LARYNGECTOMEE SPEECH ENHANCEMENT USING VOICE CONVERSION TECHNIQUES

Arantza del Pozo

Christ’s College
University of Cambridge
July 22, 2004

Submitted in partial fulfilment of the requirements for the degree of Master of Philosophy
Abstract

People who suffer from larynx cancer are often laryngectomized, losing, as a consequence, their ability to produce normal speech. Despite advances in alaryngeal speech rehabilitation and restoration, laryngectomee speech is of poorer quality and intelligibility than natural speech. This project explores the use of voice conversion as an alternative method to the enhancement of the quality and intelligibility of laryngectomee speech. An initial comparison of normal and laryngectomee glottal excitations confirms that the cause of their disorder is a voicing problem. Then, several experiments attempt to improve its perceptual quality. The first one checks the sanity of laryngectomee articulation and the rest replace its excitation with a better glottal source. We have found that residuals do not model normal and laryngectomee excitation differences and that glottal waveforms need to be mapped instead. However, the difficult task of estimating glottal waveforms conditions a straightforward continuous glottal mapping, deriving converted utterances with artefacts which need to be solved.

Declaration

I Arantza del Pozo of Christ’s College, being a candidate for M.Phil in Computer Speech, Text and Internet Technology, hereby declare that this dissertation and the work described in it are my own work, unaided except as may be specified below, and that the dissertation does not contain material that has already been used to any substantial extent for a comparable purpose. The source code can be found under ~ad371/Project.

Signed,

July 22, 2004

Acknowledgements

I would like to thank my supervisor Steve Young for his guidance and for always finding time to discuss my doubts. Many thanks to KK Hui Ye for providing me with the details and code for the analysis and synthesis of speech signals and for responding to my questions concerning its modification. Thanks to Antonia Kilcommons and Rachael Beddard for contacting the patients, organising the recording slots and providing facilities to build the laryngectomee data collection. Also special thanks to all the patients who took the time to record the utterances for the speech corpus.
LIST OF FIGURES

Fig. 1 Speech production apparatus.................................................................7
Fig. 2 Speech production apparatus before and after laryngectomy ...............8
Fig. 3 Glottal flow volume velocity waveform .............................................13
Fig. 4 Glottal waveform model.................................................................14
Fig. 5 Time-varying cross-sectional tube and Lossless Tube Concatenation models of the vocal tract ............................................................15
Fig. 6 Source/Filter model of speech production .........................................16
Fig. 7 Voice conversión framework..........................................................18
Fig. 8 Simplified speech production model ..............................................21
Fig. 9 Alignment of normal and laryngectomee voiced frames within phone g-r+iy....23
Fig. 10 Model used in practice to estimate glottal waveforms ......................26
Fig. 11 Flow diagram of the implemented glottal waveform estimation algorithm......27
Fig. 12 Glottal excitation waveforms obtained for different normal male (a) and female (b) speakers .................................................................29
Fig. 13 Glottal excitation waveforms obtained for the normal target speaker (a) and the different tracheoesophageal speakers from the data collection (b) ....................31
Fig. 14 Glottal excitation spectrums of the normal target (a) and different tracheoesophageal speakers in the speech corpus (b) .....................................................33
Fig. 15 Flow diagram of the envelope mapping experiment ................................37
Fig. 16 Flow diagram of the residual mapping experiment ..........................40
Fig. 17 Short speech segment glottal mapping flow diagram ...........................42
Fig. 18 Normal, tracheoesophageal and converted short segments of vowel /ah/........44
Fig. 19 Glottal excitation waveforms for different normal speech vowels ........45
Fig. 20 Flow diagram of the fixed excitation mapping experiment ....................46
Fig. 21 Flow diagram of the continuous glottal mapping experiment ..............48
1 INTRODUCTION

Laryngectomees are people who have had their larynx removed due to laryngeal cancer, losing as a consequence their ability to produce natural voice. After laryngectomy, they have to learn a new method for speech production in order to be able to talk again. Though vocal rehabilitation and voice restoration techniques have improved considerably over the years, the quality and intelligibility of the resulting alaryngeal speech is still significantly lower than that of laryngeal speech.

Voice conversion (VC) is the process of modifying an utterance from a particular speaker to make it sound like that of a specific different target speaker. State of the art VC systems achieve transformations mapping spectral envelope, excitation, average pitch and duration characteristics from target to source speech.

The goal of the project is to explore the enhancement of the quality and intelligibility of laryngectomee speech with voice conversion techniques which allow the transformation of the characteristics that cause the degradation in the quality of their voice.

1.1 MOTIVATION

In the UK, annually around 2,400 people are diagnosed with laryngeal cancer [1]. Its causes are strongly associated with tobacco and alcohol use. Though approximately 80% of cases occur in men, changes in smoking and drinking habits have increased laryngeal cancer rates in women in recent years. When diagnosed at an early stage, tumors can be treated with radiotherapy or laser surgery. However, larger tumors and instances of the recurrence of the disease often make a total laryngectomy necessary.

Primary functions of the larynx are breathing and protection of the lungs, i.e. closing the airway during swallowing, coughing for clearance of the airways. Voice production is a secondary, not life-saving function of the larynx. It is, nevertheless, a significant function, since voice production enables together with the vocal tract and articulators, oral communication. After laryngectomy, all these functions of the larynx are lost and need to be substituted. Breathing and protection of the lungs is achieved with surgery, which results in a complete separation between the pharynx and the trachea and thus, between the alimentary and respiratory pathways. Voice and speech
restoration, however, can’t be achieved with surgery, being hence the focus of rehabilitation after laryngectomy.

The loss of the normal voice is generally considered to be the most obvious consequence of total laryngectomy. Despite advances in alaryngeal communication, laryngectomee speech achieved with state-of-the-art voice rehabilitation and restoration techniques has lower quality and is significantly less intelligible than normal laryngeal speech [2]. As a consequence, laryngectomees often find oral communication difficult and embarrassing. Activities which were normal and easy before surgery such as speaking on the phone or chatting to friends in a pub become difficult tasks with their new speech, which explains their social autoexclusion. Therefore, alternative methods to enhance laryngectomee speech quality and intelligibility are necessary to help improve their oral communication. In such a sense, this project aims to explore the application of voice transformation techniques used in voice morphing systems to enhance the quality and intelligibility of laryngectomee speech and as a consequence, to improve their quality of life.

1.2 SPEECH PRODUCTION BEFORE AND AFTER LARYNGECTOMY

The anatomy of the speech production system, the role of the different organs and the speech production mechanism change after laryngectomy and laryngectomees can use different voice rehabilitation techniques after surgery for the production of their new voice. A good understanding of how normal and laryngectomee speech production mechanisms differ is thus important in order to gain insight into the causes of the perceptual degradation of laryngectomee speech quality.

1.2.1 NORMAL SPEECH PRODUCTION

Speech is produced by air-pressure waves that originate from the movements of the anatomical structures which constitute the human speech production system. The human speech production apparatus consists of:

- **Lungs**: source of air during speech
- **Trachea**: windpipe
- **Larynx**: organ of voice production
- **Pharyngeal cavity** and **Oral or buccal cavity**: correspond to the throat and mouth respectively and are often grouped into one unit referred to as the vocal tract
- **Nasal cavity**: nose, often called nasal tract
- **Articulators**: finer anatomical components which move to different positions to produce the different speech sounds. The main articulators are the **vocal cords**, **soft palate or velum**, **tongue**, **teeth**, **lips** and **jaw**.

![Speech production apparatus](image)

Fig. 1 Speech production apparatus

Speech production can be explained in terms of an acoustic filtering operation, where the three main cavities of the speech production system (vocal plus nasal tracts) comprise the main acoustic filter. This filter is excited by the organs below it and loaded at its output by the radiation of the lips. The change of the articulators associated with the filter itself is used to vary the properties of the system, the form of excitation and its output loading over time.

In such a model, the larynx is the main organ responsible for the production of speech. Formed by the vocal cords, epiglottis and the first two or three tracheal rings, it is connected through the pharynx with the vocal and nasal upper airways and through the trachea with the lungs. The opening between the vocal folds is often referred to as the glottis. When the air stored in the lungs passes through the larynx, the vibration of the vocal cords modulated by the shape of the pharynx and mouth which changes with the position of the tongue, lips and jaws produces the speech signal.

### 1.2.2 LARYNGECTOMEE SPEECH PRODUCTION

Most people who are diagnosed with laryngeal cancer are laryngectomized. During this surgical procedure, the entire larynx (ie. vocal cords, epiglottis and tracheal rings) is removed. After surgery, the mucosa overlying the pharyngeal muscles in the esophagus and pharynx serve as the new voice source known as neoglottis. In order to
replace the lack of a valve to separate respiratory and digestive tracts, the trachea is connected to an opening in the neck called the stoma. Laryngectomees will still eat through their mouth but will breathe, however, through the stoma. In addition, they will also have to learn a new speech production technique. Nowadays, three are the main methods used for voice restoration after total laryngectomy: esophageal speech, tracheoesophageal speech and electrolaryngeal speech.

**Fig. 2 Speech production apparatus before and after laryngectomy**

### 1.2.2.1 ESOPHAGEAL SPEECH

In esophageal speech, air is first brought from the mouth into the esophagus, and then during eructation this air is brought back into the mouth, causing vibrations of the neoglottis.

Esophageal speech is often described as a harsh voice of low pitch and loudness. This method is sometimes compared with belching although the air that creates the voice doesn’t come from the stomach but from the upper part of the esophagus. The amount of air that can be used for voice production is very small and therefore, the maximum phonation time of this type of speech is very short, limiting speech production to short segments and sentences.

Not all laryngectomized patients are able to acquire esophageal speech, and the time required to learn it varies among speakers from as little as a week to as long as a year. However, this method of voice rehabilitation doesn’t require any additional surgery or prosthesis, and as a result is adequate for patients whose anatomy after laryngectomy doesn’t meet the requirements for further surgery and implantation of a prosthesis.
1.2.2.2 TRACHEOESOPHAGEAL SPEECH

For this method of voice restoration, a voice prosthesis is inserted into a surgically created fistula between the trachea and the esophagus. This voice prosthesis enables the use of exhaled air from the lungs for voice production. When the stoma is closed, the exhaled pulmonary air is directed through the voice prosthesis into the esophagus, setting the neoglottis into vibration. Then, the voice sound produced by the neoglottis is formed into speech by the normal resonators and articulators. The voice prosthesis itself does not thus generate any voice sound, but only allows the air from the lungs to enter the esophagus.

Tracheoesophageal voice restoration has often been cited as the alaryngeal speech alternative most comparable to normal laryngeal speech in quality, fluency and ease of production. It has a perceptually better quality than esophageal speech [3], mainly due to its greater speaking rate and fluency made possible by the voice prosthesis. It is, in addition, characterized by a longer phonatory duration and louder voice. Intelligibility is also generally higher for tracheoesophageal speech than for esophageal speech.

With the development of a useful voice prosthesis two decades ago, the use of tracheoesophageal speech was initiated making it the most frequently used method of voice restoration after total laryngectomy today. The prostheses act as one-way valves, which enable air passing from the lungs into the esophagus, but prevent leakage of food or saliva from the esophagus into the lungs. Over the years, several voice prostheses have been developed. Initial prostheses required manual occlusion of the stoma with a finger or thumb. Nowadays, tracheostoma breathing valves can be used to achieve hands-free communication. These valves automatically shunt air into the esophagus for voice production, eliminating the need for manual interaction.

1.2.2.3 ELECTROLARYNGEAL SPEECH

An electrolarynx is a hand-held device with an electromagnetically vibrating membrane. When this membrane is held against the skin of the neck or the floor of the mouth, the vibrations at one fixed frequency are transmitted through the skin into the vocal tract, where they are modulated with the articulatory organs into speech.

Speech produced by an electrolarynx sounds mechanical, robot-like and monotonous. This is why, although these devices offer rapid acquisition of speech ability and easy of use, many laryngectomees prefer other alternatives that offer a more natural vocal quality and a capacity for hands-free communication.
An additional drawback is that it is sometimes difficult to position the electrolarynx to the neck. Also, in case of edema or scar tissue of the neck, transmission of the vibrations through the skin results impossible. Therefore, in general, an electrolarynx is only used in case of failure of tracheoesophageal or esophageal speech.

1.3 VOICE CONVERSION AS A SPEECH ENHANCEMENT APPROACH

Voice conversion aims to transform the voice of one speaker into that of another speaker. A considerable amount of effort has been directed at this problem in the last few years due to its wide range of possible applications. It can, for example, be used to customize text-to-speech (TTS) systems, allowing the production of many different voices in order to achieve a more natural human-machine interaction. Film dubbing applications can also benefit from voice conversion and maintain the speaker identity of the original actors when dubbing to different languages, without them having to learn new languages or be present.

The aim of voice conversion needs to be refined for speech enhancement purposes. The objective is not to make the disordered voice sound like that of the normal target speaker, but to make it sound better. Voice conversion is achieved mapping a set of parameters modelling voice characteristics from target to source speech. This technique is still valid for speech enhancement, if only the characteristics responsible for the decrease in the quality of disordered speech are mapped. As a result, the converted speech is expected to still sound like the original speech, but to have an improved quality.

Speech models allow the parameterisation of vocal tract, excitation and prosodic characteristics of speech and hence, the majority of voice morphing systems convert spectral envelopes [4], excitation residuals [5] and pitch contours [6]. With the removal of the larynx, and thus of the vocal cords, laryngectomies lose the natural voicing source used in the production of laryngeal speech. However, their vocal tract is not altered and laryngectomies speech articulation is still normal. The main cause of the reduced quality and intelligibility of laryngectomy speech is thus thought to be due to the lack of a proper excitation signal. As a consequence, a voice conversion laryngectomy speech enhancement approach should attempt to improve its perceptual quality mapping excitation signals from normal to laryngectomies speech.
The voice conversion speech enhancement approaches explored in the project are based on the observation that the main limitation of laryngectomee speech is the lack of a natural voicing source. An initial experiment grafts the vocal tract spectrum of the laryngectomee speech onto that of the normal target speech. The aim of this experiment is to determine the highest level of improvement the excitation signal mapping approach can obtain. Though this enhancement method is not applicable to real time situations, it is directed to provide a sense of the possible enhancement that can be achieved. In subsequent experiments, the alaryngeal speech excitation is substituted with a better glottal source.

1.4 OUTLINE

This document can be outlined as follows. Chapter 2 provides a background in speech modelling and voice conversion. Chapter 3 analyses the differences between normal and laryngectomee excitations and reviews the enhancement options voice conversion allows. Chapter 4 describes the data collection and the implemented voice repair experiments. Chapter 5 contains the conclusions derived from this thesis and possible further extensions.
2 SPEECH MODELLING AND VOICE CONVERSION

This chapter summarises the signal processing framework on which voice conversion depends. Section 2.1 describes the models used to represent the different subareas of the speech production process, which give rise to the Source/Filter model of speech production presented in section 2.1.4. The voice conversion system implemented for the purposes of this project is then presented in section 2.2.

2.1 MODELLING SPEECH PRODUCTION

A complete characterization of speech production requires mathematical representations describing the physical principles of air propagation in the vocal system, based on acoustic theory and air fluid mechanics. Although a detailed universal theory is not yet available, acoustic tube models provide in practice a good approximation and understanding of the physics involved.

The human speech production mechanism reveals three separate areas to model: source excitation, vocal tract shaping and lip radiation. Most modern speech modelling techniques assume that these components are linear and separable, since linearity and noncoupling assumptions facilitate computationally feasible modelling, synthesis and coding methods. The following subsections describe the characteristics of each of these areas.

2.1.1 SOURCE EXCITATION MODELLING

There are two elemental source types: voiced and unvoiced excitation. These source excitation signals classify speech sounds into:

- **Voiced sounds**: produced by the vibration of the vocal folds which periodically interrupt the subglottal airflow passing through the glottis, generating quasi-periodic puffs of air which excite the vocal tract.

- **Unvoiced sounds**: generated by forming a constriction at some point along the vocal tract, driving a turbulent noise-like excitation caused by the airflow passing through such a constriction.

Additional forms of excitation exist (plosive excitation, whisper, silence) derived from combinations and variations of voiced and unvoiced sources. Actually, a sound may be simultaneously voiced and unvoiced or may be composed of a short region of silence, followed by a region of voiced speech, unvoiced speech or both. However, only
voiced and unvoiced excitations are taken into account in practical speech production models.

2.1.1.1 VOICED EXCITATION

The glottal flow volume velocity waveform acts as the source excitation of the vocal tract in voiced speech. Its shape, illustrated in Fig. 3, is due to the repeated vibration of the glottis, periodically opening and closing in response to the subglottal air pressure from the trachea. The closed-phase of the oscillation occurs when the glottis is closed and the glottal flow volume velocity is zero. On the other hand, the open-phase is characterized by a non-zero volume velocity, in which the lungs and the vocal tract are coupled.

![Glottal flow volume velocity waveform](image)

The vibration process can be described as follows. Initially, the vocal folds are closed and the air stream from the lungs builds up pushing the vocal folds apart. Eventually, the pressure reaches a level sufficient to force the vocal cords to open and thus allow air to flow through the glottis. Then, the pressure in the glottis falls allowing the cords to come together, and the cycle is repeated.

The time between successive vocal fold openings is called the fundamental period $T_0$, while the rate of vibration is called the fundamental frequency of the phonation $F_0 = \frac{1}{T_0}$.

In order to model the voiced excitation signal, the glottal waveform is generated by a glottal shaping filter driven by a discrete-time impulse train $e(n)$, whose pulses are
spaced at the pitch period. The resulting signal corresponds to the excitation of the vocal tract during voiced speech.

![Glottal waveform model](image)

**Fig. 4 Glottal waveform model**

### 2.1.1.2 UNVOICED EXCITATION

Unvoiced excitation includes frication at a point of major constriction along the vocal tract and is normally modelled using white noise. This type of excitation theoretically has no effect on the shape of the speech spectrum, since its power density spectrum is flat over all frequencies.

### 2.1.2 VOCAL TRACT MODELLING

A sound wave in free space propagates radially outward from the point sound source. When the speech sound wave is produced due to the vibration of the vocal folds, the vocal tract prevents its radial propagation and limits it to only one direction along the vocal tract to the lips.

The vocal tract can be approximated by a tube with a time varying cross-sectional area like that of Fig. 5. The complex formulation derived from such a time-varying vocal tract shape can be simplified by representing the vocal tract as a concatenation of lossless tubes with cross-sectional areas $A_k$ of length $l_k$, chosen to approximate the vocal tract area function. If a large number of tubes with short lengths is used, the formant structure of the concatenated tubes is expected to approach that of a tube with a continuously varying cross-sectional area as in Fig. 5. The Lossless Tube Concatenation Model of the Vocal Tract is a widely used model for speech production. It can be shown [7] that the lossless tube model exhibits characteristics common to a discrete-time digital filter model and that the derived vocal tract transfer function $H(z)$ is all-pole, its poles defining the resonant or formant structure of the N tube model. As a result, the vocal tract is generally modelled by an all-pole digital filter.
It should be pointed out that some of the assumptions made regarding the vocal tract influence the outcome of this model and should be taken into account when using the lossless tube model to analyse and process speech. The assumption that the vocal tract can be appropriately represented by a lossless series of tubes ignores certain frequency domain effects that occur as a result of the vocal tract losses, which broaden the bandwidths of the lower and higher formants. Also, nasal and fricative tract configurations are not modelled by the lossless tube model, since it does not take into account side cavities and antiresonances (zeros) involved in their production.

### 2.1.3 LIP RADIATION MODELLING

The radiation at the lips can be thought of as a low-impedance load that terminates in the vocal tract and converts the volume velocity at the lips to a pressure wave. The accurate model given by [8] is shown to have the same properties as a digital differencer

\[ R(z) = Z_{\text{lips}} = 1 - z_0 \cdot z^{-1} \]

with \( z_0 \approx 1, z_0 < 1 \). Therefore, the pressure/volume velocity relation at the lips is usually approximated by a differenciatior like the one in (1).

### 2.1.4 THE SOURCE FILTER MODEL OF SPEECH PRODUCTION

Derived from the source excitation, vocal tract shaping and lip radiation modelling approaches described in the previous sections, a reasonably general linear discrete-time model for speech production can be defined.
In such a system, a vocal tract model $V(z)$ and radiation model $R(z)$ are excited by a discrete-time glottal excitation signal $u_{\text{glottis}}(n)$. During unvoiced speech production, the excitation source is a flat spectrum noise source modelled by a random white noise generator. During voiced speech periods, on the other hand, the excitation uses an estimate of the local pitch period to set an impulse train generator that drives a glottal pulse shaping filter $G(z)$. This excitation produces a glottal flow velocity waveform similar in shape to the one discussed in the previous section. The output of the system is the result of filtering the appropriate excitation by several linear, separable filters: $S(z) = E(z) \cdot V(z) \cdot R(z)$ for unvoiced speech, and $S(z) = E(z) \cdot G(z) \cdot V(z) \cdot R(z)$ for voiced speech.

It is important to note that the Source-Filter Model of Speech Production is based on the assumptions made in the modelling of the source excitation, vocal tract and lip radiation mechanisms. Regarding the excitation signal, only voiced or unvoiced source signals are modelled and mixed excitation is ignored. The all pole filter representing the vocal tract does not take into account side cavities, providing a poor model of nasal and fricative sounds. In addition, the model assumes excitation at the glottis to be decoupled from the vocal tract. Nevertheless, it has been shown to be a good approximation in practice and linear prediction techniques which allow the identification of an all pole model from a speech waveform make the source filter model of speech production potentially useful, being widely used in speech analysis and processing applications.
2.2 VOICE CONVERSION FRAMEWORK

A flow diagram of the implemented voice conversion system can be seen in Fig. 7. This section focuses on the various modules illustrated on this figure. The steps can be summarized as follows:

*Pitch mark extraction:* As a first step, pitch period marks are extracted from both normal and tracheoesophageal utterances. These pitch marks will serve as frame boundaries in the analysis and synthesis modules.

*Analysis:* Normal and tracheoesophageal utterances are analysed separately with the Quasiharmonic ABS/OLA sinusoidal model, proposed by [9]. Each signal is segmented into pitch synchronous frames and a set of parameters is generated for each frame. These parameters enable high quality modification and reconstruction of the speech signal. In addition, the spectral envelope and excitation spectrum for each pitch period are derived from a per frame linear prediction (LP) analysis and a glottal excitation extraction algorithm. These two latter parameters will allow the mapping of spectral envelopes and excitations between normal and tracheoesophageal speech.

*Alignment:* The duration and rate of articulation of a particular utterance varies among speakers. Time alignment of normal and tracheoesophageal sentences is necessary in order to map speech features that correspond to the same phone. Time aligned transcriptions were obtained by running HTK in forced alignment mode. The number of frames in each phone can also vary for different speakers. Therefore, a frame alignment procedure is also applied in this step in order to align each frame of each phone between normal and tracheoesophageal utterances.

*Mapping:* The experiments proposed to explore the suitability of the voice conversion approach as a laryngectomee speech enhancement method involve different mappings between normal and tracheoesophageal spectral envelope and excitation signals. Two different mapping algorithms have been developed, for envelope and excitation mapping respectively.

*Synthesis:* Once the desired characteristics of the processed utterances have been modified, the converted speech is resynthesised with the synthesis formula provided by the Quasiharmonic ABS/OLA Sinusoidal Model.
2.2.1 PITCH MARK EXTRACTION

Pitch marks indicate the moment of glottal closure and are required by the analysis module. Unfortunately, automatic pitch extraction is still an unsolved problem. Different pitch extraction methods exist which attempt to estimate the pitch pulse locations of voiced sounds. However, there is no 100% reliable automatic pitch extraction algorithm and pitch mark verification is necessary when accurate pitch information is required. Where possible, it is desirable to record electroglottograph (also known as laryngograph) signals at the same time as the speech samples. The electroglottograph records electrical activity in the glottis during speech. These
electrical signals can then be processed to obtain pitch pulse locations in an easier and more precise way.

Due to the lack of laryngograph signals in the speech corpus, the ESPS *epochs* algorithm [10] was used to extract the locations of the pitch pulses during voiced speech segments. The DC component of the signal under analysis was removed before applying the epoch algorithm to prevent impairing voiced/unvoiced decisions. Analysis of the extracted pitch marks demonstrated that on average the algorithm provided relatively accurate pitch pulse locations for normal utterances. Regarding tracheoesophageal speech, the epoch algorithm was more prone to errors. In some cases, incorrect pitch marks which led to very short pitch periods were detected. Such errors caused distortion effects when the signals were modified and resynthesised. Therefore, a pitch mark filtering algorithm had to be implemented to solve the distortion problem. However, undetected pitch mark errors might still exist causing artefacts in the converted speech utterances.

Nevertheless, the ESPS epoch pitch extraction algorithm was used in most of the experiments. Electroglottograph signal recording or manual pitch marking are proposed as alternatives to improve the pitch mark extraction module of the implemented system in the future.

### 2.2.2 ANALYSIS/SYNTHESIS

In order to map spectral envelopes and excitation signals between normal and tracheoesophageal speech, a sinusoidal model capable of providing a framework for modification of the speech spectrum was required. The Quasiharmonic Analysis-by-Synthesis/Overlap-and-Add (ABS/OLA) Sinusoidal Model by [9] has been used in a pitch synchronous basis to implement the analysis and synthesis modules. A software implementation of the ABS/OLA model has been provided, while some modifications were made for its adaptation to this project.

First of all, the signal under analysis is normalised to account for the differences in energy between utterances and segmented into pitch-synchronous frames. Then, the sinusoidal parameters describing each frame are obtained using the ABS/OLA model. Finally, spectral envelopes and excitation spectrums are also calculated.
2.2.2.1 NORMALISATION

Due to the appreciable difference in energy between the speech produced by some tracheoesophageal speakers and the normal speaker, an initial normalisation step has been necessary to adjust the energy of the utterances that take part in the mapping process.

2.2.2.2 FRAME SIZE

For voiced speech, the signal is divided into frames that are one pitch period long, derived from the pitch marks generated by the epoch algorithm. For unvoiced sections of speech, a fixed frame size equal to the sampling frequency divided by 100 is used to decide frame boundaries. Voiced/unvoiced decision is also derived from the pitch mark information, since the epoch algorithm only generates pitch marks for voiced segments of speech. Hence, regions in which the distance between two consecutive pitch marks is longer than twice the default unvoiced frame size will certainly belong to an unvoiced segment of speech. These regions are thus divided into fixed size frames and labelled as unvoiced.

2.2.2.3 ABS/OLA SINUSOIDAL MODEL

In the ABS/OLA model, the speech signal \( s(n) \) is represented by a summation of overlapping frames

\[
s(n) = \sigma(n) \cdot \sum_{k=\infty}^{\infty} w_s(n - k \cdot N_s) \cdot \tilde{s}_k(n - k \cdot N_s)
\]

where \( \sigma(n) \) is a modulating envelope sequence that controls the time-varying intensity of \( s(n) \), \( w_s(n) \) is a complementary synthesis window, \( N_s \) is the length of the synthesis frame and \( \tilde{s}_k(n) \) is the \( k^{th} \) synthetic contribution or analysis frame given by

\[
\tilde{s}_k(n) = \sum_{l=0}^{L_k} A^k_l \cdot \cos(w^k_l \cdot n + \phi^k_l)
\]

where \( L_k \) is the number of sinusoidal components and \( A^k_l \), \( w^k_l \) and \( \phi^k_l \) are the sinusoidal amplitudes, frequencies and phases respectively. The envelope sequence \( \sigma(n) \) can be reasonably estimated by lowpass filtering \( s(n) \). Regarding the sinusoidal parameters, an iterative analysis-by-synthesis procedure is used for their calculation. An ensemble search is carried out over a set of uniformly spaced candidate frequencies, to obtain the optimum values of the sinusoidal amplitudes, frequencies and phases.
In practice, the synthesis window is normally a symmetric window of length $2 \cdot N_s + 1$ and thus, a synthesis frame of $N_s$ samples can be expressed as

$$s(n + k \cdot N_s) = \delta(n + k \cdot N_s) \cdot [w_i(n) \cdot \tilde{s}_k(n) + w_i(n - N_s) \cdot \tilde{s}_{k+1}(n - N_s)]$$

Therefore, each synthesis frame is in practice generated by a summation of two successive overlapping analysis frames.

For the purpose of speech modification, the authors refine a quasiharmonic ABS/OLA model in which each analysis frame also reflects pitch information. $\tilde{s}_k(n)$ can be rewritten in quasiharmonic form as

$$\tilde{s}_k(n) = \sum_{i=0}^{L_k} A^k_i \cdot \cos \left( \left( l \cdot w^k_0 + \Delta^k_i \right) \cdot n + \phi^k_i \right)$$

where $w^k_0$ is the fundamental frequency of the $k$th analysis frame and $\Delta^k_i$ is its differential frequency. This refined Quasiharmonic ABS/OLA model is the one adopted for speech modification purposes in this project. Such a formulation requires the calculation of the fundamental frequency $w^k_0$ associated with every analysis frame, apart from the amplitude, frequency and phase estimation. The use of pitch synchronous analysis frames derived from the epoch pitch marks as described in section 2.2.1 allows a straightforward calculation of each frame’s fundamental frequency $w^k_0$. In fact, the fundamental frequency of an analysis frame of size $N_k$ can be obtained as

$$w^k_0 = \frac{2 \cdot \pi}{N_k}$$

and the number of harmonics $L_k$ as the integer of $\frac{\pi}{w^k_0}$.

### 2.2.2.4 SPECTRAL ENVELOPE EXTRACTION

Linear prediction can be used to estimate the parameters associated to the speech model in Fig. 8, which corresponds to a simplification of the true speech production system described in section 2.1.4.

![Simplified speech production model](image)
It can be shown [11] that the linear prediction coefficients obtained for a stationary frame of speech can be used to model the all-pole filter representing the shape of the vocal tract. The frequency response of the derived linear prediction filter corresponds to a smoothed approximation of the speech spectrum referred to as the spectral envelope.

For conversion purposes, a linear prediction analysis of each frame is necessary to extract spectral envelopes. These will then be transformed in the envelope mapping experiment. Therefore, the LP analysis module performs an autocorrelation linear prediction analysis of each frame. The resulting coefficients are used to obtain a spectral envelope of every analysed frame. These are then stored in a matrix to be further modified by the mapping module.

2.2.2.5 EXCITATION SPECTRUM EXTRACTION

In addition to spectral envelopes, excitation spectrums are also required for the mapping experiments. The aim of the excitation spectrum extraction module is thus to extract the excitation spectrum for each frame. The various excitation estimation algorithms used in the different experiments are discussed in further sections.

2.2.3 ALIGNMENT

Alignment of normal and tracheoesophageal utterances is achieved in two steps: phone alignment and frame alignment.

Initially, the HTK HCopy and HVite commands [12] are run in force alignment mode on both normal and tracheoesophageal utterances to obtain their phone aligned transcriptions. A large dictionary of English is used for the phonetic transcriptions of the words. The label files acquired this way contain the start and end time in $100\text{ns}$ units of each of the recognised phone units. Thus, the phone boundaries need to be converted to number of samples in order to identify the set of frames that correspond to each phone. The different pronunciation of speakers results in mismatches between phone transcriptions. To deal with these mismatches which result in insertions, deletions and substitutions, normal and tracheoesophageal transcriptions are compared with HResults. This tool uses a dynamic programming-based string alignment procedure to output a time-aligned transcription of the previously obtained phone alignment files. Such a transcription will then be used to align phones between utterances.

Once phones from normal and tracheoesophageal sentences have been aligned, frames within each phone need to be aligned. In general, the rate of articulation varies
among speakers and hence, the number of pitch periods corresponding to each phone will differ too. A simple frame alignment algorithm has been used to address this problem. Assuming that characteristics of articulation hardly change within each phone, the first and last frames per phone are aligned correspondingly to each other and the alignment of the frames in between varies depending on the difference in the total number of frames per phone. This can be achieved applying the following expression:

\[
n = \frac{(s - 1) \cdot sF}{nF} + 1
\]

where \( n \) is the index of the normal speech frame which corresponds to the tracheoeosophageal speech frame \( s \), and \( nF \) and \( sF \) are the number of frames in the normal and tracheoesophageal aligned phones respectively. An example of the frame alignment obtained this way for the phone g-r+iy is shown in Fig. 9.

![Graph showing alignment of normal and laryngectomee voiced frames within phone g-r+iy](image)

**Fig. 9 Alignment of normal and laryngectomee voiced frames within phone g-r+iy**

### 2.2.4 MAPPING

Speech modification is carried out in the mapping module. Two mapping functions were designed to achieve conversion in the proposed experiments. Envelope mapping is used in the first experiment to graft the tracheoesophageal vocal tract spectrum onto the normal utterance in order to check the level of improvement the voice conversion approach can obtain. Then, tracheoesophageal excitation is substituted in successive experiments with that of the normal speech using the excitation mapping function. Both mappings are applied in the frame alignment basis derived from the alignment step.
LP and glottal estimation analysis of an utterance extract the spectral envelope and excitation spectrum of each of its frames. Since frames are already aligned, the mapping process should only substitute envelope and excitation spectrums between utterances. However, this is not straightforward. The number of sinusoidal components per frame derived from the quasiharmonic ABS/OLA analysis step varies as a function of the length of the frame, which in the adopted pitch synchronous approach corresponds to its pitch period. As a result, the number and location of the harmonic frequencies used in the computation of the spectral envelope and excitation spectrums is different for each frame. Thus, an intermediate resampling and interpolation step is necessary for their mapping.

Resampling of the spectral envelope is achieved using the new set of harmonics corresponding to the frame whose envelope needs to be substituted as an input to the MATLAB `freqz` function. This configuration returns the frequency response of the all pole filter described by the LP coefficients at the designated harmonic frequencies.

For the interpolation of the excitation spectrum, the complex phasor form interpolation proposed by [9] is used. An implementation of this interpolation algorithm has also been provided. Phasor interpolation yields the excitation spectrum magnitudes and phases that correspond to the harmonics of the frame which needs to be converted.
3 EXCITATION IN NORMAL AND LARYNGECTOMEE SPEECH

The main difference between normal and laryngectomee speech production mechanisms lays in the excitation. This chapter describes the algorithm used to extract glottal excitations and compares the ones obtained for normal and laryngectomee speech.

3.1 GLOTTAL WAVEFORM ESTIMATION

The estimation of the glottal waveform is one of the most complex processes in speech analysis due to the difficulty of separating glottal and vocal tract characteristics in the speech waveform.

From the speech production model presented in section 2.1.4, the voiced speech signal can be thought of as being generated as follows. First, a set of pitch pulses \( e(n) \) excite the glottal shaping filter \( G(z) \), which results in the glottal excitation signal \( u(n) \). This signal is then used to excite the supraglottal system formed by the vocal tract transfer function \( V(z) \) and the output radiation component \( R(z) \) to produce the voiced speech signal \( s(n) \). The z-domain transfer function for the voiced speech signal can thus be written as

\[
S(z) = G(z) \cdot V(z) \cdot R(z)
\]

And the z-domain glottal excitation can be derived using

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

In order to obtain an approximation of the excitation waveform \( U(z) \), inverse filtering can be used to invert the influence of the vocal tract and radiation components. The pressure or volume velocity relation at the lips can be approximated by a differentiator modelled in discrete time with a first order zero \( R(z) = 1 - z_0 \cdot z^{-1} \), 
\(0.95 \leq z_0 \leq 1\). As shown in Fig. 10, its radiation effect is typically included in the source function. Thus, the source to the vocal tract becomes the derivative of the glottal flow volume velocity. Therefore, the main component of the inverse filter is the vocal tract component \( V(z) \) and the fundamental problem in the estimation of the glottal volume velocity waveform is to determine the parameters of the inverse filter \( \frac{1}{V(z)} \). An accurate estimation of the vocal tract spectrum is crucial to remove its effect as
completely as possible from the speech waveform in order to obtain an accurate
estimate of the glottal flow. Closed phase linear prediction analysis is generally used to
obtain accurate estimates of vocal tract formant locations and bandwidths. It is based on
the assumption that intervals in the signal can be located which correspond to periods in
which the glottis is closed and in which, therefore, there is no excitation to the vocal
tract. Once these intervals are identified, they are employed to compute estimates of the
vocal tract filter only modelling the vocal tract characteristics. These can in turn be used
to inverse filter the speech back to an accurate glottal waveform.

So, all glottal waveform estimation closed phase inverse filtering (CPIF) methods
involve three basic steps. First, the glottal closure instants in which to calculate the
closed phase inverse filter need to be determined. Then the closed phase inverse filter
needs to be computed. Finally, the speech waveform has to be inverse filtered and post-
processed to obtain an estimate of the glottal waveform. Among these steps, finding
glottal closure instants for estimating the vocal tract is the most difficult task. Different
glottal waveform estimation methods differ in how they determine such closed phase
intervals. Two basic approaches have been used in the literature to determine the
instants of glottal closure: single-channel and two-channel methods.

Single-channel approaches only use the speech signal in the analysis. Identifying
glottal closure instants directly from the speech waveform is complicated due to
numerous factors such as gender or speaking style. Female speech tends to have higher
pitch than male speech, requiring a faster movement of the glottis, which does not
always yield a complete closure even in normal speech. Stress and vocal disorders can
also affect the identification of glottal closure instants. Various studies have
investigated the identification of glottal closure instants using different techniques such
as dynamic programming [13], glottal input power [14], formant stability [15] or
residual energy [16]. Though a number of favourable experimental results have been
published in the papers cited above, large scale testing results of these single-channel
methods have not been reported.
Two-channel approaches supplement the speech signal by an electroglottographic trace as an indicator of the closed phase region, simplifying and increasing the efficiency of the glottal closure instant determination. However, in order to use these methods it is necessary to collect data from electroglottograph sensors in addition to the acoustic data, which is not possible in many applications.

Due to the lack of electroglottograph signals in the speech corpus, only single-channel methods are applicable for this project. In addition, the closed phase inverse filtering algorithm used to estimate the glottal waveforms should be robust enough as to provide reliable laryngectomee glottal waveform estimates. Section 3.1.1 describes the algorithm adopted in the project for the estimation of normal and laryngectomee glottal waveforms, which doesn’t require precise glottal closure information and has been shown to produce reliable estimates of the glottal waveform over a wide range of speaking styles.

3.1.1 GLOTTAL WAVEFORM ESTIMATION ALGORITHM

A block diagram of the algorithm proposed by [17] used to estimate the glottal waveforms required for the excitation mapping experiment is shown in Fig. 11.

![Flow diagram of the implemented glottal waveform estimation algorithm](image)

Initially, the input signal, which corresponds to four or five pitch periods of speech, is analysed with a pitch-synchronous autocorrelation linear prediction (LP) analysis of order P. The LP coefficients derived from this analysis are then used to
inverse filter the input and obtain a residual signal. The locations of the most negative peaks in the residual signal are used as initial estimates of the glottal closure instants.

The identified negative peak positions are used as midpoints of disjoint intervals corresponding to the locations of disjoint sliding windows used to iteratively compute average LP and glottal derivative estimates. The starting points (c) of the iterative procedure are determined by subtracting the model order P from the locations of the negative peaks. LP coefficients are then estimated from sliding windows of length 2P located at the points specified by c, and updated by one sample at each iteration. Due to the small window size (ie. 20 samples for a LP analysis of order 10), the covariance method is used to compute the LP analysis of each disjoint window. As a result, stability of the poles has to be verified and if unstable poles are detected, they need to be reflected. Then, LP coefficients from each disjoint window are averaged together to obtain a smoothed vocal tract estimate for every iteration. Both the averaged LP coefficients and the glottal derivative for each iteration, obtained inverse filtering the input signal with such averaged coefficients, are stored in matrices A and G respectively.

The main difference among the estimates in G is that some produce noise-like glottal derivatives while others provide relatively smooth waveforms. The next step of the algorithm thus involves implementing a method for choosing the best estimate(s). A first order autocorrelation LP analysis of each glottal derivative estimate in G is performed, and the coefficients $a_i$ for every glottal estimate are stored in a vector. The first term of the LP analysis $a_1$ represents the ratio of the autocorrelation at lag 1 to the autocorrelation at lag 0, indicating how well two consecutive samples are correlated to one another. Thus, smoother estimates will exhibit values of $a_1$ closer to 1 than more noisy estimates. The indices (e) of the top 99th percentile (estimates with values of $a_i$ closest to one and greater than 99% of the rest) are used to average the best estimates of the LP coefficients $A[e]$ and glottal derivative waveform $G[e]$ to obtain final A and G estimates (Aestimate and Gestimate respectively). Finally, the glottal waveform is obtained by integrating the resulting glottal derivative estimate.

Fig. 12 shows glottal waveform estimates obtained with this algorithm for three different normal speakers.
Fig. 12 Glottal excitation waveforms obtained for different normal male (a) and female (b) speakers

3.2 COMPARISON OF NORMAL AND LARYNGECTOMEE GLOTTAL WAVEFORMS

As described in section 2.1.1, there are two basic source signals which can excite the vocal tract deriving voiced or unvoiced sounds. Voiced excitation is produced by the vibration of the vocal cords, which causes a repeated opening and closing of the glottis, allowing puffs of air excite the vocal tract. On the other hand, unvoiced excitation is produced by air passing through a constriction along the vocal tract. Laryngectomees can still constrict the vocal tract and produce unvoiced excitations the same way normal speakers do. However, their voiced excitation production mechanism differs from that of normal speakers. Their vocal cords have been removed after laryngectomy and thus they are not capable of producing a quasiperiodic excitation with the opening and closing of the glottis. Instead, they produce voiced excitations with the vibration of their new voice source referred to as neoglottis. Discovering the site or source of vibration in laryngectomee voice production has been a subject of investigation for several years. At present, it is known that the source of vibration, the neoglottis, is situated at the level of
the muscles of the upper esophageal sphincter and the pharyngoesophageal segment. The vibratory characteristics of the neoglottis have been found to be quite irregular [3], particularly when compared to the vibration of the glottis in normal speech production and thus, voiced excitation is thought to be the main cause of reduced quality and intelligibility of laryngectomee speech.

The differences in the vibration characteristics of the glottis and neoglottis in normal and laryngectomee speech are expected to be reflected in the shape of the extracted glottal waveforms, since they model the air flow volume velocity variations occurring at the speech source which have been found to be more irregular in laryngectomee speech.

Fig. 13 shows the glottal waveform estimates obtained with the implemented glottal estimation algorithm for four or five time-aligned pitch periods of the normal target and the eight different tracheoesophageal speakers in the speech corpus. Pitch marks were set manually to avoid automatic pitch pulse estimation errors, particularly for the case of tracheoesophageal speech. As expected, there are appreciable differences between normal and laryngectomee glottal excitations. Source waves produced by the normal speaker are regular and well-defined, while the ones produced by the different tracheoesophageal speakers have a more noisy and irregular shape.
Fig. 13 Glottal excitation waveforms obtained for the normal target speaker (a) and the different tracheoesophageal speakers from the data collection (b)
If glottal excitation spectrums are compared, as in Fig. 14, the difference can also be easily appreciated. The excitation spectrum of the normal speaker consists of a set of impulses located at the fundamental frequency and its harmonics. On the contrary, laryngectomee glottal excitation spectrums have no clear harmonic structure.
3.3 POTENTIAL FOR REPAIR

The differences between the extracted normal and laryngectomee glottal waveforms clearly show that there is a problem with laryngectomee excitation. The irregular vibrations of their new voice source make glottal pulses have a noisy and less defined shape, which causes laryngectomee speech be perceived as hoarse, rough and breathy. There are several mapping experiments which can be run to explore the enhancement that can be achieved using voice morphing techniques, given the implemented glottal estimation algorithm and voice conversion framework.

As an initial sanity check, the spectral envelope of the laryngectomee utterance can be mapped to the normal one. The converted speech is expected to have an improved quality, since its excitation corresponds to a normal speaker and there is no apparent problem with laryngectomees’ articulation. The transformed utterance will also give an idea of the highest improvement that can be achieved with the voice conversion approach.
Then, different excitation mapping experiments can be tested. First, the amount of improvement linear prediction residual mapping can achieve can be checked. Linear prediction residuals can be easily derived from linear prediction errors and do not require determination of glottal closure instants for their computation. The interest of this experiment thus lies in obtaining an idea of the amount of improvement a simple linear prediction modelling of the excitation residuals can achieve, before implementing more complex approaches.

Glottal excitation mapping experiments can next be run. An ideal voice conversion excitation mapping experiment should copy the glottal excitation estimate from each voiced normal frame to its corresponding time-aligned laryngectomee one. However, glottal waveform estimation methods only report positive experimental results for isolated short segments of speech, ie. the glottal waveform extraction algorithm implemented in this project obtains a source waveform estimate of an input speech segment comprising four or five pitch periods, and successful glottal waveform estimation algorithms for continuous speech have not been reported yet. As an alternative, two different experiments can be tried once the short segment performance of the glottal excitation mapping approach is checked: fixed excitation and phone dependent continuous excitation mapping.

The fixed excitation experiment involves substituting the excitation of every voiced frame in the laryngectomee utterance by an average glottal excitation extracted from a short segment of normal speech. This experiment assumes that a fixed normal excitation is better than the glottal excitation produced by the neoglottis. The converted speech might sound robotic, but the breathiness is expected to disappear.

The last experiment is an attempt to implement a continuous glottal excitation mapping. Instead of copying a fixed excitation to every voiced frame, the glottal waveform estimation algorithm is applied to four pitch period segments in every voiced phone and the closed phase LP estimate is extended to the rest of the phone. Different excitations are thus estimated for each phone. This approach should improve on the previous experiments and the converted speech is expected to sound less mechanical.
4 VOICE REPAIR EXPERIMENTS

This chapter presents the data collection and the tracheoesophageal speech enhancement experiments implemented in this project. Section 4.2 describes the initial envelope mapping experiment and its conversion results as a motivation for the rest of the experiments. The next subsection describes the different excitation mapping experiments which have been implemented.

4.1 DATA COLLECTION

Constructing a speech corpus appropriate for the task involved recording laryngectomee speech. Eight laryngectomized tracheoesophageal patients of the Speech and Language Therapy Department at Addenbrooke’s Hospital provided the required data. Among this group, seven were male and one was female. All laryngectomized speakers used tracheoesophageal speech as their primary mode of oral communication. Still, the perceptual quality and intelligibility of the recorded tracheoesophageal speech varied among speakers. Several speakers produced fluent tracheoesophageal speech whose perceptual lower quality is due to the whispy and belching sound resulting from their voicing problem. Other speakers, however, had difficulties producing fluent utterances and the perceptual quality of their speech was further decreased because of this prosody limitation. Each subject recorded thirty six utterances extracted from text passages containing phonetically rich sentences used in vocal rehabilitation exercises.

The parallel laryngeal speech data necessary for mapping purposes was provided by a man and a woman matched within a decade to the ages of the laryngectomized speakers who recorded the same thirty six phonetically rich sentences.

4.2 ENVELOPE MAPPING

This initial experiment was implemented in order to check the viability of the adopted hypothesis and to obtain a reference idea of the maximum improvement the voice conversion approach can achieve.

As described in section 1.2, the main anatomical difference in the speech production apparatus between a normal and a laryngectomee speaker is the lack of the larynx. However, the larynx is responsible for the voice source which excites the vocal tract to produce speech and hence, the hoarse quality and reduced intelligibility of the resulting laryngectomee speech can be assumed to be due to the lack of a proper
excitation signal. A voice conversion approach which transforms the laryngectomee voicing source signal while maintaining its spectral envelope is therefore expected to fix their excitation problem and to provide a converted speech whose quality and intelligibility will be closer to that of laryngeal speech.

Before implementing the excitation mapping approach, an equivalent envelope mapping experiment was run. Instead of converting the source excitation of the laryngectomee utterance, its spectral envelope was mapped to the normal utterance. The resulting speech signal thus maintained the pitch structure, prosody and excitation of the normal speech but contained the laryngectomee spectral envelope instead of the normal one. This experiment is not applicable to real time voice conversion systems, since it is the normal utterance which is being modified. Such an implementation would require predicting the utterances the laryngectomee speaker would produce and to have them recorded by the normal speaker for its transformation, which would be unfeasible. However, it can be implemented in a time alignment basis and can be very useful to analyse the highest improvement the excitation mapping experiment can obtain.

The frequency response of the linear prediction vocal tract filter was used to estimate spectral envelopes. Since the detailed excitation is mainly modelled by the linear prediction error signal, the linear prediction filter frequency response results in a smoothed approximation to the speech spectrum and represents the formant structure of the vocal tract. This spectral envelope estimation method is the approach adopted by most voice conversion systems since it constitutes a good approximation of the resonances in the vocal tract while it exploits the advantages of the linear prediction model of speech production.

The voice conversion framework described in section 4 was used as shown in Fig. 15 for the implementation of the envelope mapping experiment.
First, the DC component of the normal and tracheoesophageal utterances under conversion was removed and the ESPS epoch algorithm was used to extract the pitch period marks corresponding to the voiced segments of speech. The obtained pitch marks were then filtered to remove the incorrect ones which would cause distortions in the converted speech. Once unvoiced speech regions were segmented into fixed size frames, a quasiharmonic ABS/OLA sinusoidal analysis was carried out in a per frame basis. The aim of this experiment was to map spectral envelopes. Thus, the next step was to extract the spectral envelope for each frame. These were derived from a LP autocorrelation analysis of order 10, applied to each of the pitch-synchronous frames. In order to map spectral envelopes between utterances, these had to be time aligned. Not only phones were aligned, but also the frames within each phone as described in section 2.2.3. At this point, we already know which laryngectomee spectral envelope corresponds to which normal frame and thus, they just need to be mapped. Spectral envelope mapping involved resampling the laryngectomee spectral envelope so that the sampled values corresponded to the fundamental frequency and harmonics of the
aligned normal frame. The sinusoidal amplitudes of the normal utterance were then recalculated and finally, the transformed utterance was resynthesised.

Experiments were run mapping spectral envelopes from each of the seven male tracheoesophageal speakers to normal male speech utterances and from the female tracheoesophageal speaker to normal female speech. Informal listening of the converted utterances shows that in all cases, the quality and intelligibility of the transformed speech is perceptually higher. However, since the quality and intelligibility of the recorded tracheoesophageal utterances varies among speakers, the amount of perceptual improvement also varies among the transformed utterances. All the converted samples lack the characteristic breathy sound produced due to the tracheoesophageal voicing problem. In addition, the fact that the mapped sentences keep the prosody of the normal utterances increases the perceived quality and intelligibility of the tracheoesophageal samples with fluency limitations. This observation suggests that intermediate silence elimination and prosody conversion might also help enhance the quality and intelligibility of the speech produced by some tracheoesophageal speakers. Nevertheless, overall results obtained with this experiment reinforce the idea that an excitation mapping voice conversion approach is appropriate for the improvement of tracheoesophageal speech quality and intelligibility.

4.3 EXCITATION MAPPING

The aim of the initial envelope mapping experiment was to test the suitability of the proposed laryngectomee speech enhancement approach. The positive results obtained from the informal perceptual evaluation of the converted utterances confirmed the hypothesis that the cause of the reduced quality and intelligibility in laryngectomee speech is their lack of a normal source excitation. The fact that the converted utterance containing normal excitation and laryngectomee spectral envelope was perceived as an improved signal led to the development of these experiments.

The goal of the excitation mapping experiments is to transform the laryngectomee utterance by substituting its excitation signal with a normal glottal source. The converted utterance will thus maintain the prosody and spectral envelope of the laryngectomee speaker but contain an improved excitation. In this way, the resulting speech is expected to be perceived as produced by the laryngectomee speaker, since the main speaker identity segmental characteristics (spectral envelope, average pitch and rate of speech) are not modified. At the same time, the excitation substitution is
expected to obtain an improvement in the perceived quality and intelligibility of the transformed speech. As opposed to the previous experiment, the excitation mapping approach can be implemented in real time applications, i.e. a codebook based voice conversion system can be developed to map the excitation signal from the normal speech to the laryngectomee speech.

Three excitation mapping experiments have been implemented which differ in the way the excitation signal is calculated and/or mapped. First, linear prediction residuals are copied from normal to tracheoesophageal frame-aligned utterances, in order to determine the amount of improvement the simple linear prediction residual estimation method not requiring a complex closed phase analysis can achieve. Then, a glottal excitation derived from the application of the glottal waveform estimation algorithm to a normal segment of speech is used as a fixed excitation of every voiced frame in the laryngectomee speech. Finally, as an attempt to achieve a continuous glottal excitation mapping, a different glottal excitation is extracted with the glottal waveform estimation algorithm from each normal voiced phone and then mapped to the corresponding time-aligned tracheoesophageal phone.

4.3.1 RESIDUAL MAPPING

As described in section 2.2.2.4, linear prediction can be used to identify the parameters associated with the all-pole estimated model for the true speech production system. Linear prediction simply but accurately extracts the coefficients of the all-pole filter which models the vocal tract and its frequency response representing the smoothed formant structure, which is used as an estimation of the spectral envelope in voice conversion systems. The linear prediction error, also called residual, can be used to model the excitation signal of the simplified Source/Filter model of Fig. 8. In such a model, the residual is expected to be approximately white noise for unvoiced speech and an approximate impulse train for voiced speech. In practice, the unvoiced approximation is quite good and the replacement of the unvoiced residual by white noise followed by the LP filter typically results in no audible difference. For voiced speech, however, the residual is far from an impulse train. Furthermore, replacing the residual by an impulse train followed by the LP filter results in speech that sounds somewhat robotic. This is partly because real speech is not perfectly periodic and also because the all-pole assumption is not altogether valid. Nevertheless, the linear prediction residual is used as a reasonable approximation to the excitation signal in
state-of-the-art voice conversion systems which convert excitations in addition to spectral envelopes. These systems report an improvement of the quality of the morphed speech when converting excitation signals besides spectral envelopes [5]. The interest of mapping excitation residuals derived from a simple linear prediction analysis lies in the observation of the amount of improvement that can be achieved avoiding the complexity of finding glottal closure instants.

The glottal residual mapping experiment substituted the residuals of the voiced tracheoesophageal speech segments by the corresponding residuals of the normal speech voiced frames. The voice conversion framework described in section 4 and used for the envelope mapping experiment was adapted this time for the mapping of residuals as shown in Fig. 16.

![Flow diagram of the residual mapping experiment](image)

Excitation residual spectrums for each frame were calculated dividing the sinusoidal amplitudes derived from the ABS/OLA analysis by the frequency response of the all-pole vocal tract filter. The filter coefficients were obtained from a linear prediction autocorrelation analysis of order 10, just as in the envelope mapping
experiment. The mapping module this time had to copy the aligned excitation residual spectrums from normal to tracheoesophageal voiced frames. After updating the sinusoidal parameters, the transformed utterances were finally resynthesised.

Informal listening of the converted sentences reported no perceptual improvement. This fact revealed that residuals are not capable of modelling the differences in the glottal waveforms between normal and laryngectomee speech which is required in order to obtain the expected quality and intelligibility enhancement. Linear prediction residuals are successfully used for pitch extraction, voiced/unvoiced decision, synthesis and coding purposes. However, a better glottal excitation estimate needs to be extracted for the laryngectomee speech enhancement purposes of this project.

4.3.2 GLOTTAL EXCITATION MAPPING

As opposed to residual excitations, glottal excitations obtained with the implemented glottal estimation algorithm have been shown to model the differences between normal and laryngectomee voice sources. Glottal excitation mapping is thus the adequate approach towards laryngectomee speech enhancement. However, the difficulties derived from the complexity of closed phase determination must be sorted out in order to implement a continuous speech glottal excitation mapping.

4.3.2.1 SHORT SEGMENT MAPPING

The difficulty in the extraction of glottal waveforms has limited the design, testing, evaluation and application of glottal waveform estimation algorithms to short stationary segments of speech, i.e. the authors of [17] apply the implemented glottal extraction algorithm to input signals of four or five pitch periods long. However, for voice conversion purposes a continuous processing of the utterances under analysis is required.

Before trying alternatives to transform laryngectomee excitations in continuous speech, the adequacy of the glottal excitation approach was tested on short segments of speech comprising four or five pitch-synchronous frames as shown in Fig. 17.
Time-aligned normal and tracheoesophageal four or five pitch period speech segments were extracted from the speech corpus. Pitch marks were set manually to avoid automatic pitch pulse extraction errors, particularly in alaryngeal speech. The ABS/OLA analysis, alignment, mapping and synthesis modules from the voice conversion framework were used to obtain sinusoidal parameters for modification purposes, frame align normal and tracheoesophageal utterances, substitute laryngectomee glottal excitation spectrums by their corresponding normal ones and resynthesise the transformed speech. In addition, the glottal waveform estimation algorithm was used to obtain glottal excitation spectrums for each frame.

Fig. 18 shows the difference between normal speech, original tracheoesophageal speech and the converted speech segments for the same phone and eight different speakers from the data collection. The main visually appreciable difference between normal and tracheoesophageal speech is the lack of periodicity. Quasiperiodic pitch periods can be easily identified in normal speech. On the contrary, periodicity in tracheoesophageal speech is not as clear. Though pitch period marks can be somewhat
identified in most cases, the shape of the signal in between does not have the periodic shape present in normal speech waveforms. The converted signal, however, presents a structure more similar to normal speech. The substitution of the irregular tracheoesophageal glottal excitation with a normal one makes pitch periods more easily identifiable and the waveform presents a more periodic shape.
FIXED EXCITATION

A straightforward alternative to the impossibility of mapping glottal excitations between time- and frame-aligned normal and laryngectomee utterances is the use of a fixed excitation.

In this experiment, the excitation of every voiced tracheoesophageal frame is substituted with a fixed glottal excitation, extracted from a short segment of normal speech. Fig. 19 shows the glottal waveforms obtained for segments of normal speech corresponding different vowels. As it can be noted, the shape of the glottal waveform varies among the different phones.
The use of a fixed excitation obtained from a particular normal phone as the excitation of all voiced frames in tracheoesophageal speech assumes that a fixed normal excitation will still be better than any tracheoesophageal excitation depending on the type of phoneme. However, due to the differences observed in the glottal excitations of the various phones, the quality of the converted speech is expected to improve if the fixed excitation is further substituted by a phone or pitch-synchronous frame varying excitation.

The flow diagram used for the implementation of this experiment is shown in Fig. 20.
The glottal estimation algorithm was applied to a four pitch period segment of normal speech from which an averaged fixed excitation spectrum was derived. Laryngectomee utterances were analysed with the ABS/OLA model and spectral envelopes for each voiced frame were obtained, using the ABS/OLA and spectral envelope estimation modules from the voice conversion framework. An alignment between utterances was not necessary for this experiment, due to the use of the same excitation for every voiced frame. Then, the fixed excitation spectrum was resampled for each frame, according to its fundamental frequency, and stored as the new excitation spectrum of each particular voiced pitch period. Sinusoidal amplitudes were recalculated in order to account for the change in the excitation spectrum and the converted signal was finally resynthesised. Experiments with automatically and manually extracted pitch marks revealed that in many cases the epoch algorithm failed to mark segments which were supposed to be voiced, reducing as a consequence the number of converted frames and the overall conversion performance. Therefore, manual
pitch extraction of the tracheoesophageal utterance is necessary before its transformation.

As expected, converted utterances obtained with this experiment sound somewhat robotic due to the lack of variation in the excitation of voiced segments. However, artefacts make it difficult to judge if the quality and intelligibility of the resulting waveform is perceptually higher than that of the original. These artefacts are probably caused by differences in the energy between the fixed excitation and the tracheoesophageal excitations and should be fixed before perceptually evaluating results obtained with this enhancement approach.

4.3.2.3 GLOTTAL EXCITATION MAPPING

The next step towards a continuous glottal excitation mapping is to apply the glottal waveform algorithm to four or five frames in each phone and extend the obtained glottal estimate to the rest of the frames in that phone. In this way, the variations observed on the glottal excitation waveforms obtained for different phonemes will be taken into account and the resulting converted speech is expected to sound more natural.

This approach was implemented as shown in Fig. 21.
Normal and tracheoesophageal utterances were analysed and aligned using the ABS/OLA analysis, spectral envelope estimation and alignment modules from the voice conversion framework. In addition, glottal excitation spectrums of the four central pitch periods in every voiced phone were extracted using the glottal waveform estimation algorithm. Based on the assumption that the vocal tract spectral envelope does not practically change within each phone, the average closed phase linear prediction coefficients (\(A_{\text{estimate}}\)) provided by the glottal extraction algorithm were used to extract the glottal excitation spectrums of the rest of the voiced frames in the phone. Then, excitation spectrums of each laryngectomee voiced frame were substituted by the ones obtained for the corresponding aligned normal frames after resampling when necessary. Once the sinusoidal amplitudes were recalculated to account for the excitation spectrum modification, the converted utterance was resynthesised. Manual
pitch marks were also found to give a better overall conversion performance and were thus used in the experiment.

The implementation of the glottal estimation algorithm in a phone basis was expected to obtain more natural utterances, since a varying excitation is being mapped rather than a fixed one. As expected, the transformed utterances sound less machine-like than the ones obtained in the previous experiment. However, they also have artefacts which make the evaluation of the improvement of the perceptual quality complicated. An analysis of the artefacts revealed that they are caused in phones which contain non-uniform frames or frames with big differences in amplitude. In these cases, the differences in the magnitude between the closed phase LP estimate and the spectrum derived from the ABS/OLA analysis produce excitation spectrums with high values of amplitude, which distort the converted waveform. These problems arise from the lack of a glottal waveform estimation algorithm which extracts accurate glottal excitations in a pitch-synchronous way.
5 CONCLUSIONS AND FURTHER WORK

This project explored the application of voice conversion techniques to the enhancement of laryngectomee speech. First, normal and laryngectomee glottal excitation waveforms were estimated. An analysis and comparison of the extracted excitations proved the hypothesis that the main cause of the degradation of the perceived quality in laryngectomee speech is their irregular voicing source. Then, several conversion experiments were implemented. An initial envelope mapping experiment checked the sanity of laryngectomee articulation, demonstrated that an effective enhancement can be achieved substituting laryngectomee excitation by a better glottal source and provided an idea of the highest improvement a voice conversion enhancement approach can achieve. A second experiment showed that simple excitation residual mapping doesn’t model accurately enough normal and laryngectomee source differences and that glottal excitation mapping is necessary to achieve the expected improvement. However, the difficulty of determining glottal closure instants limits the application of the glottal waveform estimation algorithm to short stationary segments of speech. Therefore, after checking the adequacy of the glottal excitation mapping approach on short speech segments, two more experiments towards a continuous glottal excitation mapping were implemented. The first one copied a fixed normal excitation to every voiced frame in the laryngectomee utterance. The second one extracted the normal excitation from each voiced phone in order to map it to the corresponding time-aligned laryngectomee one. Artefacts in the converted utterances obtained in these experiments point out the need of a glottal waveform estimation algorithm capable of extracting reliable closed phase spectral envelope and excitations in a per frame basis.

Several problems make laryngectomee voice conversion speech enhancement a difficult task. Laryngectomee pitch mark extraction is one of them. Automatic pitch extraction is an unsolved problem. However, several pitch extraction algorithms exist which extract relatively accurate pitch pulse locations for normal speech. The lack of a clear periodic structure in laryngectomee speech makes pitch extraction algorithms more prone to errors, causing artefacts in the converted speech. Parallel laryngograph signal recording or manual pitch mark extraction is suggested to solve this problem. However, these solutions are not applicable to real time voice conversion systems. Glottal waveform estimation is another problem. Closed phase linear prediction can be used to extract glottal excitations. Several methods have been proposed in the literature
to detect glottal closure instants. However, these algorithms have only been tested on short segments of vowel recordings. An ideal voice conversion laryngectomiee speech enhancement approach would require closed phase spectral envelope and glottal excitation estimates for each pitch-synchronous frame. More work can be done towards the development and implementation of a continuous glottal excitation mapping. The accuracy of the source waveforms obtained with different glottal estimation algorithms can be compared and their extension to continuous speech investigated. Regarding the additional fluency limitations some laryngectomees have, silence elimination, time restoration and prosody mapping are suggested as techniques to further improve the naturalness of their speech. The automatisation of the voice conversion system could also be explored, ie. a codebook based excitation conversion system to allow for the conversion of arbitrary laryngectomiee speech input could be implemented. All experiments in this thesis were run with tracheoesophageal speech samples. However, the limitation of esophageal speech is also a voicing problem and thus the same voice conversion approaches are applicable to esophageal speech enhancement and could be implemented as further extensions of this work.
6 REFERENCES


APPENDIX

A  RECORDED TEXT

The rainbow passage

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colours. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

The North Wind and the Sun

The North Wind and the Sun were arguing one day about which of them was the stronger, when a traveller came along, wrapped in a warm coat. They agreed that the one who could make the traveller take his coat off could be considered stronger than the other one. Then the North Wind blew as hard as he could, but the harder he blew, the tighter the traveller wrapped his coat around him, and at last the North Wind gave up trying. Then the Sun began to shine warmly, and right away the traveller took his coat off. And so the North Wind had to admit that the Sun was stronger than he was.

The story of Arthur the Rat

There was once a young rat named Arthur, who would never take the trouble to make up his mind. Whenever his friends asked him if he would like to go out with them, he would only answer, “I don’t know.” He wouldn’t say “Yes” and he wouldn’t say “No” either. He could never learn to make a choice. His aunt Helen said to him, “No-one will ever care for you if you carry on like this. You have no more mind than a blade of grass.” Arthur looked wise, but said nothing.

One rainy day the rats heard a great noise in the loft where they lived. The pine rafters were all rotten, and at last one of the joists had given way and fallen to the ground. The walls shook and the rats’ hair stood on end with fear and horror. “This won’t do,” said the old rat who was chief. “I’ll send out scouts to search for a new home.” Three hours later the seven scouts came back and said, “We’ve found a stone house which is just what we wanted. There’s room and good food for us all. There’s a kindly horse named Nelly, a cow, a calf and a garden with an Elm tree.” Just then the
old rat caught sight of young Arthur. “Are you coming with us?” he asked. “I don’t
know,” Arthur sighed, “the roof may not come down just yet.” “Well,” said the old rat
angrily, “we can’t wait all day for you to make up your mind. Right about face! March!”
And they went off.

Arthur stood and watched the other rats hurry away. The idea of an immediate
decision was too much for him. “I’ll go back to my hole for a bit,” he said to himself,
“just to make up my mind.”

That night there was a great crash that shook the earth, and down came the whole
roof. Next day some men rode up and looked at the ruins. One of them moved a board,
and under it they saw a young rat lying on his side, quite dead, half in and half out of his
hole.