fant. **The voice source in connected speech**. *Speech Communication*, **22**:125–139, 1997. 21

[36] G. Fant. **Some problems in voice source analysis**. *Speech Communication*, **13**:7–22, 1993. 21

[37] K. Ishizaka and J. L. Flanagan. **Synthesis of voiced sounds from a two-mass model of the vocal cords**. *Bell System Technical Journal*, **51**:1233–1268, 1972. 21

[38] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1978. 22, 23, 94

[39] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer, New York, 1976. 22

[40] I. McLoughlin. *Applied Speech and Audio Processing*. Cambridge University Press, Cambridge, 2009. 22, 26

[41] J. Makhoul. **Linear prediction: a tutorial review**. *Proceedings of the IEEE*, **63**:561–580, 1975. 23

[42] J. W. Picone. **Signal modelling techniques in speech recognition**. *Proceedings of the IEEE*, **81**:1215–1247, 1993. 23

[43] J. Benesty, J. Chen, and Y. Huang. *Springer Handbook of Speech Processing*, chapter Linear prediction, pages 121–134. Springer, New York, 2008. 23

[44] B. S. Atal and J. R. Remde. **A new model of LPC excitation for producing natural sounding speech at low bit rates**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 614–617, 1982. 24

[45] P. Kroon, E. F. Deprettere, and R. J. Sluyter. **Regular pulse excitation: a novel approach to effective and efficient multi pulse coding of speech**. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34**:1054–1063, 1986. 24

[46] M. R. Shroeder and B. S. Atal. **Code excited linear prediction (CELP): high quality speech at very low bit rates**. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, **10**:937–940, 1985. 24

[47] B. S. Atal and M.R. Schroeder. **Predictive coding of speech signals**. In *Conference on Communications and Signal Processing*, pages 360–361, 1967. 24

[48] B. S. Atal and M.R. Schroeder. **Adaptive predictive coding of speech signals**. *Bell System Technical Journal*, **49**:1973–1986, 1970. 24

[49] C. Avendano, L. Deng, H. Hermansky, and B. Gold. *Speech Processing in the Auditory System*, chapter The analysis and representation of speech, pages 63–100. Springer, New York, 2006. 24

[50] R. Viswanathan and J. Makhoul. **Quantization properties of transmission parameters in linear predictive systems**. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **23**:309–321, 1975. 25

[51] J. L. Flanagan. **Automatic extraction of formant frequencies from continuous speech**. *Journal of the Acoustical Society of America*, **28**:110118, 1956. 25

[52] R.W. Schafer and L. R. Rabiner. **System for automatic formant analysis of voiced speech**. *Journal of the Acoustical Society of America*, **47**(2):634648, 1970. 25

[53] B. S. Atal and S. L. Hanauer. **Speech analysis and synthesis by linear prediction of the speech wave**. *Journal of the Acoustical Society of America*, **50**(2):637655, 1971. 25

[54] S. McCandless. **An algorithm for automatic formant extraction using linear prediction spectra**. *IEEE Transactions on Acoustics, Speech and Signal Processig*, **22**(2):135141, 1974. 25

[55] G. Kopec. **Formant tracking using hidden Markov models and vector quantization**. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34**(4):709 – 729, 1986. 25

[56] S. W. Metz, J. A. Heinen, and R. J. Niederjohn. **Auditory modeling applied to formant tracking of noise-corrupted speech**. In *IEEE International Conference on Industrial Electronics, Control and Instrumentation*, pages 2120 – 2124, 1991. 25

[57] L. Welling and H. Ney. **A Model for efficient formant estimation**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 797–800, 1996. 25

[58] K. Mustafa and I. C. Bruce. **Robust formant tracking for continuous speech with speaker variability**. *IEEE Transactions on Speech and Audio Processing*, **14**(4):435– 444, 2006. 25, 46, 48

[59] Y. Qian, G. Chahine, and P. Kabal. **Pseudo-multi-tap pitch filters in a low bit-rate CELP speech coder**. *Speech Communication*, **14**:339–358, 1994. 26, 28

[60] R. P. Ramachandran and P. Kabal. **Pitch prediction filters in speech coding**. *IEEE Transactions on Acoustics, Speech and Signal Processig*, **37**:467–478, 1989. 27

[61] P. Kroon and W. B. Kleijn. *Speech Coding and Synthesis*, chapter Linear-prediction based analysis-by-synthesis coding, pages 79–119. Elsevier Science, Amsterdam, 1995. 28, 29

[62] J. S. Marques, I. M. Trancoso, J. M. Tribolet, and L. B. Almeida. **Improved pitch prediction with fractional delays in CELP coding**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1990. 29

[63] P. Kroon and B. S. Atal. **On the use of pitch predictors with high temporal resolution**. *IEEE Transactions on Signal Processing*, **39**:733–735, 1991. 29

[64] R. S. Weitzman, M. Sawashima, and H. Hirose. **Devoiced and whispered vowels in Japanese**. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics*, **10**:61–79, 1976. 29, 31

[65] J. H. Esling. **Laryngographic study of phonation type and laryngeal configuration**. *Journal of the International Phonetic Association*, **14**:56–73, 1984. 29

[66] V. C. Tartter. **Whats in a whisper?** *Journal of the Acoustical Society of America*, **86**:1678–1683, 1989. 30, 31, 33

[67] I. B. Thomas. **Perceived pitch of whispered vowels**. *Journal of the Acoustical Society of America*, **46**:468–470, 1969. 30, 33

[68] K. N. Stevens. *Acoustic Phonetics*. The MIT Press, Cambridge, MA, 1998. 30, 31, 32, 33

[69] J. C. Catford. *Fundamental Problems in Phonetics*. Edinburgh University Press, Edinburgh, 1977. 31, 37

[70] R. W. Morris. *Enhancement and recognition of whispered speech*. PhD thesis, Georgia Institute of Technology, 2003. 32, 63

[71] D. H. Klatt and L. C. Klatt. **Analysis, synthesis, and perception of voice quality, variations among male and female talkers**. *Journal of the Acoustical Society of America*, **87**:820–857, 1990. 33, 69

[72] H. E. Stevens. *The representation of normally-voiced and whispered speech sounds in the temporal aspects of auditory nerve responses*. PhD thesis, University of Illinois, 2003. 33

[73] I. Lehiste. *Suprasegmentals*. MIT Press, Cambridge, 1970. 33

[74] K. J. Kallail and F. W. Emanuel. **The identifiability of isolated whispered and phonated vowel samples**. *Journal of Phonetics*, **13**:11–17, 1985. 34

[75] J. Hillenbrand and R. T. Gayvert. **Vowel classification based on fundamental frequency and formant frequencies**. *Journal of Speech and Hearing Research*, **36**:694–700, 1993. 36

[76] J. D. Miller. **Auditory-perceptual interpretation of the vowel**. *Journal of the Acoustical Society of America*, **85**(5):2114–2134, 1989. 36

[77] I. V. Bele. **The Speaker's Formant**. *Journal of Voice*, **20**(4):555 – 578, 2006. 36

[78] T. Nawka, L. C. Anders, M. Cebulla, and D. Zurakowski. **The speaker's formant in male voices**. *Journal of Voice*, **11**(4):422 – 428, 1997. 36

[79] A. K. Syrdal. **Aspects of a model of the auditory representation of American English vowels**. *Speech Communication*, **4**:121–135, 1985. 36

[80] T. M. Nearey. **Static, dynamic, and relational properties in vowel perception**. *Journal of the Acoustical Society of America*, **85**(5):2088–2113, 1989. 36

[81] M. P. Gelfer and V. A. Mikos. **The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels**. *Journal of Voice*, **19**(4):544 – 554, 2005. 36

[82] G. E. Peterson and H. L. Barney. **Control methods used in a study of the vowels**. *Journal of the Acoustical Society of America*, **24**:175–184, 1952. 36, 40

[83] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler. **Acoustic characteristics of American English vowels**. *Journal of the Acoustical Society of America*, **97**(5):3099–3111, 1995. 36, 40, 99

[84] R. A. Fox and E. Jacewicz. **Analysis of total vowel space areas in three regional dialects of American English**. In *Acoustics 08 Paris*, pages 495–500, 2008. 36, 40

[85] A. D. Rubin, V. Praneetvatakul, S. Gherson, C. A. Moyer, and R. T. Sataloff. **Laryngeal Hyperfunction During Whispering: Reality or Myth?** *Journal of Voice*, **20**(1):121 – 127, 2006. 37

[86] J. Sundberg, R. Scherer, M. Hess, and F. Mller. **Whispering–A Single-Subject Study of Glottal Configuration and Aerodynamics**. *Journal of Voice*, **24**(5):574 – 584, 2010. 37

[87] S. T. Jovicic and Z. Saric. **Acoustic Analysis of Consonants in Whispered Speech**. *Journal of Voice*, **22**(3):263 – 274, 2008. 37, 42, 106, 108

[88] S. T. Jovicic. **Formant feature differences between whispered and voiced sustained vowels**. *Acta Acustica united with Acustica*, **84**:739–743, 1998. 37, 68, 79

[89] F. W. Smith. *A formant study of whispered vowels*. PhD thesis, University of Oklahoma, 1973. 37

[90] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi. **Reconstruction of Normal Sounding Speech for Laryngectomy Patients through a Modified CELP Codec**. *IEEE Transactions on Biomedical Engineering*, **57**(10):2448–2458, 2010. 37

[91] R. W. Morris and M. A. Clements. **Reconstruction of speech from whispers**. *Medical Engineering and Physics*, **24**:515 – 520, 2002. 37, 63

[92] Q. Yan and S. Vaseghi. **Analysis, modeling and synthesis of formants of British, American and Australian accents**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 712– 715, 2003. 37, 38

[93] S. D'Arcy. *The effect of age and accent on automatic speech recognition performance*. PhD thesis, University of Birmingham, 2007. 38, 41, 45, 46, 47

[94] Q. Yan and S. Vaseghi. **A comparative analysis of UK and US English accents in recognition and synthesis**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 413– 416, 2002. 38

[95] J. Harringtona, F. Coxa, and Z. Evansa. **An acoustic phonetic study of broad, general, and cultivated Australian English vowels**. *Australian Journal of Linguistics*, **17**(2):155 – 184, 1997. 38

[96] C. Watson, J. Harrington, and Z. Evans. **An acoustic comparison between New Zealand and Australian English vowels**. *Australian Journal of Linguistics*, **18**(2):185 – 207, 1998. 38

[97] J. C. Wells. *Accents of English, Volume 2: The British Isles*. Cambridge University Press, Cambridge, 1982. 38, 50, 54

[98] A. Hughes and P. Trudgill. *English Accents and Dialects*. Edward Arnold, London, 1987. 38, 52

[99] B. Kortmann and C. Upton. *Varieties of English 1: The British Isles*. Mouton de Gruyter, Berlin, 2008. 38

[100] U. Clark. *Varieties of English 1: The British Isles*, chapter The English West Midlands: phonology, pages 145–177. Mouton de Gruyter, Berlin, 2008. 38, 39, 54

[101] P. Trudgill. *The Dialects of England (2nd Edition)*. Wiley-Blackwell, 1999. 39

[102] J. Harrington. **An acoustic analysis of happy-tensing in the Queens Christmas broadcasts**. *Journal of Phonetics*, **34**:439 – 457, 2006. 39

[103] C. H. Wu, Y. H. Chiu, C. J. Shia, and C. Y. Lin. **Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs**. *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(1):266 – 276, 2006. 41

[104] F. Ramus and J. Mehler. **Language identification with suprasegmental cues: A study based on speech resynthesis**. *Journal of the Acoustical Society of America*, **105**(1):512–521, 1999. 41

[105] M. A. Zissman. **Comparison of four approaches to automatic language identification of telephone speech**. *IEEE Transactions on Audio, Speech, and Language Processing*, **4**(1):31–44, 1996. 41

[106] A. G. Adami and H. Hermansky. **Segmentation of speech for speaker and language recognition**. In *Eurospeech*, 2003. 41

[107] L. R. Rabiner. **A tutorial on hidden Markov Models and selected applications in speech recognition**. *Proceedings of the IEEE*, **77**(2):257–286, 1989. 43, 44

[108] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon. **Acoustic modeling for large vocabulary speech recognition**. *Computer Speech & Language*, **4**(2):127–165, 1990. 43

[109] D. Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 2003. 45

[110] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993. 45

[111] S. D'Arcy, M. J. Russell, S. R. Browning, and M. J. Tomlinson. **Accents of the British Isles (ABI) Corpus**. In *Modelisations pour l'Identification Des Langues (MIDL)*, 2004. 47

[112] R. A. Kazi, V. M. Prasad, J. Kanagalingam, C. M. Nutting, P. Clarke, P. Rhys-Evans, and K. J. Harrington. **Assessment of the formant frequencies in normal and laryngectomized individuals using linear predictive coding**. *Journal of Voice*, **21**:661–668, 2007. 62

[113] A. V. McCree and T. P. Barnwell. **A mixed excitation LPC vocoder model for low bit rate speech coding**. *IEEE Transactions on Speech and Audio Processing*, **3**(4):242 – 250, 1995. 63

[114] N. Sugie and K. Tsunoda. **A Speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production**. *IEEE Transactions on Biomedical Engineering*, **32**(7):485–490, 1985. 64

[115] **ITU-T Recommendation G.729 Annex B: A silence compression scheme for G.729 optimized for terminal conforming**, 1996. 65, 66

[116] Q. Li, J. Zheng, A. Tsai, and Q. Zhou. **Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition**. *IEEE Transactions on Speech and Audio Processing*, **10**(3):146–157, 2002. 65

[117] J. Sohn, N. S. Kim, and W. Sung. **A statistical model-based voice activity detection**. *IEEE Signal Processing Letters*, **6**(1):1–3, 1999. 66

[118] K. Woo, T. Yang, K. Park, and C. Lee. **Robust voice activity detection algorithm for estimating noise spectrum**. *Electronics Letters*, **36**(2):180 –181, 2000. 66

[119] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano. **Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors**. *IEEE Signal Processing Letters*, **9**:85–88, 2002. 66

[120] B. Chen and P. C. Loizou. **Formant frequency estimation in noise**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 581–584, 2004. 68

[121] Q. Zhao, T. Shimamura, and J. Suzuki. **A robust algorithm for formant frequency extraction of noisy speech**. In *IEEE ISCAS*, pages 534–537, 1998. 68

[122] G. E. Peterson. **Parameters of vowel quality**. *Journal of Speech and Hearing Research*, **4**:10–29, 1961. 68

[123] A. Goalic and S. Saoudi. **An intrinsically reliable and fast algorithm to compute the line spectrum pairs (LSP) in low bit rate CELP coding**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 728–731, 1995. 69, 70

[124] I. V. McLoughlin. **Review: Line spectral pairs**. *Signal Processing*, **88**(3):448–467, 2008. 69, 72

[125] P. Kabal and R. P. Ramachandran. **The computational of line spectral frequencies using Chebyshev polynomials**. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34**(6):1419 – 1426, 1986. 70

[126] S. Saoudi, J. M. Boucher, and A. Guyader. **A new efficient algorithm to compute the LSP parameters for speech coding**. *Signal Processing*, **28**(2):201–212, 1992. 70

[127] I. V. McLoughlin and R. J. Chance. **LSP-based speech modification for intelligibility enhancement**. In *13th international conference on DSP*, pages 591–594, 1997. 70

[128] H. R. Sharifzadeh, F. Ahmadi, and I. V. McLoughlin. **Analysis-by synthesis method for whisper-speech reconstruction**. In *IEEE APCCAS*, pages 1280–1283, 2008. 72

[129] H. Kuwabara and K. Ohgushi. **Contributions of vocal tract resonant frequencies and bandwidths to the personal perception of speech**. *Acoustica*, **63**:120–128, 1987. 73

[130] J Steinier, Y. Termonia, and J. Deltour. **Smoothing and differentiation of data by simplified least square procedure**. *Analytical Chemistry*, **44**(11):1906– 1909, 1972. 77

[131] Y. S. Hsiao and D. G. Childers. **A new approach to formant estimation and modification based on pole interaction**. In *30th Asilomar Conference on Signals, System and Computers*, pages 783–787, 1997. 78

[132] H. Mizuno, M. Abe, and T. Hirokawa. **Waveform-based speech synthesis approach with a formant frequency modification**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 195–198, 1993. 78

[133] A. M. Kondoz. *Digital Speech: Coding for Low bit Rate Communication Systems*. John Wiley & Sons, 1995. 80

[134] B. S. Atal. **Predictive coding of speech at low bit rates**. *IEEE Transaction on Communications*, **30**(4):600–614, 1982. 80

[135] J. P. Campbell, T. E. Tremain, and V. C. Welch. **The proposed federal standard 1016 4800 bit/s voice coder: CELP**. *Speech Technology*, **5**:58–64, 1990. 80

[136] W. T. Fitch and J. Giedd. **Morphology and development of the human vocal tract: A study using magnetic resonance imaging**. *Journal of the Acoustical Society of America*, **106**(3):1511–1522, 1999. 85

[137] P. F. Assmann and T. M. Nearey. **Relationship between fundamental and formant frequencies in voice preference**. *Journal of the Acoustical Society of America*, **122**(2):EL35–EL43, 2007. 85

[138] P. F. Assmann, S. Dembling, and T. M. Nearey. **Effects of frequency shifts on perceived naturalness and gender information in speech,Proceedings of the 9th ,**. In *International Conference on Spoken Language Processing*, pages 889–892, 2006. 85

[139] **ITU-T Recommendation P.800: Methods for subjective determination of transmission quality**, 1996. 90, 100

[140] L. Cai, T. Ronghui, Z. Jiying, and M. Yongyi. **Speech quality evaluation: a new application of digital watermarking**. *IEEE Transactions on Instrumentation and Measurement*, **56**(1):45–55, 2007. 90

[141] A. W. Rix. **Perceptual speech quality assessment - a review**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004. 90, 92, 96

[142] S. Voran. **Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique**. *IEEE Transactions on Speech and Audio Processing*, **7**(4):371–382, 1999. 92, 93

[143] A. W. Rix, J. G. Beerends, D-S. Kim, P. Kroon, and O. Ghitza. **Objective assessment of speech and audio quality- technology and applications**. *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(6):1890 – 1901, 2006. 92, 96, 97

[144] S. Wang, A. Sekey, and A. Gersho. **An objective measure for predicting subjective quality of speech coders**. *IEEE Journal on Selected Areas in Communications*, **10**(5):819–829, 1992. 93

[145] M. Hansen and B. Kollmeier. **Using a quantitative psycho-acoustical signal representation for objective speech quality measurement**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1387–1390, 1997. 93

[146] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends. **Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part ITime-delay compensation**. *Journal of Audio Engineering Society*, **50**(10):755–764, 2002. 93

[147] A. De and P. Kabal. **Auditory distortion measure for speech coder evaluationDiscrimination information approach**. *Speech Communication*, **14**(3):205–229, 1994. 93

[148] **ITU-T Recommendation P.861: Objective quality measurement of telephone-band (300–3400 Hz) speech codecs**, 1998. 93

[149] **ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs**, 2001. 93

[150] A. W. Rix and M. P. Hollier. **The perceptual analysis measurement system for robust end-to-end speech quality assessment**. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1515–1518, 2000. 93

[151] Y. Hu and P. Loizou. **Evaluation of objective quality measures for speech enhancement**. *IEEE Transactions on Audio, Speech, and Language Processing*, **16**(1):229–238, 2008. 94, 103

[152] S. P. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality. Englewood Cliffs, NJ: Prentice-Hall, 1988.* Prentice Hall, Englewood Cliffs, NJ, 1988. 94, 95

[153] D-S. Kim. **ANIQUE: An auditory model for single-ended speech quality estimation**. *IEEE Transactions on Speech and Audio Processing*, **13**(5):821–831, 2005. 96

[154] P. Gray, M. P. Hollier, and R. E. Massara. **Non-intrusive speech-quality assessment using vocal-tract models**. *IEE Proceedings - Vision, Image and Signal Processing*, **147**(6):493–501, 2000. 96

[155] **ITU-T Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications**, 2004. 96

[156] **ITU-T Recommendation P.562: Analysis and interpretation of INMD voice-service measurements**, 2000. 97

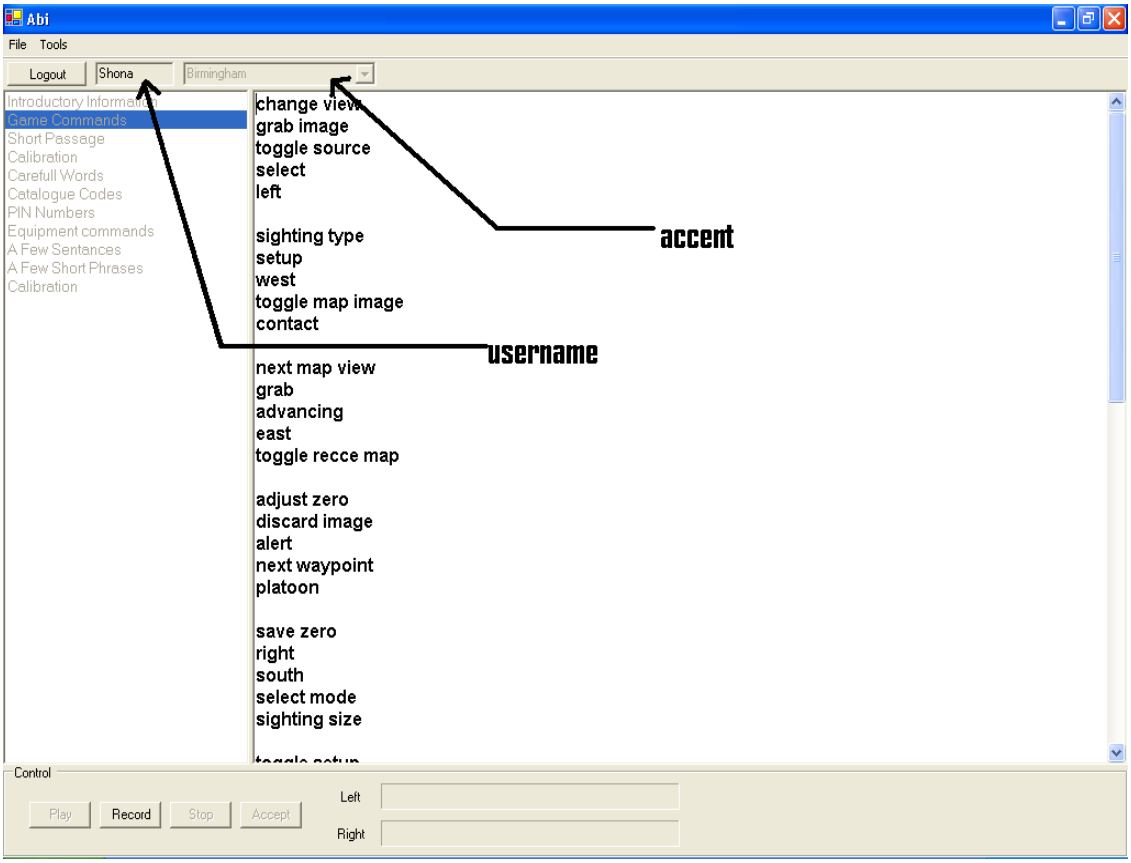
[157] J. Hansen and B. Pellom. **An effective quality evaluation protocol for speech enhancement algorithms**. In *ICSLP*, **7**, pages 2819–2822, 1998. 103

[158] T. Ito, K. Takeda, and F. Itakura. **Analysis and recognition of whispered speech**. *Speech Communication*, **45**(2):139–152, 2005. 106

[159] H. Kuwabara and Y. Sagisaka. **Acoustic characteristics of speaker individuality: control and conversion**. *Speech Communication*, **16**:165–173, 1995. 107

[160] I. Schofield. **Pioneering speech-generating software for MND sufferers**. *Engineering & Technology*, **5**(16):33–37, 2010. 114

Appendix A: ABI Interface



Appendix B: Ethics Statement

Checklist for Ethics Approval

NANYANG TECHNOLOGICAL UNIVERSITY

Principal Applicant: Dr Ian McLoughlin
Project Title: Bionic voice - technology for voice replacement in post-laryngectomised patients

Type of grant: (Project)/Programme/Co-operative/Core Competence Amount requested: $ 180,700

A	Please indicate if the protocol of the application involves any of the following for bio-ethics consideration:	Yes/No
1	human experimentation/trial	YES
2	human tissues/organs	NO
3	collection of patient biodata	NO
4	cell lines derived from human tissues (except ATCC)	NO
5	experimentation which requires Class 2 and above containment	NO

B	Please indicate if the protocol of the application involves any of the following for animal ethics consideration:	Yes/No
1	animal experimentation	NO
2	animal tissues/organs	NO
3	cell lines derived from animal tissues (except ATCC)	NO

If yes to any of the above items, please give details on source(s) of research materials and explain how ethical issues relating to their use will be dealt with (use additional page if space is insufficient):
(Recommended reference: Bioethics Advisory Committee, Singapore - www.bioethics-singapore.org)

I confirm that the above statements are true and correct.

Signature: [signature] Date: 9/11/08

. APPENDIX B: ETHICS STATEMENT

NIE
NATIONAL INSTITUTE OF EDUCATION

National Institute of Education
1 Nanyang Walk Singapore 637616
Tel: (65) 6790 3888
NTU Reg. No. 200604393R

BERC 010/2008

24 January 2008

Associate Prof Ian McLoughlin
School of Computer Engineering

NTU BIOETHICS REVIEW COMMITTEE APPROVAL
NMRC PROJECT TITLE: Bionic voice - technology for voice replacement in post laryngectomised patients (NMRC/EDG/0004/2007. Grant amount: $180,700)

I refer to your application for ethics approval in respect of the above project.

The Committee has deliberated on your application and noted from the Checklist for Ethics Approval form submitted (copy attached) that informed consent will be obtained from the human subjects. The Committee also noted that none of the tests to be conducted at NTU will require any invasive procedures, and the risk to the human subjects will be minimised.

The Committee is therefore satisfied with the bioethical considerations for this project and approves the ethics application.

Prof Lee Sing Kong
Chair, NTU Bioethics Review Committee
enc.

cc Director, Research Support Office
Chair, School of Computer Engineering
Members, NTU Bioethics Review Committee

I AWARD
INNOVATION SINGAPORE
Innovation of the year 2001

An Institute of NANYANG TECHNOLOGICAL UNIVERSITY