Faculdade de Engenharia da Universidade do Porto

Artificial Voicing of Whispered Speech

Patrícia Cristina Ramalho de Oliveira

For Jury Evaluation

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Prof. Dr. Aníbal João de Sousa Ferreira

Co-Supervisor: Dr. Ricardo Jorge Teixeira de Sousa

July 30, 2015

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES **2014/2015**

A Dissertação intitulada

"Artificial Voicing of Whispered Speech"

foi aprovada em provas realizadas em 23-07-2015

o júri

Eurico Carrapatoso

Presidente Professor Doutor Eurico Manuel Elias Morais Carrapatoso
Professora Auxiliar do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professor Doutor Salviano Filipe Silva Pinto Soares
Professor Auxiliar do Departamento de Engenharias da Escola de Ciências e Tecnologias da Universidade de Trás-os-Montes e Alto Douro

Aníbal João de Sousa Ferreira

Professor Doutor Aníbal João de Sousa Ferreira
Professor Associado do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Patrícia Cristina Ramalho De Oliveira

Autor - Patrícia Cristina Ramalho de Oliveira

Faculdade de Engenharia da Universidade do Porto

Resumo

O sussurro é uma importante forma de comunicação que é vulgarmente utilizada em alternativa à denominada fala vozeada. Infelizmente, em algumas situações não se trata de uma alternativa, mas sim a única forma de dialogar. Tal é o caso, por exemplo, de pacientes laringectomizados, com paralisia bilateral das pregas vocais ou disfonia espasmódica. Consequentemente, as diferenças entre as características dos dois tipos de fala, tanto a nível fisiológico como a nível das propriedades acústicas associadas, têm despertado interesse na comunidade científica, tendo por isso sido exploradas ao longo dos últimos anos. No entanto, a conversão de fala sussurrada para fala vozeada ainda não foi inteiramente conseguida, já que as soluções propostas sofrem ainda de algumas debilidades, como incapacidade de operação em tempo real ou falta de naturalidade da fala sintetizada.

Assim, esta proposta de dissertação tem como objetivo o estudo e desenvolvimento de um algoritmo capaz de realizar tal conversão em tempo real, através da criação ou implante de vozeamento em regiões selecionadas da fala sussurrada. Para isto, será necessário também identificar de forma automática, na fala não-vozeada, as regiões candidatas a vozeamento. Adicionalmente, deverá ser assegurada uma sonoridade natural da voz sintetizada e incorporando, o mais possível, elementos da assinatura vocal específica do orador.

Neste documento é apresentado primeiramente o estado da arte, o qual contém a descrição de conceitos base, assim como soluções desenvolvidas no âmbito do tema proposto. Seguidamente, é proposto um sistema para a conversão da fala sussurrada em fala vozeada, assim como a descrição de todo o seu desenvolvimento. Isto é dividido em duas partes: algoritmo para a identificação automática das regiões candidatas a vozeamento na fala não-vozeada e algoritmo para implante de vozeamento artificial. Todas as operações e análises efetuadas durante o desenvolvimento das soluções, tiveram em conta amostras dos dois modos de fala de uma base de dados anotada de um paciente com disfonia espasmódica. Por fim, são apresentados os resultados obtidos da conversão automática da fala não-vozeada em fala artificialmente vozeada. Além disso, são realizados testes subjetivos com o objetivo de avaliar o impacto percetivo da técnica de vozeamento artificial adotada, em três parâmetros - inteligibilidade, naturalidade e identidade. Conclui-se tecendo considerações sobre futuros refinamentos do trabalho exposto.

Abstract

Whispering is an important form of communication commonly used as an alternative to the so-called voiced speech. Unfortunately, sometimes it is not an alternative, but rather the only means of conveying information. Such is the case, for instance, of laryngectomised patients, patients suffering from bilateral paralysis of the glottal folds or patients suffering from spasmodic dysphonia. Consequently, the differences between the characteristics of the two speech modes, both at physiological and acoustic levels, have drawn attention among the scientific community and thus, have been researched over the years. However, an effective whisper-to-speech conversion has not yet been fully accomplished, since the proposed solutions have some shortcomings, such as their inappropriate nature for real-time operation or the lack of naturalness in the synthesized speech.

This dissertation proposal aims for the research and implementation of an algorithm for real-time whisper-to-speech conversion, by implanting artificial voicing on selected regions of the whispered speech. To that end, it is necessary to perform an automatic identification of the regions that are candidates to artificial voicing. Additionally, a natural sounding synthesized voice should be ensured as well as the incorporation of as much vocal signature elements as possible.

This document starts by presenting the fundamental concepts as well as proposed solutions for whisper-to-speech conversion so as to provide an insight on the state of the art in this field. Afterwards, a system for whisper-to-speech conversion is proposed, along with the description of all its implementation steps. This consists of two parts: the algorithm for the automatic identification of the candidate regions to artificial voicing in the whispered speech and the algorithm to implant artificial voicing. All the analysis performed during the development of the solutions were based on an annotated speech database containing samples provided by a patient suffering from spasmodic dysphonia. Finally, the results of the automatic whisper-to-speech conversion are presented. Additionally, subjective tests are conducted in order to assess the perceptual impact of the adopted artificial voicing technique by using three parameters - intelligibility, naturalness and speaker's identity. The document concludes with considerations on future refinements of the described work.

Agradecimentos

Primeiramente, gostaria de agradecer ao Professor Aníbal Ferreira pela oportunidade de investigação neste tema tão desafiante, pelo apoio prestado e conhecimento partilhado durante o desenvolvimento do trabalho. Igualmente agradeço ao Dr. Ricardo Sousa, por toda a disponibilidade e perseverança sempre demonstradas.

Pelo auxílio nas questões de fonética, essenciais para o sucesso desta dissertação, agradeço à Professora Susana Freitas. À professora Maria Adelina Guimarães pela verificação ortográfica e a todos aqueles que contribuíram para a elaboração de uma grande parte do Capítulo 6 um grande obrigado.

Não poderei deixar de agradecer aos meus grandes companheiros durante todo este percurso, pelo apoio moral prestado e por todos os momentos bem passados; às minhas "meninas" e todos aqueles, que tendo igual importância, não menciono.

Finalmente, aos meus pais e irmãos, pelo apoio incessante e incondicional ao longo de todos estes anos.

Patrícia Oliveira

“ *It’s the time that you spent on your rose*
that makes your rose so important.”

Antoine de Saint-Exupéry

Contents

List of Figures

List of Tables

Abbreviations

AR	Autoregressive
CELP	Code Excited Linear Prediction
EP	European Portuguese
GMM	Gaussian Mixture Model
HAS	Human Auditory System
HMM	Hidden Markov Model
HTS	HMM-based speech synthesis system
IPA	International Phonetic Alphabet
LPC	Linear Prediction Coding
LSP	Line Spectrum Pair
MELP	Mixed Excitation Linear Prediction
MFCCs	Mel-Frequency Cepstral Coefficients
MLSA	Mel Log Spectrum Approximation
NRD	Normalized Relative Delay
PLP	Perceptual Linear Prediction
PSC	Perceptual Spectral Cluster
VC	Voice Conversion

Chapter 1

Introduction

1.1 Context

Whispering is an important mechanism in human vocal communication, as it provides a softer form of speech. This ability is very useful in several situations as in the exchange of private information or in quiet places, in which the loudness of the voiced speech is not desirable. In fact, it is interesting to note that even some animals benefit from this powerful skill, howbeit the purpose of such behavior is not entirely understood.

Unfortunately, whispers are in some cases the only means of conveying information in a human dialogue. Aphonic individuals, such as people who have been submitted to partial or full laryngectomy, or individuals with spasmodic dysphonia are not able to produce natural sounding speech but are capable of producing whispered speech [1] without much effort. Therefore, the problem of returning normal speech to patients unable to produce phonated speech has been mitigated over the years. Currently, the existing solutions involve an invasive medical procedure. Additionally, these solutions have inherent inconveniences that comprise usage difficulties, risk of infection and an unnatural sounding output speech. The most common techniques used by post-laryngectomised patients are esophageal speech, tracheo-esophageal puncture (TEP) [9] and the electrolarynx (EL) [10]. The clear disadvantages of these methods prompted the research for best solutions, ideally real-time and noninvasive alternatives. Consequently, some methods emerged based on an analysis-by-synthesis approach. These methods, rely on signal processing techniques, which encompass spectral modifications, in an attempt to perform a real-time whisper-to-speech conversion by transposing the whisper features to normal sounding speech features. The algorithm would later be inserted in an external device and thus, no medical surgery would be required to return to these patients a voice similar to the healthy voice. However, these techniques either do not synthesize a natural speech or are not yet suitable for everyday usage, e.g not appropriated

[1] In fact the concept of whispered speech comprises two types of speech, soft whispers and stage whispers. The latter implies partial phonation and consequently some vibration of the vocal folds. Since the most part of laryngectomised persons are not able to produce any degree of phonation, soft whispers are the most similar to those produced by laryngectomy patients [8].

for real-time operation. Hence, the need to continue developing such solutions and deepen the existing techniques.

1.2 Motivation

As mentioned above, the current solutions that provide these type of patients with a better quality of life are not entirely satisfactory. Therefore, it is extremely important to continue searching for solutions capable of overcoming this limitation. The innovative nature and clear impact that an effective answer to such a problem would have, provide undoubtedly a strong impetus for the research of this subject.

Actually, there is an additional reason that motivates this study, namely a real case of a patient with spasmodic dysphonia. The patient has provided samples of his voice to perform the proof-of-concept of this dissertation. Therefore, a successful solution would have a direct impact on this person's life.

1.3 Objectives

In this dissertation a research for an analysis-by-synthesis approach for a real-time whisper-to-speech conversion is undertaken. It follows the work in [11] aiming to improve the results by the research of new methods. The purpose is to implement an algorithm that allows for a reconstruction of natural sounding speech from an in-depth analysis of the whispered speech so as to perform a proof-of-concept. The algorithm is expected to implant artificial voicing by the insertion of a periodic pattern in selected regions of the whispered speech, taking into consideration the characteristics of both speech modes. The spectral modifications should be performed ensuring a coherent synthesized speech, improving intelligibility first of all. Also, temporal smoothness should be applied to ensure the best possible naturalness. Finally, the implemented algorithm would be included in an external prosthesis, providing a noninvasive alternative.

In particular, a specific stage of the overall conversion approach must be accurately exploited, namely automatic segmentation strategies of whispered speech. The success of this segmentation is essential for the implementation of a real-time system.

Therefore, the main challenges of the dissertation proposal are the implementation of an algorithm for an automatic segmentation of the whispered speech, by identifying the candidate regions to artificial voicing; to perform a whisper-to-speech conversion with satisfactory levels of intelligibility and naturalness of the synthesized speech; and to reach real-time operation.

1.4 Structure

The document consists of six more chapters. In Ch. 2 we describe the state of the art, which comprises the discussion of some useful fundamental concepts, the description of some implemented solutions about the overall procedure of whisper-to-speech conversion and the presentation of a statistical parametric model for future use in the design of an automatic segmentation strategy of whispered speech. In Ch. 3, we describe the speech corpora that are used throughout the study, as well as some considerations on the samples characteristics. The proposed approach for automatic segmentation of whispered speech is described in detail in Ch. 4, while the proposed algorithm for whisper-to-speech conversion is presented in Ch. 5. Finally, in Ch. 6 the results of the overall system are exhibited and in Ch. 7 the document is concluded along with final considerations on future work.

Chapter 2

State of the Art

2.1 Overview

This chapter discusses some of the existing methods for conversion of continuous whispered speech to natural sounding speech. To that end, the first goal is to comprehend what whispered speech is and what makes it different from the normal speech. This comprises a study of how a voice sound is generated and shaped by the human vocal system, how can the underlying features be represented from a signal processing point of view and what sounds should be taken into consideration in the analysis of European Portuguese. Therefore, the second section of the chapter describes in detail these required concepts. The second stage encompasses the investigation of how to use those differences and relate them so that a conversion is possible. Thus, in the third section of this chapter, several approaches are discussed regarding the overall procedure of normal speech reconstruction, wherein methods of normal speech coding are used and some modifications are performed so as to adapt to whispered voice.

Finally, in the last section a particular part of the reconstruction process is investigated, the automatic segmentation of whispered speech. This part deserves particular attention, since a successful automatic segmentation would be essential for the implementation of a real-time system, i.e. the inclusion of an algorithm in an external assistive device.

2.2 Main Differences Between Voiced Speech and Whispered Speech

So as to better understand the systems performing whispered speech to voiced speech conversion, some fundamental concepts explaining the main differences between these two speech modes need to be introduced. In this section, specific topics are briefly explained such as characterization of speech production mechanisms, introduction to European Portuguese phonetics, representation and characterization of speech signals, as well as a final consideration about the main spectral differences between the two types of speech.

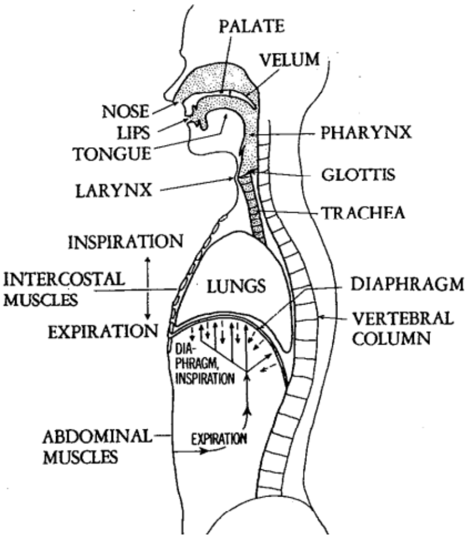


Figure 2.1: Speech Production System [1]

2.2.1 Speech Production

2.2.1.1 Human Vocal Apparatus

The human vocal apparatus is a very complex system responsible for the speech production. Thus, its study should be the starting point in our discussion as it allows to identify the main signal processing steps involved in the production of voiced and unvoiced speech, as well as to identify the characteristics and behavior of source signals and filter configurations in both cases.

The speech production system, in Figure 2.1, is divided into three parts [1]: subglottal system (system below the larynx), larynx and its surrounding components and supraglottal system (system above the larynx).

The subglottal system contains the main energy source, which is provided by the lungs in the form of an airflow. The subglottal pressure during the exhalation is responsible to provide the airstream necessary for speech production and can be controlled so as to adjust the speech volume, stress pattern and speech duration.

The larynx contains the vocal folds and the glottis. The latter is the space between the vocal folds and defines the dividing line between subglottal and supraglottal systems. The shape of the space between the vocal folds changes during phonation and breathing and is controlled by the arytenoid cartilages. During normal breathing, arytenoid cartilages are far apart, i.e. in abduction position, consequently so are the vocal folds (the glottis forming a V shape), thus the air passes freely.

While speaking, during the exhalation, the arytenoid cartilages become close to each other, closing the vocal folds (adduction position) and forming a constriction. Therefore, when the air from the lungs (subglotal pressure) passes through the closed vocal folds they vibrate according to a periodic pattern that modulates the airflow. The fundamental frequency of this periodic pattern

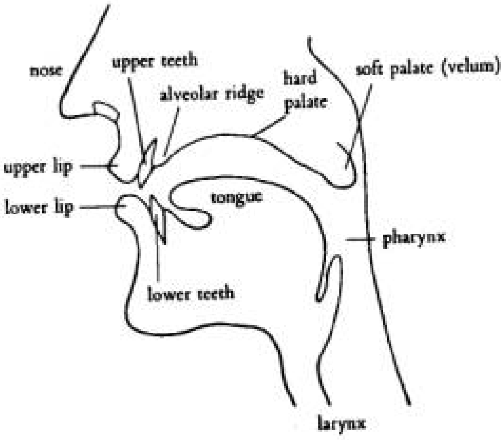


Figure 2.2: Articulators of the Vocal Tract. [2]

corresponds to the perceived pitch of the voice. This periodic pattern is further shaped in frequency by the influence of the supra-laryngeal structures as explained later on in this chapter.

The different types of voice are due to different kinds of vibration of the vocal folds. Phonated sounds are produced by the entire vibration of the vocal folds while unvoiced sounds are produced without any vibration, i.e. the vocal folds remain in the abduction position, resulting in a continuous airstream with no periodic excitation.

The supraglottal system acts as a modulator of the created sound, as it behaves as a variable ressonator, consisting of the pharynx and the vocal tract [4]. The latter can be divided into oral cavity and nasal cavity, which are responsible for the oral sounds and nasal sounds respectively. This is the system responsible for the timbre shaping and formation of the vowels and consonants through articulatory movements. Thus, the different parts of the vocal tract are called the articulators and consist of the active articulators, such as lips, tongue, velum (soft palate) and jaw and the passive articulators, such as teeth (upper and lower), alveolar ridge and hard palate. The different articulators can be seen in Figure 2.2. These articulators control the production of speech in such a way that different speech sounds are determined by the manner and place of articulation, as will be described in the next section.

2.2.1.2 Source-filter Model

The source-filter model describes the human vocal system as a combination of sound sources and filters, relating the articulation of the speech sounds with the features of acoustic signals [4] and relies on the simplifying assumption that source and the filter are independent, i.e. it ignores coupling effects.

Thus, there are two types of sound sources, a periodic and a non periodic one, representing the vocal folds vibration and noise, respectively. The filters represent the supraglottal system, namely the vocal tract, whose resonances represent the filter formants (this concept will be further explained later), and the radiation characteristics. The model is represented in Figure 2.3.

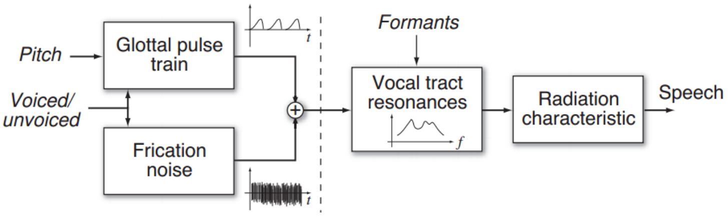


Figure 2.3: Source-Filter Model of Speech Production [3]

The common implementation of the source-filter model of speech production assumes that the sound source, or excitation signal, is modelled as a periodic impulse train, for voiced speech, or white noise for unvoiced speech. Moreover, the vocal tract filter is, in the simplest case, approximated by an all-pole filter, where the coefficients are obtained through the use of linear prediction minimizing the mean-squared error in the speech signal to be reproduced. This is mentioned in the literature as the basis of Linear Prediction Coding [12].

2.2.2 European Portuguese Phonetics

As this work is developed assuming the characteristics of the European Portuguese dialect, it is important to first identify results in this area of study, in order to take them into consideration in all future conclusions.

The first step is to find a common way to represent the different types of speech sounds, since the use of graphemes by itself are not enough (one grapheme may represent several phones [1] and a phone can be orthographically represented by several distinct graphemes). The IPA (International Phonetic Alphabet) is an alphabet of phonetic notation, which associates each existing pronounceable sound in all existing dialects to a single specific representation, through the use of letters and diacritics. In short, IPA establishes a univocal relation between the sound and the symbol. Each language only uses a subset of the IPA, thus the one regarding the European Portuguese is described in Appendix A.

As stated before, resonances in the vocal tract modify the sound waves according to the position and shape of the articulators, creating formant regions and thus different qualities of sound, leading to the distinction of vowels and consonants. Therefore, the classification of the vowels and consonants of the EP, according to the place of articulation and position of the articulator, is now discussed.

2.2.2.1 EP Vowels Articulatory Classification

The EP vowels are produced without significant constrictions to the airflow in the vocal tract, which explains that they are very resonant, and always involve vibration of the vocal folds. This means that, typically all vowels are voiced sounds. There are nine oral vowels and five nasalized

[1] A phone is a unit of speech sound, a speech segment that holds distinct physical or perceptual properties.

vowels[2] in EP [4], that are classified according to two parameters: height (high, medium, low) and backness (front, central, back) of the tongue and lips position. In addition, there are two semivowels, j and w (see Appendix A), which are different from the vowels, because of their lower energy. A semivowel is always followed by a vowel, thus giving rise to a diphthong.

2.2.2.2 EP Consonants Articulatory Classification

Unlike the vowels, the consonants are produced with significant constrictions to the airflow in the vocal tract, caused by the movement of the articulators. Thus, their sound can be affected by noise.

Consonants can be classified according to the place of articulation, manner of articulation and phonation (voiced or unvoiced), which together gives the consonant its distinctive sound.

The place of articulation is the point of contact where an obstruction occurs in the vocal tract depending on the location of an active articulator and a passive one. The possible classifications are [4]: Bilabial, Labio-Dental, Dental, Alveolar, Palatal, Velar and Uvular.

The manner of articulation describes the way the airflow is expelled depending on its perturbation during the passage in the vocal tract. It can be classified as [4]:

- Plosive/stop (oral or nasal) - results from a total constriction to the airflow because of the blocking of the vocal tract.

- Fricative - results from a partial constriction to the airflow by forcing air through a narrow space when two articulators are close together.

- Lateral - results from a central constriction to the airflow, forcing the air to proceed along the sides of the tongue.

- Trill - results from a partial constriction that causes tongue vibration.

The consonants classification is shown in table 2.1.

Finally, the speech sounds can be divided into obstruents and sonorants. The obstruents are produced with total or partial constriction to the airflow in the vocal tract (oral plosive consonants and fricative consonants), while the sonorants are produced without any constriction to the airflow (nasal plosive consonants, lateral and trill consonants, vowels and semivowels). Therefore, there are only six unvoiced consonants, three oral plosive and three fricative.

It is important to mention that the above classification of vowels and consonants according to their articulatory characteristics, was performed considering each sound produced individually. However, during speech, these sounds are produced sequentially, leading to a superposition of the individual segments and modifying the speech sound. This phenomenon known as coarticulation, should be taken into account.

[2] In the nasalized vowels, the airflow passes both through the oral and nasal cavities.

Place of Articulation	Manner	Plosive		Fricative	Lateral	Trill
		Oral	Nasal			
Bilabial	Voiced	b	m			
	Unvoiced	p				
Labio-dental	Voiced			v		
	Unvoiced			f		
Dental	Voiced	d		z		
	Unvoiced	t		s		
Alveolar	Voiced		n		l	ɾ
	Unvoiced					
Palatal	Voiced		ŋ	ʒ	ʎ	
	Unvoiced			ʃ		
Velar	Voiced	g				
	Unvoiced	k				
Uvular	Voiced					
	Unvoiced					R

Table 2.1: Articulatory classification of the EP consonants. Adapted from [4]

2.2.3 Acoustic Phonetics

2.2.3.1 Fundamental Concepts for Representation and Analysis of Speech Signals

Speech sound waves can be periodic, as in the case of vowels, continuous aperiodic, as fricative consonants or non-continuous aperiodic, such as in the explosion of the plosive consonants. Furthermore, speech signals are usually classified as voiced or unvoiced, but they can consist of the two types of sounds. According to this feature, they can be analyzed differently:

- Voiced sounds - They are characterized by a fundamental frequency (F0), which is the lowest frequency, and its harmonic components produced by the vocal folds. The vocal folds generate complex periodic sound waves with fundamental frequency values between 50 and 500Hz. Furthermore, the signal is characterized by its formants (poles) and sometimes antiformants (zeros) frequencies, caused by modification of the excitation signal by the vocal tract. Each formant frequency has also an amplitude and bandwidth that should be considered.

- Unvoiced sounds - There is no fundamental frequency in the excitation signal and thus no harmonics. Moreover, the excitation signal is non-periodic and resembles white noise. Some unvoiced sounds are characterized by stoppage of airflow followed by a sudden release in the vocal tract.

  In the special case of whispering a voiced sound, there is no fundamental frequency, as expected, and the first formant frequencies produced by the vocal tract are perceived. This particular case will be discussed later in more detail.

The fundamental frequency and formant frequencies are probably the most important concepts in speech synthesis, therefore we will address them in detail along with other important concepts that will be used throughout the dissertation.

Formants

Formants represent the acoustic resonances of the vocal tract. Each area of the vocal tract has its own resonance frequency, thus the amplitude and harmonics of the signal are modified during the passage through the supraglottal cavities, depending on the area. Formants are often measured as amplitude peaks in the envelope of the magnitude spectrum of the sound.

Cepstral Analysis

As for the determination of the fundamental frequency cepstral analysis is often used. It provides a method for separating the vocal tract information from excitation as it allows the conversion of signals obtained by convolution (such as source and filter) into sums of their cepstra, thus providing linear separation. Therefore, as the fundamental frequency represents the pitch (excitation signal) it is possible to determine it from the cepstrum.

As a particular case, the real cepstrum is obtained by first windowing and performing Discrete Fourier Transform (DFT) of the signal, determining the logarithm of the magnitude spectrum and finally taking the Inverse Discrete Fourier Transform (IDFT), i.e. transforming it back to the time-domain. Thus, the cepstrum we are considering in this dissertation is the real cepstrum although a more general complex cepstrum also exists.

Mel-Frequency Cepstral Coefficients (MFCCs)

In cepstral analysis, the spectrum is usually first transformed using the Mel Scale, resulting in a Mel-frequency cepstrum (MFC), whose coefficients are denominated Mel-Frequency Cepstral Coefficients (MFCCs). The MFC is a representation of the short-term spectrum of a sound, having its envelope representing the shape of the vocal tract.

Mel frequency scale represents the spectrum coefficients taking into account the natural frequency resolution of the human auditory system. Humans are more sensitive to small changes in pitch at low frequencies than they are at high frequencies, so the Mel filterbank uses this information to size and space its filters properly. The first filter is very narrow, but as the frequencies get higher the filters become wider because of the human lower sensitivity in those regions.

In short, to obtain the MFCCs, firstly the speech signal is split into small frames. This is done because speech is a non-stationary signal, thus the need arises to split the signal into short-term stationary segments. The samples of the speech signal presented in the frame are weighted by a Hamming window. Then, the FFT is applied to each frame to obtain the magnitude spectrum, and this is followed by the Mel scale transformation through a bank of triangular filters uniformly spaced in the mel scale. This scale is defined as [13]:

$$m = 2595\log_{10}(1+\frac{f}{700}) \tag{2.1}$$

where m is the mel frequency and f represents frequency in Hertz.

Finally, the logarithm is applied to the filter outputs (the resulting mel frequencies) aiming the compression of the dynamic range. This is due to the fact that the human ear does not perceive the loudness in a linear scale, but rather in a logarithmic approximated manner. In the final step, the DCT of the logarithm energies is computed in order to de-correlate them (because of the previous overlapping of the filterbank) and smooth the spectrum by eliminating the higher DCT coefficients, which represent fast transitions in the filterbank energies.

As a summary, the advantage of MFCCs is that they allow the representation of the spectra using logarithmic frequency resolutions similar to that of the human ear, which involves a higher resolution at low frequencies.

2.2.3.2 Acoustic Segmental Properties of some EP Speech Sounds

As mentioned above, the different phonemes can be distinguished by the properties of their source(s) and their spectral shape. Wherein the vocal folds vibration represents the fundamental frequency, the supraglottal cavities configuration represents the spectral structure, the exhaling force is represented by the amplitude and finally the exhaling duration is represented by the time. All these features can be observed in the spectrogram of the sounds.

The relation between the acoustic properties, observed in the spectrogram, and the articulation of the sounds is briefly addressed in this section, regarding the EP speech sounds.

Vowels

It was previously mentioned that vowels consist of voiced sounds without constrictions in the vocal tract. Therefore, they can be characterized by having a source mostly due to periodic glottal excitation, which can be approximated by an impulse train in the time domain and by harmonics in the frequency domain, and a filter that depends on, for instance, tongue position and lip protrusion.

As the vowels are produced with a great amount of energy, they have a well defined formant structure and are therefore, highly visible in the spectrogram. The first three formants (represented upwards in the spectrogram) of a vowel are the most important, although usually the first two provide enough information for the identification of the vowel. The first formant (F1) is related to the height of the tongue during the vowel, namely a low F1 corresponds to a high vowel and a high F1 corresponds to a low vowel. As for the second formant (F2), it is related to the backness of the tongue, wherein a low F2 corresponds to back vowels and a high F2 to front vowels.

The vowel triangle diagram, shown in Figure 2.4, does the matching between each EP oral stressed vowel and its formant pair.

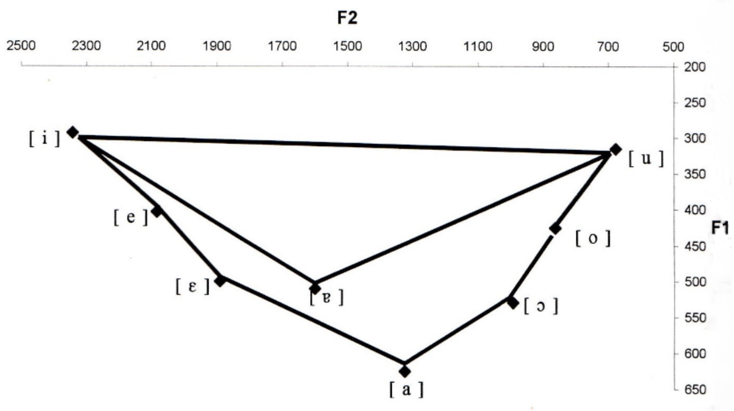


Figure 2.4: Acoustic triangle diagram of the EP oral stressed vowels [4]. It illustrates the matching between each vowel and its formant pair, as well as the relative positions between vowels.

With regard to semivowels, they also exhibit a well defined formant structure, although there is a reduction in the amount of energy. This is observed comparing Figure 2.5a, which represents the spectrogram of the oral vowel [a], to Figure 2.5b, which represents the spectrograms of the words "pai" and "pau". In these cases, vowel [a] precedes the semivowel [j] and [w]. It should be noticed that despite the fact that the vowel [a] is common to all three spectrograms, the formant structure differs. This is due to the fact that in the second case the sound of the vowel receives and produces influences in the adjacent sounds.

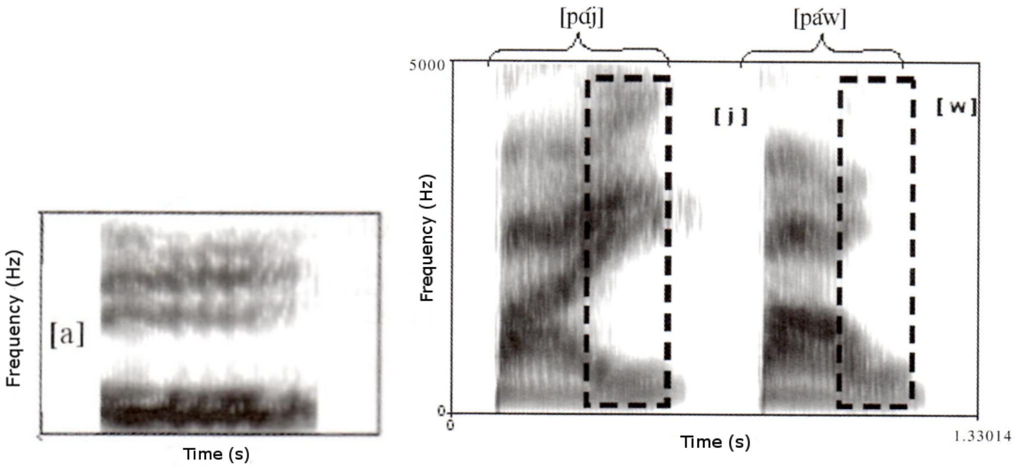
Plosive Consonants

As explained above, the stop consonants are characterized by two distinct moments: a stoppage of airflow (occlusion), corresponding to a silent moment and a sudden release of the airstream, corresponding to an explosion moment. The spectrogram provides information about these two moments, allowing to differentiate the stop consonants.

The case illustrated in Figure 2.6a, corresponding to a possible spectrogram of the word "ata", includes an area with a well defined formant structure, representing the vowels, and a silent area between the two vowels followed by an explosion bar, representing an unvoiced stop consonant, [t].

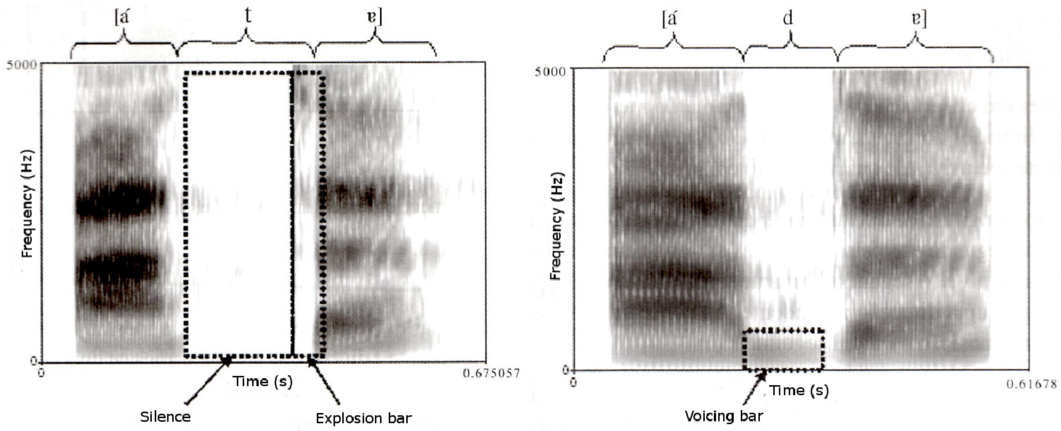
The unvoiced stop consonants ([p, t, k]) have a similar spectrogram, but different articulation places. Thus, to identify each of them it is possible to use the spectrogram characteristics in the transitions between adjacent sounds, which contains the information about the place of articulation. Besides, the spectrum of the explosion bar can be also used to get this information.

Concerning the voiced plosive consonants, the main differences when compared to the unvoiced ones, is the presence of a voicing bar in the lower frequencies of the silence period of the spectrogram and a mitigation of the explosion noise. This is shown in Figure 2.6b, representing the spectrogram of the word "ada", containing the voiced stop consonant [d]. Particularly in the case of the voiced stop consonants, the place of articulation can be determined by the analysis of



(a) (b)

Figure 2.5: (a) Spectrogram of the oral vowel [a]. (b) Spectrograms of the words "pai" and "pau". Adapted from [4]



(a) (b)

Figure 2.6: (a) Spectrogram of the word "ata" containing the unvoiced plosive [t]. (b) Spectrogram of the word "ada" containing the voiced plosive [d]. Adapted from [4]

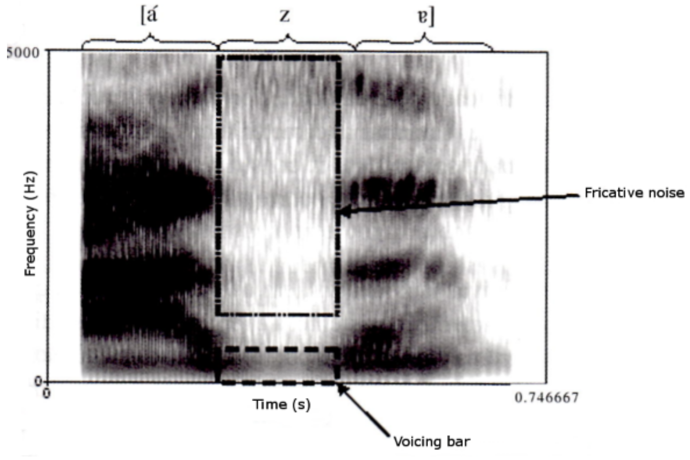


Figure 2.7: Spectrogram of the word "aza" containing the fricative [z]. Adapted from [4]

the transition between the consonant and the following vowel, namely analysing the directions of the second and third formants of the vowel [3].

Finally, in the nasal plosives the period of the occlusion might be longer and the explosion is very brief, because the air can escape through the nasal cavity. In this case, the first formant can be clearly visible and a decrease of energy might be observed in the rest of the spectrogram.

Fricative Consonants

The features of turbulent noise in fricatives are dictated by the place and shape of the constriction in the vocal tract and by the airflow properties. This noise is observed in Figure 2.7.

The differences between the spectrograms of voiced and unvoiced fricative are not very clear regarding the area where energy is concentrated. The difference lies in the existence of a voicing bar and a greater intensity in lower frequencies in the case of voiced fricative consonants.

Furthermore, as unvoiced fricative consonants are produced with a stronger airflow [4], they are more visible in the spectrogram.

According to the area of energy concentration in the spectrogram, it is possible to conclude on the place of articulation of each fricative. It should be noted that, although such conclusion can be easily drawn by a human, an automatic procedure for the same purpose would not be straightforward. The larger the cavity after the place of constriction, the lower the frequency range where the energy is concentrated, meaning that energy is concentrated in higher frequencies. This is due to the fact that the fricative noise is modified depending on the shape of the cavity that follows the place of constriction. Thus, in the case of labio-dental fricative consonants, for instance, as they have an almost non-existent cavity after the place of constriction, the energy of its spectrogram is spread over several frequencies starting at the lower ones, as shown in the leftmost spectrogram in Figure 2.8. However, in the case of palatal fricative consonants, the area after the

[3] Details about the correspondence between these directions and the place of articulation of the stop consonant can be found in [4].

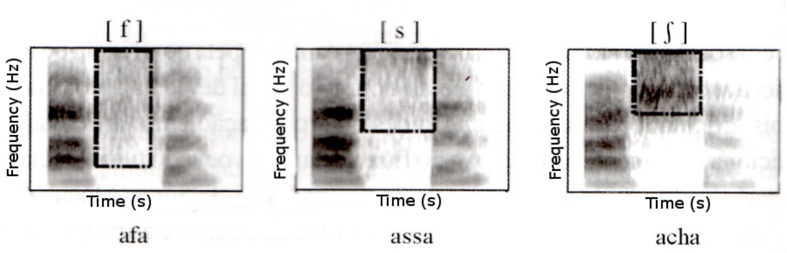


Figure 2.8: Spectrogram of the words "afa", "assa" and "acha". Adapted from [4]

place of constriction is large, thus the energy is concentrated in higher frequencies, as observed in the rightmost spectrogram in Figure 2.8.

Lateral and Trill Consonants

Lateral and trill consonants, also called liquid, are very similar to vowels in the sense that they are also very resonant and have a well defined formant structure. As shown in Figure 2.9 the formant structure is not so evident in lateral consonants as it is in the vowels but it is also stable.

The trill consonants spectrograms shown in Figure 2.10, demonstrate a great difference between the duration of each type of consonant, which is related to the articulation of each of them.

However, the liquid consonants have a variable acoustic pattern that hinders their identification.

2.2.4 Spectral Differences between Normal and Whispered Speech

Throughout the previous sections several differences were mentioned between voiced and unvoiced sounds, namely regarding the human vocal system, the source characterization according to the Source-Filter Model and the acoustic features in EP speech sounds.

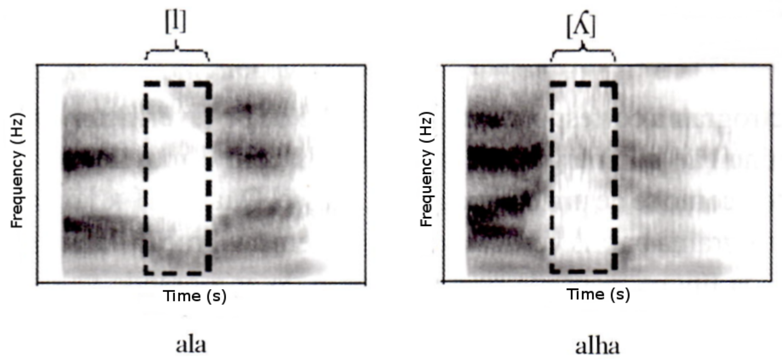


Figure 2.9: Spectrogram of the words "ala" and "alha", containing the lateral consonants [l] and [ʎ] respectively. Adapted from [4]

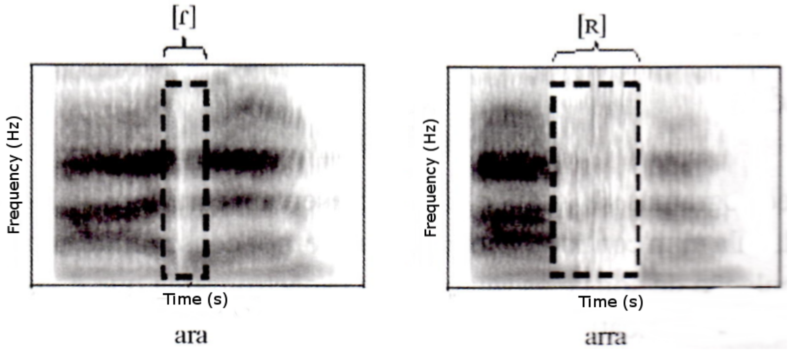


Figure 2.10: Spectrogram of the words "ara" and "arra", containing the trill consonants [ɾ] and [R] respectively. Adapted from [4]

However, the special case of whispering a voiced sound has not yet been clearly discussed. Thus, in this section the differences between normal speech and its unvoiced counterpart at the spectral level will be discussed. It is important to recall that normal speech does not entirely consist of voiced sounds. In fact, there are both voiced and unvoiced sounds in normal speech, due to the existence of sounds that are naturally unvoiced (as mentioned before, the consonants [p, t, k, f, s, ʃ] are naturally unvoiced).

Firstly, as mentioned previously, there is no periodic excitation or harmonic structure in whispered speech [14], although the sensation of pitch still exists. Its changes are related to the speech intensity level as well as the formant frequencies and bandwidths [15].

Generally, the whispered speech is approximately 20dB lower in power than its phonated counterpart [15] and has a flatter spectral slope [14]. In addition, formant shifts to higher frequencies occur in whispered speech, which are more prominent in the first formant [15]. The amplitude of the vowels and voiced consonants (naturally voiced sounds in normal speech) is greatly reduced in whispered speech, as expected, since there is no vibration of the vocal folds [16, 14]. As for the unvoiced consonants, they have a similar amplitude in both spectra [16], as might be expected. Furthermore, in whispered speech, the amplitude of vowels is lower than the amplitude of consonants, unlike the normal speech, as observed in the spectrograms of the previous section [16].

In the spectrum of the vowels in whispered speech, the following aspects are observed: an upward shift in the formants compared to normal speech, the lower the formants frequencies, the larger the shift amount [16]; there is no periodic structure but the formant structure still exists [16].

Regarding the spectrum of the voiced consonants in whispered speech, it has lower energy at low frequencies and is flatter comparing to normal speech [16]. Besides, it proved to be similar to the spectrum of the unvoiced consonants (also in whispered speech) with the same place of articulation [16]. The duration of consonants in all phonetic subclasses in whispered speech is longer than their duration in phonated speech [17].

As a summary, as might be expected, characteristics of vowels and voiced consonants change more significantly than those of unvoiced consonants [16, 17].

It is important to note that the information gathered applies to studies undertaken to different languages. Consequently, the information herein serves as a guideline to future research, since some similarity is expected with EP. The provided information does not include quantitative measures, but rather an indication of the nature of the similarities.

2.3 Methods for Reconstruction of Natural Speech from Whispered Speech

Despite not being a much researched area, some methods were already explored regarding the reconstruction of normal speech from whispered speech. Once the fundamental concepts about these two types of speech have already been described in the previous section, we are now able to address such methods. Some of them will be discussed in this section.

In speech production, there are three types of speech coders: waveform coders, parametric coders and hybrid coders. The parametric coders estimate some input signal parameters so they can synthesize the speech signal using very low bit-rate coding strategies. This means that only the parameters of a speech model are transmitted and a decoder is used to regenerate speech with the same perceptual characteristics as the input speech waveform. Some known parametric models are Linear Prediction Coding (LPC) and Mixed Excitation Linear Prediction (MELP). Hybrid coders also rely on a speech production model, but at the same time, attempts are made to approximate the decoded signal to the original waveform in the time domain, as in the case of waveform coders. An example of a hybrid coder is the Code Excited Linear Prediction (CELP) model.

These voice coders (or vocoders) were originally designed for code speech for transmission. However, they are now used for another purpose and in this chapter they are introduced as a way to relate the whispered and voiced signals. Before proceeding, it is important to mention that all vocoders work only a short section of speech, because they exploit the idea that for short periods of time the quasi-stationarity assumption is valid.

2.3.1 MELP Approach

In LPC [5], current speech samples can be approximated from linear combination of past samples. Within a frame, the weights used for the linear combination, can be determined by minimizing the squared error between a current sample and its prediction from past samples. These weights, represent the filter coefficients (linear prediction coefficients) in the LPC model of speech production, production, as shown in Figure 2.11.

The excitation signal is represented by an impulse train or random noise, depending on the voiced or unvoiced state of the signal respectively. Thus, the switch, in Figure 2.11, is set to the proper location according to the voicing parameter. The synthesis filter represents the combination of spectral contributions of the glottal flow, vocal tract and lips radiation and its coefficients are estimated by the encoder. Therefore, each frame of speech can be reduced to a set of LPC

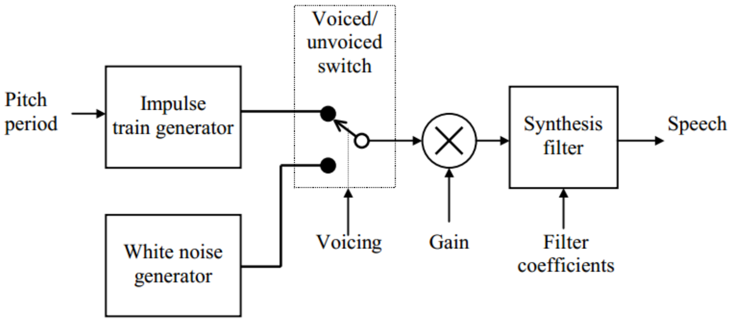


Figure 2.11: LPC model of speech production [5]

coefficients (that are transformed to LSP frequencies [4]), the pitch value which is used for the voiced/unvoiced decision and the gain value (mainly related to the energy level of the frame). In the decoder the speech signal is reconstructed by a synthesizer based on a time varying all-pole filter.

As a summary, LPC is a technique that allows the identification of the parameters of a system. It is assumed that the speech can be modeled as an AR (autoregressive) signal, meaning that the output variable depends linearly on its own previous values.

The MELP speech production model [5], is illustrated in Figure 2.12, and arose to overcome the LPC problems, using a mixed-excitation model that can produce more natural sounding speech since it represents a larger set of speech characteristics. In short, this model combines periodic excitation (impulse train, whose shape is extracted from the input speech signal) with noise excitation, through the use of two shaping filters so as to form the mixed excitation signal. The frequency responses of these shaping filters are controlled by a set of parameters, the voicing strengths, that are estimated from the input speech signal and represent voiced/unvoiced decision. In fact, the speech signal is split into five frequency bands (0–500 Hz, 500–1000 Hz, 1000–2000 Hz, 2000–3000 Hz and 3000–4000 Hz), and for each band an excitation based on pulse train and noise is generated. In the decoding process, an LPC synthesis filter processes the combined excitation signal so as to generate the synthetic speech.

Existing System

The seminal work in [15] uses one of the best-known approaches to whisper-to-speech conversion, the previously mentioned Mixed Excitation Linear Prediction vocoder. It aims to develop a real-time synthesis solution of normal speech, proposing methods for estimating the parameters of MELP model from whispered speech. Thus, it contains methods for compensating spectral differences between whispered and voiced speech, as well as methods for determining the parameters of the excitation signal. These estimates allow the synthesis of normal speech using the MELP model vocoder.

[4] Line Spectrum Pairs (LSP) are used to represent linear prediction coefficients because of their quantization properties, useful in speech coding.

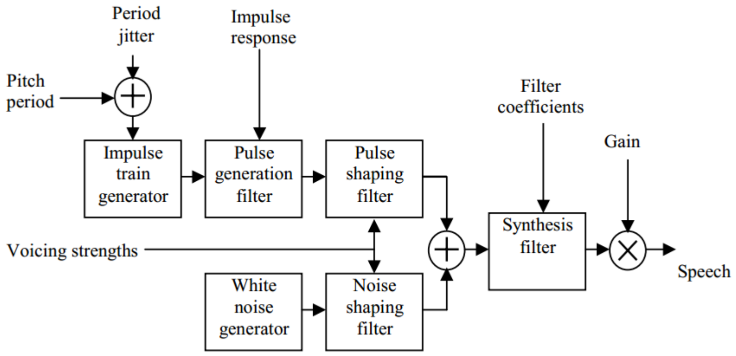


Figure 2.12: MELP model of speech production [5]

The need of a spectral modification is due to the fact that the linear prediction spectrum used by MELP model is different from the one obtained considering whispered features. This is because of the aforementioned differences in spectral excitation, formant shifts and the different variance of the linear prediction spectrum estimate, due to the completely noise excited speech in whispering.

So, the proposed spectral modifications include: design of a static linear filter to make up for the long-term source spectrum differences between normal and whispered speech; spectral smoothing, to eliminate abrupt variations in the spectrum (most evident in whisper speech since it is represented by a noise-like excitation signal) in undesirable situations, like during steady vowels, that might lead to an unnatural spectrum of the synthesized speech; formant shifting, which is made downwardly according to each formant location, i.e. according to the formant frequency difference to the equivalent formant in the phonated counterpart.

Particularly, for spectral smoothing three methods are proposed and discussed, namely linear filtering (Kalman filter) and non linear filtering (a median filter and a Gaussian mixture based filter).

Regarding the estimation of the excitation signal parameters, while MELP coder determines the voicing excitation using five frequency bands, the proposed method in this article states that the four lower bands (0-3kHz) represent voiced speech and only the upper band (3-4kHz) represents unvoiced speech. The pitch estimation is done by filtering the gain parameter, considering the relation between the speech intensity level and the perceived pitch in whisper. This estimation method might not offer a very fair approximation of pitch contour in normal speech but it allows a modulation of the whisper intensity according to the desired pitch in real time synthesis.

At the end of the algorithm the linear prediction spectra of voiced speech and whisper speech were compared and a great similarity in unvoiced parts of speech was observed. However, phonated parts showed some differences, namely in formants. Conclusions about the spectral estimates in what regards the three different filters used for spectral smoothing showed that, although being successful in preserving the abrupt transitions and formant bandwidths, the median filter proved not to be suited for real-time systems as it introduces signal delays. The use of a Kalman filter filter did not prove to be so effective in smoothing as the median filter, since it caused some distor-

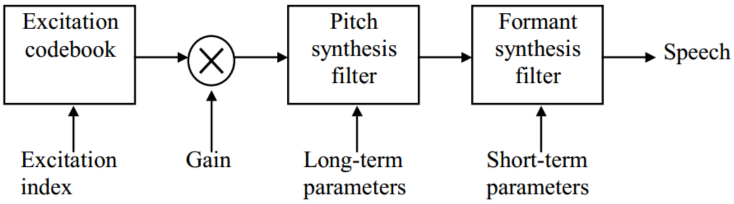


Figure 2.13: CELP model of speech production [5]

tion in rapid transitions. At last, the Gaussian mixture based filter, although introducing also some distortion in rapid transitions and requiring more computational effort, provided better smoothing with an almost non existent delay.

To summarize, the proposed technique shows reasonable results, although it is not suited to real-time operation and requires the knowledge of the characteristics of the normal sounding voice. This is due to the fact that it uses a comparison of normal and whispered speech samples from the same speaker to train a jump Markov linear system (JMLS) to estimate pitch and voicing parameters. Thus, it is not recommended to situations where there is no information about the original voice.

2.3.2 CELP Approach

Code Excited Linear Prediction (CELP) is a hybrid vocoder [5], that also attempts to overcome the problems introduced by LPC vocoder, avoiding its rigid voiced/unvoiced classification, through the use of long-term and short-term linear prediction models, and introducing the use of an excitation codebook (hence the name "code excited"). This codebook contains the "code" to "excite" the synthesis filters, i.e. it is a vector quantization approach, where the whole sequence is encoded using a single index, and thus the encoder uses it to search for the best excitation sequence. The CELP model of speech production is represented in Figure 2.13.

In the decoder, the excitation sequence is extracted from the codebook through a received index, it is then scaled and finally filtered by a pitch synthesis filter and formant synthesis filter. The former provides a signal periodicity related to the fundamental pitch frequency and the latter creates a spectral envelope (by first splitting the original speech into analysis frames and then performing the LPC analysis, generating a set of coefficients used in a short-term predictor). The use of the two synthesis filters is an advantage of the CELP model over the other speech synthesis models, as it allows an effective modeling of transition frames with smoothness and continuity, thus synthesizing much more naturally sounding speech.

The codebook contains deterministic pulses or random noise and might be classified as fixed or adaptive. A transmitted index concerns the excitation sequence determined as the best, i.e. the sequence capable of generating the synthetic speech most similar to the original one. The similarity is often measured by a perceptually weighed error signal. Furthermore, the synthetic speech is not only a result of a match in the magnitude spectrum domain but also in the time domain, capturing some phase information, unlike the LPC model. This feature provides a more natural synthetic speech. The capture of phase information is performed through a closed-loop analysis-by-synthesis method, meaning that the encoder fine-tunes the encoding parameters through the analysis of the synthesized signal so as to generate the most accurate reconstruction. Thus, the excitation sequences are selected from the codebook according to a closed-loop method.

Existing System

The work described in [18] proposes a CELP-based system for whisper-to-speech reconstruction. Unlike the previously mentioned work [15], that required the availability of the phonated speech samples, this approach proposes a method whose target applications focus on laryngectomy patients and thus, works without the use of normal speech samples.

Along with the modified CELP codec, there are three main proposed additions for whisper-to-speech reconstruction: a whisper-activity detector (WAD), a whisper phoneme classifier (WPC) and a novel spectral enhancement of the reconstructed speech. The CELP algorithm can operate with only a few modifications if plosive or unvoiced fricatives are detected, otherwise the speech frame is considered to be voiced, but with missing pitch, requiring significant modifications in the algorithm, as gain adjustment and spectral enhancements. The detection is performed by estimating the sound frame average energy.

The whisper-activity detector and whisper phoneme classifier are part of the preprocessing necessary for whispered speech analysis. The former uses two detection mechanisms: a classifier based on the signal power, that identifies whisper, noise and silence regions, comparing time-domain energy with two adaptive thresholds; a classifier based on zero crossing rate, that helps the decision of the previous classifier through the use of a differential adaptive threshold crossing rate. Regarding the WPC module, it performs a phoneme classification assigning weights to the previously detected whisper speech segments. The weight of unvoicing is high when a plosive or an unvoiced fricative is detected, and is low if a vowel is detected. This weighing is then used by the pitch insertion algorithm. The detection of fricatives is based on the comparison of the power of whispered frames in bandwidths above and below a certain frequency value; the signal energy ratios in small bands of low and high frequency are used to identify vowels and plosives. Additionally, the latter is confirmed if a small silence is detected, represented by low energy, before a burst of energy. It is worth mentioning that the thresholds used for the classifications are speaker dependent, meaning that they are determined manually.

The enhancement of the spectral characteristics includes reconstructing the disordered and unclear formants caused from the noisy-like whispered signal. For this purpose, the algorithm applies a formant-track smoother to avoid large variations in the formant trajectories between adjacent frames. The formant tracking is based on the method of LP coefficient root solving and a novel approach aiming the generation of the formant trajectory uses the probability mass-density function (PMF). Finally, the modified formants are smoothed and shifted. Additionally, a bandwidth improvement of the final formants is performed, based on the spectral energy.

Furthermore, the proposed system contains a pitch template to estimate the pitch.

Results show that the WPC module provides a robust method of phoneme classification, having a significant importance in the overall system performance; more naturalness of the synthesized speech could be achieved by improving pitch, but the modified formants show reasonable results. Although the proposed methods have worked well for phonemes or single words, the results were not so good in the case of regenerating complete sentences, i.e. continuous speech.

2.3.3 Other Approaches

The method described in [19] has a similar reasoning to the MELP and CELP-based source-filter approaches, but has reduced computational complexity. Furthermore, it does not require a priori or speaker-dependent information and is designed for real-time operation. Unlike the previous approaches, this method proposes a whisper-to-speech conversion by reversing the LTI source-filter speech production model, i.e. it synthesizes individual formants, representing an excitation source, before modulating them using an artificial glottal signal. Thus, this is a different approach, since it operates in the opposite sequence of the human speech production process and the previously mentioned MELP/CELP reconstruction methods.

The artificial formants are obtained by a first formant extraction followed by a refinement mechanism. Specifically, they consist of a frequency translated version of oversampled and time-domain extracted, smoothed whispered speech formants. In this method there is no voiced/unvoiced decision process of the frames. A scalar gain is added to the derived formants, allowing the inclusion of high frequency wideband distribution from the original whisper in the reconstructed speech. The result is modulated by a raised cosine glottal signal which is harmonically related to F1. In fact, this method synthesizes pseudo-F0 by taking into consideration the harmonic relationship between pitch and formants. During low energy analysis frames, the depth of modulation is reduced.

This work showed modest results regarding the reconstructed speech, although the latter exhibited improved quality and intelligibility over the original whispers.

2.3.4 Previous Work

This dissertation follows the work in [11] aiming to improve the results by the research and implementation of new solutions.

The work described in [11] performs the proof-of-concept of the possibility to implant artificial voicing in whispered speech and proposes two different solutions denoted by: dependent version and independent version. The former uses information provided by the original normal speech signal, namely the frequency bins of the harmonics and the energy of each frame. The latter, only uses the information about the average value of the fundamental frequency throughout the frames of the original normal speech signal and computes its multiples - harmonics. Both solutions perform a manual segmentation of the candidate regions to artificial voicing in whispered speech. Although a solution for the automatic segmentation of the whispered speech is also proposed, which uses Neural Networks, it is not used in the overall whisper-to-speech conversion system.

The algorithm consists of the following steps:

1. The whispered signal is manually segmented into regions where artificial voicing should be implanted (naturally voiced regions) and regions that are naturally unvoiced;
2. The spectral envelope of the frames of naturally voiced regions are extracted in order to estimate the vocal tract filter;
3. The estimated vocal tract filter is modified in order to shift the formants to their locations in the normal speech signal;
4. The original normal speech signal is also manually segmented into voiced and unvoiced regions. The harmonic structure is extracted from the frames classified as voiced and the amplitudes are normalized;
5. A glottal decay filter is applied to the harmonics;
6. A synchronization between the voiced frames of the original signal and the previously selected frames of the whispered signal is performed, since both the whispered and the original normal speech sentences are used;
7. Finally the harmonic structure is multiplied by the estimated vocal tract filter and the artificial speech signal is obtained by multiplying the result by a lips radiation filter - according to the source-filter model;

Therefore, the difference between the two versions relies on the method used to obtain the harmonics, i.e. in the fourth step.

The speech database of the algorithm consists of utterances of speakers that do not suffer from voice disorders.

2.4 Automatic Segmentation Strategies of Whispered Speech

2.4.1 Hidden Markov Model

Hidden Markov Models (HMMs) [6] are statistical models commonly used in speech processing, namely in speech recognition and speech synthesis systems, as stochastic signal models. These models are representable by finite state machines, which generate a sequence of discrete time observations.

Unlike the basic Markov Models, in which each state corresponds to an observable output event, in the HMMs an observation is a probabilistic function of the state. Thus, a state changes according to the state transition probability distribution (as in Markov Model) generating an observable output according to the output probability distribution of the current state. Therefore, since in the case of HMM the state is "hidden" because there is not a direct correspondence between the output and the state, the latter is determined by the generated sequence of observations (the observable outputs).

An HMM is characterized by the number of states, the number of distinct observation symbols (outputs) per state, the state transition probability distribution, the observation symbol probability distribution in each state and the initial state distribution. Having these values, the HMM can generate a sequence of observations.

There are three problems to solve for HMM design. The first problem evaluates, for an observation sequence and a given model, the probability (likelihood) that the observed sequence was generated by the model, i.e. it allows to evaluate the matching between a sequence of observations and a model. It is solved by the forward-backward algorithm [6]. The second problem aims to determine, given a sequence of observations and a model, the corresponding sequence of states that better "explains" the observation sequence, i.e. find the optimal state sequence. This sequence can be found using the Viterbi algorithm. The third problem is how to adjust the model parameters so that the probability of an observation sequence given the model is maximized, which is solved by the Baum-Welch algorithm. In this last problem, a training sequence is used to train the HMM, allowing to optimally adapt the model parameters. This training sequence is an observation sequence.

Application of HMM in Speech Recognition

So as to better understand the HMM application in speech, an example of the relation between these three problems and an isolated word speech recognizer is presented as follows [6]. Having a W word vocabulary [5] it is intended to model an HMM for each word. Thus, first a temporal sequence of coded spectral vectors is generated representing each speech signal (word). These spectral vectors together create a codebook, wherein the index are the observable outputs. The codebook is created through a training set of repetitions of each spoken word (spoken by one or more talkers). Therefore, a first step involves finding a solution to the third problem as described above, since it is necessary to determine the optimal model parameters for each word model, i.e. to optimize the likelihood of the training set observation vectors for the word. The next task is to solve the second problem, transforming each of the word training sequences into states and then studying the features of the spectral vectors that lead to the observable outputs in each state. Thus, it is possible to improve the model by e.g. adding states or changing the codebook size to better model the words. Finally, the problem to be solved is the first one, since it aims to recognize an unknown word, i.e. to use the observation sequence corresponding to the features of the unknown word in order to find which of the models is most likely to have generated such a sequence. An illustrated example of such a recognizer is shown in Figure 2.14.

Thus, it becomes clear that the usefulness of HMMs in speech processing lies in the fact that they allow to follow the temporal evolution of the phonetic changes of speech. In fact, there is a particular type of HMM, known as left-right model, which is suitable for modeling signals whose properties change over time. In left-right models, the states change from left to right, hence the

[5] It is important to note that the word is just an example of a speech recognition unit. Several basic speech units could be chosen as phones, diphones, demissyllables, syllables, among others.

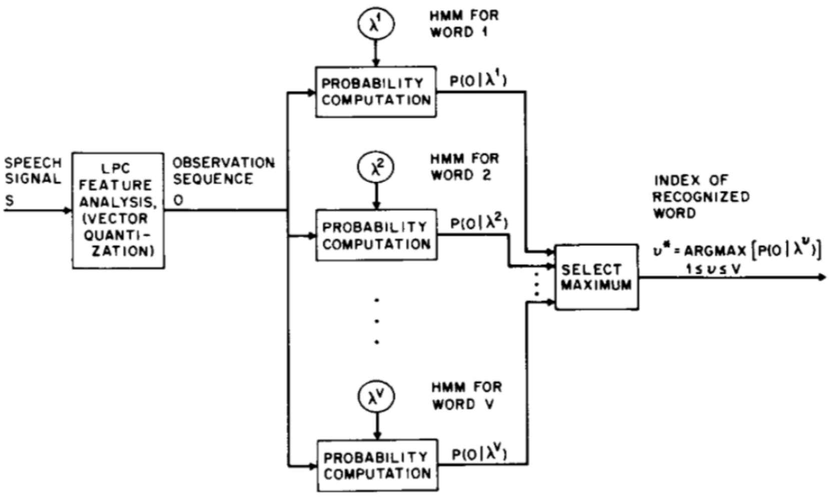


Figure 2.14: Example of an isolated word HMM recognizer [6]

name. This implies that the state index increases (or stays the same) as time increases, i.e. there are no transitions from the current state to states with lower indices.

Application of HMM in Speech Synthesis

Since HMMs have achieved successful results in speech recognition, their application in speech synthesis has also been investigated. The main advantage of statistical parametric speech synthesis, as is the case of HMM-based speech synthesis system (HTS), is the ability to synthesize speech with various voice characteristics, such as speaking styles, emotions and speaker individualities, by transforming the model parameters. A typical statistical parametric speech synthesis system starts with the extraction of speech signal parameters, such as spectral and excitation parameters, from a speech database and their subsequent modeling by a set of generative models, such as HMMs. The estimation of the model parameters is often performed through the use of a maximum likelihood criterion, i.e. the selected parameters are those which maximize the probability of observing the set of training data. As such, the model parameters estimated by this criterion are given by [20]:

$$\hat{\lambda} = \arg\max_{\lambda}\{p(\mathbf{O}|W,\lambda)\}, \tag{2.2}$$

where λ represents the set of model parameters, **O** represents the set of training data and W the set of word sequences corresponding to **O**.

Then, given a word sequence to be synthesized (w), the speech parameters (**o**) are generated using the previously estimated models ($\hat{\lambda}$), such that their output probabilities are maximized as [20]:

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}}\{p(\mathbf{o}|w,\hat{\lambda})\}, \tag{2.3}$$

The synthesized speech waveform is finally generated.

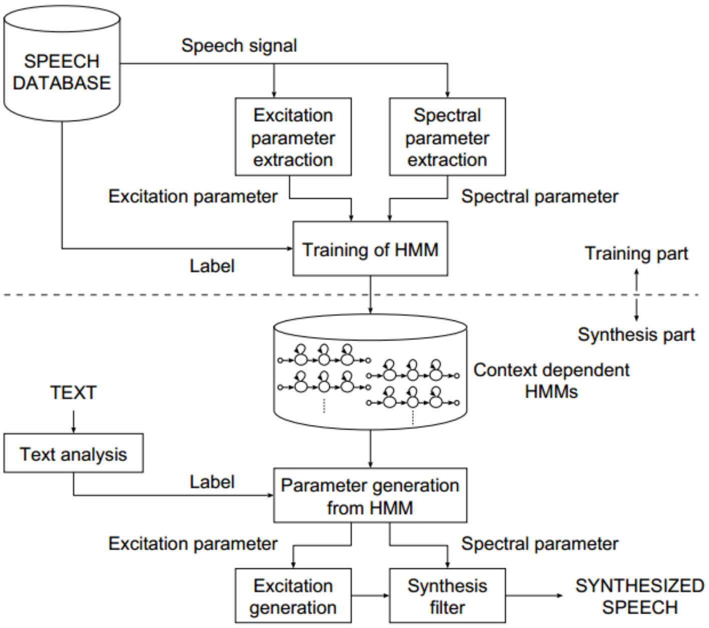


Figure 2.15: Example of an HMM-based speech synthesis system (HTS) [7]

Particularly in the case of statistical parametric speech synthesis using HMMs (HTS), the overall procedure consists of two parts, the training part and the synthesis part. In the former, the maximum likelihood criterion of Eq. 2.2 is performed by using the Expectation-Maximization (EM) algorithm, similarly to the case of speech recognition. The main difference lies in the fact that in HTS, the output vector of HMM consists of spectrum part and excitation part. Therefore, both spectrum and excitation parameters are extracted from a speech database and modeled by context dependent HMMs [7, 21]. Context dependent HMMs, as the name implies, take into consideration contextual factors (e.g. linguistic, prosodic and phonetic contexts) for proper interpretation. For instance, the contexts used in HTS English [7] are related to phoneme (e.g. preceding, current and succeeding phoneme), syllable (e.g. numbers of phonemes within preceding, current and succeeding syllables or stress of preceding, current, and succeeding syllables), word (e.g. number of syllables in preceding, current and succeeding word), phrase (e.g. number of syllables in preceding, current and succeeding phrase) and utterance (numbers of syllables, words, and phrases in utterance). The synthesis part is a sort of inverse operation of speech recognition, performing the maximization of the Eq. 2.3. First, a given word sequence is converted into a context-dependent label sequence. Second, context-dependent HMMs are concatenated according to the label to construct a sentence HMM. Then, using a speech parameter generation algorithm, sequences of spectral and excitation parameters from the sentence HMM are generated, in such a way that their output probabilities for the HMM are maximized. Finally, the speech waveform is synthesized from the generated spectral and excitation parameters by using an excitation generation module and a speech synthesis filter. An example of an HMM-based speech synthesis system is shown in Figure 2.15.

Some solutions have been proposed [7], [22], [23], [24] wherein the parameters of the spectrum part consist of mel-cepstral coefficients and their dynamic features, i.e. delta and delta-delta parameters, and the excitation part consists of the logarithm of the fundamental frequency and its delta-delta parameters. Furthermore, the HMMs incorporate state duration using duration densities to control the temporal structure of speech. The speech waveform is synthesized directly from the generated mel-cepstral coefficients and values of the logarithm of the fundamental frequency, through the use of the Mel Log Spectrum Approximation (MLSA) filter.

2.4.2 Considerations on Whispered Speech Segmentation Strategies

The use of HMM in speech processing has contributed significantly to advances in both speech recognition and speech synthesis systems in spite of its limitations [6]. Therefore, the potential of this statistical parametric model should be further developed, for instance, for the implementation of an automatic segmentation strategy of whispered speech.

An automatic segmentation algorithm aims to identify the regions in whispered speech where artificial voicing should be implanted. Since the unvoiced sounds have similar features in both speech modes (whispered and normal speech), as discussed above, it is intended to detect only the parts corresponding to voiced sounds. Therefore, these regions correspond to selected parts of the whispered speech that would be naturally voiced. The automatic nature of such an algorithm would be essential for the implementation of a real-time system.

2.5 Conclusions

In order to convert whispered speech into normal sounding speech, it is necessary to implant voicing in the so-called naturally voiced sounds in whispered speech. For this purpose, some modifications have to be performed, such as insertion of a fundamental frequency, formant shifts and changes in magnitude.

Actually, whispers have much lower acoustic power than normal speech and a relatively flat spectrum, being inherently noise-like and thus, highly sensible to acoustic interference. This is a disadvantage for systems which analyse whispers to determine both time-domain and frequency-domain information, as they have to be robust to errors [19].

The solutions implemented so far for whisper-to-speech conversion use the concepts of parametric coders to synthesize normal sounding speech. Particularly, they perform spectral modifications and estimate the excitation signal parameters, according to the whispered speech features, in order to use speech production models, such as MELP or CELP. These models assume that both the pitch glottal component and the vocal tract component can be represented as linear time invariant (LTI) systems and are mutually independent. Although the assumption is not true, it

[6] One of the limitations of HMMs is the Markov assumption that the probability of being in a given state at time t only depends on the previous state, i.e. the state at time t-1. This is unsuited for speech sounds, since they often depend on several states. [6]

is considered to be a reasonable approximation [19]. However, these solutions have some weaknesses. The MELP based system is unsuitable for real-time operations and relies on the knowledge of the original voice (normal speech). The solution based on CELP vocoder is unsuitable for continuous speech, since it struggles to regenerate complete sentences. Finally, neither of the solutions produce a natural-sounding artificial voice.

The theory of Hidden Markov Model applied to speech processing was studied because of its statistical parametric nature, which is suitable for the modeling of speech signals. Thus, two applications of this theory were illustrated, namely in speech recognition and in speech synthesis systems, both exhibiting good results. The intent of this brief discussion is to emphasize the great potential of HMMs for characterizing the basic processes of speech production and to infer how the illustrated techniques could be applied to more speech-related problems, such as an algorithm for automatic segmentation of whispered speech. This could be accomplished through an appropriate statistical and parametric modeling of the whispered speech, namely a proper identification of the speech signal features and a correct characterization of the states.

Chapter 3

Speech Corpus

A speech corpus is an organized collection of recorded speech data and supporting files. Its purpose is to provide the necessary data for several stages of a speech system. In the case of this dissertation, the speech corpus will provide information that will be used for training and testing the algorithm for automatic segmentation of whispered speech, as well as the algorithm to implant artificial voicing. A speech corpus should have an appropriate phonetic coverage so as to ensure a good characterization of the speech sounds of a target language or a target speaker if there is any. However, in our case, the database has little variety of each type of phoneme, which hinders the training procedure and consequently, the success of the different stages of the overall whisper-to-speech conversion algorithm.

The collected samples pertain to a male patient suffering from spasmodic dysphonia, aged around 40, who has shown interest in supporting this research. There are two main types of speech samples both provided by the same individual:

- Normal speech samples [1];
- Whispered speech samples.

Thus, the database was first organized in these two speech modes. In addition, for each speech mode three types of speech corpora were distinguished:

- Read speech - which includes the reading of book excerpts and utterance of sequences of numbers;
- Spontaneous speech - which includes dialogues;
- Sustained vowels - vowels uttered in a sustained way, with silence before and after the vowel sound.

This division is important because the same sentence may have different prosody characteristics depending on how it is uttered. However, there are no recorded samples from Spontaneous Speech database, thus the speech samples are limited to those of Read Speech and Sustained Vowels corpora.

[1] These samples were collected when the patient could still produce normal sounding speech, albeit with effort.

3.1 Speech Material

The sentences from the Read Speech database consist of different types of representative phonemes of the EP. There are three sentences and each comprises a normal speech version and a whispered speech version:

- "A Sofia saiu cedo da sala para conversar com o Aurélio.";
- "O vento norte e o sol discutiam qual dos dois era o mais forte.";
- "O vento norte e o sol discutiam qual dos dois era o mais forte quando sucedeu passar um viajante envolto numa capa." - extended version of the previous sentence.

These sentences consist of different types of vowels sounds, such as [a], [ɛ], [e], [i], [ɔ], [u], [ɐ], [ɨ], [o]; plosive voiced consonants (e.g. [m] in the word "mais"); plosive unvoiced consonants (e.g. [t] in the word "norte"); fricative voiced consonants (e.g. [v] in the word "vento"); fricative unvoiced consonants (e.g. [s] in the word "Sofia"); lateral consonants (e.g. [l] in the word "envolto") and trill consonants (e.g. [ɾ] in the word "Aurélio").

The collected sentences were recorded with 44,1kHz sampling frequency and 16 bit resolution and saved in a lossless format. However, their sampling frequency was converted to 22,05kHz, by using the Cool Edit software [25], since it is known that the most relevant spectral content is concentrated below 10kHz. Furthermore, the sentences have a duration between 5 and 11 seconds.

3.2 Annotated Database

The sentences of the whispered speech database were annotated through the use of the corresponding feature of the Praat software [26]. This tool allows to segment the sound file into words or into phonemes, by manually adding boundaries, and to label each interval (the space between the assigned boundaries). The labelling data is stored in a TextGrid file, which has a particular structure - a start and an end time, in seconds, assigned to each label.

An example of a manual segmentation is depicted in Figure 3.1 for the sentence "A Sofia saiu cedo da sala para conversar com o Aurélio.". Two tiers of labels are illustrated, where the lower one concerns a phoneme segmentation and the upper one distinguishes between five segment types - silence ("Sil" label), unvoiced plosive consonants ("UP" label) [2], unvoiced fricative consonants ("UF" label), voiced consonants ("Vc" label) and vowels ("Vv" label).

The manual segmentation is a very important stage of the whisper-to-speech conversion process. It should be performed accurately, since the automatic nature of the algorithm depends on a right sound labelling. Some basic criteria were used for segmentation:

- Speech waveform - used to detect silences, pauses, frication noises (noise in fricative consonants) and periodic regions corresponding to vowels or naturally voiced consonants;

[2] This label is not visible in the figure since its corresponding interval/segment is too narrow, i.e. it has a very short duration.

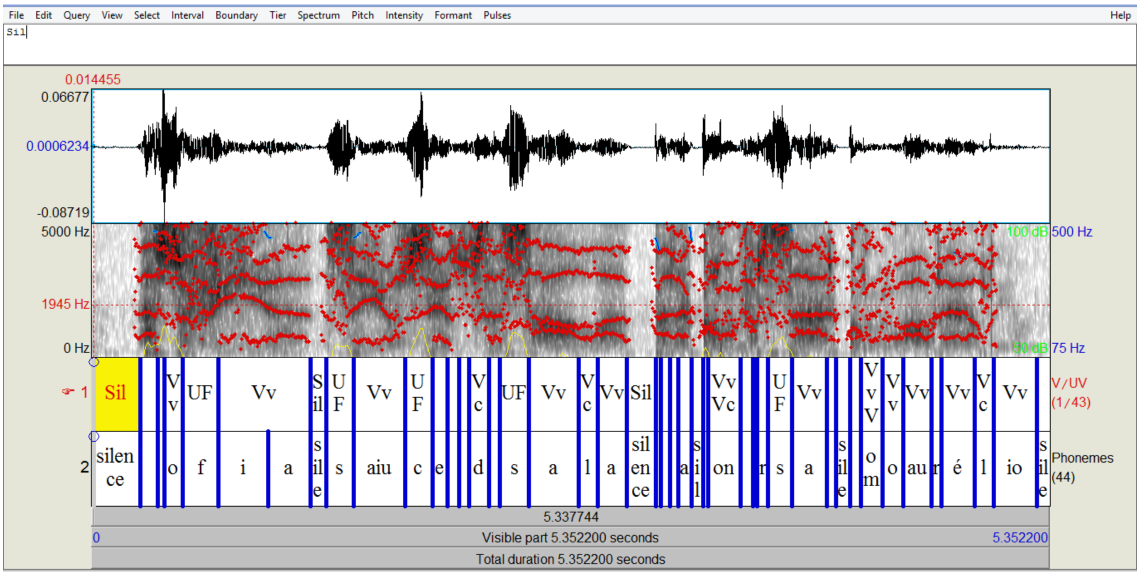

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
Sil
0.014455
0.06677
0.0006234
-0.08719
5000 Hz
1945 Hz
0 Hz
100 dB 500 Hz
50 dB 75 Hz
1 Sil
V/UV (1/43)
2 silence
Phonemes (44)
Vv UF Vv Sil UF Vv UF Vc UF Vv Vc Vv Sil Vv Vc UF Vv Vv Vv Vv Vc Vv
o f i a sil s aiu c e d s a l a silence a s il on r s a sil o m o au r é l io sil
5.337744
0
Visible part 5.352200 seconds
5.352200
Total duration 5.352200 seconds

Figure 3.1: Example of a manual segmentation of a sentence using Praat software. The time waveform of the whispered sentence, the corresponding spectrogram and the manual labelling are illustrated, where the blue vertical lines represent the boundaries.

- Formants temporal trajectory in the spectrogram - used to detect boundaries (transitions) between phonemes and steady-state vowel regions;
- Perceived sound (perceptual impression).

However, it is very difficult to perform an accurate segmentation, since in some cases, detecting the right boundaries between two types of phonemes is a hard process for the human ear, mostly due to coarticulation phenomena. In fact, some issues were encountered during the manual labelling of the three whispered speech sentences:

- Vowels between two naturally unvoiced consonants are perceptually difficult to distinguish;
- In some cases, when naturally voiced plosive and fricative consonants are whispered, they are perceived as the naturally unvoiced plosive and fricative consonants that have the same place of articulation (see Table 2.1). This problem is illustrated in Table 3.1.

Uttered Consonant	Perceived Consonant
v	f
z	s
ʒ	ʃ
d	t
b	p
g	k

Table 3.1: Correspondence between the actually uttered consonants and the perceived consonants in whispered speech.

Furthermore, in this particular case, these problems are more prominent both due to the patient's pronunciation and voice disorder, which result in a noisier and less intelligible speech. For instance, the word "dois" is perceived as "doiz", which means that the unvoiced fricative consonant [s] should be labeled as the voiced fricative [z].

Chapter 4

Proposed Approach for Automatic Segmentation of Whispered Speech

The implementation of a real-time whisper-to-speech conversion system requires an automatic solution for the identification of the regions in whispered speech where artificial voicing should be implanted. The proposed approach relies on a machine learning technique, specifically Hidden Markov Model. The HMM is used as a classifier, to identify the frames that should remain unvoiced, i.e. frames corresponding to silence periods or to unvoiced consonants regions, and the frames where voicing should be implanted, i.e frames corresponding to vowels or naturally voiced consonants.

4.1 Approach Overview

An overview block diagram for the proposed system is shown in Figure 4.1. The approach encompasses two main stages: training stage and classification stage. In the training part, the system uses the collection of whispered speech utterances to extract parameters from the speech signal as well as the labels to train the HMM. After training has been completed, a classification part is performed. At this stage, new whispered speech utterances are used to test the classification procedure. These utterances are analysed and their extracted parameters together with the HMM are used to classify whether it is necessary to implant artificial voicing or not in the utterance region (frame) under analysis.

4.1.1 Analysis

Feature extraction is the process of computing a sequence of feature vectors, which represent the speech signal in a parametric way. The analysis of the signal is performed on a frame-by-frame basis, thus for each speech frame a feature vector is computed. The set of feature vectors pertaining to all frames is used to build a database.

The first step is to chose the features that are most likely to identify different parts of the whispered speech. A speech signal can be represented by different features and different features

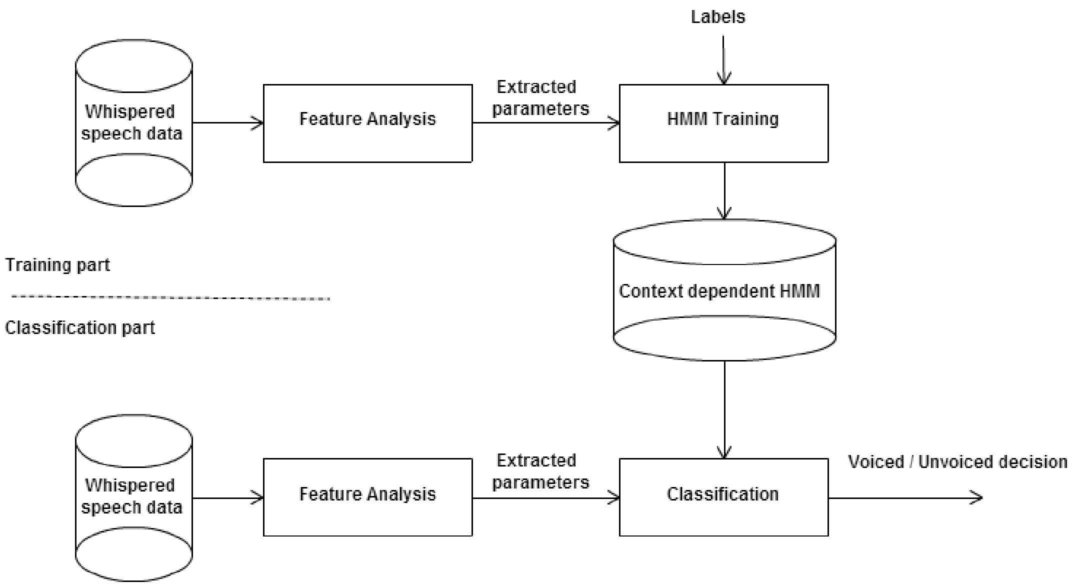


Figure 4.1: Block diagram for the proposed system

emphasize different spectral characteristics. Thus, we may use more than one type to describe the signal and train the HMM. The selected features were MFCCs, LSPs and PLPs (Perceptual Linear Prediction) [27], which are the most widely used features in the area of speech processing, namely in speech recognition systems [28]. However, few works regarding the identification and analysis of whispered speech were encountered. The works [29, 16] use MFCC coefficients and waveform power information to represent the whispered speech signal. Particularly, the latter work uses HMM to recognize whispered speech. Herein, besides using MFCCs to represent the spectral features of whispered speech signals, the possibility of using LSPs and PLPs will also be considered and studied in this dissertation.

As mentioned in the previous chapter, the speech data was converted to 22,050kHz sampling frequency. Each speech segment, i.e. frame, has a duration of 10ms with an overlap of 8ms, which implies a 2ms stepsize. By performing windowing with overlap, one is artificially increasing the time resolution which is especially useful in this case, where unvoiced plosive should be detected. These consonants are characterized by a explosion period after silence. This explosion lasts for about 2ms and is followed by noise as observed in Figure 4.2, hence the chosen overlap of 8ms.

Each frame was weighted by Hamming window with preemphasis to compute a 13th order MFCC , a 13th order LSP and a 13th order PLP analysis. Therefore, a feature vector consisting of 13 MFCC coefficients, 13 LSF coefficients and 13 PLP coefficients was computed for each speech frame.

4.1.2 Training

There are two situations in which artificial voicing is not applied: frames corresponding to silence periods and frames corresponding to unvoiced consonants. Thus, four speech sound categories

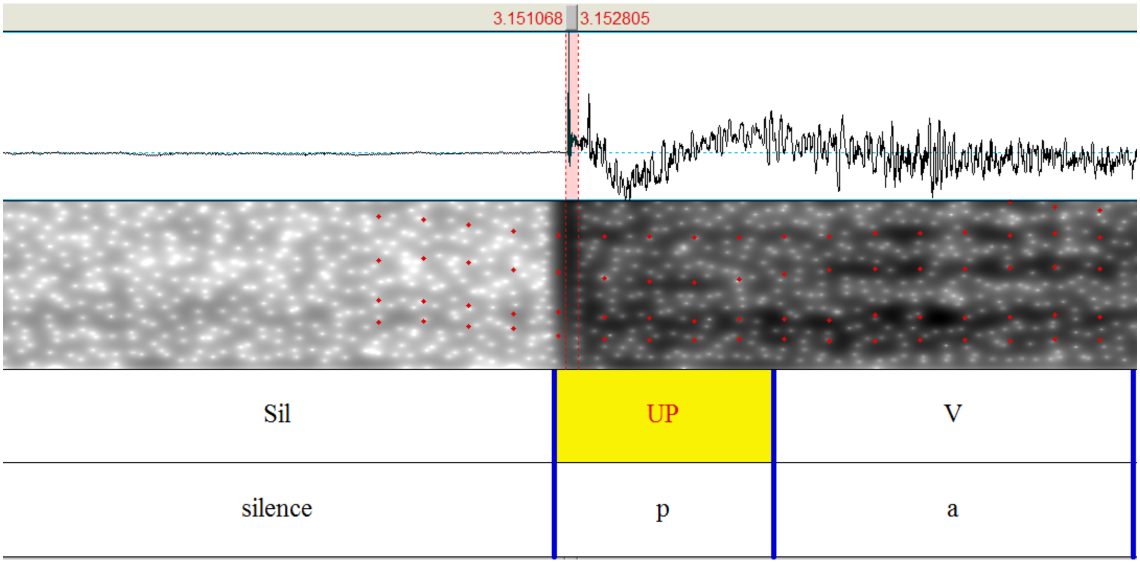


Figure 4.2: Duration of the explosion period of the unvoiced plosive [p] in a whispered speech sentence. The temporal span is indicated by the red bar.

were selected for identification: silence, unvoiced plosive, unvoiced fricative and naturally voiced sounds, which includes vowels and voiced consonants. For this purpose a 4-state ergodic HMM was implemented as depicted in Figure 4.3. An ergodic model is one where every state can be reached from any other state. However, in case some transition between two states never occurs, the HMM training will automatically detect it and assign the corresponding transition probability to zero.

Each state has multiple continuous observation values (continuous-valued emissions) consisting of the elements of the feature vector mentioned in the previous section, i.e. 13 MFCC, 13 LSP and 13 PLP coefficients. Since each element of the observation vector (coefficient) follows a single Gaussian distribution this is a HMM whose observations follow a multivariate normal distribution.

In order to train the HMM each frame must be associated with a feature vector and a label. The train function receives these as parameters and computes the transition probabilities between the states as well as the output probability of each feature in each state.

The labels were obtained by manual segmentation of the three whispered speech sentences as described in Ch. 3, considering only the four categories - "Sil", "UP", "UF" and "V". Only two of the three sentences were used for the training part since the third one was used for the classification stage.

4.1.3 Classification

The classification stage corresponds to the test part, i.e. uses the previously trained HMM model to classify each frame of a whispered speech sentence. The classification function receives a

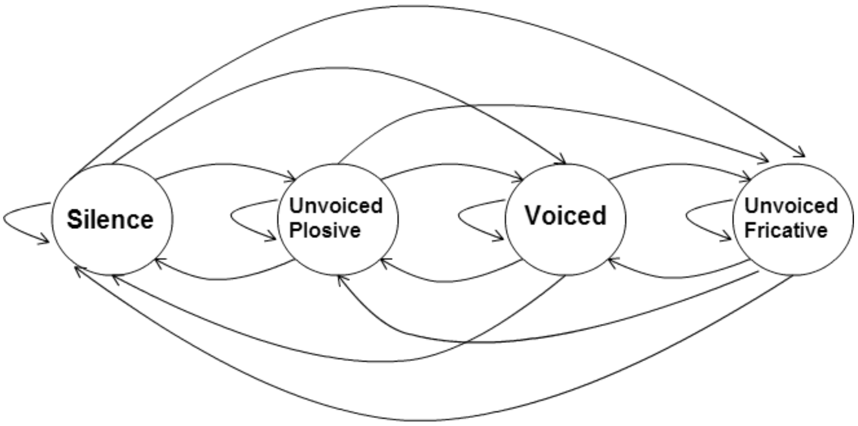


Figure 4.3: HMM state diagram

sequence of feature vectors from the test sentence and generates a state sequence. To that end, this function uses the Viterbi algorithm to find the most likely state associated with each input vector.

A cross-validation was performed at a sentence level since frames can not be scrambled due to the context-dependent nature of the HMM, i.e. the fact that it relies on the information about the sequence of states (transitions between states) to perform the classification. Since we only have three sentences, three tests were made.

The open-source Matlab [30] Toolbox, HMM Toolbox [31], has been used to perform the train and classification functions.

4.2 Evaluation

As mentioned before, three experiments were conducted in order to test the algorithm. One of these will be presented, namely:

- Training sentences: "A Sofia saiu cedo da sala para conversar com o Aurélio."; "O vento norte e o sol discutiam qual dos dois era o mais forte.".
- Test sentence: "O vento norte e sol discutiam qual dos dois era o mais forte quando sucedeu passar um viajante envolto numa capa.".

By using 10ms-long frames with an overlap of 8ms, the test sentence has 4732 frames and the set of training sentences has 4817 frames.

The performance of the algorithm was evaluated using three error measures computed from the confusion matrix. The confusion matrix, $C(i,j)$, shows the number of observations (speech frames) known to be in the state i but predicted to be in state j. It is represented in Table 4.1.

The error probability is given by:

$$Pe(\%) = \frac{\text{number of frames incorrectly classified}}{\text{total number of frames}} * 100 \tag{4.1}$$

	Sil	UF	UP	V
Sil	1412	12	49	205
UF	1	441	1	74
UP	4	15	87	87
V	41	225	179	1899

Table 4.1: Confusion matrix. It shows in each row the number of speech frames pertaining to the state indicated in the first column that have been classified as silence (Sil), unvoiced plosive (UP), unvoiced fricative (UF) or voiced (V).

which is equivalent to sum all the elements of the confusion matrix where $i \neq j$ and divide for the total number of elements. In this case the probability of error is 18,87%.

The intermediary performance measure False Negative Rate (FNR) was considered [32]. A false negative (FN) occurs when an output is incorrectly predicted as negative when it is positive. The FNR is given by:

$$FNR(\%) = \frac{\text{number of frames incorrectly classified as a type different from type X}}{\text{total number of frames of type X}} * 100 \tag{4.2}$$

This is equivalent to say that FNR represents the probability of not detecting a frame of type X. It is obtained by summing the elements of the row of the confusion matrix corresponding to type X where $i \neq j$ and divide for the total number of elements in that row. The results are illustrated in Figure 4.4.

Finally, a final performance measure was considered, the $F-Measure$, [32] defined as:

$$F-Measure(\%) = \frac{2*TP}{2*TP+FP+FN} * 100 \tag{4.3}$$

where TP, FP and FN denote, respectively, the number of True Positives, False Positives and False Negatives. A true positive is represented by the number of correctly classified frames of type X and is obtained by the element of the corresponding row of the confusion matrix where $i = j$. Regarding false positives, they occur when an output is incorrectly predicted as positive when it is negative. For a type X frame, false positives are obtained by summing the elements of the column of the confusion matrix corresponding to type X where $i \neq j$. The $F-Measure$ scores are depicted in Figure 4.5.

As might be observed the unvoiced plosive consonants have a very high probability of not being detected and they are often classified as voiced sounds, which may be related to:

- **The very short duration of the unvoiced plosive.** These consonants have the shortest duration of all speech sounds thus they require a great frame precision;
- **Great influence of coarticulation phenomenon.** The fact that unvoiced plosive have a short time span makes their detection more sensitive to the context they are inserted in, i.e.

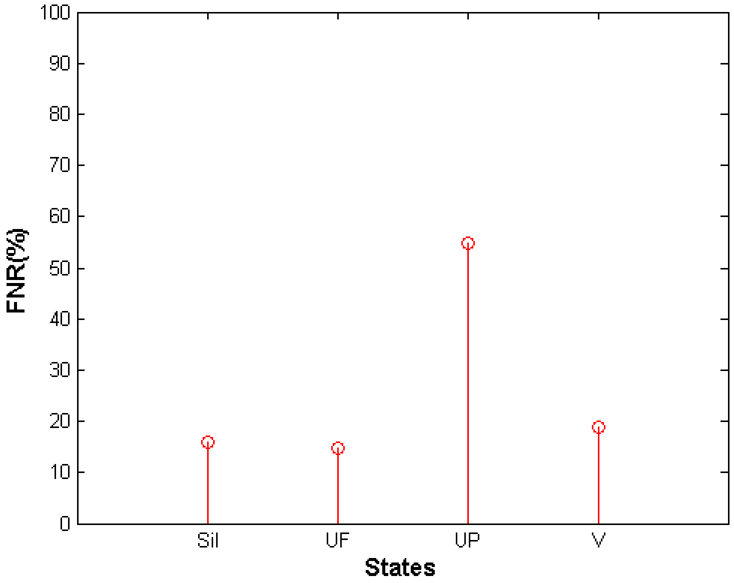


Figure 4.4: FNR results for the automatic classification of the frames. Silence (Sil) - 15,85%; Unvoiced fricative (UF) - 14,70%; Unvoiced plosive (UP) - 54,92% ; Voiced (V) - 18,98%.

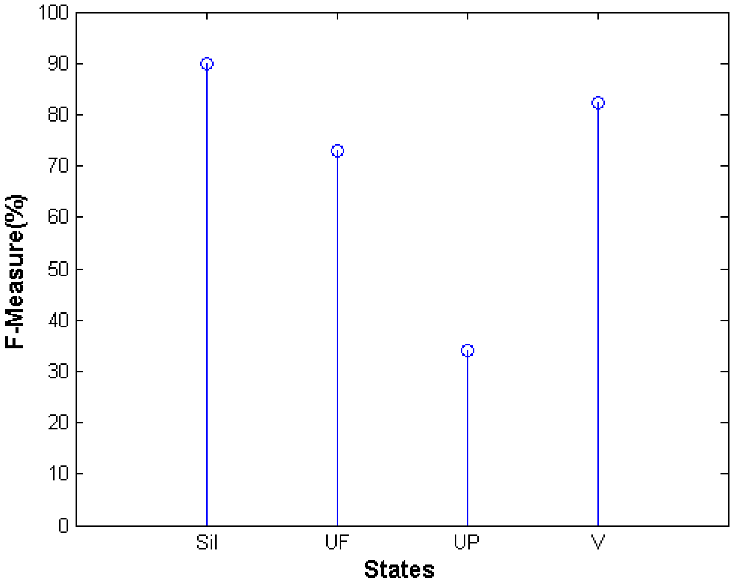


Figure 4.5: $F - measure$ scores for the automatic classification of the frames. Silence (Sil) - 90,05%; Unvoiced fricative (UF) - 72,89%; Unvoiced plosive (UP) - 34,18% ; Voiced (V) - 82,40%.

	Sil	UF	UP	V
Sil	1317	1	6	139
UF	6	604	26	62
UP	0	13	106	23
V	25	117	167	2120

Table 4.2: Confusion matrix obtained after fine-tuning of the results. It shows in each row the number of speech frames pertaining to the state indicated in the first column that have been classified as silence (Sil), unvoiced plosive (UP), unvoiced fricative (UF) or voiced (V).

from the preceding and following speech sounds. Actually, the preceding sounds are not quite harmful as the plosive are usually preceded by silence. The following sounds in turn, are usually voiced sounds which result in higher detection errors as can be seen in Table 4.1;

- **Very little information of the database.** The information of the database encompasses few examples of unvoiced plosive in several coarticulation contexts to train the HMM.

The problem identified in the second point could be overcame through the increased performance due to some modifications to the manual labelling. In fact, it was concluded that some errors could be directly related to the manual labelling as will be described in the next section.

Another encountered issue was related to the contents of Table 3.1. In fact, after some experiments with different versions of the labeled data, it was concluded that in the particular case of the studied patient, not all the depicted transformations occurred.

4.2.1 Fine-tuning of the Results

With the conclusions of the first experiments in mind, the following modifications were made in the manual labeling of the speech data:

- Shortening of the unvoiced plosive duration to a time span of [20;28] ms so as to reduce the mentioned coarticulation phenomenon;
- Decrease of the amplitude threshold below which signal segments are labeled as silence;
- Modification of the labelling according to the transformations depicted in Table 3.1 that actually occur in the case of the studied patient.

The results obtained after performing the modifications are shown in the confusion matrix represented in Table 4.2. Now, we achieved an error probability of 12,36% and FNR and $F-Measure$ scores as illustrated in Figures 4.6 and 4.7 respectively.

As might be observed, the FNR scores have decreased for all the types of speech sounds. Particularly in the case of unvoiced plosive it has decreased by 50%. In fact, almost all unvoiced plosive are detected. As mentioned before, these consonants are preceded by silence and the HMM can successfully detect the transition between these states. However, this not apply to the following speech sounds, which are usually voiced. Thus, the HMM successfully classifies the first frames of the unvoiced plosive but misses the last frames, hence the high error rate.

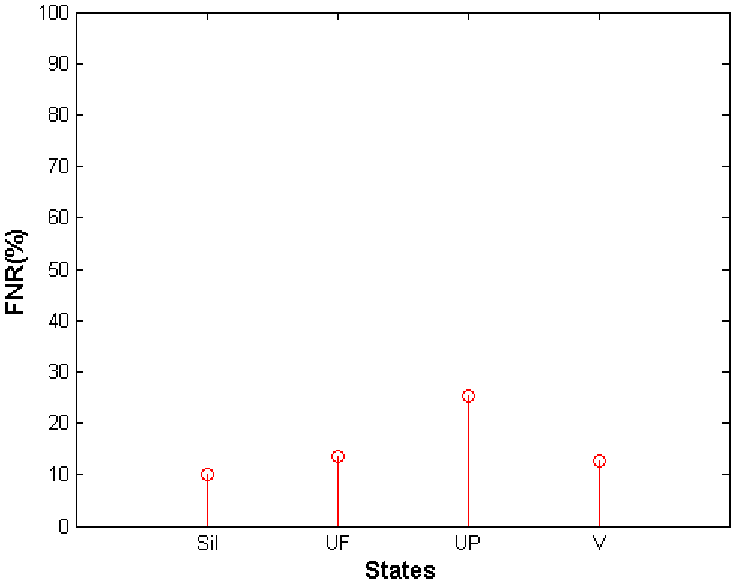


Figure 4.6: *FNR* results, obtained after fine-tuning of the results, for the automatic classification of the frames. Silence (Sil) - 9,98%; Unvoiced fricative (UF) - 13,47%; Unvoiced plosive (UP) - 25,35%; Voiced (V) - 12,72%.

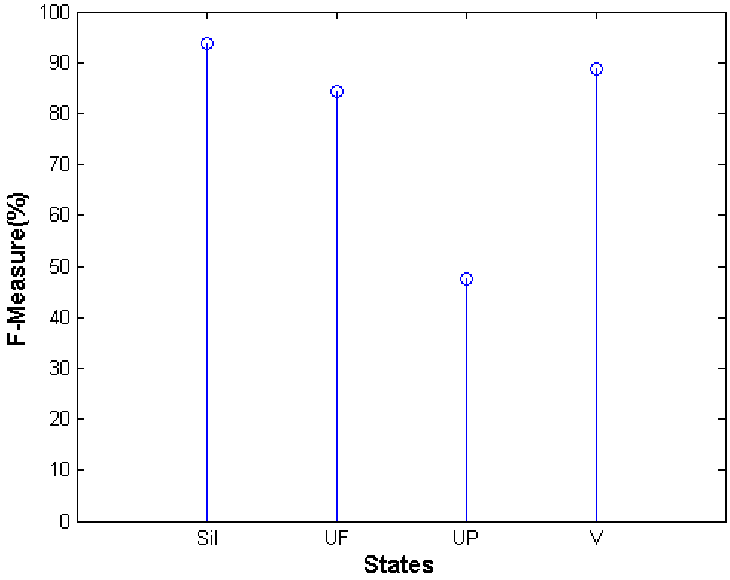


Figure 4.7: $F - measure$ scores, obtained after fine-tuning of the results, for the automatic classification of the frames. Silence (Sil) - 93,70%; Unvoiced fricative (UF) - 84,30%; Unvoiced plosive (UP) - 47,43% ; Voiced (V) - 88,83%.

4.3 Final Considerations

In spite of the small variety and quantity of phonemes available in the speech corpus, the HMM classifier produced promising results. Even so, only a perceptual evaluation of an actual synthesized sentence would allow for a definitive assessment of the performance of the HMM. In fact, for the chosen values of overlap and frame size, it is possible that the impact of this number of errors in the final result is perceptually negligible.

Additionally, it is important to note that the aim is to obtain a binary automatic segmentation, i.e. a voiced/unvoiced frame classification. Thus, all the frames of type "silence", "unvoiced plosive" and "unvoiced fricative" will be identified as unvoiced frames. The frames of type "voiced" will be identified as frames where artificial voicing should be implanted. This will certainly reduce the classification errors.

Chapter 5

Proposed Algorithm for Whisper-to-Speech Conversion

In this chapter the proposed algorithm for whisper-to-speech conversion is presented along with the explanation of its main stages.

5.1 Algorithm Overview

An overview block diagram for the proposed system is shown in figure 5.1 which operates on a frame-by-frame basis with 1024 samples and 50% overlap.

The frame of the whispered signal, $x[n]$, is first multiplied by the window $h[n]$ and transformed into the ODFT domain [33]. The spectral envelope of the signal is computed through its cepstrum and the cepstral parameters are transformed through the use of an estimated transformation function. The aim is to perform the projection of the whispered envelope so as to generate the envelope of the target voiced frame. Thus, it is possible to extract the magnitudes of the harmonic partials from the projected envelope. In order to extract the magnitudes, one has to start by computing the locations of the partials, which are represented by $\boldsymbol{\ell}$ and $\Delta\boldsymbol{\ell}$, which are, respectively, the integer part and the fractional part of a DFT bin [33]. The locations are obtained by computing the integer multiples of the fundamental frequency, F_0. Regarding the value of F_0, it was obtained by simply computing the average of the fundamental frequency value of the original sentences of the normal sounding speech database and adding a cumulative random normal distributed value, with a mean of 0 and a standard deviation of 2,7Hz, in each frame transition. The F_0 average obtained was 110Hz. The standard deviation value was obtained through fine-tuning of the perceived variation of pitch, after several experiments.

The next stage is to synthesize the phases of the harmonic sinusoids, ϕ_l. For this purpose, we first predict the phase of the fundamental frequency on the current frame, ϕ_0, and estimate a model that describes the relative delays between the partials (NRDs model). Together, these two steps provide with the necessary information to compute the phases.

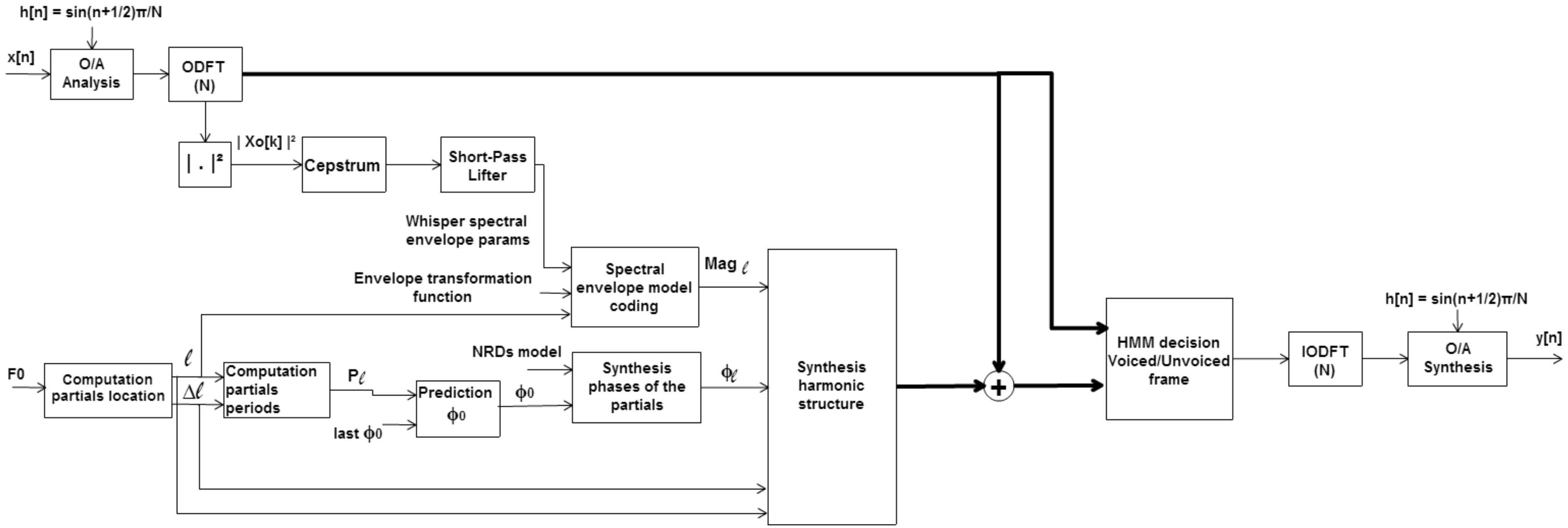

h[n] = sin(n+1/2)π/N
x[n]
O/A Analysis
ODFT (N)
| . |²
| Xo[k] |²
Cepstrum
Short-Pass Lifter
Whisper spectral envelope params
Envelope transformation function
Spectral envelope model coding
Mag ℓ
F0
Computation partials location
ℓ
Δℓ
Computation partials periods
Pℓ
Prediction ϕ0
last ϕ0
ϕ0
NRDs model
Synthesis phases of the partials
ϕℓ
Synthesis harmonic structure
+
HMM decision Voiced/Unvoiced frame
IODFT (N)
h[n] = sin(n+1/2)π/N
O/A Synthesis
y[n]

Figure 5.1: Block diagram of the proposed algorithm for whisper-to-speech conversion.

After obtaining the locations, the magnitudes and the phases of the partials, we can finally synthesize the harmonic sinusoids and generate a harmonic structure. The artificially voiced frame is transformed back to the time domain through IODFT and overlap-added to generate the final reconstructed signal, $y[n]$.

These operations only apply in the case of frames classified as voiced by the HMM. The frames classified as unvoiced are obtained directly from the whisper signal, i.e. are overlap-added to the final artificially voiced signal, $y[n]$.

As might be observed, this is a parametric approach, where the required information to synthesize the harmonic partials is obtained through the estimation of models that are then applied to extracted parameters of the whispered signal in order to convert them into those of normal sounding speech.

The methods used to compute the mentioned stages will be described in detail in the following sections.

5.2 Transformation of the Spectral Envelope

The spectral envelope is the amplitude spectrum of the filter that models the vocal tract and glottal source spectrum and in spectral transformations it is used to manipulate the amplitudes of the harmonics. The purpose of this study, is to convert the spectral envelope characteristics of a whispered sentence in those of a voiced sentence on a frame-by-frame basis. The final objective is to extract the magnitudes of the harmonic partials, to be used in the synthesis algorithm, from the converted envelope.

5.2.1 Estimation

The estimation of the spectral envelope of a speech frame was performed through the cepstrum method. As described in a previous chapter, the cepstrum allows for the separation of the source spectrum and the filter, which is a good estimation of the spectral envelope, according to the source-filter model of speech production. The underlying assumption is that the source spectrum (excitation signal) contains rapid variations and the vocal tract filter contains the smooth part of the speech signal. Thus, the filter contribution is concentrated in the lower regions of the cepstrum and therefore, only the first p cepstral coefficients are kept, where p represents the order of the cepstrum. These first p cepstral coefficients are obtained through the use of a Short-Pass Lifter in the quefrency domain, which acts as a low pass filter in the frequency domain. Therefore, this operation results in smoother signal when converted from the quefrency domain to the frequency domain, representing the spectral envelope.

In order to accurately estimate the spectral envelope of the voiced frames [1], some modifications have been introduced in the cepstrum method of spectral envelope estimation. The envelope

[1] In this section, voiced frames refer to those pertaining to natural sounding speech and unvoiced frames to those pertaining to whispered speech.

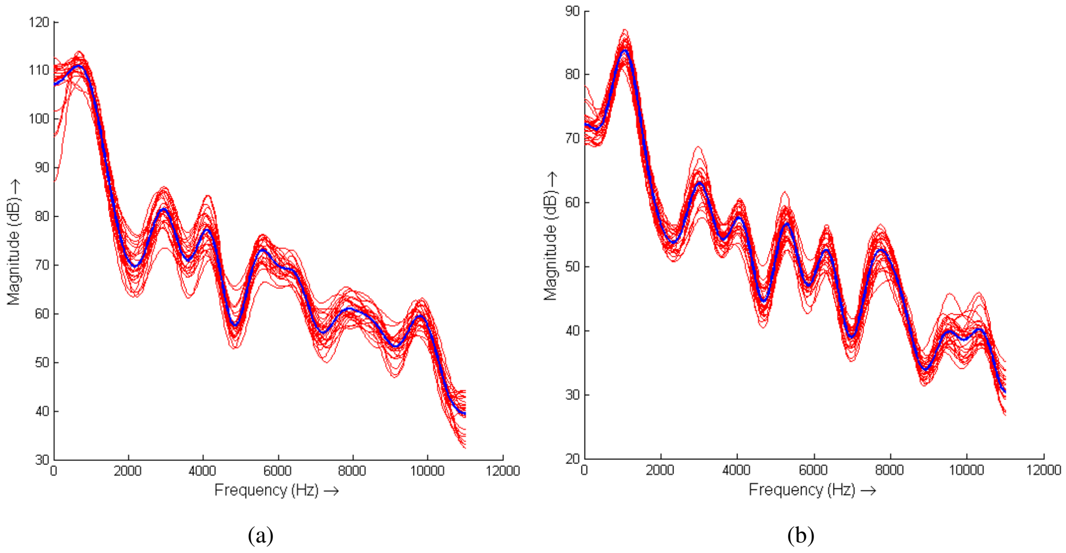


Figure 5.2: (a) Spectral envelopes for each frame of the voiced oral vowel [a] (thin red lines). The blue thick solid line represents the average spectral envelope. (b) Spectral envelopes for each frame of the whispered oral vowel [a] (thin red lines). The blue thick solid line represents the average spectral envelope.

should pass through the estimated amplitudes of the harmonics of the signal, however the cepstrum computation acts as a low-pass filtering of the spectrum and thus, smooths some important irregularities and amplitudes of the harmonics. Besides, when the spectral peaks, which are most often the harmonic partials, are spaced too far from each other, the spectral envelope will descend down to the residual noise level, which is not a good estimation. The problem is addressed by first computing the magnitudes of the found spectral peaks and then performing an interpolation to link them. The cepstral analysis is finally applied to the interpolated envelope.

The same reasoning can not be applied to the case of the unvoiced frames since they do not have a harmonic structure and thus, the cepstrum method can be computed straightforwardly.

The estimation of the spectral envelopes was computed only for 8 oral vowels of the EP ([a], [ɐ], [ɛ], [ɨ], [i], [ɔ], [o], [u]), in both speech modes,[2] due to the small amount of information in the speech database. The vowels were uttered in sequence and in a sustained way, with silence before and after each vowel sound. It is important to note that the two modes of each vowel - voiced and whispered - were produced consecutively in the same sentence to ensure that the recording conditions, as well as the articulatory gesture were the same. The recorded vowel sounds, with 22,05kHz sampling frequency, were manually segmented in order to extract only samples from a stationary region, i.e. a region with stable time plot and sound. The analysis involves frames with 1024 samples and 50% overlap.

The aim was to conclude on common spectral envelope behaviours among different vowels and establish a general solution for the transformation.

[2] These samples are uttered by the patient suffering from spasmodic dysphonia mentioned throughout this dissertation.

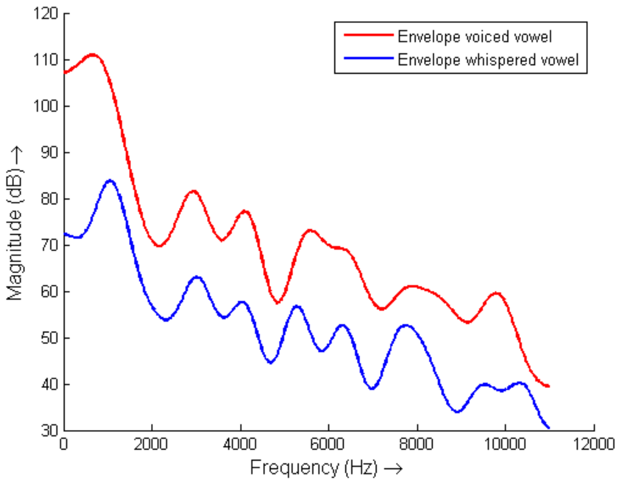


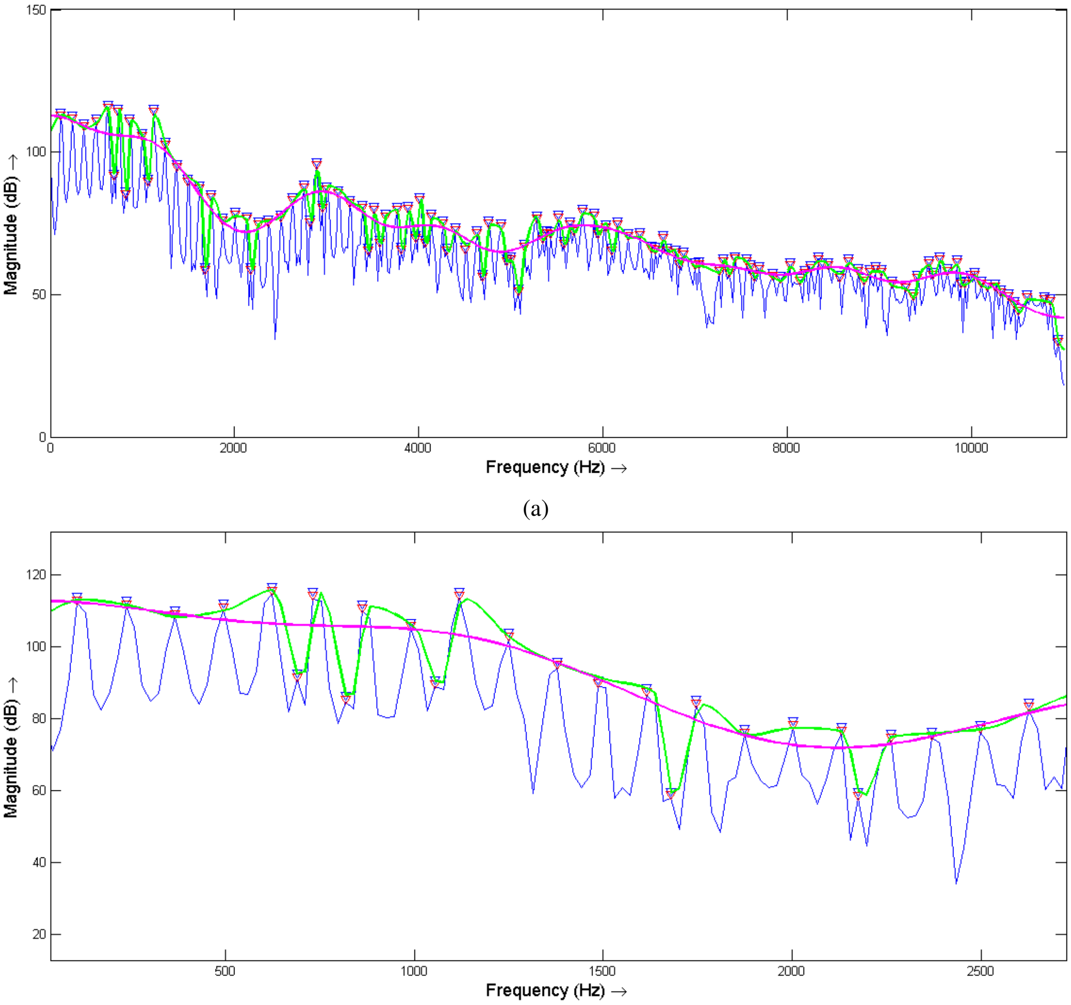
Figure 5.3: Average spectral envelope of the voiced oral vowel [a] (red line) and the whispered counterpart (blue line) derived from 22th-order cepstral analysis.

Figure 5.2 depicts the spectral envelopes for each frame of the vowel [a], uttered both in voiced and whispered modes, as well as the average spectral envelopes. The results were obtained using 22 cepstral coefficients. Figure 5.3 depicts the two average spectral envelopes jointly in order to better understand their relationship.

However, the resulting envelope presented some problems: it provides a very smooth representation of the speech signal and it passes through all the found spectral peaks and not only through those pertaining to the harmonic structure. Therefore, spectral envelope anomalies may arise. This is observed in Figure 5.4a, which illustrates the result of the algorithm for one frame of the voiced vowel [a]. It shows the short-time power spectral density of the signal, the interpolated envelope based on the magnitudes of all the found spectral peaks and the final smooth spectral envelope model derived from a 22th-order cepstral analysis. The observed blue triangles identify the magnitudes and frequency bins of the found spectral peaks and the red triangles identify the spectral peaks considered in the envelope computation. In this approach the blue and red triangles always coincide. Figure 5.4b is a zoomed version of the Figure 5.4a in order to illustrate the problem of using all the spectral peaks in the computation of the signal envelope. The small peaks observed do not pertain to the harmonic structure of the voiced vowel, resulting in an envelope (magenta line) which does not respect the magnitudes of the harmonic partials.

To address these problems, the search for the harmonic partials was improved and thus, only the spectral peaks pertaining to the harmonic structure of the voiced vowel were considered. Besides, a linear interpolation and a 50th-order cepstral analysis were performed in order to reduce the smoothing. As observed in Figure 5.5, the blue and red triangles no longer coincide, where only the red ones are used in the computation of the spectral envelope. Comparing the Figure 5.4b and Figure 5.5b the impact of the small peaks on the accuracy of the envelope is evident, since the latter now respects the magnitudes of the harmonic partials.

Figures 5.6 and 5.7a illustrate the resulting average spectral envelopes, in both speech modes,



(a)

(b)

Figure 5.4: (a) Short-time power spectral density of the voiced vowel [a] (blue line), interpolated envelope based on the magnitudes of all the found spectral peaks (green thick line), final spectral envelope model derived from 22 cepstral coefficients (magenta thick line), found spectral peaks (blue triangles) and considered spectral peaks (red triangles). In this case, red and blue triangles coincide. (b) Zoomed version of the previous figure depicting the problem of considering all spectral peaks.

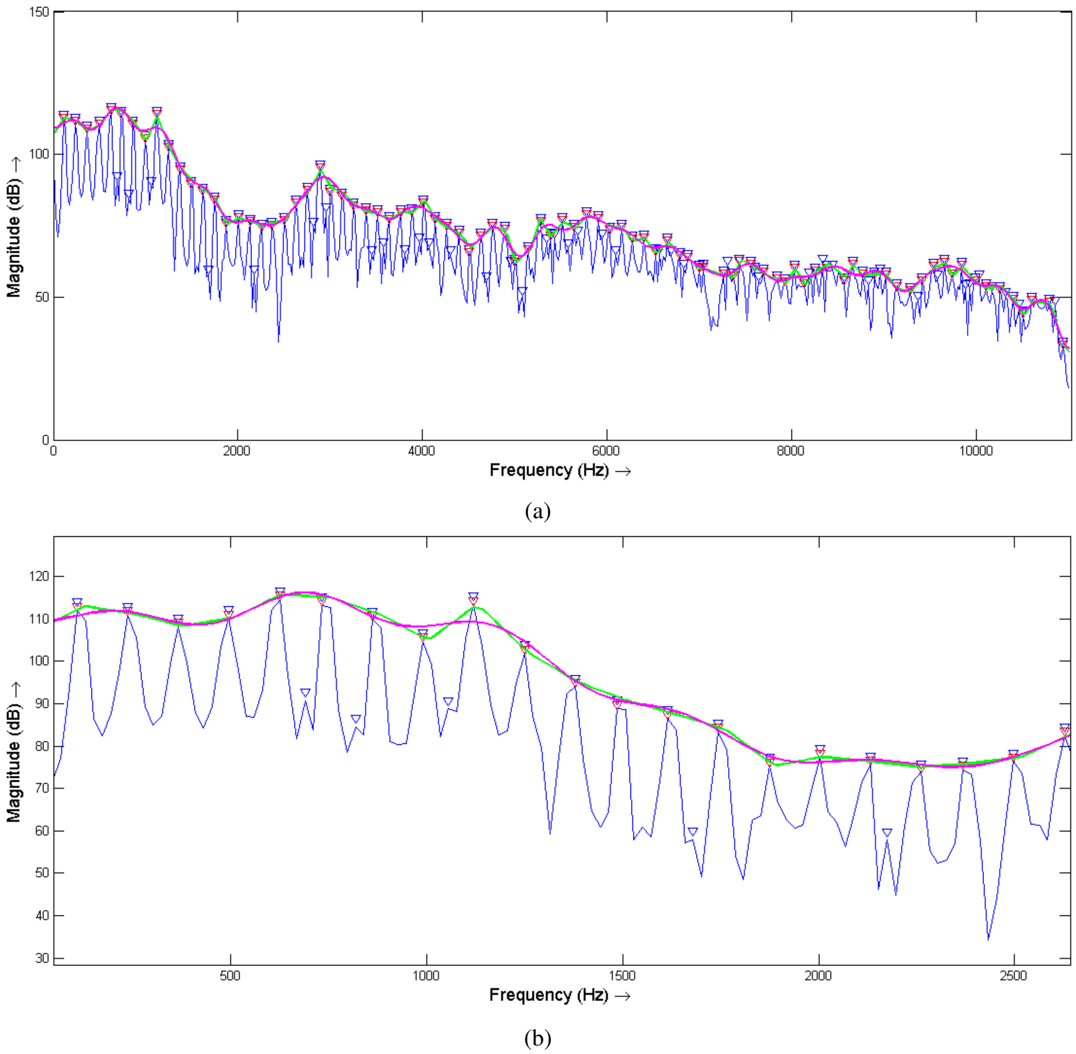


Figure 5.5: (a) Short-time power spectral density of the voiced vowel [a] (blue line), interpolated envelope based on the magnitudes of the spectral peaks pertaining to the harmonic structure (green thick line), final spectral envelope model derived from 50 cepstral coefficients (magenta thick line), found spectral peaks (blue triangles) and actually considered spectral peaks (red triangles). (b) Zoomed version of the previous figure depicting the improvement on the spectral envelope computation.

of the oral vowel [a] using the improved method of spectral envelope estimation. Figure 5.7b shows the average spectral envelopes of the oral vowel [ɨ]. These average spectral envelopes are the ones which will be used in the search for a transformation method, as well as its computation.

As might be observed, the envelope of the whispered vowel presents a similar shape to that of the voiced counterpart, preserving the formants (prominences in the spectral envelope) especially at low frequencies, which are the most relevant to the human ear. In particular, the magnitudes of the frequencies falling below about 2kHz in the voiced case can basically be obtained by a shift of the whispered ones. Furthermore, in the range of 2kHz to 4kHz, there is a visible preservation of the formants in the whispered envelope with respect to the voiced one. In fact, these observations are supported by [34], which introduces the concept of perceptual spectral clusters (PSC) of harmonic partials as a feature for isolated vowel identification. The underlying concept is that the human auditory system (HAS) discriminates a voiced sustained vowel from its pitch and timbre. The latter is characterized by the spectral power of the partials and thus, an identification of the harmonic structure is required. The PSC concept assumes that the HAS performs a perceptual integration of partials pertaining to the same harmonic structure and therefore, searches for clusters of harmonic partials responsible for the identification of each vowel. In other words, it is assumed that HAS differentiates among vowels by carrying out a perceptual clustering of partials in the harmonic structure of the vowel, hence the name PSC. The work concluded that only two spectral clusters (with highest average magnitude) of adjacent harmonic partials are necessary to clearly recognize a vowel, i.e. are relevant to preserve vowel identity [3]. In fact, the first PSC, falling below about 2kHz, is the most relevant in vowel identification; the second PSC is situated between 2kHz and 4kHz and its average magnitude represents the most important feature in vowel identity.

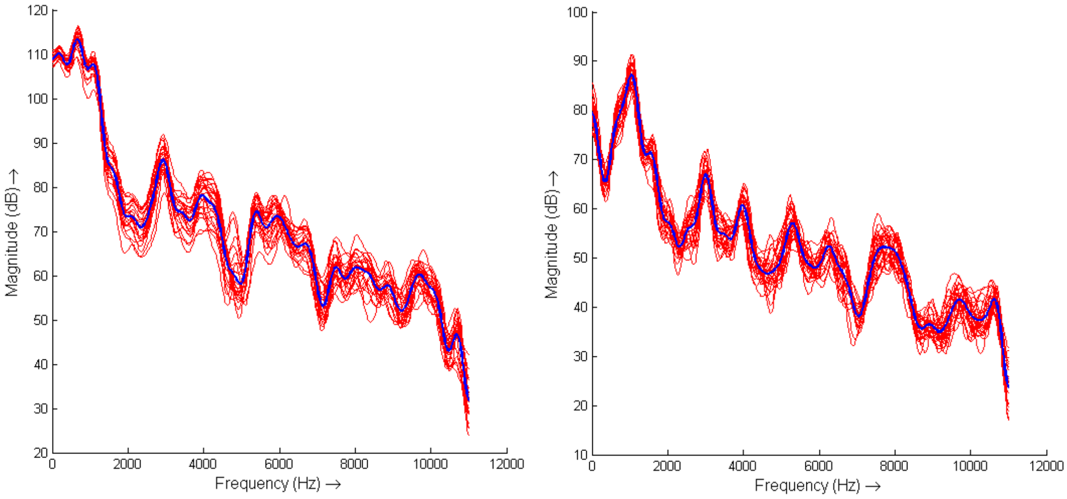
Both in Figure 5.3 and Figure 5.7 it is observed that the first formant of the whispered vowel is higher in frequency than that of the voiced vowel, which is supported by [35].

Finally, in order to illustrate the spectral similarities of naturally unvoiced consonants when they are uttered in whispered and voiced speech modes, Figure 5.8 depicts the two spectral envelopes of the unvoiced fricative [f], as manually segmented from the Portuguese word "Sofia". It shows a small magnitude shift and a very similar shape. Since the unvoiced plosive consonants have a too short duration it is not possible to extract information from a stationary region and thus, this analysis is not valid.

5.2.2 Envelope Projection - Transformation Function

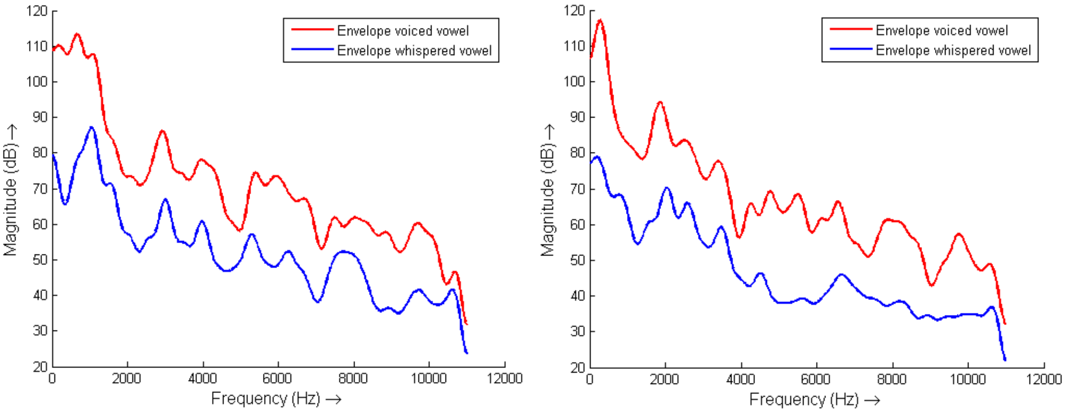
The discussion will focus on the design of a model, namely a transformation function, to convert an original spectral envelope (pertaining to whispered speech frames) into a target one (pertaining to normal sounding speech frames). To that end, 13 pairs of average spectral envelopes (whispered and voiced counterpart) of vowel sounds, obtained as described above, were used - two pairs pertaining to each of the oral vowels [a], [ɛ], [i], [ɔ], [u] extracted from different audio files, one pair pertaining to each of the oral vowels [ɐ], [ɨ], [o].

[3] The study was carried out using five EP vowels.



(a) (b)

Figure 5.6: (a) Spectral envelopes for each frame of the voiced oral vowel [a] (thin red lines). The blue thick solid line represents the average spectral envelope. (b) Spectral envelopes for each frame of the whispered oral vowel [a] (thin red lines). The blue thick solid line represents the average spectral envelope.



(a) (b)

Figure 5.7: (a) Average spectral envelope of the voiced oral vowel [a] (red line) and the whispered counterpart (blue line) derived from 50th-order cepstral analysis. (b) Average spectral envelope of the voiced oral vowel [i] (red line) and the whispered counterpart (blue line) derived from 50th-order cepstral analysis.

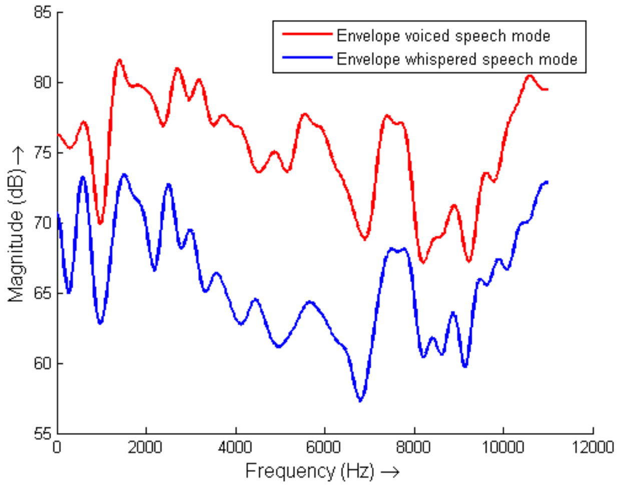


Figure 5.8: Average spectral envelope of the naturally unvoiced fricative consonant [f] uttered in voiced speech mode (red line) and uttered in whispered speech mode (blue line) derived from 50th-order cepstral analysis.

The first proposed approach consisted of the following steps:

1. Compute the absolute difference between the whispered and the voiced average spectral envelopes for each vowel sound;

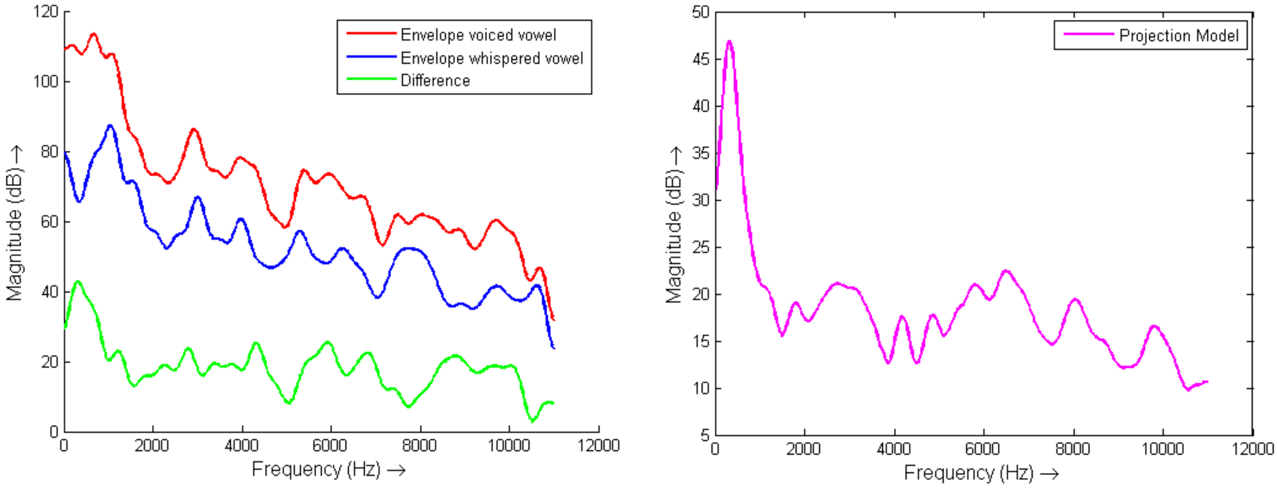
2. Compute the average of the obtained differences, which represents the projection model;

3. Sum the model to the average spectral envelopes of each whispered vowel and obtain the projection.

Each of these steps is depicted in Figures 5.9a, 5.9b and 5.10 respectively. The first step is illustrated only for the case of the oral vowel [a] and the last step for the cases of the oral vowels [a] and [i].

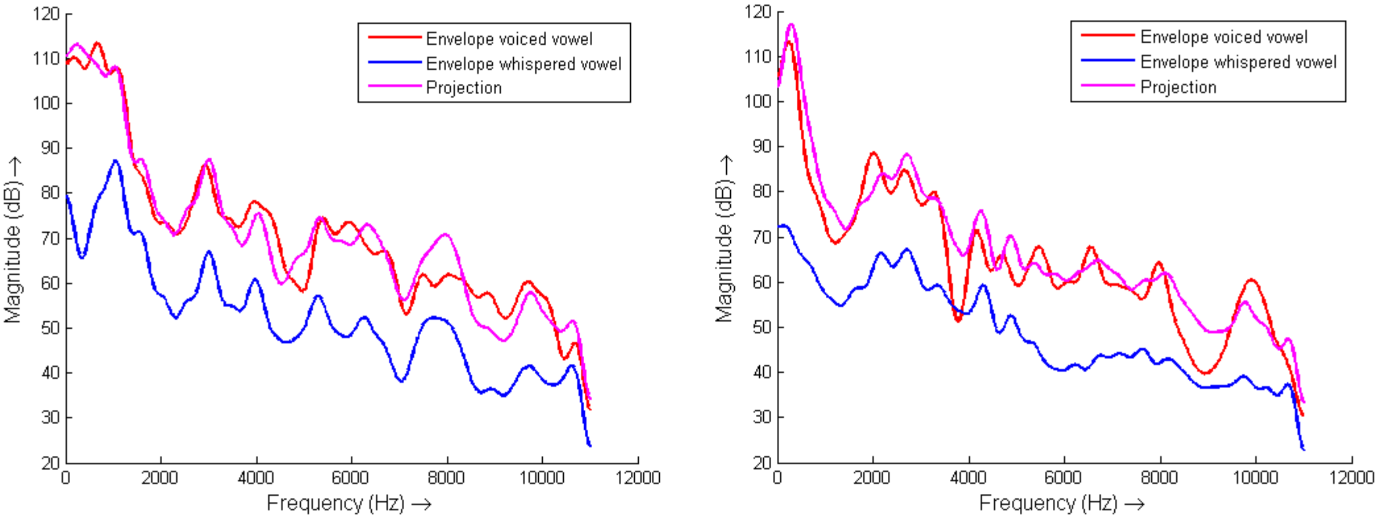
Despite the simplicity of the approach the results were not negligible. However, a more accurate method was searched, namely Statistical Modeling through the use of Gaussian Mixture Model (GMM). This method is usually used in voice conversion systems [36, 37, 38] to modify the speaker voice. In this case, it aims to represent the statistical relations between the two spectral envelopes of a sound uttered in the two speech modes - whispered and voiced - by an appropriate model (transformation function). This model is trained from experimental data, which consists of two sets of spectral vectors, $X_{pxN} = [x_1, x_2, ..., x_N]$ and $Y_{pxN} = [y_1, y_2, ..., y_N]$, from source (whispered) and target (voiced) signals respectively. Each set has N vectors (spectral envelopes of the frames of the vowel sounds) and each vector has p dimension (50 cepstrum coefficients that represent the spectral envelope). The source vectors should be aligned with the target ones to describe the same acoustic content.

Thus, a function ***F()*** is desired that better converts each vector of the source data set into its counterpart in the target data set.



(a) (b)

Figure 5.9: (a) Result of the absolute difference between the two average envelopes of the oral vowel [a] (green line). (b) Envelope projection model.



(a) (b)

Figure 5.10: (a) Result of the projection of the average spectral envelope of the whispered oral vowel [a] (magenta line). (b) Result of the projection of the average spectral envelope of the whispered oral vowel [i] (magenta line).

Hereinafter, some basic concepts about the proposed method will be explained. For a deep understanding about this study the reader should refer to [36, 37].

The probability density function of a random variable x follows a Q-order GMM model is given by [36]

$$p(x) = \sum_{i=1}^{Q} \alpha_i N(x;\mu_i,\Sigma_i), \tag{5.1}$$

with

$$\sum_{i=1}^{Q} \alpha_i = 1 \quad \text{and} \quad \alpha_i \geq 1, \tag{5.2}$$

where i=1,...,Q are the component densities, α_i are the mixture weights and $N(x;\mu_i,\Sigma_i)$ is the probability density of the normal distribution with p-by-1 mean vector μ_i and p-by-p covariance matrix Σ_i.

The approach described herein performs a regression using GMM. A regression is a statistical process that measures the relationship among variables and searches for a mathematical relation to describe it using a least squares error criterion. After computing the regression model, it is possible to predict the value of a variable given the known values. This regression model is the transformation function responsible for relating the source and target vectors. Therefore, in order to estimate this function, the GMM is used to model the joint probability density of the source and target vectors. The following conversion function is proposed [36, 37]:

$$\hat{y} = F(x) = E(Y|X=x) = \sum_{i=1}^{Q} p(q_i|x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x-\mu_i^x)], \tag{5.3}$$

where x and y are the source and target spectral vectors respectively, $p(q_i|x)$ is the posterior probability that x is generated by the i^{th} component density, μ_i^y and Σ_i^{yx} are respectively the mean target vector and the cross-variance matrix of the source and target vectors, Σ_i^{xx} is the covariance matrix of the source vectors and μ_i^x is the mean source vector.

This function is determined by minimizing the mean square error:

$$E(F) = \sum_{i=1}^{N} ||y_i - F(x_i)||^2, \tag{5.4}$$

where N is the number of vectors.

However, in the methodology adopted, the GMM was reduced to a single component density and thus, the transformation function becomes [37]:

$$F(x) = \mu_y + \Sigma_{yx}(\Sigma_{xx})^{-1}(x-\mu_x), \tag{5.5}$$

The performance of the proposed function was judged from the spectral differences between the transformed spectral envelopes and the target envelopes on a frame-by-frame basis:

$$E = \frac{1}{N}\sum_{i=1}^{N}\frac{||y_i - \hat{y}_i||}{||y_i||}, \tag{5.6}$$

where E is the mean relative error, N the number of vectors (frames), y the target vectors and $\hat{y}$ the vector of the transformed spectral envelopes.

The results of the envelope conversion using the transformation function in Equation 5.5 are illustrated in Figure 5.11 for the oral vowels [a], [ɔ], [ɨ], [u]. As might be observed, the results are surprisingly good with an estimated mean relative error of 4,32%, which is equivalent to -13.65 dB.

5.3 Normalized Relative Delays Model

5.3.1 NRD Concept

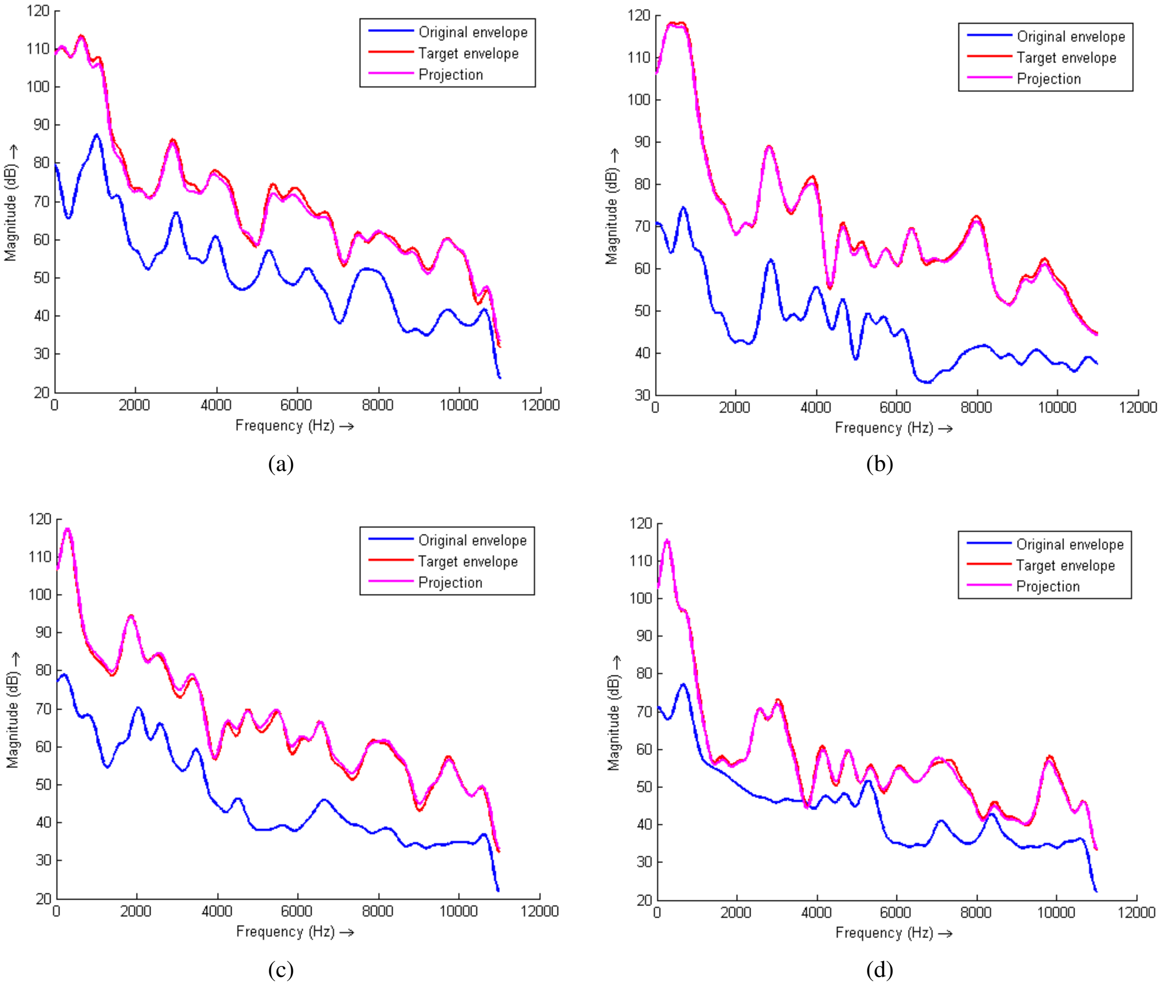
The Normalized Relative Delays (NRDs) of harmonic partials are introduced in [39] as a phase-related feature that is obtained from the harmonics of a speech signal. The importance of the phase in preserving the acoustic signature of a speaker and the quality of speech is emphasized. The NRDs represent the relative delays between the harmonic partials of a periodic sound allowing implementation of phase synchronization, independently of the overall time delay of the waveform and of its fundamental frequency. Thus, they characterize the shape of the speech waveform, i.e. they insure shape invariance.

The analysis presented hereinafter follows the discussion in [39]. Therefore, the interested reader should refer to this document for further details.

Assuming the quasi-periodic signal $x[n]$ with fundamental frequency ω_0 and its L harmonics:

$$\begin{aligned} x[n] &= A_0 sin(n\omega_0 + \phi_0) + \sum_{\ell=1}^{L-1} A_\ell\, sin(n\omega_\ell + \phi_\ell) \\ &= A_0 sin[\omega_0(n+n_0)] + \sum_{\ell=1}^{L-1} A_\ell\, sin[\omega_\ell\,(n+n_\ell)], \end{aligned} \tag{5.7}$$

where A_ℓ, ϕ_ℓ and n_ℓ represent, respectively, the magnitude, phase and time delay of the ℓ^{th} harmonic sinusoid relative to a reference point in n. The latter usually corresponds to the center of the window used before the transformation of the signal to the frequency domain. Therefore, let $X[k]$ denote the complex transform of $x[n]$ after windowing, its phase represents the time delay n_k



(a)

(b)

(c)

(d)

Figure 5.11: (a) Result of the projection of the average spectral envelope of the whispered oral vowel [a] (magenta line), obtained by regression. (b) Result of the projection of the average spectral envelope of the whispered oral vowel [ɔ] (magenta line), obtained by regression. (c) Result of the projection of the average spectral envelope of the whispered oral vowel [ɨ] (magenta line), obtained by regression. (d) Result of the projection of the average spectral envelope of the whispered oral vowel [u] (magenta line), obtained by regression.

relative to the center of the window. If n is omitted the NRD concept arises as follows:

$$\begin{aligned} x &= A_0 sin\omega_0 n_0 + \sum_{\ell=1}^{L-1} A_\ell \, sin\omega_\ell \, n_\ell \\ &= A_0 sin2\pi\frac{n_0}{P_0} + \sum_{\ell=1}^{L-1} A_\ell \, sin2\pi(\frac{n_0}{P_\ell} + \frac{n_\ell - n_0}{P_\ell}) \\ &= A_0 sin2\pi\frac{n_0}{P_0} + \sum_{\ell=1}^{L-1} A_\ell \, sin2\pi(\frac{n_0}{P_\ell} + NRD_\ell), \end{aligned} \tag{5.8}$$

where P_ℓ denotes the period of the ℓ^{th} harmonic sinusoid and NRD_ℓ represents the delay difference between the ℓ^{th} sinusoid and the fundamental frequency, relative to the period of that sinusoid.

The period P_ℓ and the time delay n_ℓ of the ℓ^{th} sinusoid are given by:

$$\begin{aligned} P_\ell &= \frac{2\pi}{\omega_\ell} \\ n_\ell &= \frac{\phi_\ell}{\omega_\ell}. \end{aligned} \tag{5.9}$$

Thus, the NRDs are defined by the magnitude and phase information provided by $X[k]$, as the periods of the different sinusoids are estimated from the magnitude spectrum and the time delays are obtained from the phase spectrum. Since the estimated relative delays are normalized, i.e. $0.0 \leq NRD_\ell < 1.0$, they are obtained as follows:

$$\begin{aligned} NRD_\ell &= \frac{n_\ell - n_0}{P_\ell} + \lfloor \frac{n_0}{P_\ell} \rfloor \\ \text{if} \quad NRD_\ell < 0.0, \quad NRD_\ell &= NRD_\ell + 1 \end{aligned} \tag{5.10}$$

where $\lfloor . \rfloor$ denotes the largest integer.

The delay of the fundamental frequency, n_0, is relative to the overall time-analysis reference, represents the overall time shift of a periodic pattern of the speech waveform. This is the only variable time information, as the delays of the harmonic sinusoids are relative to it. The NRD concept allows to preserve the shape invariance since the magnitude and frequency relations among the harmonic partials are preserved. Therefore, the NRD denotes the time waveform of a periodic signal. In fact, two different signals may have the same magnitude spectrum but have different NRD profiles because they have different waveform shapes (time envelopes) [39].

The speech signal quality may be improved by combining spectral magnitude information and phase information provided by NRDs as NRDs denote idiosyncratic information.

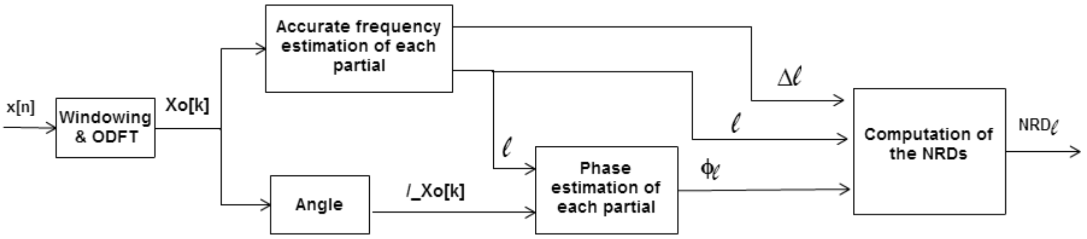


Figure 5.12: Block diagram of the algorithm for the estimation of the NRD parameters.

5.3.2 Model Estimation

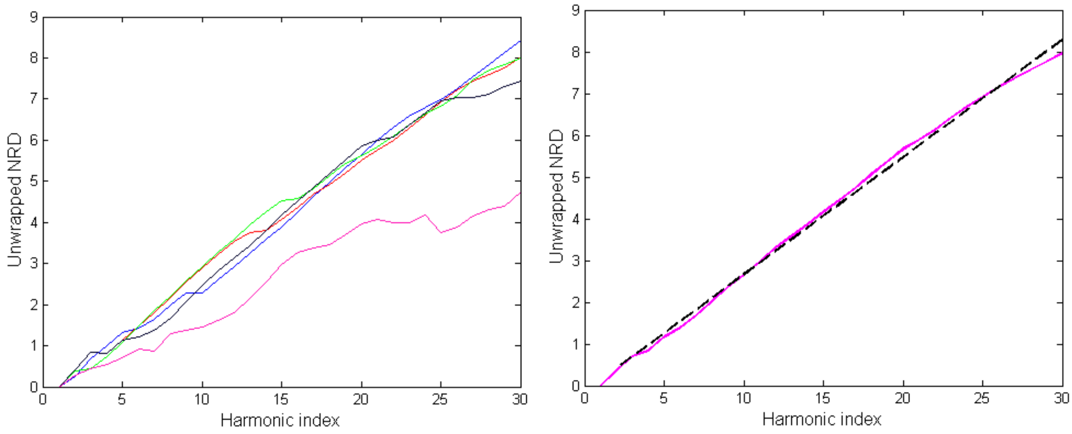
The NRD model was obtained through the estimation of the NRD parameters pertaining to the five EP oral vowels [a], [ɛ], [ɨ], [ɔ], [u] uttered in a sustained way in voiced speech mode by the patient under study. The sampling frequency of the signals is 22050Hz and the NRD analysis concerns frames with 1024 samples and 50% overlap. Each frame generates a NRD feature vector consisting of the first 30 NRD coefficients. The algorithm for their estimation is depicted in Figure 5.12. The parameters ℓ and $\Delta\ell$ are, respectively, the integer part and the fractional part of a DFT-type frequency bin scale of each partial. The integer parts are obtained from the frequency bins of the spectral peaks of the most prominent harmonic structure of the signal. The estimation of the fractional part, as well as the estimation of the phase of each partial, are performed as explained in [33].

The estimation of the NRD model consists of the following steps:

1. Compute de NRD feature vectors for each vowel as illustrated in the block diagram of the Figure 5.12;
2. Compute the average of the obtained NRD feature vectors for each vowel (see Figure 5.13a);
3. Compute the total average of the NRD coefficients of the vowels (see magenta line in Figure 5.13b);
4. Obtain the NRD model by performing a linear regression of the total average (see black dotted line in Figure 5.13b).

Since the NRDs have the same properties of phase, the concepts of periodicity, wrapping and unwrapping also apply [39] and thus, in Figure 5.13 unwrapped NRDs are represented.

In fact, the total average of the NRD coefficients in the second step, was computed considering only the vowels [a], [ɛ], [ɨ], [ɔ]. As can be observed in Figure 5.13a, the vowel [u] has a very different NRD pattern from the rest of the vowels. Indeed, this behaviour was expected due to the different nature of the spectrum of this vowel, which is mostly concentrated in the lower frequencies and and thus noise affects significantly higher order partials, namely the phases (see spectral envelope in Figure 5.11). Thus, the vowel [u] is not representative of the normal configuration of the NRD parameters.



(a) (b)

Figure 5.13: (a) Average of the NRD coefficients of each vowel. Vowel [a] - blue line; vowel [ɛ] - red line; vowel [ɨ] - green line; vowel [ɔ] - black line; vowel [u] - pink line . (b) Total average of the NRD coefficients (magenta line). Estimated NRD model (black dotted line).

The NRD model obtained is then used to predict the phases of the partials to be synthesized by computing their relative delays as:

$$NRD(i) = im + b \quad \text{with} \quad i = 1, ..., N \tag{5.11}$$

where *NRD* is the vector containing the delays of the *N* harmonic partials, *i* is the harmonic index, *m* and *b* are, respectively, the slope and the bias of the line that represents the model. The number of harmonic partials to be synthesized, N, corresponds to the maximum number of partials that fit in the available bandwidth.

The information of the NRDs together with the delay of the fundamental frequency, which is computed as will be explained in the following sections, allows to synthesize the phase of each partial.

5.4 Prediction of the Delay of the Fundamental Frequency

In this section we will describe a method for the prediction of the delay of the fundamental frequency sinusoid, n_0, whose conversion to ϕ_0 is straightforward (see Eq. 5.9). Actually, as mentioned in the previous sections, our aim is to find the phase of the fundamental frequency on a frame-by-frame analysis in order to compute the phase of the partials.

5.4.1 Frame-to-frame F_0 Continuation

The prediction of the phase of the fundamental frequency in the current frame is based on the assumption of a linear variation of the frequency between consecutive frames and is given by

$$f(t) = \frac{d\phi}{2\pi dt} = F_0 + (F_1 - F_0)\frac{t}{T}, \tag{5.12}$$

where the continuous time variable t was chosen for simplicity purposes. F_0 and F_1 denote, respectively, the instantaneous fundamental frequency of the previous frame (at its center) and the instantaneous fundamental frequency of the center of the current frame, and T is the temporal distance between the frames.

This is the simplest form of variation that we might assume, which is very realistic if the frame-to-frame frequency variations are sufficiently slow.

5.4.2 Frame-to-frame Phase Continuation

Starting from the above equation Equation 5.12, we can write it in the form

$$\phi = 2\pi \int f(t)\mathrm{d}t = 2\pi(F_0 t + (F_1 - F_0)\frac{t^2}{2T}) + \phi_0, \tag{5.13}$$

which implies that

$$\begin{aligned} \text{if} \quad t = 0 \quad \text{then,} \quad \phi &= \phi_0, \\ \text{if} \quad t = T \quad \text{then,} \quad \phi_1 &= 2\pi(\frac{F_0 + F_1}{2})T + \phi_0, \end{aligned} \tag{5.14}$$

where ϕ_0 and ϕ_1 are, respectively, the phase of the fundamental frequency at the center of the previous frame and at the center of the current frame.

Since our analysis is performed in the discrete-time domain, where T corresponds to $\frac{N}{2}$ samples, we have

$$\phi_1 = \frac{2\pi}{N} \frac{\boldsymbol{\ell}_0 + \Delta\boldsymbol{\ell}_0 + \boldsymbol{\ell}_1 + \Delta\boldsymbol{\ell}_1}{2} \frac{N}{2} + \phi_0, \tag{5.15}$$

where

$$\phi_0 = \frac{2\pi}{N}(\boldsymbol{\ell}_0 + \Delta\boldsymbol{\ell}_0)n_0 = \frac{2\pi}{N}(\boldsymbol{\ell}_1 + \Delta\boldsymbol{\ell}_1)n_{0new}, \tag{5.16}$$

and ℓ and $\Delta\ell$ are, respectively, the integer part and the fractional part of a DFT bin of the fundamental frequency. Thus,

$$\begin{aligned} n_{0new} = n_0 \frac{\ell_0 + \Delta\ell_0}{\ell_1 + \Delta\ell_1} = n_0 \frac{P_1}{P_0}, \\ \text{with} \quad P_i = \frac{N}{\ell_i + \Delta\ell_i}. \end{aligned} \tag{5.17}$$

As for the first part of the equation Eq. 5.15, which is due to the increment from $F0$ to $F1$, it can be written as

$$\begin{aligned} \frac{2\pi}{N} \frac{\frac{\ell_0 + \Delta\ell_0}{N} + \frac{\ell_1 + \Delta\ell_1}{N}}{2} \frac{N}{2} = 2\pi \frac{P_0 + P_1}{2P_0P_1} \frac{N}{2} \\ \text{where} \quad \frac{P_0 + P_1}{2P_0P_1} = \frac{1}{P}, \end{aligned} \tag{5.18}$$

P is the average period. However, we will consider $P = \frac{P_0 + P_1}{2}$ as it allows for a smoother frequency transition.

Finally, we get the new phase of the fundamental frequency in the current frame, ϕ_1, by computing

$$\frac{N}{2} + n_{0new} \mod P, \tag{5.19}$$

where mod denotes the remainder after the integer division.

5.5 Final Considerations

The proposed solution for artificial voicing of the whispered speech is completely parametric. It relies on the extraction of representative parameters from the whispered signal and in the estimation of models capable of converting these parameters into those of normal sounding speech thus, obtaining an artificially voiced signal. Since the models are obtained from training of experimental data extracted from voiced utterances of the patient under study, they should preserve some elements of his vocal signature necessary to ensure the speaker's identity of the reconstructed sentence.

The results of the algorithm will be described shortly in this dissertation.

Chapter 6

Results of Automatic Whisper-to-Speech Conversion

In this chapter we finally describe the results of the automatic whisper-to-speech conversion algorithm, as well as the results of subjective listening tests that were conducted in order to assess its performance.

The sentence that will be analysed herein is "A Sofia saiu cedo da sala para conversar com o Aurélio.".

6.1 Automatic Segmentation

Since the HMM classification was applied to 10ms-long frames with 8ms-overlap, which results in frames with 220 samples (as the sampling frequency is 22050Hz), it is necessary to convert the HMM classification results to frames with 1024 samples and 50% overlap, that are used in the conversion algorithm. This conversion of the vectors is performed after the classification because using frame lengths of 1024 samples to train the HMM would not be viable since frame precision is required.

Furthermore, the results of the HMM classifier were converted to binary classifications, namely "Unvoiced frame" and "Voiced frame". The former encompasses all the frames identified as "silence", "unvoiced plosive" and "unvoiced fricative" and the latter represents all the frames identified as "voiced" as described in Ch. 4.

The confusion matrix obtained after all these adjustments to the HMM classification is represented in Table 6.1. The test sentence is the one to be converted into normal sounding speech and

	UV	V
UV	62	9
V	28	131

Table 6.1: Confusion matrix. It shows in each row the number of speech frames pertaining to the state indicated in the first column that have been classified as unvoiced (UV) or voiced (V).

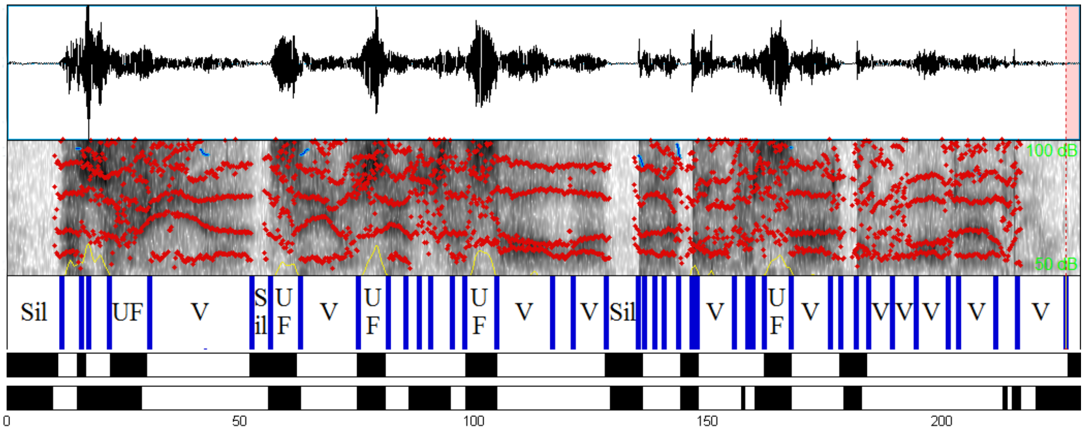


Figure 6.1: Representation of the time waveform of the whispered sentence "A Sofia saiu cedo da sala para conversar com o Aurélio.", the corresponding spectrogram and the manual labelling, where the blue vertical lines represent the boundaries. The frame vectors representing the ground truth (upper) and the HMM classification (lower) are illustrated, where black corresponds to "Unvoiced frame" and white corresponds to "Voiced frame".

now consists of 230 frames; the training sentences are the remaining two of the whispered speech database and together they consist of 591 frames.

Figure 6.1 depicts the time waveform of the whispered sentence, the corresponding spectrogram, the manual labelling and the vectors of the frames after the conversion to 1024 samples and to a binary classification, where black represents "Unvoiced frame" and white represents "Voiced frame". The upper frame vector represents the ground truth and the lower represents the HMM classification.

In spite of the observable classification errors and the problems mentioned in Table 3.1, these are actually difficult to perceive in the final reconstructed sentence and thus, it can be concluded that voice synthesis using automatic segmentation of the whispered sentence is feasible.

6.2 Artificial Voicing

The frames classified as "Voiced frame" by the HMM classifier, were artificially voiced and the remaining frames were directly extracted from the whispered sentence.

However, a problem was encountered in the magnitudes of the synthesized sentence due to the spectral envelope modeling, i.e. the estimated envelope transformation function, used to compute the magnitudes of the harmonic partials. Particularly, some regions of the time waveform of the signal exhibited an excessive amplitude as the example of Figure 6.2 illustrates. In fact, as described in Ch. 5, the training of the model is performed based on the spectral envelopes of 8 sustained vowels, which probably are not representative of the whole range of spectral content that may arise. Therefore, we can conclude that the approach can not achieve the desired prediction capacity of the cepstral coefficients of the target envelope, quite likely, due to the reduced training

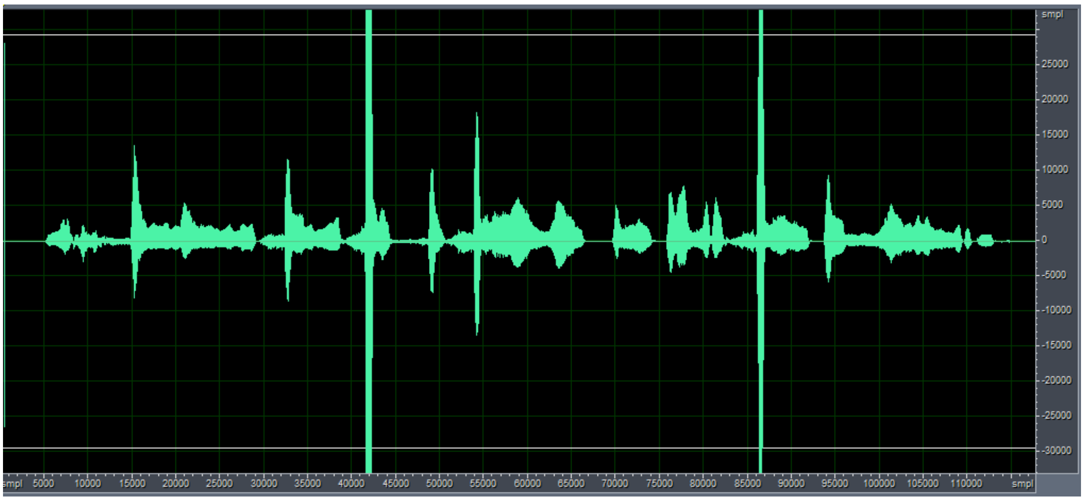


Figure 6.2: Time waveform of the synthesized sentence, i.e. after implanting artificial voicing, generated by using the estimated spectral envelope model. The magnitude errors are illustrated.

data set. Indeed, the results of the envelope transformation are clearly inaccurate when the input vector is possibly quite different from those used for training.

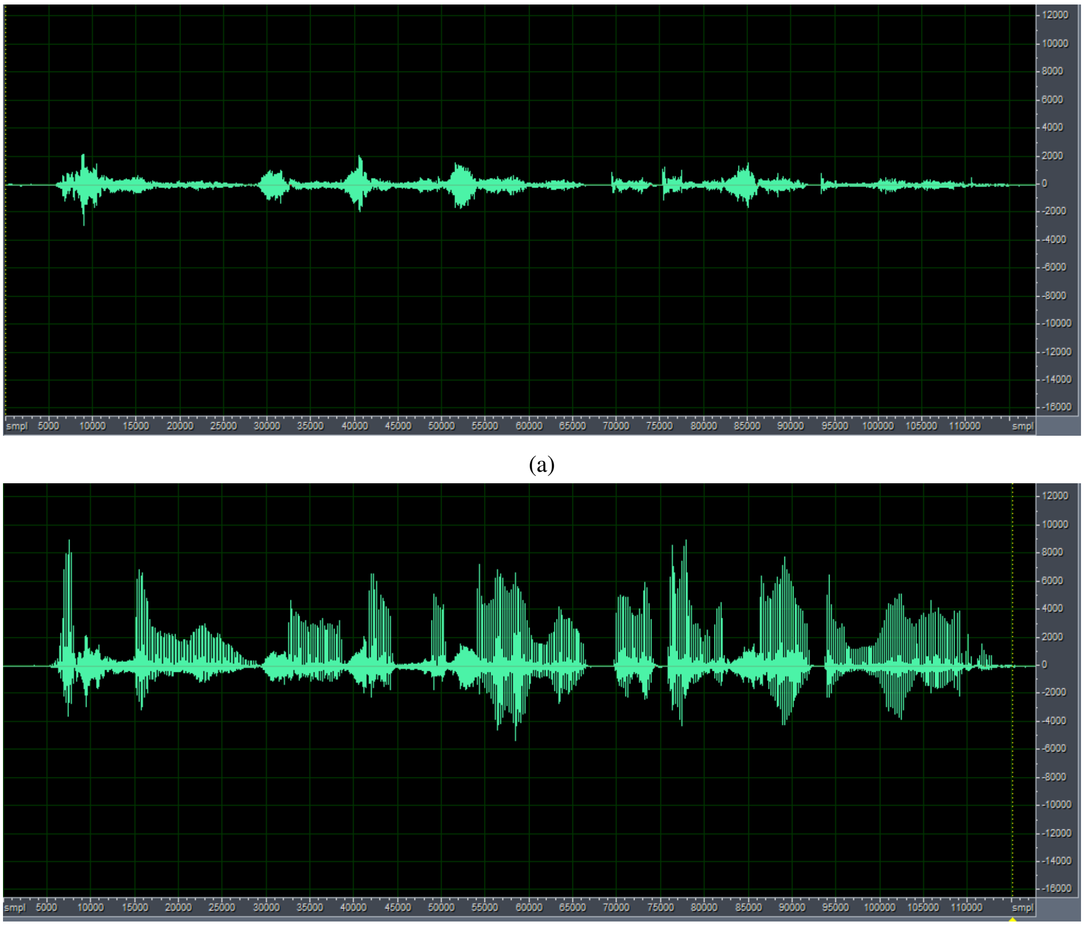
In order to avoid this problem, we use the spectral envelope of the whispered sentence, computed as described in the previous chapter, and transform the magnitudes of the harmonic partials by summing a set of values that vary linearly from 25dB (for the first partial) to 10dB (for the last partial). These values were obtained through fine-tuning, according to the results of some experiments. This is a simple approach that turned out to give rise to interesting results. Figure 6.3 depicts the time waveform of the whispered sentence and the result of the conversion to normal sounding speech. In Figure 6.4, the corresponding spectrograms are illustrated, where a well defined harmonic structure is observable in the spectrogram of the artificially voiced signal.

6.3 Subjective Listening Tests

In order to evaluate the performance of the automatic whisper-to-speech conversion algorithm, a subjective listening test was conducted and involving 28 Portuguese native listeners, 16 male and 12 female, with ages ranging from 19 to 52. Seven of the listeners (three speech therapists, three researchers and a Professor/researcher in the voice field) had previous experience in the evaluation of speech signals, which ranges from 4 to 16 years. Besides, thirteen listeners were engineering students and the remaining were students from a different area or workers.

Three parameters of perceived quality of the synthesized speech were evaluated:

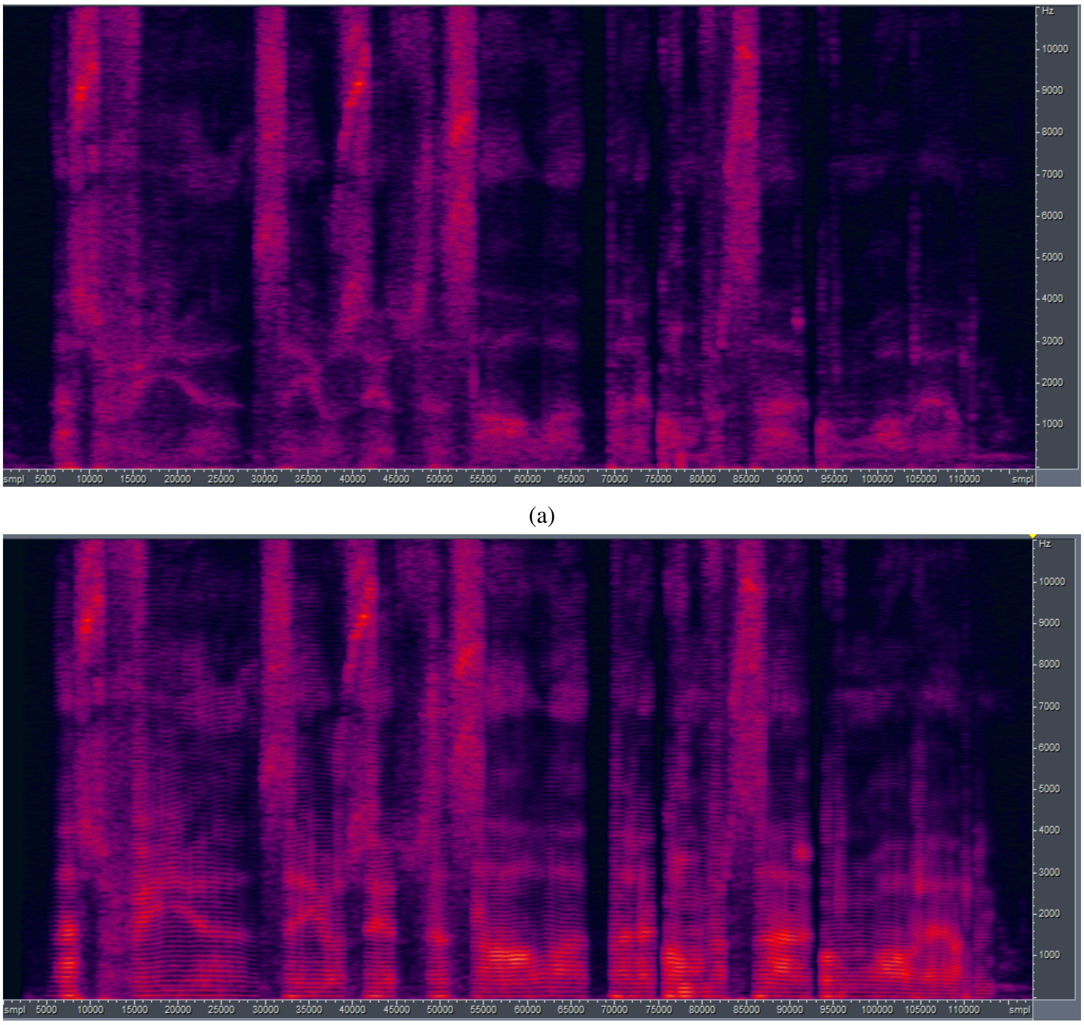
- **Intelligibility** - degree to which the sentence content can be understood;
- **Naturalness** - degree of similarity with a human voice, i.e. how closely it sounds like human speech. In other words, absence of perceptual artificiality;



(a)

(b)

Figure 6.3: (a) Time waveform of the whispered sentence "A Sofia saiu cedo da sala para conversar com o Aurélio.". (b) Time waveform of the synthesized sentence, i.e. after implanting artificial voicing.



(a)

(b)

Figure 6.4: (a) Spectrogram of the whispered sentence "A Sofia saiu cedo da sala para conversar com o Aurélio.". (b) Spectrogram of the synthesized sentence.

Intelligibility

Score	**Description**
1	Much less intelligible than the whispered version.
2	Slightly less intelligible than the whispered version (up to four sounds / words not correctly perceived).
3	Comparable with the whispered version.
4	Slightly more intelligible than the whispered version.
5	Much more intelligible than the whispered version.

Naturalness

Score	**Description**
1	Completely robotic.
2	It presents some, although very few components that sound like a human voice.
3	It presents as many components that sound like a robotic voice as those that sound like a human voice.
4	It is very similar to a human voice, however it still presents some robotic voice components.
5	Quite natural, it sounds like human speech.

Identity

Score	**Description**
1	Can not identify the speaker or his gender.
2	It is possible to identify the gender of the speaker.
3	It appears to have some characteristics of the speaker's vocal signature, however it is not conclusive.
4	It presents some characteristics of the speaker's vocal signature, however it is not completely accurate.
5	It is possible to identify the speaker with accuracy.

Figure 6.5: Score tables for the parameters: (a) intelligibility; (b) naturalness; (c) identity.

- **Identity** - degree of similarity with the vocal signature of the speaker, i.e. the possibility of identifying the speaker.

The intelligibility of the synthesized signal was assessed through a comparison with the whispered signal. As for the naturalness, two synthesized versions were evaluated without any sound file to serve as a comparison: a version with a variable F_0 (the one used in the assessment of the intelligibility and identity) and a version with a fixed F_0. Finally, the speaker's identity of the synthesized speech was assessed through a comparison with the original normal speech version.

A score was assigned by the listeners to each parameter over a five-point scale. The descriptions of the scores depend on the corresponding parameter as described in the tables of Figure 6.5.

Two sentences were assessed, namely "A Sofia saiu cedo da sala para conversar com o Aurélio." and "O vento norte e o sol discutiam qual dos dois era o mais forte.", which will be denoted hereinafter as $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively, for graphical visualization purposes. Regarding to the second versions of each sentence, i.e. those with a fixed F_0, we will denote them as $sentence_{\upsilon 2}$ and $sentence_{\omega 2}$.

The three measures selected to describe the data set resulting from the subjective listening tests were: mean, variance and mode. Their values for each synthesized sentence are illustrated in Figure 6.6.

Concerning intelligibility, it was assessed with:

- **A mean value of 3,42 and 2,81** for $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively. According to the corresponding table in Figure 6.5, these values mean that, in general, the degree to which the synthesized sentence content can be understood is roughly comparable to that of the whispered version, with a slight improvement. However, it is slightly less intelligible than the whispered version in the case of $sentence_{\omega 1}$.
- **A variance value of 0,81 and 0,96** for $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively.
- **A mode value of 4 and 2** for $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively.

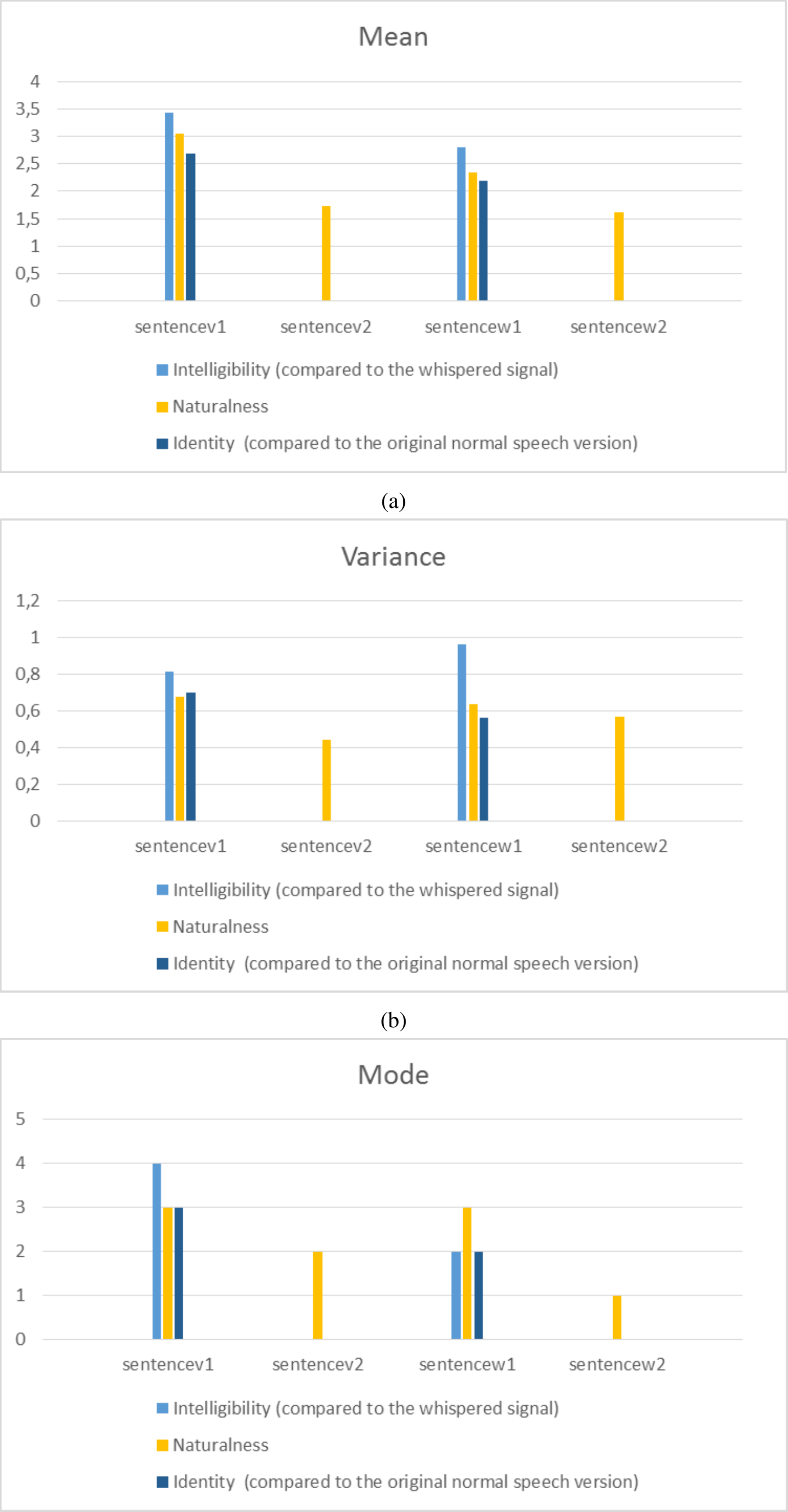
The naturalness was assessed with:

- **A mean value of 3,04, 1,73, 2,35 and 1,62** for $sentence_{\upsilon 1}$, $sentence_{\upsilon 2}$, $sentence_{\omega 1}$ and $sentence_{\omega 2}$ respectively. These values denote that $sentence_{\upsilon 1}$ has as many components that sound like a robotic voice as those that sound to a human voice. However, the synthesized sentence $sentence_{\omega 1}$ is slightly closer to a robotic voice. As expected, $sentence_{\upsilon 2}$ and $sentence_{\omega 2}$ sound almost completely robotic.
- **A variance value of 0,68, 0,44, 0,63 and 0,57** for $sentence_{\upsilon 1}$, $sentence_{\upsilon 2}$, $sentence_{\omega 1}$ and $sentence_{\omega 2}$ respectively.
- **A mode value of 3, 2, 3 and 1** for $sentence_{\upsilon 1}$, $sentence_{\upsilon 2}$, $sentence_{\omega 1}$ and $sentence_{\omega 2}$ respectively.

Finally, the identity was assessed with:

- **A mean value of 2,69 and 2,19** for $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively. These values denote that $sentence_{\upsilon 1}$ some listeners can perceive some characteristics of the speaker's vocal signature but many of them can only identify the speaker's gender. However, in the synthesized sentence $sentence_{\omega 1}$ in general, the listeners only can identify the gender of the speaker.
- **A variance value of 0,7 and 0,56** for $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively.
- **A mode value of 3 and 2** for $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$ respectively.

As expected, the second versions of each synthesized sentence, i.e. $sentence_{\upsilon 2}$ and $sentence_{\omega 2}$, demonstrate much lower degrees of naturalness. This fact proves the importance of pitch variation



(a)

(b)

(c)

Figure 6.6: Results from the subjective listening tests of each parameter - intelligibility, naturalness, identity - for each synthesized sentence: (a) mean; (b) variance; (c) mode.

throughout the frames for the synthesis of human-like sounding speech. Despite the simplicity of the solution for obtaining a varying F_0, it had an important impact on the degree of naturalness of the synthesized sentence, which means that this feature warrants further research.

In general, the results were worse in the case of the sentence $sentence_{\omega 1}$. This fact might be related to different existing coarticulation problems between $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$.

In some cases, the values of the variance are relatively high, which denotes that the opinions of the listeners were not unanimous. Such is the case of the variance value of the parameter intelligibility for $sentence_{\upsilon 1}$.

6.4 Final Considerations

We have shown the potential of HMM for identifying the candidate regions that should be artificially voiced, as well as the capacity of the whisper-to-speech conversion algorithm to synthesize speech with the incorporation of some elements of the speaker's vocal-signature. However, the results are still preliminary and require future developments.

We also proved the importance of F_0 variation to generate human-like sounding speech. In fact, [40] demonstrates, through the use of audio examples, that the normal speech is never steady in pitch. It contains frequency micro-modulations that affect the voice perception.

Co-articulation problems are observable in several aspects, such is the case of the classification errors of the HMM classifier or the differences observed in the values of the parameters intelligibility, naturalness and identity between $sentence_{\upsilon 1}$ and $sentence_{\omega 1}$. Therefore, these problems require future research.

Chapter 7

Conclusions

In this chapter, we conclude on the described work, analyse the fulfilment of the proposed objectives and suggest refinements for future developments. Besides, some differences with respect to previous work are also pointed out.

7.1 Achievements

The dissertation proposal consists of the implementation of a real-time whisper-to-speech conversion system, capable of achieving satisfactory levels of intelligibility and naturalness of the synthesized speech. Additionally, the incorporation of vocal signature elements would be desirable.

In fact, the proposed automatic whisper-to-speech conversion system has shown promising results both in the algorithm for automatic identification of the candidate regions that should be artificially voiced in whispered speech and in the algorithm to implant artificial voicing, showing that it is suitable for real-time operation. The models used to convert the whispered signal features into those of normal speech allowed for a preservation of some characteristics of the speaker's vocal signature.

In spite of the classification errors of the Hidden Markov Model classifier, these were not strongly perceived and thus, a good degree of intelligibility in the final reconstructed sentence was obtained.

The solution proposed herein is quite different from the one proposed in work [11], as the algorithm is completely parametric in our case. It uses representative features of the whispered signal either to identify the candidate regions that should be artificially voiced or to synthesize the voiced signal by transforming these features through the use of estimated models. However, there are some differences [1] between the previous work and the one presented herein that are important to mention:

[1] For comparative purposes with the work described in this document, we will only refer to the solution denoted as independent version by [11], since it is the only one that falls in the scope of this dissertation.

- **Speech database** - The speech database used in the previous work regarded healthy voices, i.e. information about speakers without voice disorders. However, as we have mentioned throughout the dissertation, the characteristics of the whispered signal are affected by the patient's disorder.

- **Manual segmentation of whispered speech** - Although an algorithm for automatic segmentation of whispered speech has been proposed, it did not fulfil the expectations and thus, a manual segmentation of the whispered signal had to be performed, which prevents real-time operation.

- **Low-level degree of naturalness** - The naturalness of the synthesized speech was classified as "Completely robotic" by the conducted subjective listening tests.

Finally, the work described herein warrants further research and development, since we believe that after some refinements a real-time and non-invasive solution may be implemented.

7.2 Considerations

Although the automatic classification through the use of Hidden Markov Model proved to achieve good results for the case of study, we can not completely infer on its performance since a richer speech database, i.e. a database with a wider variety and quantity of phonemes, is required. In fact, we observed that the classification process strongly depends on the coarticulation phenomenon. Actually, the small database had a negative impact also in the whisper-to-speech conversion algorithm, since we believe that the magnitude errors caused by the estimated spectral envelope model arose from that fact.

Additionally, the speech database should contain information about more than one speaker, in order to validate the results of the subjective listening tests, otherwise they are not completely conclusive.

Finally, the short time available for the study of all the required theoretical concepts and the development of the dissertation did not allow for the refinement of the proposed solutions, such is the case of the estimation of pitch throughout the frames.

7.3 Future Research

Future developments of the work described herein would involve solutions to the aforementioned problems. Therefore, some refinements are proposed:

- **Increase the speech database** - Samples from different speakers, suffering from voice disorders, and involving several coarticulation contexts should be collected.

- **Estimate a new spectral envelope model** - We believe that the magnitude problems of the synthesized speech encountered when using the estimated transformation function, can be overcame by increasing the training data set used to estimate the model, i.e. using a wider variety of phonemes.

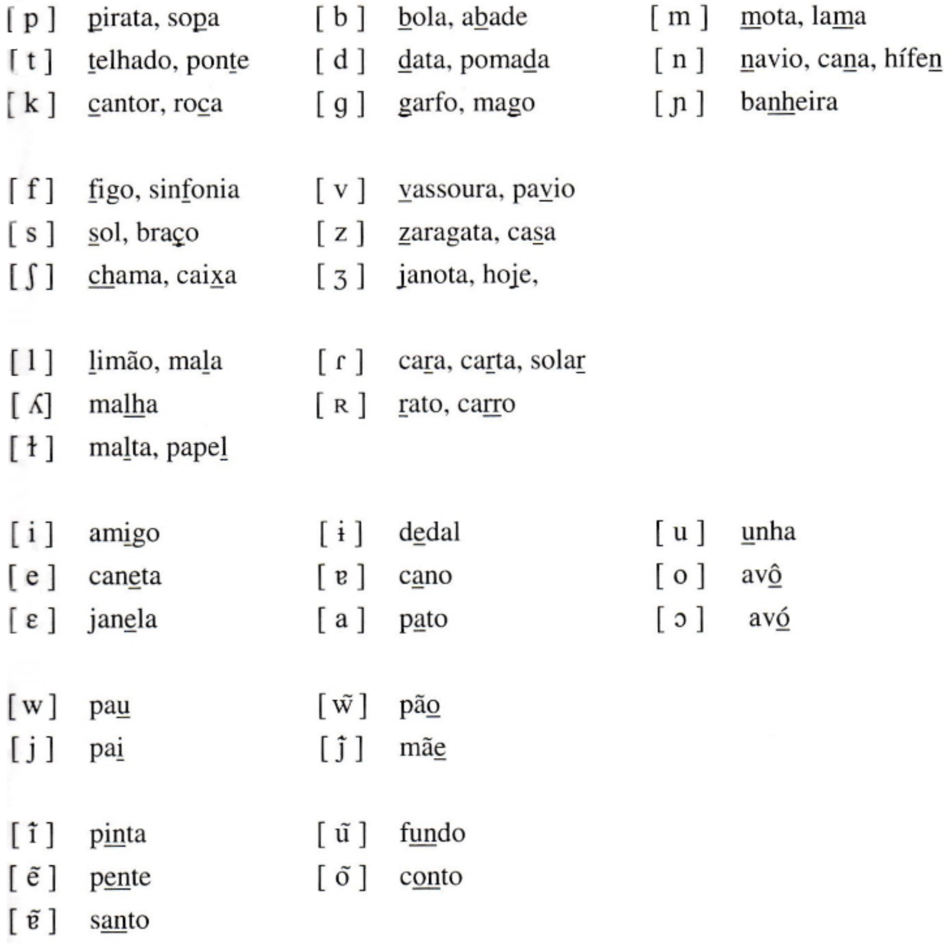
- **Estimate pitch (F_0)** - An accurate model to predict the variation of the pitch value throughout the frames is the next most challenging step. The F_0 variation is the one responsible for providing a human-like sounding voice to the synthesized speech. A possible approach to follow is described in [36], which uses a similar reasoning to the one used in our work for the estimation of the spectral envelope model, i.e. it performs Statistical Modeling through the use of GMM.

These are only some research directions, however more refinements can be performed in order to improve the algorithm and implement it in an external assistive device.

Appendix A

Representation of speech sounds in EP - IPA

In this section are exposed the subset of symbols from IPA used to describe the several speech sounds in European Portuguese [4].



[p]	pirata, sopa	[b]	bola, abade	[m]	mota, lama
[t]	telhado, ponte	[d]	data, pomada	[n]	navio, cana, hífen
[k]	cantor, roca	[g]	garfo, mago	[ɲ]	banheira
[f]	figo, sinfonia	[v]	vassoura, pavio		
[s]	sol, braço	[z]	zaragata, casa		
[ʃ]	chama, caixa	[ʒ]	janota, hoje,		
[l]	limão, mala	[ɾ]	cara, carta, solar		
[ʎ]	malha	[ʀ]	rato, carro		
[ɫ]	malta, papel				
[i]	amigo	[ɨ]	dedal	[u]	unha
[e]	caneta	[ɐ]	cano	[o]	avô
[ɛ]	janela	[a]	pato	[ɔ]	avó
[w]	pau	[w̃]	pão		
[j]	pai	[j̃]	mãe		
[ĩ]	pinta	[ũ]	fundo		
[ẽ]	pente	[õ]	conto		
[ɐ̃]	santo				

Figure A.1

References

[1] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[2] Articulators of the vocal tract, Accessed: 2014-11-30. URL: http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics_files/image002.jpg.

[3] Source-filter model of speech production, Accessed: 2014-12-15. URL: http://www.ee.columbia.edu/~dpwe/e4896/lectures/E4896-L06.pdf.

[4] Maria Helena Mira Mateus, Isabel Falé, and Maria João Freitas. *Fonética e Fonologia do português*. 2005.

[5] Wai C Chu. *Speech coding algorithms: foundation and evolution of standardized coders*. John Wiley & Sons, 2004.

[6] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[7] Keiichi Tokuda, Heiga Zen, and Alan W Black. An hmm-based speech synthesis system applied to english. In *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*, pages 227–230. IEEE, 2002.

[8] Hamid Reza Sharifzadeh and Ian Vince McLoughlin. Voiced speech from whispers for post-laryngectomised patients. *IAENG International Journal of Computer Science*, 36(4):367–377.

[9] Vincent Callanan, Paul Gurr, David Baldwin, Morwenna White-Thompson, Jane Beckinsale, and Jane Bennetf. Provox$^{\mathrm{TM}}$ valve use for post-laryngectomy voice rehabilitation. *The Journal of Laryngology & Otology*, 109(11):1068–1071, 1995.

[10] James H Brandenburg. Vocal rehabilitation after laryngectomy. *Archives of Otolaryngology*, 106(11):688–691, 1980.

[11] Paulo Jorge Proença de Azevedo. Vozeamento artificial de fala não vozeada. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2012.

[12] Jeremy Bradbury. Linear predictive coding. *Mc G. Hill*, 2000.

[13] Zbynek Tychtl and Josef Psutka. Speech production based on the mel-frequency cepstral coefficients. In *EUROSPEECH*, volume 99, pages 2335–2338, 1999.

[14] Xing Fan and John HL Hansen. Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. *Speech communication*, 55(1):119–134, 2013.

[15] Robert W Morris and Mark A Clements. Reconstruction of speech from whispers. *Medical Engineering & Physics*, 24(7):515–520, 2002.

[16] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45(2):139–152, 2005.

[17] Slobodan T Jovičić and Zoran Šarić. Acoustic analysis of consonants in whispered speech. *Journal of voice*, 22(3):263–274, 2008.

[18] Hamid R Sharifzadeh, Ian Vince McLoughlin, and Farzaneh Ahmadi. Reconstruction of normal sounding speech for laryngectomy patients through a modified celp codec. *Biomedical Engineering, IEEE Transactions on*, 57(10):2448–2458, 2010.

[19] Ian Vince McLoughlin, Jingjie Li, and Yan Song. Reconstruction of continuous voiced speech from whispers. In *INTERSPEECH*, pages 1022–1026, 2013.

[20] Heiga Zen, Keiichi Tokuda, and Alan W Black. Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064, 2009.

[21] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku. Hmm-based speech synthesis utilizing glottal inverse filtering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):153–165, 2011.

[22] Keiichi Tokuda, Takashi Masuko, Tetsuya Yamada, Takao Kobayashi, and Satoshi Imai. An algorithm for speech parameter generation from continuous mixture hmms with dynamic features. 1995.

[23] Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech synthesis using hmms with dynamic features. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 389–392. IEEE, 1996.

[24] Raphael Meyer. Source excitation generation for a hmm based synthesizer. 2006.

[25] Adobe audtion (cool edit pro)., Accessed: 2015-03-15. URL: `https://creative.adobe.com/products/audition`.

[26] Paul Boersma e David Weenink. *Praat: doing phonetics by computer.* URL: `http://www.fon.hum.uva.nl/praat/`.

[27] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[28] Namrata Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *International Journal for Advance Research in Engineering and Technology*, 1, 2013.

[29] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9):2505–2517, 2012.

[30] MATLAB. *Matrix Laboratory*. The MathWorks Inc. URL: `http://www.mathworks.com/`.

[31] Kevin Murphy. Hidden markov model (hmm) toolbox for matlab, 1998. URL: `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`.

[32] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[33] Aníbal JS Ferreira. Accurate estimation in the odft domain of the frequency, phase and magnitude of stationary sinusoids. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 47–50. IEEE, 2001.

[34] Aníbal JS Ferreira. Static features in real-time recognition of isolated vowels at high pitch. *The Journal of the Acoustical Society of America*, 122(4):2389–2404, 2007.

[35] Ken J Kallail and Floyd W Emanuel. Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *Journal of Speech, Language, and Hearing Research*, 27(2):245–251, 1984.

[36] Taoufik En-Najjari. *Conversion de voix pour la synthèse de la parole*. PhD thesis, Rennes 1, 2005.

[37] Ioannis Stylianou. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[38] Alexander Blouke Kain. *High resolution voice transformation*. PhD thesis, Oregon Health & Science University, 2001.

[39] Ricardo Sousa and Aníbal Ferreira. Importance of the relative delay of glottal source harmonics. In *Audio Engineering Society Conference: 39th International Conference: Audio Forensics: Practices and Challenges*. Audio Engineering Society, 2010.

[40] Albert S Bregman and Pierre A. Ahad. *Demonstrations of Auditory scene analysis: The perceptual organization of sound*. MIT press, 1996.

[41] Klára Vicsi and György Szaszák. Automatic segmentation of continuous speech on word level based on supra-segmental features. *International Journal of Speech Technology*, 8(4):363–370, 2005.